MABS
2023, VOL. 15, NO. 1, 2168470
https://doi.org/10.1080/19420862.2023.2168470

Taylor & Francis
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

Selection of target-binding proteins from the information of weakly enriched phage display libraries by deep sequencing and machine learning

Tomoyuki Ito[a†], Thuy Duong Nguyen[b†], Yutaka Saito[b,c,d,e], Yoichi Kurumida[b], Hikaru Nakazawa[a], Sakiya Kawada[a], Hafumi Nishi[f,g,h], Koji Tsuda[d,e,i], Tomoshi Kameda[b,e], and Mitsuo Umetsu[a,e]

[a]Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan; [b]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan; [c]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Tokyo, Japan; [d]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan; [e]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; [f]Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Japan; [g]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan; [h]Faculty of Core Research, Ochanomizu University, Tokyo, Japan; [i]Research and Services Division of Materials Data and Integrated Systems, National Institute for Materials Science, Tsukuba, Japan

ABSTRACT
Despite the advances in surface-display systems for directed evolution, variants with high affinity are not always enriched due to undesirable biases that increase target-unrelated variants during biopanning. Here, our goal was to design a library containing improved variants from the information of the “weakly enriched” library where functional variants were weakly enriched. Deep sequencing for the previous biopanning result, where no functional antibody mimetics were experimentally identified, revealed that weak enrichment was partly due to undesirable biases during phage infection and amplification steps. The clustering analysis of the deep sequencing data from appropriate steps revealed no distinct sequence patterns, but a Bayesian machine learning model trained with the selected deep sequencing data supplied nine clusters with distinct sequence patterns. Phage libraries were designed on the basis of the sequence patterns identified, and four improved variants with target-specific affinity (EC_{50} = 80–277 nM) were identified by biopanning. The selection and use of deep sequencing data without undesirable bias enabled us to extract the information on prospective variants. In summary, the use of appropriate deep sequencing data and machine learning with the sequence data has the possibility of finding sequence space where functional variants are enriched.

ARTICLE HISTORY
Received 24 July 2022
Revised 28 December 2022
Accepted 10 January 2023

KEYWORDS
Machine learning; antibody mimetics; directed evolution; deep sequencing analysis; phage display

Introduction

Accumulation of information on native proteins that is obtained by genomics and proteomics analyses may enable the discovery of proteins with desired functions,[1] and mutagenesis assists in generating novel functions not found in native proteins.[2,3] In the mutagenesis approach, several amino acid residues in a selected native protein are randomly altered to make a variant library, and variants with desirable phenotypes are selected under evolutionary pressure. However, the number of possible sequences generated by mutagenesis (sequence space) becomes dramatically expanded with the increase in the number of mutated residue positions, so the sequence space is often too large to experimentally prepare the variant library or screen all the variants.[3]

Surface-display systems linking genotype and phenotype have been used to generate proteins with molecular recognition function. In this system, a genetic library is prepared from an immune library or a library prepared by random mutagenesis of a fragment of a gene encoding a scaffold protein, and the encoded variants are displayed on phage,[4] yeast,[5] ribosome,[6] or mRNA.[7] Nonfunctional variants in a display library are eliminated during selection called biopanning, and the remaining variants are screened for their target binding affinity. At present, libraries containing 10^{9-13} variants can be prepared,[3] and degenerate codons[4] and trinucleotide cassettes[8] are used to limit the expansion of library size, so that those variants with high affinity for the target can be efficiently obtained. However, the functional variants obtained are not always the best in the designed sequence space.

Recently, machine learning has been combined with directed molecular evolution.[9–19] For improving or changing the function of protein interested, the sequences and functions of the variants in the initial mutagenesis library were evaluated and used as training data to construct a machine learning model that predicts the function from the sequence. By using the model, a second-round library that contains predicted variants to have improved or changed functions is generated. This method has successfully enabled us to design a library with high enrichment of desirable variants in the directed

CONTACT Tomoshi Kameda kameda-tomoshi@aist.go.jp Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan; Mitsuo Umetsu mitsuo@tohoku.ac.jp Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan

[†]Both authors contributed equally to this work.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2023.2168470

evolution of various proteins, including fluorescent proteins,[11–13] enzymes,[8,14–17] and others.[18,19] In the field of antibody engineering, deep sequencing data has been used as training data of machine learning. Deep sequencing analysis of the surface-display system supplies a large number of sequences with their antigen-binding properties, and then, a machine learning model trained with the deep sequencing data has proposed sequences with higher target affinity than that of the experimentally selected variants.[20–24]

For machine learning to successfully predict high functional variants, the experimental data where truly functional variants are distinguished is appropriate. In the case of a phage display approach combined with deep sequencing analysis, the variants with high affinity would be preferred to be more enriched than nonfunctional variants. However, the phages bearing target-unrelated variants are often propagated in biopanning, which may inhibit propagation of the phages bearing target-related variants.[25,26] This situation causes weak enrichment of functional variants and truly functional variants are hardly distinguished in the variant libraries after biopanning.

Antibody mimetics are small target-binding proteins with non-immunoglobulin (Ig) scaffolds.[27–30] These small antibody mimetics can be prepared by means of *Escherichia coli* (*E. coli*) expression and they are advantageous to fuse with other protein domains.[28,30] For example, they can be fused with full-length antibodies to generate multi-specific antibodies, which is a modality for the next generation of immunotherapy.[31–33] To functionalize a small protein, several amino acid residues are randomized to generate a variant library, and surface-display approaches are used to select the variants with target-specific affinity.[27] However, an immune library cannot be prepared, and scaffold proteins are prone to destabilization by randomization.[34] Consequently, functional variants are less likely to be enriched in the panned libraries of antibody mimetics than those of antibodies.

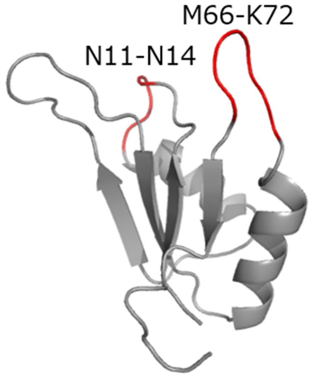
Here, we present a possibility of generating functional non-Ig scaffold proteins from the information of the "weakly enriched" libraries, i.e., where functional variants are weakly enriched, by deep sequencing and machine learning. We used a series of phage pools displaying a mutated non-Ig scaffold protein in which no prospective variants with target-specific affinity were experimentally identified (weakly enriched library). The phage pools have been obtained in the biopanning process against galectin-3, which can be a potential therapeutic target for cancer treatment and diagnostic biomarkers for several diseases, including heart failure and cancers.[35,36] By using deep sequencing analysis to the weakly enriched library, we evaluated sequence frequencies at various timepoints in the biopanning, and appropriate data was chosen for clustering analysis. The clustering analysis revealed no distinct sequence patterns, but a Bayesian machine learning model trained with the selected deep sequencing data supplied several clusters with distinct sequence patterns. Selection from the phage libraries based on the patterns led to the discovery of improved variants with target-specific affinity. This study shows the possibility of deep sequencing and machine learning for designing a refined library with prospective variants in the surveyed sequence space.

Results

Deep sequencing analysis of biopanned phage display libraries

Here, we used a series of phage pools displaying mutated second domain of human RNA-binding protein (Protein Data Bank ID: 2u2f), which have been previously biopanned.[34] Two adjacent loop structures (N11–N14 and M66–K72) were randomized (Figure 1) with degenerate codons designed to mimic an amino acid frequency of antibody complementarity-determining regions (CDRs),[37] and the variants were displayed on phages.[34] To select the functional variants, the phage bearing 2u2f variants (library size: ~10^9) were biopanned against galectin-3 in four rounds. Of ~200 clones screened from the phage pools in the last two rounds, only one variant with low target specificity was obtained from the third pool (Figure S1a). When produced in *E. coli*, this variant tended to form soluble aggregates (Figure S1b), and the monomeric form was partially denatured (Figure S1c). From the result that no target-specific variants were discovered from the screening of 200 clones, we considered that this panned library was weakly enriched.

The flow chart of the phage display biopanning is shown in Figure 2. In each round, we 1) selected target-bound phages, 2) infected *E. coli* with selected phages, and 3) amplified phages in *E. coli*. Besides the initial phage library, the sub-libraries of eluted phages, infected *E. coli*, and amplified phages (Figure 2) were used in deep sequencing analysis. Raw sequences were



```
1         10        20        30        40
AHKLFIGGLPNYLNDDQVKELLTSFGPLKAFNLVKDSATG

          50        60        70        80
LSKGYAFAEYVDINVTDQAIAGLNGMQLGDKKLLVQRASV

    85
GAKNA
```

Figure 1. Three-dimensional structure of the entire sequence of 2u2f. The two randomized loops are in red.

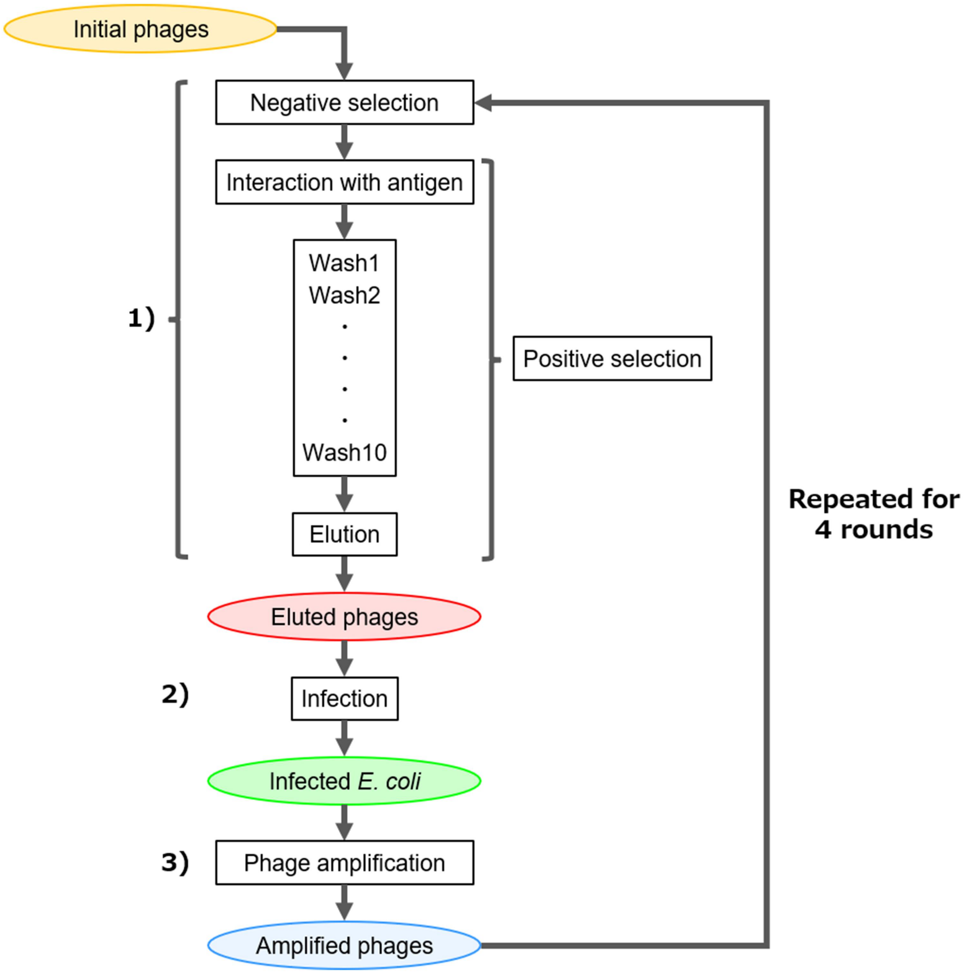


Figure 2. Workflow of biopanning. At each round, 1) target-bound phages were selected, 2) *E. coli* was infected with selected phages, and 3) phages were amplified in *E. coli*. Sub-libraries are surrounded by colored ellipses.

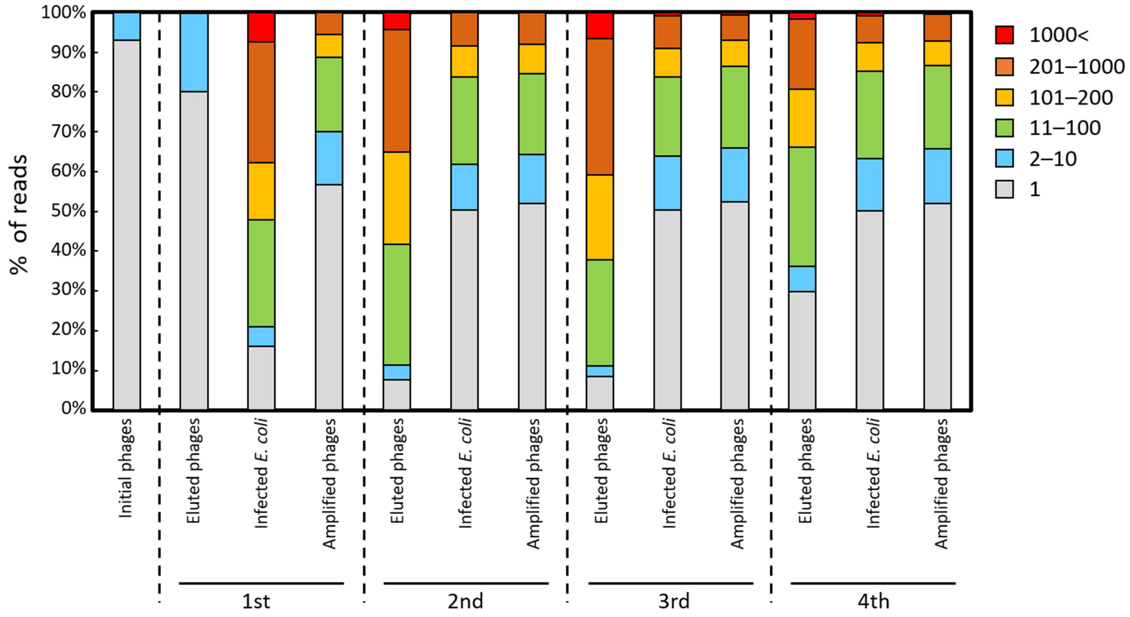


Figure 3. Distribution of unique sequences in each sub-library. The frequency of unique sequences is shown for single reads in gray, 2–10 reads in blue, 11–100 reads in green, 101–200 reads in yellow, 201–1000 reads in brown, and >1000 reads in red.

filtered and trimmed according to their quality, and the forward reads were merged with the corresponding reverse reads. The remaining 417,000–582,000 reads were translated, and sequences with any mismatches in the framework regions were excluded. These steps yielded 260,000–365,000 sequences per sub-library (Table S1). The number of sequences in the library of initial phages was 318,894, which is 0.03% of the initial library, and the coverages of deep sequencing data for the sub-libraries (eluted phages, infected *E. coli*, and amplified phages in each round) were 0.012–60, assuming that the diversity of sub-libraries corresponds to the number of output phages in Table S2.

The frequency distributions of variants in the sub-libraries in each round are shown in Figure 3. All the unique sequences in each sub-library were grouped by their read counts (1 read, 2–10 reads, 11–100 reads, 101–200 reads, 201–1,000 reads, and over 1,000 reads), and the frequencies of each group were calculated. The frequency of multiple-read unique sequences increased in the sub-library of eluted phages in the second round in comparison with that of amplified phages in the first round, indicating the enrichment of certain unique sequences by the selection of target-bound phages. A similar distribution change was observed in the sub-libraries between eluted phages in the third round and amplified phages in the second round.

However, the distributions of unique sequences at the infection steps differed from those of eluted phages: the frequency of multiple-read sequences increased in the first round, whereas those of single-read sequences increased in the second–fourth rounds. In the appearance frequency changes of amino acids at the infection steps (Figure S2a), the appearance frequency at all the residue positions were noticeably changed in the first round, and stop codons were gradually enriched in the second and third rounds. The infection of phages into *E. coli* caused the bias unrelated to target-binding selection. In particular, the infection of phages induced the enrichment of the phages bearing no variants. In the comparison between the sub-libraries of infected *E. coli* and amplified phages, the frequency distribution of unique sequence and the frequency appearance changes of amino acids showed undesirable bias in the first round, but not in other rounds (Figure 3 and S2b). Consequently, undesirable bias was caused at the infection and amplification steps in the first round and at the infection steps in other rounds. Notably, the phages bearing no variants tended to be enriched in the undesirable bias after the first round.

Round-to-round correlation plots between sequence frequencies at the infection steps in the adjacent rounds have been used to assess library enrichment during biopanning.[20,24] We made two types of correlation plots: conventional round-to-round comparison of the infection steps (Figure S3a) and the round-to-round comparison before and after the selection step without the influence of infection (Figure S3b). We observed no apparent enrichment in the former and a slight enrichment in the latter. In the correlation plot of the infection steps between first and second rounds, the sequences whose frequencies were decreased in the second round were present, but the decrease was not observed in other plots (Figure S3a): in the first round, the phages containing their sequences survived at the target-binding and wash steps, and they were infected, but they disappeared in the sub-library of infected *E. coli* in the second round. The phages with the sequences appearing in the unenriched area in the first round may have disadvantages for amplification in *E. coli*. These results indicate that the slight enrichment at the selection step was too low to overcome the sequence distribution in the undesirable bias caused by the infection of phages into *E. coli*. We considered that the enrichment at the selection step is correlated with target-binding function.

The clustering analysis with the deep sequencing data of the input and output sub-libraries (i.e., amplified and eluted phages, respectively) was tried to characterize the enrichment. Because no positive variants had been previously identified in the fourth round,[34] we used the sequence data of the second and third rounds. The blastp[38] was used to compare the top 10,000 sequences, and they were clustered by Cytoscape[39] on the condition that the cluster size is more than three. Consequently, they were clustered to 748 (second) and 775 (third), the largest cluster contains only 84 (second) and 72 (third) sequences, and the number of unclustered sequences was 4,323 (second) and 3,433 (third). The number of the clusters containing either of the top 1,000 sequences was 194 (second) and 302 (third), indicating that the clusters contain only 2 ~ 3 top 1,000 sequences on average.

Machine learning with training data

We constructed a machine learning model that predicts the binding affinity of 2u2f variants from their amino acid sequences (see Methods). A variant with higher enrichment in both the second and third rounds was postulated to have higher binding affinity. Accordingly, the performance score of each variant was defined based on its sequence frequencies in the second and third rounds and used as the regression label (see Methods). To calculate sequence frequencies reliably, we only used sequences with at least two reads in all input and output sub-libraries in the second and third rounds, which yielded 3,925 sequences for the training data. In the clustering analysis on the cluster size of more than three, the top 300 sequences according to the regression label, whose performance score was more than 0.5, were clustered to seven. The largest cluster of them contained only four sequences and the number of unclustered sequences was 277. Recently, we applied a Gaussian process to the directed evolution of proteins and a library with high enrichment of desirable variants was successfully designed.[8,11] The Gaussian process model was trained on this dataset and used to predict high score variants. To save calculation time, the sequence space size for prediction (prediction space) was limited by defining the amino acids appearing at each position: the amino acids whose frequencies increased in both the second and third rounds at each position were applied in the prediction space (Figure S4). The size of the prediction space was defined as 9.2×10^8. Consequently, in the training data, no sequences composed of only the amino acids applied for the prediction spaces were present, and there were three sequences where the amino acids were used at 10 of 11 positions. Aromaphilicity index[40] was applied to the amino acid descriptor for training the Gaussian process model.

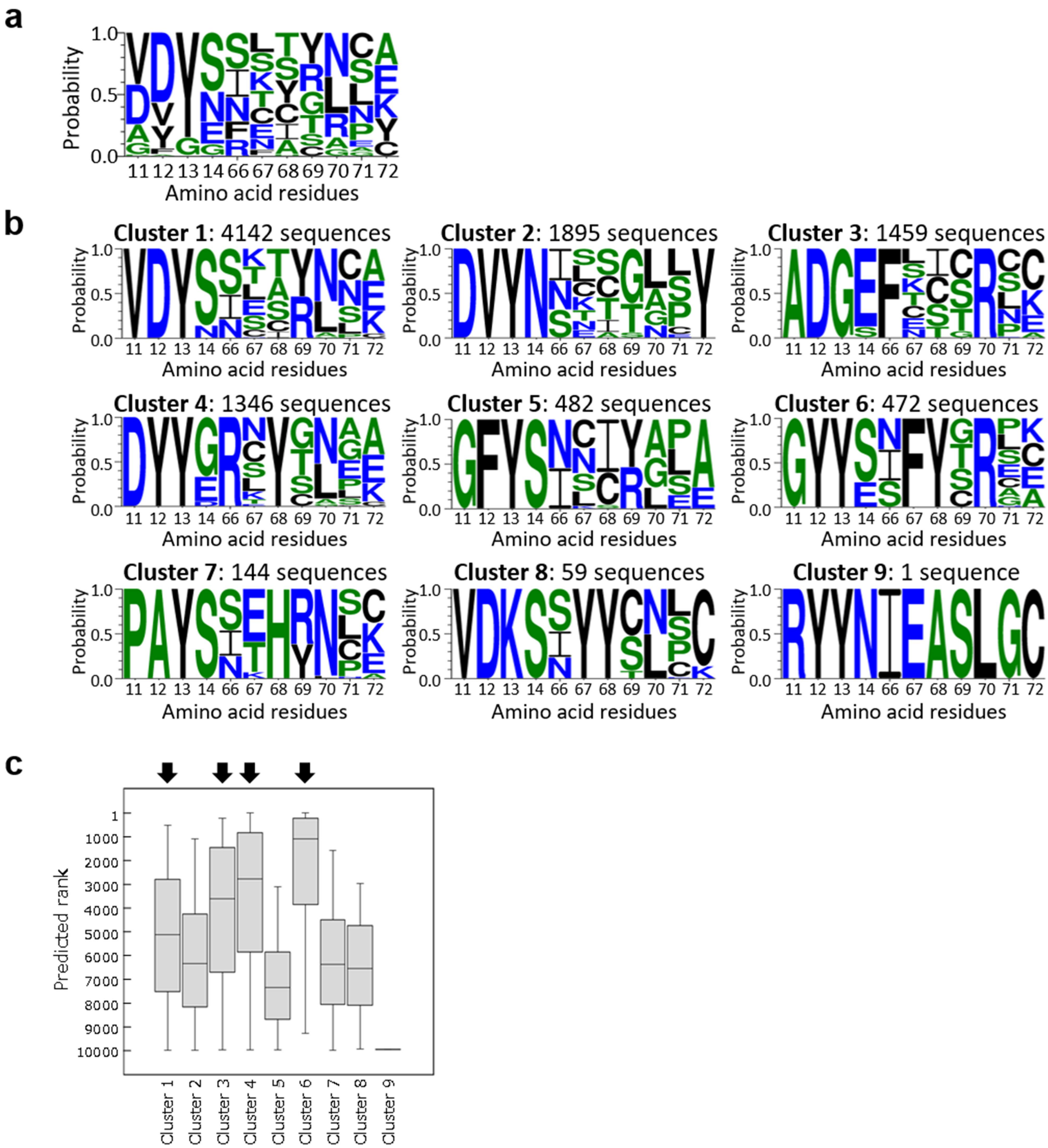


Figure 4. Amino acid frequencies and rank distribution of the sequences predicted by machine learning. (a) Amino acid frequencies of top 10,000 sequences predicted by machine learning, visualized by WebLogo.[41] (b) Amino acid frequencies of clustered sequences. (c) Rank distribution of each cluster. Black arrows indicate clusters containing the top 1,000 sequences.

By using the trained model, we ranked all the variants in the prediction space except for the variants used as training data by the probability-of-improvement score (Data Table S1). A variety of amino acids were proposed in the top 10,000 sequences at each position, except for the thirteenth position where tyrosine and glycine dominated (Figure 4a). The comparison of the top 1,000 predicted sequences with the training data by blastp (with the e-value threshold of 0.1) showed that 375 predicted sequences resembled 13 sequences in the training data: 62% of the top 1,000 predicted sequences was not comparable to the training data. To characterize the highly scored sequences, we used blastp to compare the top 10,000 sequences (Figure 4a) and identified nine clusters by Cytoscape. Intriguingly, each of the clusters had nearly unique amino acid sequences in the loop of N11–N14, showing their distinct sequence patterns (Figure 4b). The prediction ranks of the sequences in the clusters were not evenly scattered (Figure 4c), but each cluster had a distribution with different averages. Since the clusters 1, 3, 4, and 6 contained the top 1,000 sequences, we decided to design machine-learning-guided libraries based on these clusters.

Construction of a machine-learning-guided mutagenesis library

We constructed the phage libraries with the 2u2f variants in clusters 1, 3, 4, and 6. We designed degenerate codons to reflect the amino acid frequency at each position in each cluster. Amino acids with less than 5% appearance frequency were eliminated (Figure S5), but the constraints of degenerate codon design led to the appearance of some amino acids absent in the clusters. Consequently, each sequence space created by the amino acids appearing in the designed library (~10^5) was more than 100 times the size of that of the amino acids appearing in the corresponding clusters (~10^3). Therefore, phage libraries with a number of variants (> ~10^8) sufficient to cover the sequence spaces of the designed library were prepared (Table S3).

Each library was biopanned against galectin-3 with three rounds. The number of phages recovered from each round was relatively large in the first round (Table S4, Figure S6). We considered that this result showed more enrichment of functional variants in the machine-learning-guided libraries than in the initial library. Subsequently, 88 clones were isolated in the third round. Some phages bound the target in the phage enzyme-linked immunosorbent assay (ELISA) (20 in cluster 1, 14 in cluster 3, 20 in cluster 4, and 9 in cluster 6). The 2u2f variants displayed on the phages were separately expressed in *E. coli* and named individually, e.g., variant 1A1 from cluster 1 was in well A1 of a 96 deep-well plate. On Blue Native PAGE, several variants showed bands with higher molecular weight than that of the monomer, indicating aggregation, but 12 of them (6 in cluster 1, 2 in cluster 3, and 4 in cluster 4) appeared to be monomeric (Figure. S7). These 12 variants were purified by immobilized metal ion affinity chromatography (IMAC) and size exclusion chromatography (SEC). The SEC confirmed that they were mainly in a monomeric form (Figure S8) and the fractions of monomers were collected and the purified monomers were used for the further experiments. In ELISA, four variants of 1E2, 1H2, 3B5, and 4H5 bound specifically to galectin-3, but not to NeutrAvidin, which was used for immobilizing galectin-3 on an ELISA microplate (Figure 5a, Table S5). Further, the four candidate variants showed little binding to other negative targets: streptavidin (pI = 5.5), lysozyme from chicken egg (pI = 11), bovine serum albumin (pI = 4.7), and receptor-binding domain of SARS-CoV-2 (pI = 8.9) (Figure S9). This result indicates that the binding of 2u2f variants was not driven by nonspecific charge interaction with negatively charged galectin-3 (pI = 9.4). The EC_{50} values were 93 nM for 1E2, 80 nM for 1H2, 277 nM for 3B5, and 201 nM for 4H5 (Figure 5b). The circular dichroism (CD) spectra of the four variants were similar to that of wild-type 2u2f (Figure 6), indicating similarity of the secondary structures. Thus, a machine learning model trained with deep sequencing information produced from a weakly enriched library, where no functional variants had been experimentally identified, resulted in the discovery of several functional and correctly folded variants with target selectivity. For a more quantitative binding analysis, we applied SPR measurements, but little binding response was observed for all the selected variants, potentially because the immobilized form of the target on the sensor chip interferes with the binding to the target.

The four variants with specific affinity to the target were not contained in the top 10,000 sequences predicted by machine learning because 2 ~ 4 residue positions in the variants had the amino acids that do not appear in the sequence space of each cluster predicted by machine learning. However, the variants where the amino acids at the positions are altered to those that appear in the identified clusters (left in Figure S5) were contained in the top 10,000, and they showed relatively high performance scores (Table S6, Figure S10). The five variants highly ranked among the variants resembling 4H5 (predicted rank: 1, 4, 9, 20, and 42, Figure S11a) could be prepared in monomeric form (Figure S11b), and they bound to the target, but less specific affinities than 4H5 were observed (Figure S11c). These results suggest that variants with less specific affinity might be predominant in the sequence space predicted by machine learning. Enlargement of the sequence space of the designed library by the constraints of degenerate codon design resulted in the discovery of the variant with specific affinity to the target. In the correlation plot between measured affinity strength and top % in the performance score ranking for all the variants in the sequence space of the designed libraries (Figure S12), the four variants did not show high performance score, but the affinity strength showed a positive correlation with the order of performance score. The machine learning approach with deep sequencing data may be influenced by the factor of nonspecific binding, but the scores of the variants that have specific binding to targets might show a correlation with affinity strength.

Discussion

Library size is critical for the probability of discovering a variant with high target affinity in a surface-display system.[3] Depending on the system, 10^{9-13} phenotypes can be prepared in a library, so a variant library can cover all possible variants with 7–10 residues being randomized. The use of degenerate codons and trinucleotide cassettes in genetic library preparation enables us to decrease the number of amino acids at mutagenized positions,[4,9] which limits library size expansion by excluding undesirable variants.

These advances increase the probability of obtaining a variant with high affinity for the target. However, the prepared library does not always cover the size of the surveyed sequence space, so the selected variant may not be the one with optimal affinity, or no variant may be selected at all. Machine learning has been applied to surface-display systems to predict the sequence–function landscape.[10,11] This approach can potentially discover the variant with optimal affinity in the surveyed sequence space. The round-to-round enrichment data from the biopanning of a phage pool displaying Fab fragments with a randomized CDR3 region in the heavy chain has been used to train an ensemble of neural network models.[24] The use of training data with 10-fold enrichment led to the identification of a target-specific variant with an EC_{50} of 0.29 nM, which was 1.7-fold stronger than the affinity of

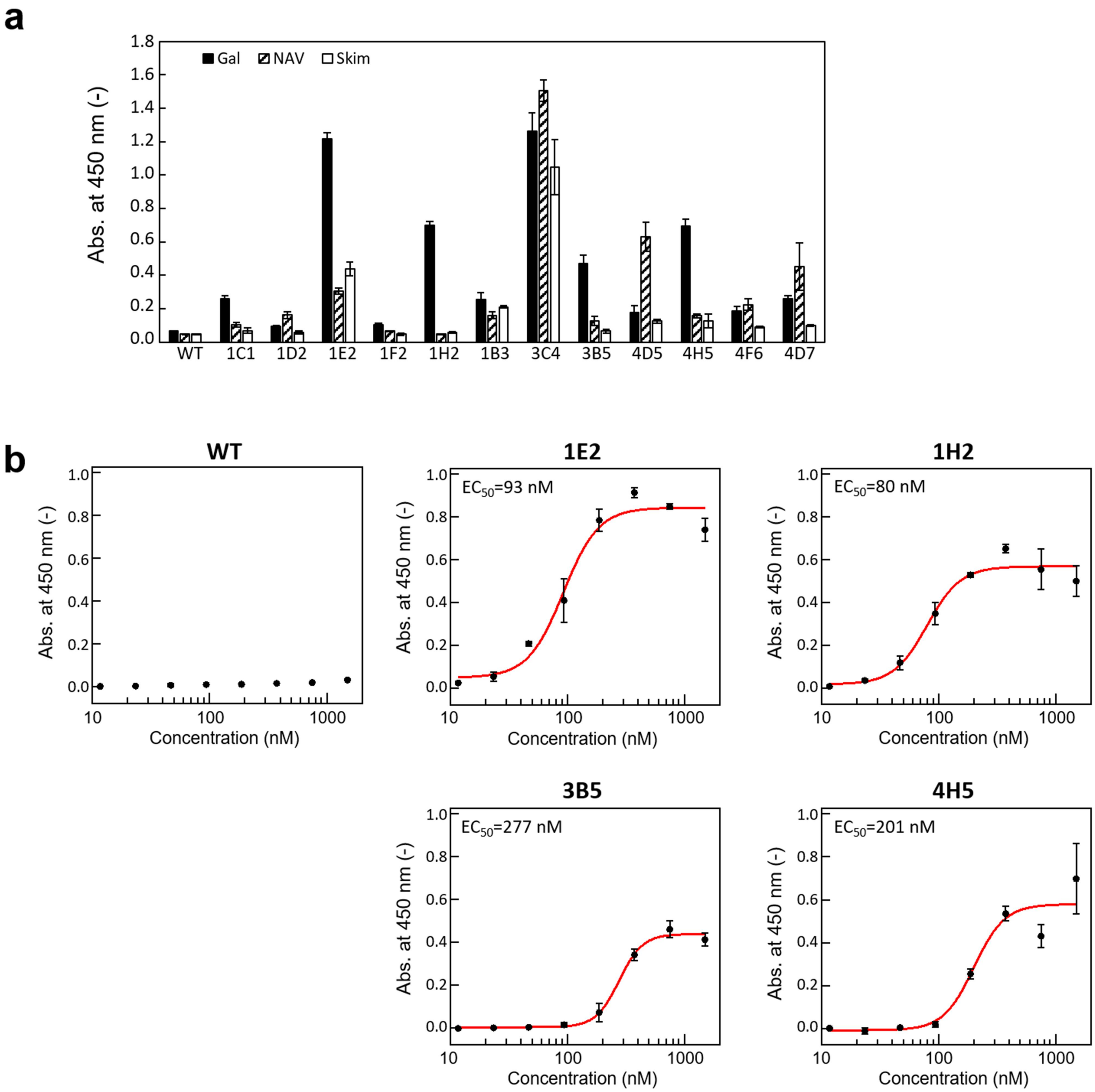


Figure 5. Binding function of wild-type 2u2f and obtained 2u2f variants. (a) Enzyme-linked immunosorbent assay of the candidate 2u2f variants after purification on galectin-3 (Gal), NeutrAvidin (NAV), or blocking buffer (Skim). (b) EC$_{50}$ values of wild-type 2u2f and four functional variants with affinity to galectin-3. The plots show the absorbance of galectin-3 minus that of NAV. The EC$_{50}$ values were determined by using Hill equation.

a variant obtained from deep sequencing analysis. A Fab fragment with low target affinity has been matured by using a long short-term memory network trained with the deep sequencing data of round-to-round comparison;[21] machine learning discovered variants with affinity 10 times that of a matured variant found experimentally by using only deep sequencing analysis.

Although an appropriately enriched library offers potential for discovering the optimal variant, in our study undesirable variants were amplified at the infection and amplification steps (Figure 3); consequently, no enrichment in the round-to-round comparison was observed (Figure S3) and no functional variants were identified (Figure S1). In this situation, the variants with binding function may be less efficiently enriched during selection than the undesirable variants that tend to be strongly enriched at the infection step. Here, we used the round-to-round comparison before and after the selection step to avoid the influence of infection. The enrichment observed in the comparisons between selection steps was so low that it was masked by that at the infection and amplification steps. Although machine learning with the training data based on the comparisons between selection steps did not show a strong correlation between affinity strength and performance score, it successfully led to the design of a library containing functional variants. Our deep sequencing analysis showed gradual enrichment of stop codons (Figure S2). The use of variant library excluding stop codons may supply the training data where the influence of undesirable bias is

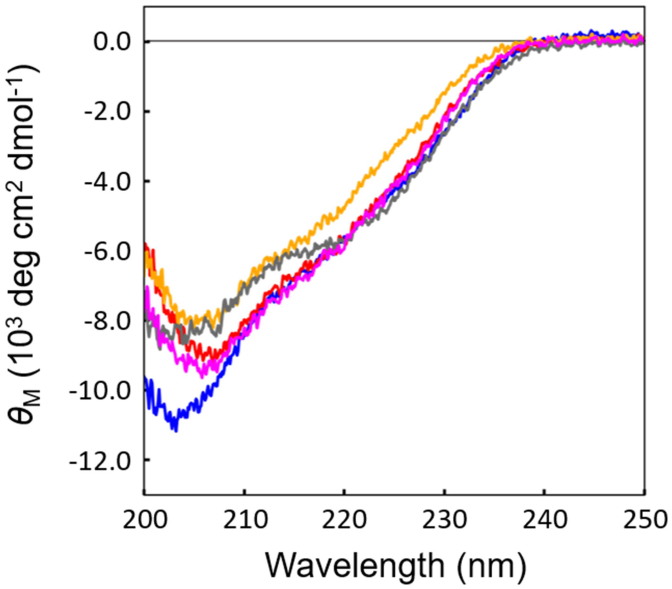


Figure 6. CD spectra of the functional 2u2f variants. Wild-type 2u2f is shown in blue, 1E2 in Orange, 1H2 in red, 3B5 in gray, and 4H5 in magenta.

decreased. The result that the 4H5-resembling variants highly ranked in the machine learning result showed less specific affinity than 4H5 might indicate that the factor of nonspecific affinity influences the prediction by machine learning. Although a complete removal method for nonspecific binding has not been reported, the use of an improved removal method should be considered.

The five variants (predicted rank: 1, 4, 9, 20, and 42, Figure S11a), which are highly ranked in the prediction space and resemble 4H5, bound to targets, but they showed less specific affinity. In contrast, the four variants (1E2, 1H2, 3B5, and 4H5), which were not contained in the clusters predicted by machine learning but were found in the sequence spaces expanded by degenerate codon design, bound to target and showed specific affinity. The library designed by degenerate codon design was needed to allow for more sequence variation in a focused sequence space. This suggests that the introduction of purposeful variation in the prediction of machine learning possibly better informs future library designs.

There are reports on machine learning-assisted directed evolution for discovering optimized or suboptimized variants in the affinity landscape.[10,42] In this study, the top 10,000 sequences predicted by our machine learning model produced nine clusters (Figure 4); as a result, four functional variants were identified from three of the clusters. This result suggests that machine learning is able to explore more than one local region containing functional fitness in a multi-peak landscape.[10,42]

In conclusion, we tried to generate functional variants from the sequence information from a series of weakly enriched libraries where few functional variants were experimentally identified. The selection and use of deep sequencing data from appropriate steps enabled us to extract the information on functional variants, so four functional variants with target-specific affinity (EC_{50} = 80–277 nM) were obtained from the libraries designed by machine learning. Despite the advances in surface-display systems, desirable variants are often not obtained. Our machine learning approach increases the possibility of obtaining a functional variant that may be matured to high-affinity variants.

Materials and methods

Preparation of galectin-3

E. coli BL21(DE3) cells harboring the plasmid coding biotin ligase BirA were transformed with pET22b vector that contains the gene coding Avi-tag and His_6-tag labeled galectin-3. Cells were grown overnight at 28°C on LB agar media containing 100 µg/ml of ampicillin and 34 µg/ml chloramphenicol. With five colonies grown on the plates, 50 ml of LB broth containing ampicillin and chloramphenicol was inoculated and cultured overnight at 28°C. Five mL of the culture was transferred to 500 mL of 2 × YT broth containing 100 µg/mL of ampicillin and 34 µg/ml chloramphenicol. Once the optical density of the culture reached OD_{600} = 0.8, IPTG and biotin were added to the flask to a final concentration of 1 mM and 50 µM, respectively. The cells were shaken at 160 rpm at 20°C overnight. The cells were harvested by centrifugation, resuspended in 50 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA (pH 8.0), and sonicated. The insoluble matter was removed by centrifugation. Variants were purified from the supernatants by IMAC (Ni Sepharose™ 6 Fast Flow; Cytiva, IL, USA) and SEC (HiLoad 26/600 Superdex 75 pg; Cytiva, IL, USA) (Figure S13). Purified Galectin-3 was dialyzed in phosphate-buffered saline (PBS).

Biopanning with phage display

The biopanning procedure was described previously.[34] Briefly, N11–N14 and M66–K72 in 2u2f were randomized using degenerate codons reflecting an amino acid frequency of antibody CDRs[37] for training data. M13 phage libraries displaying 2u2f variants with a size of ~10^9 were prepared. Colony-forming units (5.0 × 10^{11}) from an M13 phage library displaying 2u2f variants were exposed to magnetic beads (Dynabeads MyOne Streptavidin T1 or C1; Thermo Fisher Scientific, MA, USA) for 60 min at room temperature (Negative selection in Figure 2). For target preparation, 2 µM galectin-3 in PBS was incubated with magnetic beads for 60 min at room temperature such that the amount of targets on beads was 9 µg, which was calculated from the amount of supernatants measured by means of BCA assay using bovine serum albumin as a standard (Pierce™ BCA Protein Assay Kit; Thermo Fisher Scientific, MA, USA). The supernatant containing unbound phages was collected and incubated with galectin-3–immobilized magnetic beads for 60 min at room temperature. The beads were washed 10 times with PBS with 0.05% Tween-20 for 5 min each wash. Bound phages were eluted with 100 µL of triethylamine and neutralized with 300 µL of 1 M Tris–HCl (pH 6.8). Log-phase *E. coli* TG-1 cells were incubated overnight at 37°C with 200 µL of the eluted phages in 2× YT agar medium containing 100 µg/mL ampicillin and 1% (w/v) glucose. Cells grown on the plates were used to prepare phage particles for the next round. For the training data, phage pools were collected at each step (eluted phages, infected *E. coli*, and amplified phages) for deep sequencing analysis.

For machine-learning-guided libraries, N11–N14 and M66–K72 were randomized using degenerate codons reflecting an amino acid frequency in each machine learning proposed cluster. M13 phage libraries displaying 2u2f variants with a size of ~10^8 were prepared and biopanned as described above. The targets were immobilized on magnetic beads at the same amount as be applied for the first biopanning.

Sample preparation and deep sequencing

After each round of biopanning, polyclonal plasmid DNAs were prepared by using a GenElute Plasmid Miniprep Kit (PLN350; Sigma Aldrich, MO, USA) for sub-libraries of infected *E. coli* and then extracted with phenol–chloroform from sub-libraries of eluted and amplified phages. The extracted plasmids were used for the first polymerase chain reaction (PCR) to amplify 2u2f library fragments with the primers containing an annealing region for the second PCR primers. The PCR products were purified by using 1.5% agarose gel and a Qiaex II Gel Extraction Kit (20051; Qiagen, Hilden, Germany) and subjected to the second PCR to attach adapter sequences containing TruSeq DNA CD Indexes. The resulting fragments were purified as above, quantified by using a Qbit^{TM} 1× dsDNA HS Assay Kit (Q33231; Thermo Fisher Scientific) and pooled in equal amounts. The quality of the libraries was checked by using an Agilent 2100 Bioanalyzer (G2939B; Agilent Technologies, CA, USA). The prepared sample was sequenced on a MiSeq platform (Illumina, CA, USA) by using a MiSeq Reagent Kit v3 (15043895; Illumina) with 2 × 300 bp paired-end reads.

Evaluation of performance scores

The performance score of each 2u2f variant was evaluated on the basis of its round-to-round enrichment (sub-libraries of eluted phages in the second and third rounds vs. those of amplified phages in the first and second rounds, respectively). The enrichment ratio of a variant i (ER_i) was defined as

$$\mathrm{ER}_i = \log_2 \frac{\mathrm{Freq}_{\mathrm{Elutedphage\_2nd}}(i)}{\mathrm{Freq}_{\mathrm{Amplifiedphage\_1st}}(i)} + \log_2 \frac{\mathrm{Freq}_{\mathrm{Elutedphage\_3rd}}(i)}{\mathrm{Freq}_{\mathrm{Amplifiedphage\_2nd}}(i)} \quad (1)$$

where $\mathrm{Freq}_j(i)$ is the frequency of the variant i in the sub-library j. By using the ER, the performance score was defined as:

$$\mathrm{Score}_i = a \times \mathrm{ReLU}(\mathrm{ER}_i) \quad (2)$$

where ReLU $(\cdot)$ is a rectified linear unit and a is a normalizing constant to make the highest value in each sub-library equal to one.

Machine learning model and clustering analysis

We used a machine learning method based on COMBO, a fast implementation of Bayesian optimization, as described previously.[8,11] We defined the feature vector of a 2u2f variant by concatenating the precomputed feature vectors of amino acids at the 11 mutated sites. For the feature vector of each amino acid, we tested a variety of amino acid descriptors based on physicochemical properties or structural topology and found that the Aromaphilicity index[40] achieves the best accuracy for our problem by benchmark experiments (Figure S14). The dimensionality of the Aromaphilicity descriptor is one per a residue. Thus, the number of features used in our final model was 11 (i.e., 1 dimension × 11 mutated residues). By using the trained model, all the variants in the prediction space except for the variants used as training data were ranked by the probability-of-improvement score as described in the previous report.[11] The top 10,000 sequences were compared with each other by all-versus-all blastp search with the e-value threshold of 0.1. We constructed the similarity network where two sequences have edges when one sequence was listed in the other sequence's blastp search result. Sequence clusters were identified as connected components in this similarity network. The connected component extraction (cluster identification) was performed by Cytoscape 3.7.2.

Enzyme-linked immunosorbent assay (ELISA)

ELISA was performed as described by Ito et al.[34] Target-bound phages were detected with horseradish peroxidase–conjugated mouse anti-M13 monoclonal antibody (1:1000; sc-53,004, Santa Cruz Biotechnology, TX, USA). Purified proteins were detected with horseradish peroxidase–conjugated mouse anti-FLAG monoclonal antibody (1:10,000; A8592, Sigma Aldrich). The EC_{50} values were determined by using Hill equation. In the case of measuring target specificity shown in Figure S9, 50 µL of 4 µg/mL galectin-3, streptavidin, lysozyme, bovine serum albumin, and receptor-binding domain of SARS-CoV-2 were incubated in the wells of a 96-well polystyrene ELISA microplate before skim milk blocking and addition of 2u2f variants.

Small-scale protein expression

The gene fragments of the 2u2f variants selected from the machine-learning-guided libraries were amplified by PCR, and the products were ligated into the pET22b vector. Each plasmid was transformed into *E. coli* BL21(DE3), and the cells were incubated overnight at 28°C on LB agar plates (100 µg/mL ampicillin). The colonies were randomly transferred to 1 mL of LB broth (100 µg/mL ampicillin) in deep-well plates (Axygen, CA, USA) and incubated overnight. Harvested culture (100 µL) was inoculated into 900 µL of 2× YT broth (100 µg/mL ampicillin) in deep-well plates and incubated at 28°C with shaking. Isopropyl-β-D-thiogalactopyranoside was added at a final concentration of 1 mM, and the cells were incubated for 6 h. The cultures were centrifuged, the harvested cells were resuspended in 150 µL of PBS, sonicated, and centrifuged to remove insoluble matter.

Blue native PAGE

Cell lysates prepared by sonication and a NativePAGE^{TM} Sample Prep Kit (BN2008; Thermo Fisher Scientific) were used. Electrophoresis was performed in an SDS-free gel (HON-150-13; Oriental Instruments Co., Ltd., Japan) by using NativePAGE^{TM} Running Buffer (BN2001; Thermo Fisher Scientific) for 60 min at 150 V. Cathode buffer contained 0.02%

Coomassie Brilliant Blue G-250 (dark blue buffer) for the first 15 min and 0.002% Coomassie Brilliant Blue G-250 (light blue buffer) for the following 45 min. Proteins were transferred to PVDF membranes at 20 V for 7 min in an iBlot 2 Blotting System (Thermo Fisher Scientific). Membranes were blocked with 5% skim milk and incubated with horseradish peroxidase–conjugated mouse anti-FLAG tag monoclonal antibody (1:15,000; A8592, Sigma Aldrich) in PBS with 0.05% Tween-20 for 30 min at room temperature. Chemiluminescence was detected with a LAS 4000 instrument (Cytiva, IL, USA).

Preparation of candidate proteins

E. coli BL21(DE3) cells were transformed with the pET22b vectors carrying fragments encoding 2u2f variants, grown overnight at 28°C on LB agar, and then cultured in 2× YT broth; both media contained 100 µg/mL ampicillin. Isopropyl-β-D-thiogalactopyranoside was added to a final concentration of 1 mM at $OD_{600} = 0.8$, and the cells were shaken at 160 rpm for 6 h at 28°C. The purification of the 2u2f variants from the cell was the same as galectin-3.

Circular dichroism spectra

CD spectra were measured with a J-820 CD spectrometer (Jasco, Japan) in a 1.0-mm-long quartz cuvette, as follows: band width 1.0 nm, resolution 0.1 nm, response 8 s, scan speed 2 nm/min. The concentrations of purified 2u2f variants were 10 µM.

Abbreviations

CDR: Complementarity-determining region
CD: Circular dichroism
E. coli: Escherichia coli
ELISA: Enzyme-linked immunosorbent assay
ER: Enrichment ratio
Ig: Immunoglobulin
IMAC: Immobilized metal ion affinity chromatography
PCR: Polymerase chain reaction
SEC: Size exclusion chromatography

Acknowledgments

We thank Ms Miho Hosoya, Hiromi Ogata, and Yuri Ishigaki for experimental support. This work was partly supported by a Scientific Research Grant from the Japan Society for the Promotion of Science research fellowship (M.U., JP20H00315); by the project “Development of the Key Technologies for the Next-Generation Artificial Intelligence/Robots” of the New Energy and Industrial Technology Development Organization (M.U.); and by Support for Pioneering Research Initiated by the Next Generation (SPRING) from the Japan Science and Technology Agency (T.I., JPMJSP2114). The computations were performed using supercomputers in AI Bridging Cloud Infrastructure (ABCI), National Institute of Genetics (NIG), Academic Center for Computing and Media Studies (ACCMS, Kyoto University), Human Genome Center (SHIROKANE; The University of Tokyo), and Research Institute for Information Technology (RIIT, Kyushu University).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Scientific Research Grant from the Japan Society for the Promotion of Science research fellowship [JP20H00315]; the project “Development of the Key Technologies for the Next-Generation Artificial Intelligence/Robots” of the New Energy and Industrial Technology Development Organization; Pioneering Research Initiated by the Next Generation (SPRING) from the Japan Science and Technology Agency [JPMJSP2114].

ORCID

Tomoshi Kameda http://orcid.org/0000-0001-9508-5366
Mitsuo Umetsu http://orcid.org/0000-0003-4390-0263

Author contributions

M.U. and T.K. conceived of the study and directed the project; T.I., T.D. N., H.Nishi, Y.S. and M.U. developed the methodology; T.I., T.D.N., S. K. and H.Nakazawa conducted an investigation process; M.U., Y.S., T. K. and K.T. designed the experimental strategy and supervised analysis; T. I., T.D.N. and Y.K. visualized the experimental data; T.I., M.U. and H. Nakazawa wrote the original draft of the manuscript; T.D.N., Y.S., T. K. and K.T. reviewed and edited the manuscript.

References

1. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–D9. doi:10.1093/nar/gkaa1100.
2. Packer MS, Liu DR. Methods for the directed evolution of proteins. Nat Rev Genet. 2015;16(7):379–94. doi:10.1038/nrg3927.
3. Wang YJ, Xue P, Cao MF, Yu TH, Lane ST, Zhao HM. Directed evolution: methodologies and applications. Chem Rev. 2021;121(20):12384–444. doi:10.1021/acs.chemrev.1c00260.
4. Qi H, Ma ML, Lai DY, Tao SC. Phage display: an ideal platform for coupling protein to nucleic acid. Acta Biochim Biophys Sin (Shanghai). 2021;53(4):389–99. doi:10.1093/abbs/gmab006.
5. Linciano S, Pluda S, Bacchin A, Angelini A. Molecular evolution of peptides by yeast surface display technology. Medchemcomm. 2019;10(9):1569–80. doi:10.1039/C9MD00252A.
6. Contreras-Llano LE, Tan CM. High-throughput screening of biomolecules using cell-free gene expression systems. Synth Bio. 2018;3:1. doi:10.1093/synbio/ysy012.
7. Kamalinia G, Grindel BJ, Takahashi TT, Millward SW, Roberts RW. Directing evolution of novel ligands by mRNA display. Chem Soc Rev. 2021;50(16):9055–103. doi:10.1039/d1cs00160d.
8. Yan JR, Li GH, Hu YH, Ou WJ, Wan YK, Gajewski T, Wang Y, Wongchenko M, Choong N, Ribas A. Construction of a synthetic phage-displayed Nanobody library with CDR3 regions randomized by trinucleotide cassettes for diagnostic applications. J Transl Med. 2014;12(S1):12. doi:10.1186/1479-5876-12-S1-O12.
9. Saito Y, Oikawa M, Sato T, Nakazawa H, Ito T, Kameda T, Tsuda K, Umetsu M. Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration. ACS Catal. 2021;11(23):14615–24. doi:10.1021/acscatal.1c03753.
10. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods. 2019;16(8):687–94. doi:10.1038/s41592-019-0496-6.
11. Saito Y, Oikawa M, Nakazawa H, Niide T, Kameda T, Tsuda K. Machine-Learning-Guided UM. Mutagenesis for directed evolution of fluorescent proteins. ACS Synth Biol. 2018;7(9):2014–22. doi:10.1021/acssynbio.8b00155.

12. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16(12):1315–22. doi:10.1038/s41592-019-0598-1.
13. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with data-efficient deep learning. Nat Methods. 2021;18(4):389–96. doi:10.1038/s41592-021-01100-y.
14. Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J. Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol. 2007. 7.
15. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C, Tam S, Muley S, et al. Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol. 2007;25(3):338–44. doi:10.1038/nbt1286.
16. Cadet F, Fontaine N, Li GY, Sanchis J, Chong MNF, Pandjaitan R, Vetrivel I, Offmann B, Reetz MT. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. Sci Rep. 2018. 8.
17. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences. 2019;116(18):8852–58.
18. Giguere S, Laviolette F, Marchand M, Tremblay D, Moineau S, Liang XX, Biron E, Corbeil J, Kim PM. Machine learning assisted design of highly active peptides for drug discovery. PLoS Comp Biol. 2015;11:4. doi:10.1371/journal.pcbi.1004074.
19. Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V, Arnold FH. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat Methods. 2019;16(11):1176–84. doi:10.1038/s41592-019-0583-8.
20. Yoo DK, Lee SR, Jung Y, Han H, Lee HK, Han J, Kim S, Chae J, Ryu T, Chung J. Machine learning-guided prediction of antigen-reactive in silico clonotypes based on changes in clonal abundance through bio-panning. Biomolecules. 2020;10:3. doi:10.3390/biom10030421.
21. Saka K, Kakuzaki T, Metsugi S, Kashiwagi D, Yoshida K, Wada M, Tsunoda H, Teramoto R. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. Sci Rep. 2021;11(1):1. doi:10.1038/s41598-021-85274-7.
22. Parkinson J, Hard R, Ainsworth RI, Li N, Wang W. Engineering a histone reader protein by combining directed evolution, sequencing, and neural network based ordinal regression. J Chem Inf Model. 2020;60(8):3992–4004. doi:10.1021/acs.jcim.0c00441.
23. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, Ehling RA, Bonati L, Dahinden J, Gainza P, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng. 2021;5(6):600–12. doi:10.1038/s41551-021-00699-9.
24. Liu G, Zeng H, Mueller J, Carter B, Wang ZH, Schilz J, Horny G, Birnbaum ME, Ewert S, Gifford D. Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics. 2020;36(7):2126–33. doi:10.1093/bioinformatics/btz895.
25. Thomas WD, Golomb M, Smith GP. Corruption of phage display libraries by target-unrelated clones: diagnosis and countermeasures. Anal Biochem. 2010;407(2):237–40. doi:10.1016/j.ab.2010.07.037.
26. Menendez A, Scott JK. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. Anal Biochem. 2005;336(2):145–57. doi:10.1016/j.ab.2004.09.048.
27. Gebauer M, Skerra A. Engineering of binding functions into proteins. Curr Opin Biotechnol. 2019;60:230–41. doi:10.1016/j.copbio.2019.05.007.
28. Vazquez-Lombardi R, Phan TG, Zimmermann C, Lowe D, Jermutus L, Christ D. Challenges and opportunities for non-antibody scaffold drugs. Drug Discov Today. 2015;20(10):1271–83. doi:10.1016/j.drudis.2015.09.004.
29. Yu XW, Yang YP, Dikici E, Deo SK, Daunert S. Beyond antibodies as binding partners: the role of antibody mimetics in bioanalysis. Annu Rev Anal Chem. 2017;10(1):293–320. doi:10.1146/annurev-anchem-061516-045205.
30. Skrlec K, Strukelj B, Berlec A. Non-immunoglobulin scaffolds: a focus on their targets. Trends Biotechnol. 2015;33(7):408–18. doi:10.1016/j.tibtech.2015.03.012.
31. Weidle UH, Auer J, Brinkmann U, Georges G, Tiefenthaler G. The emerging role of new protein scaffold-based agents for treatment of cancer. Cancer Genom Proteom. 2013;10:155–68.
32. Brinkmann U, Kontermann RE. The making of bispecific antibodies. Mabs. 2017;9(2):182–212. doi:10.1080/19420862.2016.1268307.
33. Fujii H, Tanaka Y, Nakazawa H, Sugiyama A, Manabe N, Shinoda A, Shimizu N, Hattori T, Hosokawa K, Sujino T, et al. Compact seahorse-shaped T cell-activating antibody for cancer therapy. Adv Therap. 2018;1(3):3. doi:10.1002/adtp.201700031.
34. Ito T, Nishi H, Kameda T, Yoshida M, Fukazawa R, Kawada S, Nakazawa H, Umetsu M. Combination informatic and experimental approach for selecting scaffold proteins for development as antibody mimetics. Chem Lett. 2021;50(11):1867–71. doi:10.1246/cl.210443.
35. Girard A, Magnani JL. Clinical trials and applications of galectin antagonists. Trends in Glycoscience and Glycotechnology. 2018;31(172):Se211–Se20.
36. Dong R, Zhang M, Hu QY, Zheng S, Soh A, Zheng YJ, Yuan H. Galectin-3 as a novel biomarker for disease diagnosis and a target for therapy. Int J Mol Med. 2018;41(2):599–614. doi:10.3892/ijmm.2017.3311.
37. Kruziki MA, Bhatnagar S, Woldring DR, Duong VT, Hackel BJ. A 45-amino-acid scaffold mined from the PDB for High-Affinity Ligand Engineering. Chem Biol. 2015;22(7):946–56. doi:10.1016/j.chembiol.2015.06.012.
38. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. doi:10.1093/nar/25.17.3389.
39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. doi:10.1101/gr.1239303.
40. Hirano A, Kameda T. Aromaphilicity index of amino acids: molecular dynamics simulations of the protein binding affinity for carbon nanomaterials. Acs App Nano Mat. 2021;4(3):2486–95. doi:10.1021/acsanm.0c03047.
41. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90. doi:10.1101/gr.849004.
42. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol. 2009;10(12):866–76. doi:10.1038/nrm2805.