Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration

Yutaka Saito,^{1,2,4,5} Misaki Oikawa,³ Takumi Sato,³ Hikaru Nakazawa,³ Tomoyuki Ito,³ Tomoshi Kameda,^{1,5} Koji Tsuda,^{4,5,6,*} and Mitsuo Umetsu^{3,5,*}

¹Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.

²AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.

³Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, 6-6-11 Aoba, Aramaki, Aoba-ku, Sendai 980-8579, Japan.

⁴Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan.

⁵Center for Advanced Intelligence Project, RIKEN, 1-4-1 Nihombashi, Chuo-ku, Tokyo 103-0027, Japan.

⁶Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan.

*Corresponding authors:

Koji Tsuda (tsuda@k.u-tokyo.ac.jp)

Mitsuo Umetsu (mitsuo@tohoku.ac.jp)

25 **Abstract**

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

43

44

45 **Introduction**

46 Proteins are a class of biopolymer in which amino acids with various side chains are linked by
 47 peptide bonds. The diversity of possible amino acid sequences generates various biogenic functions,
 48 and allows us to create novel functions not found in nature. While genomics studies have provided
 49 vast information on natural proteins, they cover only a small subset of sequence space (defined as
 50 the set of all possible amino acid sequences of a protein). Thus, mutagenesis-based protein
 51 engineering has been used to create novel protein variants with desired functions by introducing
 52 amino acid mutations into natural proteins, which is called directed evolution [1].

53 Current experimental approaches for directed evolution often fail to obtain desirable variants
 54 due to the difficulty in exploring sequence space. Iterative saturation mutagenesis (ISM) is a
 55 conventional method where saturation mutagenesis (substitution of one or a few residues for all
 56 possible amino acids) and selection of the optimal variant are iteratively conducted [2]. In this
 57 method, the explored region of sequence space at each iteration may be too small to contain
 58 desirable variants with high probability. This probability can be increased by preparing an
 59 extremely large variant library with mutations in many residues while the library size would be
 60 beyond the limit of comprehensive screening [3-5]. Therefore, the success of directed evolution
 61 crucially depends on preparing a small library with high enrichment of desirable variants.

62 Recently, machine learning (ML) has been used to accelerate directed evolution of proteins
 63 [6,7]. In this method, saturation mutagenesis and/or random mutagenesis are performed to generate
 64 an initial library. The variants in the library are experimentally evaluated to obtain their sequences
 65 and functions, and then used as training data to construct a ML model that predicts the function
 66 from the sequence. By using the ML model, a second-round library that contains variants predicted
 67 to have improved functions is generated. This method enables to design a library with high
 68 enrichment of desirable variants, and thus has been successfully applied to directed evolution of
 69 various proteins including fluorescent proteins [8-10], enzymes [11-13], and others [14,15].

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 [16]. 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* [17-21]. Although the wild-type SrtA has poor reaction kinetics that limit its application, several variants with superior activity have been reported [18,22,23]. 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 [22] 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 [8], 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

95 mutated to make a variant library for training a Bayesian ML model, and then the ML model was
 96 used to rank all possible variants according to their predicted probability of having improved
 97 fluorescence performance. Here, we extend this method to enzymes with iterative design of libraries
 98 (Figure 1; Methods). In brief, certain residues in a target enzyme are mutated to make an initial
 99 library, and functional data (sequence, expression level, and enzyme activity) for approximately 100
 100 variants in this library are obtained. These data are then used to train a ML model, and top-ranked
 101 variants predicted by ML are used to design the second-round library. The functional data for
 102 dozens of variants in the second-round library are obtained, and used as additional training data for
 103 the ML model to design the third-round library. We applied this ML-guided library design cycle to
 104 SrtA for improving enzyme activity.

105

106 **Preparation of the initial library**

107 To select the residues that should be mutated in SrtA, we referred to the reported high activity
 108 variant called 5M [22] that has five mutated residues: P94R, D160N, D165A, K190E, and K196T
 109 (Figure 2). At these five residues, point mutagenesis and site-directed random mutagenesis were
 110 applied to make the initial library. The expression vectors for this library were prepared as
 111 described in our previous study [8], and used to transform *Escherichia coli*. Transformants were
 112 cultured in a 96-deep-well plate, and the sequence of each variant was determined. Then, each
 113 variant was purified by immobilized metal affinity chromatography (IMAC), and the expression
 114 level was measured. Variants with sufficient expression levels were used for enzyme activity assay.
 115 In this assay, two substrates—cycle3 green fluorescent protein (GFP) with an LPETG sequence at
 116 the C-terminus and Venus yellow fluorescent protein (YFP) with a GGG sequence at the
 117 N-terminus—were conjugated with each other by a SrtA variant [18] (Supplementary Figure S1),
 118 and the enzyme activity was estimated by SDS-PAGE analysis. In total, 80 variants were prepared
 119 (56 from site-directed random mutagenesis and 24 from point 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 Supplementary Table S1. Most variants had low expression levels and enzyme activity (Figure 3); only four of the 45 measured variants showed higher enzyme activity than the wild type, and all variants 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 a SrtA variant from its sequence (Methods; Equations 1 and 2). The enzyme performance score is a sigmoidal function of enzyme activity while it takes zero when the expression level is insufficient for measuring enzyme activity (e.g., the 35 low-expressed variants in Supplementary 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 [24] and position-specific score matrix (PSSM) [25] achieved the best prediction accuracy (Methods; Supplementary Figure S4; Supplementary Table S4). The dimensionality of this feature vector was six per residue. Thus, the number of features used in our final model was 30 (i.e., 6 dimensions $\times$ 5 mutated residues).

To train the ML model, two different datasets 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 datasets reflected practical scenarios in protein engineering: the 5M+ dataset 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[−] dataset simulates the situation where no such variant is available. We constructed two ML models by using the 5M⁺ and 5M[−] datasets, 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 (Supplementary 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 [22] 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 [22], had a slightly low rank within the top 5%.

Screening the variants 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 (Supplementary 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 level (Methods; Equation 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 (Supplementary Figure S2). These results indicate that ML successfully designed the second-round libraries with high enrichment of high activity variants.

Iteration of ML-guided library design

The results of the second-round libraries were used as additional training data for the ML models, and the third-round libraries were prepared by the same method as used for the second-round libraries (Supplementary Data S3-S4; Supplementary Table S3). Among these, only four variants showed insufficient expression levels for measuring enzyme activity (Supplementary 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 (Supplementary 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. 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 μ M).

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 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 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.

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 [8], 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 4 variants with higher enzyme activity than WT 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 [8,11-15]; 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 [9,10,26-28]. 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 function 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 [22]. 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.

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 [8]. Briefly, point saturation mutagenesis at each residue, and site-directed random mutagenesis at all five residues were performed by means of 22-c trick method [29]. The gene fragments of the SrtA variants were generated from the plasmid containing the wild-type sequence by overlap extension PCR [30] using external and 22-c trick primers. The second- and third-round libraries proposed by ML were generated in a similar way using appropriate mix primers instead of 22-c 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 μ g/mL ampicillin were randomly picked up, incubated overnight at 37 °C in 1 mL of LB broth containing 100 μ g/mL ampicillin in a

deep-well plate (Axygen, CA, USA), and used for gene sequence analysis. A 100- μ L aliquot of each cell culture was added into 900 μ L of 2 $\times$ YT broth supplemented with 100 μ g/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 μ L of Bugbuster Master Mix (Merk 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 μ M of the two substrates (LPETG-fused GFP and GGG-fused YFP) were reacted in buffer A containing 5 mM CaCl_2 and 3 ng/ μ L 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- μ L 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 Image Quant TL software.

To evaluate the enzyme activity of selected SrtA variants in detail, the transformed cells were cultivated in a 1-L shake flask containing 500 mL of 2 $\times$ YT broth supplemented with 100 μ g/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 μ M GGG-fused YFP was mixed with 0–120 μ M LPETG-fused GFP in buffer A containing 5 mM CaCl_2 and 3 ng/ μ L 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 [31] to conduct ML-guided directed evolution of fluorescent proteins [8]. 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:

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

where y is the enzyme performance score of the protein (defined in the next section), $\mathbf{x}$ is a feature vector of the protein, $\varphi(\mathbf{x})$ is a random feature map, and ε is an error term. Given a training dataset $\{(y, \mathbf{x})\}$, COMBO fits a weight vector $\mathbf{w}$ so that the enzyme performance score y can be predicted from the feature vector $\mathbf{x}$. For each variant not included in the training dataset, COMBO computes the acquisition function by Thompson sampling [31], which evaluates the probability that the enzyme performance score of the variant is higher than any variant in the training dataset. These values were used to rank all possible variants in sequence space.

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

319 **Enzyme performance score**

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

$$323 \quad y = \sigma(\text{activity}_n - 1), \quad (2)$$

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

330

331 **Feature vector**

332 We defined a feature vector $\mathbf{x}$ of the SrtA protein by concatenating the feature vectors of amino
333 acids at the five mutated residues. For a feature vector of each amino acid, we considered a variety
334 of amino acid descriptors including Z-scale [24], T-scale [32], ST-scale [33], FASGAI [34],
335 MS-WHIM [35], ProtFP [36], VHSE [37], and BLOSUM-based features [38]. In addition, we used
336 PSSM as a feature vector to incorporate the evolutionary information from SrtA homologs.
337 Specifically, PSSM was constructed by homology search using PSI-BLAST [25] with the wild-type
338 SrtA sequence as a query against the nr database. The homology search was iterated five times with
339 200 homologs used for updating PSSM per iteration. This produced PSSM $\{s_{ij}\}$ for each residue
340 i and amino acid j (Supplementary Data S5). We constituted a PSSM-based feature vector of the
341 SrtA protein by concatenating the PSSM values at the five mutated residues and their amino acids.

342 To select the optimal feature vector, we performed a feature selection procedure like that in our
343 previous study [8]. Briefly, we performed a benchmark experiment where COMBO was set to find

5M from the initial library by a Bayesian optimization procedure (Supplementary Figure S4). 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 (Supplementary 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.

References

1. Zeymer C, Hilvert D. Directed Evolution of Protein Catalysts. *Annu Rev Biochem.* 87:131-157 (2018). doi: 10.1146/annurev-biochem-062917-012034
2. Reetz MT, Wang LW, Bocla M. Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew Chem Int Ed Engl.* 45(8):1236-1241 (2006). doi: 10.1002/anie.200502746
3. Bunzel HA, Garrabou X, Pott M, Hilvert D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr Opin Struct Biol.* 48:149-156 (2018). doi: 10.1016/j.sbi.2017.12.010
4. Karamitros CS, Konrad M. Fluorescence-Activated Cell Sorting of Human l-asparaginase Mutant Libraries for Detecting Enzyme Variants with Enhanced Activity. *ACS Chem Biol.* 11(9):2596-2607 (2016). doi: 10.1021/acschembio.6b00283
5. Gianella P, Snapp EL, Levy M. An in vitro compartmentalization-based method for the selection of bond-forming enzymes from large libraries. *Biotechnol Bioeng.* 113(8):1647-1657 (2016). doi: 10.1002/bit.25939

- 368 6. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein
369 engineering. *Nat Methods*. 16(8):687-694 (2019). doi: 10.1038/s41592-019-0496-6
- 370 7. Chowdhury R, Maranas CD. From directed evolution to computational enzyme
371 engineering—A review. *AIChE J*. 66:e16847 (2020). doi: 10.1002/aic.16847
- 372 8. Saito Y, Oikawa M, Nakazawa H, Niide T, Kameda T, Tsuda K, Umetsu M.
373 Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS*
374 *Synth Biol*. 7(9):2014-2022 (2018). doi: 10.1021/acssynbio.8b00155
- 375 9. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein
376 engineering with sequence-based deep representation learning. *Nat Methods*. 16(12):1315-1322
377 (2019). doi: 10.1038/s41592-019-0598-1
- 378 10. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with
379 data-efficient deep learning. *Nat Methods*. 18(4):389-396 (2021). doi:
380 10.1038/s41592-021-01100-y
- 381 11. Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J.
382 Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol*. 27:16
383 (2007). doi: 10.1186/1472-6750-7-16
- 384 12. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C,
385 Tam S, Muley S, Grate J, Gruber J, Whitman JC, Sheldon RA, Huisman GW. Improving
386 catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol*. 25(3):338-44 (2007).
387 doi: 10.1038/nbt1286
- 388 13. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed
389 protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A*. 216(18):8852-8858
390 (2019). doi: 10.1073/pnas.1901979116

- 391 14. Giguère S, Laviolette F, Marchand M, Tremblay D, Moineau S, Liang X, Biron É, Corbeil J.
392 Machine learning assisted design of highly active peptides for drug discovery. *PLoS Comput*
393 *Biol.* 11(4):e1004074 (2015). doi: 10.1371/journal.pcbi.1004074
- 394 15. Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V, Arnold FH. Machine
395 learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat*
396 *Methods.* 16(11):1176-1184 (2019). doi: 10.1038/s41592-019-0583-8
- 397 16. Ton-That H, Liu G, Mazmanian SK, Faull KF, Schneewind O. Purification and characterization
398 of sortase, the transpeptidase that cleaves surface proteins of *Staphylococcus aureus* at the
399 LPXTG motif. *Proc Natl Acad Sci U S A.* 96(22):12424-12429 (1999). doi:
400 10.1073/pnas.96.22.12424
- 401 17. Pishesha N, Ingram JR, Ploegh HL. Sortase A: A Model for Transpeptidation and Its Biological
402 Applications. *Annu Rev Cell Dev Biol.* 34:163-188 (2018). doi:
403 10.1146/annurev-cellbio-100617-062527
- 404 18. Chen L, Cohen J, Song X, Zhao A, Ye Z, Feulner CJ, Doonan P, Somers W, Lin L, Chen PR.
405 Improved variants of SrtA for site-specific conjugation on antibodies and proteins with high
406 efficiency. *Sci Rep.* 6:31899 (2016). doi: 10.1038/srep31899
- 407 19. Sellmann C, Doerner A, Knuehl C, Rasche N, Sood V, Krah S, Rhiel L, Messemer A,
408 Wesolowski J, Schuette M, Becker S, Toleikis L, Kolmar H, Hock B. Balancing Selectivity and
409 Efficacy of Bispecific Epidermal Growth Factor Receptor (EGFR) × c-MET Antibodies and
410 Antibody-Drug Conjugates. *J Biol Chem.* 291(48):25106-25119 (2016). doi:
411 10.1074/jbc.M116.753491
- 412 20. Xiaolin D, Böker A, Glebe U. Broadening the scope of sortagging. *RSC Adv.* 9:4700-4721
413 (2019). doi: 10.1039/c8ra06705h
- 414 21. Matsumoto T, Furuta K, Tanaka T, Kondo A. Sortase A-Mediated Metabolic Enzyme Ligation
415 in *Escherichia coli*. *ACS Synth Biol.* 5(11):1284-1289 (2016). doi: 10.1021/acssynbio.6b00194

- 416 22. Chen I, Dorr BM, Liu DR. A general strategy for the evolution of bond-forming enzymes using
417 yeast display. *Proc Natl Acad Sci U S A*. 108(28):11399-11404 (2011). doi:
418 10.1073/pnas.1101046108
- 419 23. Piotukh K, Geltinger B, Heinrich N, Gerth F, Beyermann M, Freund C, Schwarzer D. Directed
420 evolution of sortase A mutants with altered substrate selectivity profiles. *J Am Chem Soc*.
421 133(44):17536-17539 (2011). doi: 10.1021/ja205630g
- 422 24. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant
423 for the design of biologically active peptides. A multivariate characterization of 87 amino acids.
424 *J Med Chem*. 41(14):2481-91 (1998). doi: 10.1021/jm9700575.
- 425 25. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped
426 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*
427 *Res*. 25(17):3389-3402 (1997). doi: 10.1093/nar/25.17.3389
- 428 26. Liu G, Zeng H, Mueller J, Carter B, Wang Z, Schilz J, Horny G, Birnbaum ME, Ewert S,
429 Gifford DK. Antibody complementarity determining region design using high-capacity
430 machine learning. *Bioinformatics*. 36(7):2126-2133 (2020). doi:
431 10.1093/bioinformatics/btz895.
- 432 27. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, Ehling RA, Bonati L,
433 Dahinden J, Gainza P, Correia BE, Reddy ST. Optimization of therapeutic antibodies by
434 predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng*.
435 5(6):600-612 (2021). doi: 10.1038/s41551-021-00699-9.
- 436 28. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, Church GM, Colwell LJ, Kelsic
437 ED. Deep diversification of an AAV capsid protein by machine learning. *Nat Biotechnol*.
438 39(6):691-696 (2021). doi: 10.1038/s41587-020-00793-4.

- 439 29. Kille S, Acevedo-Rocha CG, Parra LP, Zhang ZG, Opperman DJ, Reetz MT, Acevedo JP.
 440 Reducing codon redundancy and screening effort of combinatorial protein libraries created by
 441 saturation mutagenesis. *ACS Synth Biol.* 2(2):83-92 (2013). doi:10.1021/sb300037w
- 442 30. Sato K, Tsuchiya M, Saldanha J, Koishihara Y, Ohsugi Y, Kishimoto T, Bendig MM.
 443 Humanization of a mouse anti-human interleukin-6 receptor antibody comparing two methods
 444 for selecting human framework regions. *Mol Immunol.* 31(5):371-381 (1994).
 445 doi:10.1016/0161-5890(94)90115-5
- 446 31. Ueno T, Rhone TD, Hou Z, Mizoguchi T, Tsuda K. COMBO: an efficient Bayesian
 447 optimization library for materials science. *Materials Discovery.* 4:18-21 (2016). doi:
 448 10.1016/j.md.2016.04.001.
- 449 32. Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and
 450 its application in QSARs of peptides. *J Mol Struct.* 830(1):106-15 (2007). doi:
 451 10.1016/j.molstruc.2006.07.004.
- 452 33. Yang L, Shu M, Ma K, Mei H, Jiang Y, Li Z. ST-scale as a novel amino acid descriptor and its
 453 application in QSAM of peptides and analogues. *Amino Acids.* 38(3):805-16 (2010). doi:
 454 10.1007/s00726-009-0287-y.
- 455 34. Liang G, Li Z. Factor analysis scale of generalized amino acid information as the source of a
 456 new set of descriptors for elucidating the structure and activity relationships of cationic
 457 antimicrobial peptides. *QSAR Comb Sci.* 26(6):754–63 (2007). doi: 10.1002/qsar.200630145.
- 458 35. Zaliani A, Gancia E. MS-WHIM scores for amino acids: a new 3D-description for peptide
 459 QSAR and QSPR studies. *J Chem Inf Comput Sci.* 39(3):525–33 (1999). doi:
 460 10.1021/ci980211b.
- 461 36. van Westen GJ, Swier RF, Wegner JK, Ijzerman AP, van Vlijmen HW, Bender A.
 462 Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative

study of 13 amino acid descriptor sets. J Cheminform. 5(1):41 (2013). doi: 10.1186/1758-2946-5-41.

37. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers. 80(6):775-86 (2005). doi: 10.1002/bip.20296.

38. Georgiev AG. Interpretable numerical descriptors of amino acid space. J Comput Biol. 16(5):703-23 (2009). doi: 10.1089/cmb.2008.0173.

Acknowledgements

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, 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 National Institute of Advanced Industrial Science and Technology (AIST), and the supercomputer of ACCMS at Kyoto University.

Author Contributions

YS conducted the computational analysis. MO, TS, HK, and TI conducted the experiments. YS, MO, TK, KT and MU participated in the data interpretation. YS and MU wrote the paper. KT and MU conceived of the study and directed the project. All authors read and approved the final version of the manuscript.

488

489 **Competing Interests**

490 The authors declare that they have no competing interests.

491

492 **Supplementary Material**

493 Supplementary Information contains Supplementary Tables S1-S4, Supplementary Figures S1-S4,
494 and the legends for Supplementary Data S1-S5. The bodies of Supplementary Data S1-S5 are
495 available as separate Excel files.

496

497

Tables and Figures

Table 1. Summary of the initial library. 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 Supplementary Table S1 for the sequence information.

	Initial 5M+ library	Initial 5M- library
Activity measured	45	45
Low expression	35	35
Wild type	1	1
5M	1	0
Total	82	81

Table 2. Ranks of the reported high activity SrtA variants by ML prediction. 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).

Variants [22]	Prediction 5M+		Prediction 5M-	
	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	2,121	0.1
SNAKK	34,895	1.1	22,245	0.7
PNDKT	27,149	0.8	5,798	0.2
SDAKK	42,412	1.3	19,104	0.6

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

519

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

522

523

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 (Supplementary Table S1) are shown.

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 one-sided Mann-Whitney *U* test.

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.

554 corresponding positions. The first and the second principal components (PC1 and PC2) are shown
555 with their contribution rates in parentheses.
556