Data-efficient protein mutational effect prediction with weak supervision by molecular simulation and protein language models

Teppei Deguchi[1,2], Yoichi Kurumida[3], Shinji Iida[3], Kaito Kobayashi[2], Yutaka Saito[1,2,3*]

[1] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-0882, Japan.
[2]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.
[3]School of Frontier Engineering, Kitasato University, 1-15-1 Kitazato, Minami-ku, Sagamihara, Kanagawa 252-0373, Japan.

* Corresponding author: Yutaka Saito (saito.yutaka@kitasato-u.ac.jp)

Abstract

Machine learning-based protein mutational effect prediction is widely used in protein engineering and pathogenicity prediction, but training data scarcity remains a major challenge due to high costs of experimental measurements. A previous study proposed data augmentation using computational estimates by molecular simulation (1). However, this approach has been limited to predicting mutational effects on thermostability. Here, we present a new data augmentation method that combines molecular simulation with zero-shot prediction computed by protein language models. These computational estimates serve as "weak" training data to supplement experimental training data. Our method dynamically adjusts the weight and inclusion of weak training data based on available experimental training data. This reduces potential negative impacts of weak training data while extending applicability to diverse protein properties such as binding affinity and enzymatic activity. Benchmark tests demonstrate that our method improves prediction accuracy particularly when experimental training data are scarce. These results indicate the capability of our approach to advance protein engineering and pathogenicity prediction in small data regimes.

Introduction

Machine learning (ML)-based protein mutational effect prediction has been successfully applied to protein engineering and pathogenicity prediction (2-7). The task takes a mutant sequence as input and predicts its mutational effect, defined as the change of protein property of interest such as thermostability, binding affinity, and enzymatic activity compared to the wild type. ML models are trained using mutant sequences and their experimentally measured mutational effects (8-10).

To make accurate ML models, a large amount of training data is desired while it is often unavailable due to high costs of experimental measurements (9, 10). Therefore, data-efficient approaches are necessary to achieve accurate prediction with limited training data.

Augmentation using computationally generated data is an efficient way to virtually expand the amount of training data. In a previous study (1), molecular simulation was used as a means of data augmentation for thermostability prediction. Their method, mGPfusion, uses Rosetta molecular simulation software (11) to compute thermostability changes for all possible single-residue mutants. These computational estimates are integrated with experimentally measured thermostability to expand training data for ML models, thereby improving prediction accuracy. This method is regarded as data augmentation using computational estimates by molecular simulation, which could be a versatile approach for data-efficient mutational effect prediction. However, the previous study (1) was limited to thermostability prediction, and applications to other protein properties have not been explored.

More recently, protein language models (pLM) have been used to estimate the mutational effect based on the likelihood ratio of mutant and wild-type sequences (12, 13). This method is called zero-shot prediction since it works without using experimentally measured training data. Mutational effects estimated by zero-shot prediction are reported to correlate with diverse protein properties (13) although they have not yet been applied for data augmentation.

Here, we present a new data augmentation method that combines molecular simulation and zero-shot prediction using pLM. These computational estimates serve as "weak" training data to supplement experimental training data. Unlike the previous study (1) limited to thermostability, we aim to improve the prediction accuracy of mutational effects on diverse protein properties. Using Rosetta and ESM-2 (12) as pLM, we design a hybrid score that combines both estimates to enhance accuracy and robustness.

A key challenge in using weak training data is that inaccurate computational estimates can sometimes degrade ML model performance. To mitigate this issue, we develop algorithms that dynamically adjust the weight of weak training data and decide whether or not to include them based on available experimental training data.

Our method successfully improves the prediction accuracy on various proteins and their properties including protein abundance, binding affinity and enzymatic activity, surpassing ML models trained sorely with experimental data. The accuracy improvement is especially remarkable when experimental training data are scarce. Furthermore, we demonstrate that our method is capable of predicting the effects of double-residue mutations using single-residue mutants as training data.

Results

Methods overview

In this study, we developed a new method for protein mutational effect prediction with data augmentation using computational estimates by molecular simulation and pLM's zero-shot prediction (Figure 1a). We employed Rosetta for molecular simulation and ESM-2 as pLM. Mutational effects estimated by Rosetta and ESM-2 are regarded as weak training data integrated with experimental training data. Using these integrated training data, the ML model is constructed to predict the mutational effect from a mutant sequence.

Computational estimates of mutational effects involve errors in estimation as well as differences in scales and biases compared to experimental data. Thus, directly integrating them may cause a decrease in the prediction accuracy of resultant ML models. To prevent this, we developed the weight adjustment algorithm (Figure 1b) and the decision algorithm for data inclusion (Figure 1c). In the weight adjustment algorithm, the weight of computational estimates is reduced when the amount of experimental training data is large relative to the sequence length (Methods). In the decision algorithm, a subset of available experimental training data is used as validation data, and computational estimates are included only if their addition improves prediction accuracy (Methods).

To obtain computational estimates by Rosetta, we used different protocols depending on protein properties of interest (Methods). For protein abundance, the change in folding free energy upon mutations was calculated and converted to the protein abundance estimate using the biophysics-based equations. For binding affinity, the change in binding free energy with the ligand upon mutations was calculated and, together with the folding free energy, similarly converted to the binding affinity estimate. For enzymatic activity, due to the difficulty in simulating enzymatic reactions by Rosetta, we employed the folding free energy change upon mutations as used for protein abundance. To obtain computational estimates by ESM-2, we consistently used the log-likelihood ratio of mutant and wild-type sequences (Methods). The computational estimates by Rosetta and ESM-2 were further combined to calculate the hybrid score (Methods) that was used as weak training data in our method.

Our ML model takes as input a mutant sequence, calculates its feature values using ESM-2 (here used for embedding rather than zero-shot prediction), and feeds them into a downstream regression model to predict the mutational effect. The regression model was selected from support vector machine, random forest, gaussian process, and linear model for each dataset. We also tested models using amino acid descriptors for embedding instead of ESM-2-based embedding.

We evaluated our method on the datasets of experimentally measured mutational effects, which consist of various proteins and their properties including protein abundance, binding affinity, and enzymatic activity (Table 1). The datasets contain both single-residue and double-residue mutants. For each dataset, we first selected the regression model and its hyperparameters by nested cross validation on the whole experimental data (Supplementary Table 1). Then, the prediction accuracy on test data

was compared before and after adding weak training data under the same model configuration to evaluate the impact of weak training data. We carefully devised the evaluation procedure so that the training process does not look at test data (Methods; Supplementary Figure 3). As the baseline, we also evaluated the prediction accuracy of Rosetta or ESM-2's zero-shot prediction alone.

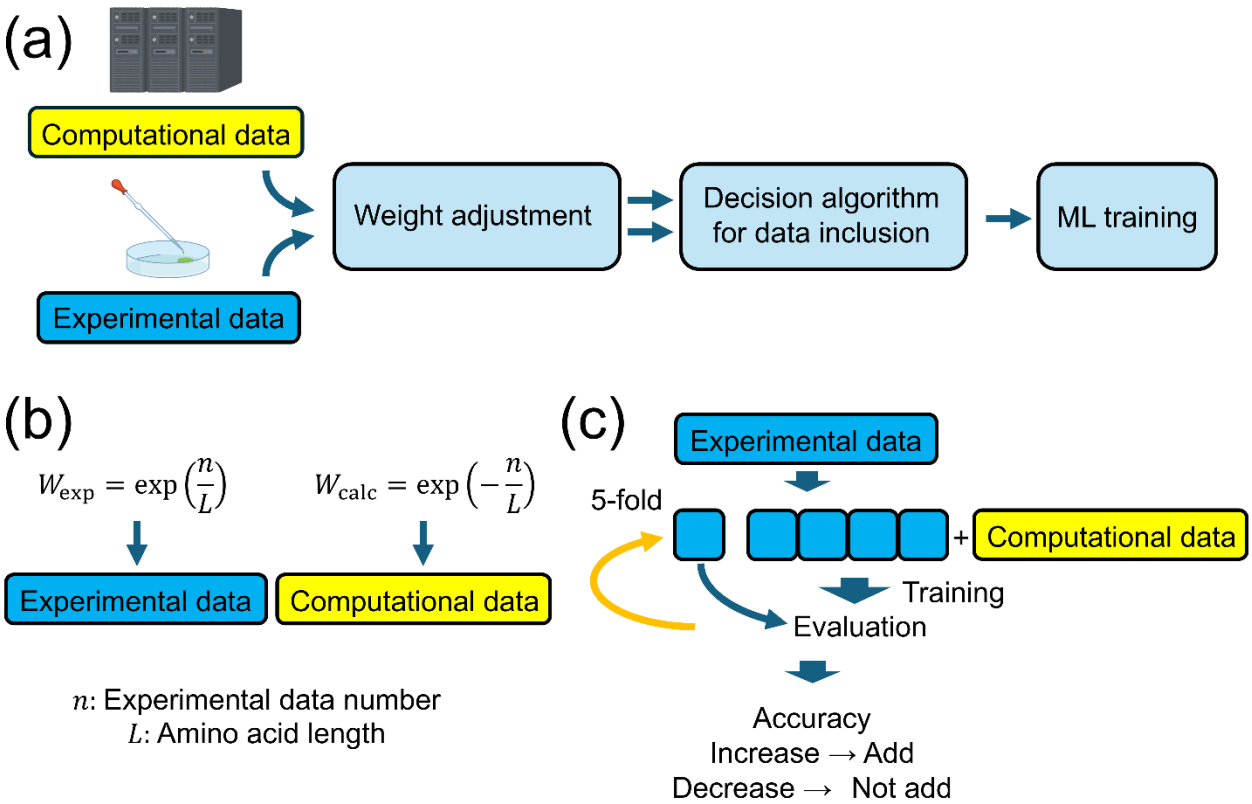

Figure 1. (a) Overview of the data-efficient protein mutational effect prediction method. Mutational effects are estimated by molecular simulation and pLM's zero-shot prediction. These computational estimates are used as weak training data integrated with experimental training data for ML models. (b) The weight of weak training data relative to experimental training data are determined by the weight adjustment algorithm. (c) The decision algorithm determines whether or not to include weak training data based on cross validation.

Table 1. Benchmark dataset.

Protein	**Sequence length**	**Ligand**	**Property**	**Single-residue mutants**	**Double-residue mutants**
GRB2-SH3 (14)	56	GAB2 (peptide)	Binding affinity	654	33442
			Protein abundance	734	62333
PSD95-PDZ (14)	84	CRIPT (peptide)	Binding affinity	1440	5697
			Protein abundance	1262	6811
GB1 (14)	55	IgG-Fc (protein)	Binding affinity	1045	495532
APH(3')II (15)	255	Kanamysin (compound)	Enzymatic activity	4845	N/A
CYP2C9 (16)	465	Warferin (compound)	Enzymatic activity	6524	N/A

Weak supervision improves prediction accuracy in data-scarce conditions

We first evaluated the impact of weak supervision using the single-residue mutant dataset for each protein (Figure 2). In most protein properties tested, incorporating weak training data improved the prediction accuracy compared to using experimental training data alone (Figure 2a-f). The improvement was more pronounced when the amount of experimental training data was small (e.g. <200). With the increase of experimental training data, the augmentation provided little additional benefit. When very few experimental training data were available (e.g. <50), predictions based solely on computational estimates (Rosetta or ESM-2) outperformed ML models trained with weak supervision. However, as the amount of experimental data increased, our method outperformed these computational estimates alone.

We also evaluated ML models using amino acid descriptors for embedding instead of ESM-2-based embedding (Supplementary Figure 1). As in the case of ESM-2-based embedding, the incorporation of weak training data improved prediction accuracy compared to the ML models trained solely with experimental training data. Our method also exceeded the computational estimates alone with the increase of experimental training data, with few exceptions of CYP2C9 enzymatic activity and GRB2-SH3 binding affinity.

These results demonstrate that our method with weak supervision enhances prediction accuracy in data-scarce conditions while adapting to the different availability of experimental training data.

Combining molecular simulation and pLM enables applicability to diverse protein properties

Molecular simulation based on biophysics and pLM pretrained on natural protein sequences may provide complementary estimates of mutational effects due to their different principles. To investigate this, we assessed prediction accuracy when using either Rosetta or ESM-2 alone as weak training data versus when combining both as the hybrid score ("Hybrid" in Figure 2). For GB1 binding affinity (Figure 2g), accuracy improved by adding Rosetta value, but not by ESM-2. In contrast, for CYP2C9 enzymatic activity (Figure 2b), accuracy increased the most when adding ESM-2 value. For the other datasets, the hybrid score achieved the highest accuracy or at least comparable accuracy to Rosetta or ESM-2 alone. (Figure 2a, c-f).

When using amino acid descriptors for embedding, adding Rosetta value increased accuracy for GB1 binding affinity (Supplementary Figure 1g). For the other datasets, the hybrid score improved prediction accuracy the most (Supplementary Figure 1a-f).

Overall, these results indicate that combining molecular simulation with pLM-based zero-shot prediction enhances the applicability of our method across diverse protein properties.

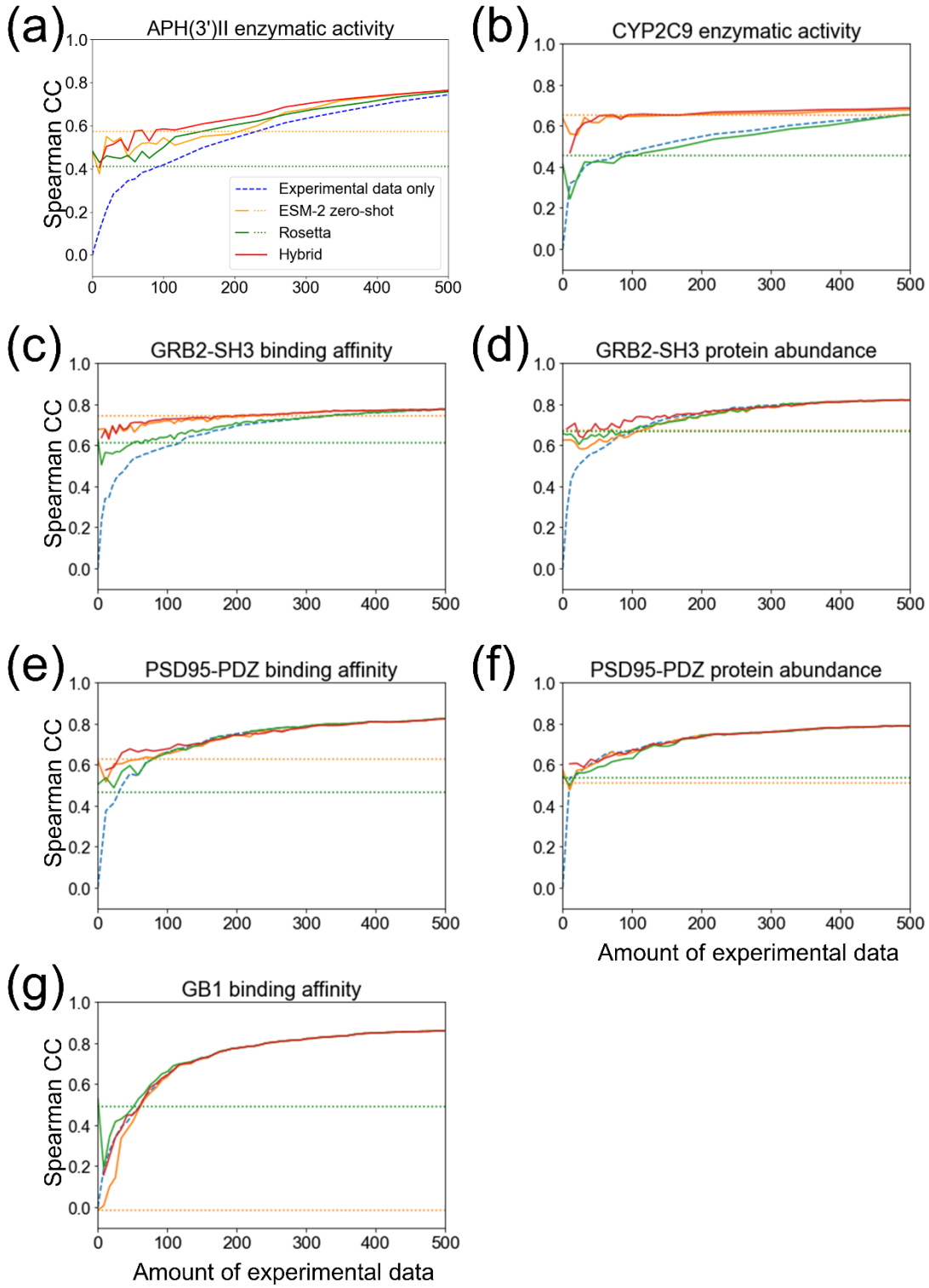

Figure 2. Impact of weak supervision with differing amounts of available experimental training data. Prediction accuracy measured by Spearman's correlation coefficient (CC) was compared between ML models trained using experimental data only (dashed line), those integrated with computational estimates by Rosetta, ESM-2, and their hybrid score (solid lines). As the baseline, prediction accuracy obtained by computational estimates alone was shown (dotted lines). esm2_t6_8M_UR50D was used

for embedding. Single-residue mutants were used for both training and test splits. (a) APH(3')II enzymatic activity. (b) CYP2C9 enzymatic activity. (c) GRB2-SH3 binding affinity (d) GRB2-SH3 protein abundance. (e) PSD95-PDZ binding affinity. (f) PSD95-PDZ protein abundance. (g) GB1 binding affinity.

The weight adjustment algorithm and the decision algorithm for data inclusion provide robust prediction accuracy

To assess the impact of the weight adjustment algorithm and the decision algorithm for data inclusion, we conducted a comparison between using and not using these algorithms. We tested the two representative datasets, APH(3')II enzymatic activity and PSD95-PDZ binding affinity (Figure 3). The weight adjustment algorithm substantially improved accuracy for APH(3')II enzymatic activity (Figure 3a). However, accuracy increase was not notable with the decision algorithm (Figure 3a). To the contrary, the decision algorithm increased accuracy for PSD95-PDZ binding affinity, but weight adjustment did not (Figure 3b). By combining the two methods, accuracy was successfully increased in both datasets (Figure 3a, b). These results suggest that these algorithms are important for robust prediction accuracy, preventing potential negative impacts of weak training data.

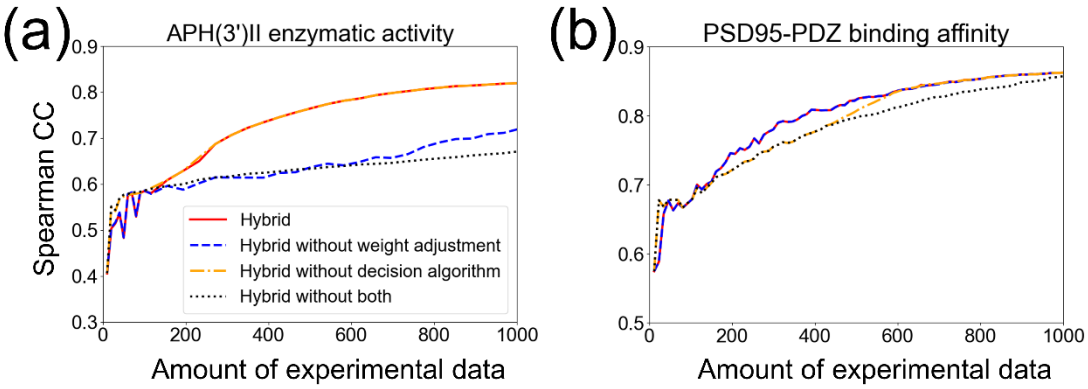

Figure 3. Ablation experiments for the weight adjustment algorithm and the decision algorithm for data inclusion. The prediction accuracy of ML model was compared among the hybrid score with both methods (Hybrid), without the weight adjustment algorithm, without the decision algorithm, and without both methods. esm2_t6_8M_UR50D was used for embedding. Single-residue mutants were used for both training and test splits. (a) APH(3')II enzymatic activity. (b) PSD95-PDZ binding affinity.

Predicting double-residue mutational effects using single-residue mutants as training data

In protein engineering, predicting the combinatorial effect of multiple-residue mutations is important to design functional proteins distant from the wild-type. Thus, we evaluated our method using double-residue mutants as test data while using single-residue mutants as training data. Remarkably, weak training data improved prediction accuracy in most of the protein properties compared to using experimental data alone (Figure 4). For PSD95-PDZ binding affinity and protein abundance (Figure 4c, d), the impact of weak supervision was relatively small while it still enhanced accuracy when experimental data was fewer than 50. For GB1 binding affinity (Figure 4e), accuracy improved the most by adding Rosetta value. For the other datasets, the hybrid score achieved the highest accuracy or at least comparable accuracy to that obtained by Rosetta or ESM-2. (Figure 4a-d).

We also evaluated the models using amino acid descriptors for embedding (Supplementary Figure 2). As with ESM-2-based embedding, incorporating weak training data improved prediction accuracy compared to using experimental training data alone (Supplementary Figure 2a-e). For the datasets other than GRB2-SH3 binding affinity and GB1, the hybrid score improved accuracy the most (Supplementary Figure 2a-d). For GRB2-SH3 binding affinity (Supplementary Figure 2a), accuracy increased the most when adding ESM-2 value. For GB1, accuracy increased the most by adding Rosetta value (Supplementary Figure 2e).

Overall, these results indicate that our method enhances prediction accuracy not only for single-residue mutational effects, but also double-residue mutations in data-scarce conditions.

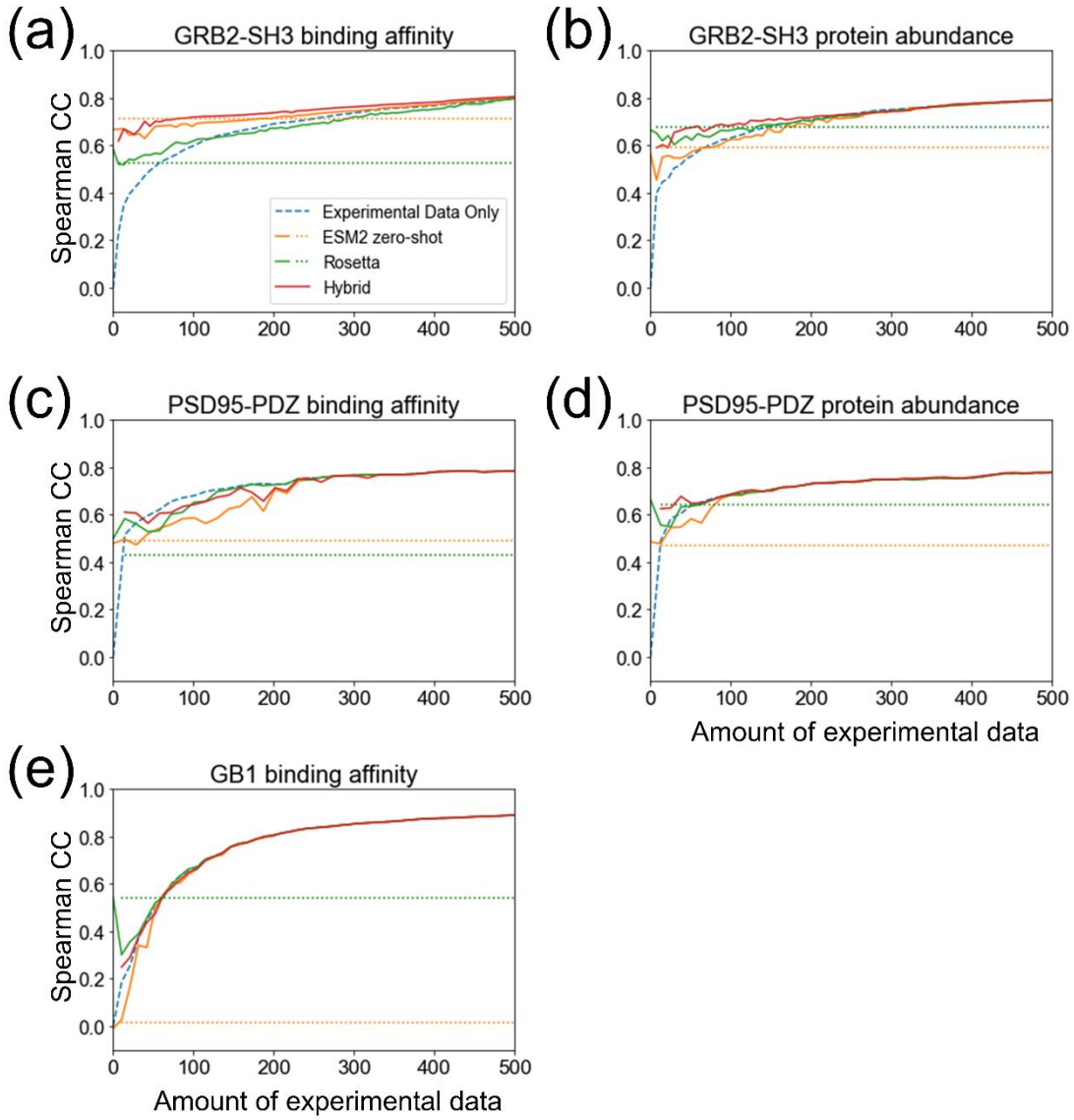

Figure 4. Prediction accuracy for double-residue mutational effects only using single-residue mutants as training data. ML models trained using experimental data only (dashed line), those integrated with computational estimates by Rosetta, ESM-2, and their hybrid score (solid lines) were compared. As the baseline, prediction accuracy obtained by computational estimates alone was shown (dotted lines). esm2_t6_8M_UR50D was used for amino acid embedding. (a) GRB2-SH3 binding affinity (b) GRB2-SH3 protein abundance. (c) PSD95-PDZ binding affinity. (d) PSD95-PDZ protein abundance. (e) GB1 binding affinity.

Discussion

In this study, we demonstrated that weak supervision using computational estimates by molecular simulation and pLMs can improve the accuracy of ML models for protein mutational effect prediction. We developed the weight adjustment algorithm and the decision algorithm for data inclusion, which successfully mitigated potential negative impacts caused by weak training data. Our results showed that integrating weak training data was particularly beneficial in data-scarce conditions where experimental data alone were insufficient to train accurate ML models. By combining Rosetta and ESM-2, our hybrid approach enhanced accuracy across diverse protein properties including protein abundance, binding affinity, and enzymatic activity.

Despite its effectiveness, weak supervision sometimes did not improve accuracy. This happened when computational estimates were quite inaccurate, even worse than ML models trained with small experimental data (e.g. Figure 2g, Figure 4e). We speculate possible reasons for inaccurate computational estimates by Rosetta and ESM-2 as follows. Rosetta-based estimates possibly become inaccurate for proteins with e.g. complex conformational changes that are not well-represented in molecular simulation. In addition, EMS-2's zero-shot prediction may not correlate with mutational effects for proteins with limited representation in its pre-training dataset.

Compared to previous studies, our approach extends the applicability of computationally augmented mutational effect prediction. A notable prior work, mGPfusion, leveraged Rosetta-based estimates for weak supervision but was limited to thermostability prediction. Our method generalizes this concept by incorporating pLM-based estimates, allowing for the prediction of a wider range of protein properties, including binding affinity and enzymatic activity. In addition, we newly introduced the weight adjustment algorithm and decision algorithm for data inclusion, preventing unreliable computational estimates from negatively impacting model performance.

Overall, this study highlights the potential of weak supervision to address data scarcity in mutational effect prediction. By combining molecular simulation and pLM's zero-shot prediction, we provide a widely applicable framework for improving ML-based protein engineering and pathogenicity prediction. Future work could explore additional computational estimation methods such as other AI-based methods (17, 18) or biophysics-based methods (19, 20). The ability to improve prediction accuracy without increasing experimental costs has significant benefits for accelerating protein design and understanding genotype-phenotype relationships.

Methods

Benchmark datasets

We collected benchmark datasets consisting of experimentally measured mutational effects on diverse protein properties, including protein abundance, binding affinity, and enzymatic activity (Table 1). The binding affinities of GRB2-SH3, PSD95-PDZ, and GB1, as well as the protein stabilities of GRB2-SH3 and PSD95-PDZ, were measured using yeast growth. The enzymatic activities of APH(3')II and

CYP2C9 were assessed based on drug resistance in *E. coli* and yeast, respectively. To systematically evaluate the impact of weak supervision under varying data availability, we selected datasets generated from high-throughput experiments, each containing at least 500 single-residue mutants per protein.

Mutational effect values were extracted from the original publications as follows. For the APH(3')II enzymatic activity dataset, we used measurements obtained under the 1:8 dilution condition of the kanamycin ligand (15). The original study reported *ΔMut* and *ΔAA*, representing the changes in observed mutant frequencies at each site and for each amino acid, respectively. We defined the mutational effect for each amino acid substitution at each site as log(*ΔMut*ΔAA*). For the CYP2C9 enzymatic activity dataset, we directly used the site- and amino acid-specific mutational effects reported in the original study (16). For the GRB2-SH3, PSD95-PDZ, and GB1 datasets, mutational effects on protein abundance and binding affinity were obtained from a previous study (14). In addition, this study also reported folding free energy and binding free energy changes upon mutation, which were computed using biophysics-based conversion equations. For single-residue mutants, we included only those with both experimentally measured mutational effects and corresponding free energy changes available. All double-residue mutants were retained without filtering.

Molecular simulation

We used Rosetta ddG monomer (21) to calculate the change in folding free energy upon mutations ($\Delta\Delta G_{\mathbf{f}}$). Rosetta Flex ddG (22) was used to calculate the change in binding free energy with the ligand upon mutations ($\Delta\Delta G_{\mathbf{b}}$). We used the tertiary structure of each protein from PDB (23) (PDB IDs, GRB2-SH3: 2VWF (24), PSD95-PDZ: 1BE9 (25), GB1: 1FCC (26), APH(3')II: 1ND4 (27), and CYP2C9: 1R9O (28)). For $\Delta\Delta G_{\mathbf{f}}$ calculations, ligands were removed from the structure. In the ddG monomer calculations, high-resolution backrub-based protocol 16 was used for GRB2-SH3, PSD95-PDZ, GB1, and APH(3')II, while the low-resolution side-chain flexibility based protocol 3 was applied to CYP2C9 (21).

To obtain the computational estimates of mutational effects from Rosetta-derived $\Delta\Delta G_{\mathbf{f}}$ and $\Delta\Delta G_{\mathbf{b}}$, we applied the biophysics-based conversion equations inspired by (14):

$$F_{\mathbf{f}} = \frac{1}{1+e^{x_{\mathbf{f}}}},\ F_{\mathbf{fb}} = \frac{1}{1+e^{x_{\mathbf{b}}}(1+e^{x_{\mathbf{f}}})},$$

$$x_{\mathbf{f}} = \frac{\Delta G_{\mathbf{f}}}{RT},\ x_{\mathbf{b}} = \frac{\Delta G_{\mathbf{b}}}{RT}.$$

$\Delta G_{\mathbf{f}}$ and $\Delta G_{\mathbf{b}}$ represent Gibbs free energies of folding and binding, respectively, and R is gas constant and T is the temperature in Kelvin. As proposed in (14), $F_{\mathbf{f}}$ indicates the probability that the protein is in the folded state while $F_{\mathbf{fb}}$ indicates the probability that the protein is both in the folded state and bound to the ligand. Therefore, we used $F_{\mathbf{f}}$ and $F_{\mathbf{fb}}$ as the computational estimates for protein abundance and binding affinity, respectively. For enzymatic activity, we used $F_{\mathbf{f}}$ as the computational estimates. This choice was motivated by the difficulty of simulating enzymatic reactions with Rosetta and the observation that folding stability was often strongly correlated with enzymatic activity.

Zero-shot prediction with pLM

We used ESM-2 to perform zero-shot prediction and obtain computational estimates of mutational effects (12, 13). Specifically, we employed the 3-billion parameter model (esm2_t36_3B_UR50D) and adopted the wild-type marginal probability scheme for scoring mutant sequences relative to the wild type.

Hybrid score

In addition to using Rosetta- and ESM-2-derived estimates independently, we also constructed a hybrid score by linearly combining the two values. This approach aims to improve robustness by integrating predictions derived from different principles: Rosetta being biophysics- and structure-based, and ESM-2 being deep learning- and sequence-based. The hybrid score was defined as follows:

$$\text{Hybrid} = a\text{Rosetta} + b\text{ESM-2} + \text{c}. \qquad (1)$$

The coefficients a, b, and c were optimized using linear regression between the hybrid score and experimental mutational effects in training data (Supplementary Figure 3). This hybrid score was then used as weak training data in the same manner as the individual estimates.

Scaling

Since the calculated values of $\Delta\Delta G_{\mathbf{f}}$, $\Delta\Delta G_{\mathbf{b}}$, $F_{\mathbf{f}}$, $F_{\mathbf{fb}}$, and ESM-2 values may differ in scales from their corresponding experimental values, we applied linear scaling using mutants in training data. For the GRB2-SH3, PSD95-PDZ, and GB1 datasets, $\Delta\Delta G_{\mathbf{f}}$ was scaled using the following equation:

$$\Delta\Delta G_{\mathbf{f_scaled}} = a\Delta\Delta G_{\mathbf{f}} + b. \qquad (2)$$

The coefficients a and b were determined by fitting $\Delta\Delta G_{\mathbf{f_scaled}}$ to folding free energy available in training data (Supplementary Figure 3). After scaling, the change in folding free energy ($\Delta\Delta G_{\mathbf{f}}$) was transformed into folding free energy ($\Delta G_{\mathbf{f}}$), which was used for the calculation of $F_{\mathbf{f}}$ and $F_{\mathbf{fb}}$. The same scaling procedure was applied to $\Delta\Delta G_{\mathbf{b}}$, $F_{\mathbf{f}}$, and $F_{\mathbf{fb}}$, and ESM-2 values. For the APH(3')II and CYP2C9 datasets, where folding free energy and binding free energy changes were not reported in the original studies, we performed empirical scaling. Specifically, $\Delta\Delta G_{\mathbf{f}}$ and $\Delta\Delta G_{\mathbf{b}}$ were normalized to have a mean of 0 and a standard deviation of 3.0.

Weight adjustment algorithm

To mitigate the potential negative impact of inaccuracies in computational estimates, we applied different weights to experimental training data and weak training data during model training. The weights were defined as follows:

$$W_{\mathbf{exp}} = \exp\left(\frac{n}{L}\right), W_{\mathbf{calc}} = \exp\left(-\frac{n}{L}\right),$$

where n is the number of experimental training data, and L is the number of residues in the protein sequence where mutations can occur. This formulation increases the relative importance of

experimental data as more experimental measurements become available while reducing the influence of computational estimates.

The rationale behind this approach is illustrated in Supplementary Figure 4. When experimental data are scarce, many regions in the sequence space are only covered by computational estimates (yellow area). However, as experimental data accumulate, the regions influenced by high-quality data (dashed circle) expand. Without proper reweighting, computational estimates may interfere with learning in these regions, leading to reduced accuracy. The weight adjustment algorithm addresses this by shifting learning emphasis away from computational estimates and toward experimental data in these regions, while still leveraging computational estimates in unexplored areas. This strategy balances model performance across the sequence space and contributes to overall prediction accuracy.

Decision algorithm for data inclusion

In addition to the weight adjustment algorithm, we developed a decision algorithm to determine whether computational estimates should be included in model training.

The available experimental data were split into five subsets for cross validation. In each fold, four subsets were used for training, and the remaining subset was used as validation data. Computational estimates were added to the training set, and a model was trained on the combined dataset. The prediction accuracy on the validation set was then compared to that obtained using only experimental data. If the inclusion of computational estimates reduced validation accuracy, the algorithm rejected the addition of computational estimates as weak training data (Figure 1c).

In cases where experimental data was too small and the validation accuracy resulted in NaN, the algorithm defaulted to including computational estimates. This ensures that weak supervision can still be applied in data-scarce settings, where experimental guidance is limited.

Machine learning

For each mutant, an amino acid sequence was embedded using either ESM-2 embedding (esm2_t6_8M_UR50D) or predefined amino acid descriptors. Letting L denotes the sequence length, the matrix of shape (L, 320) was generated by esm2_t6_8M_UR50D, which was flattened along the L-axis into a $320L$-dimensional vector. The descriptor was selected from among BLOSUM-based features (29), FASGAI (30), MS-WHIM (31), T-scale (32), ST-scale (33), Z-scale (34), VHSE (35), and ProtFP (36). Regression models were selected from random forest (37), support vector machine (38), linear regression, and Gaussian process regression (39). For each protein property, the embedding method, regression model, and its hyperparameters were first selected by nested cross-validation on the dataset of all single-residue mutants (Supplementary Table 1). These configurations were then fixed and consistently used for the benchmark tests to evaluate the impact of data augmentation.

All model training and prediction were performed using the scikit-learn library in Python (40). Due to large computational costs, random forest was excluded for the APH(3')II and CYP2C9 datasets when ESM-2 embeddings were used.

Training and evaluation

For evaluations on single-residue mutants, we first randomly selected 20% of the mutants as hold-out test data. The remaining 80% served as the pool from which training data were sampled (see below). This random split was repeated 10 times with different random seeds, and the average performance was reported.

For evaluations on double-residue mutants, all single-residue mutants in each dataset were used as the training data pool. All available double-residue mutants were used as test data, except for the GB1 dataset with ESM-2 embeddings where 10,000 mutants were randomly selected for each evaluation due to large computational costs.

To simulate different levels of experimental data availability, we constructed training sets by randomly sampling n% (n = 1, 2, ..., 99) from the above training data pool (Supplementary Figure 3). The remaining data were labeled with computational estimates from Rosetta and ESM-2, which were used for data augmentation. For the sampled training data, Rosetta and ESM-2 values were also computed and used to determine the coefficients for the hybrid score (Equation 1) and scaling function (Equation 2). Prediction accuracy was evaluated on the hold-out test data using Spearman's correlation coefficient, with and without data augmentation. Importantly, test data were never used during the model training and the fitting of the hybrid score and scaling functions to prevent data leakage. This procedure was repeated 10 times per sampling rate n using different random seeds, and the average performance was reported.

For the APH(3')II and CYP2C9 datasets, which contain a large number of mutants, we also evaluated under training data points fewer than 100 to more closely examine performance in low-data conditions.

To assess the performance of Rosetta and ESM-2 computational estimates alone (i.e., without ML models), we used the same test data as in the ML evaluation for single-residue mutants. For double-residue mutants, 5,000 variants were randomly sampled as test data. From these, 500 variants were randomly selected for 10 times to calculate the average performance. No scaling (Equation 2) was applied in these baseline evaluations.

Data Availability

The source code and the benchmark datasets used in this study are available at https://github.com/Teppei-Deguchi/prj-weak-supervision

Supporting Information

Supplementary Figures 1- 4 and Supplementary Table 1.

Author Contributions

TD and YS developed the method and wrote the paper. TD wrote the code and performed the computational experiments. YK, SI, and KK participated in the data interpretation. All authors read and approved the final version of the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by JSPS KAKENHI (22H03691), AMED grant (JP19ak0101122), NEDO grant (JPNP14004), JACI Prize for Encouraging Young Researcher (to YS), and JST SPRING (JPMJSP2108) (to TD). The computations were partly performed on the NIG supercomputer at ROIS National Institute of Genetics, the ABCI supercomputer at AIST, and the SQUID supercomputer at Cybermedia Center, Osaka University through the HPCI System Research Project (hp230057, hp240075). TD gratefully acknowledges financial support from Hattori international scholarship foundation and G-7 scholarship foundation.

References

1. Jokinen E, Heinonen M, Lähdesmäki H. mGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion. Bioinformatics. 2018;34(13):i274-i83.
2. 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.
3. Saito Y, Oikawa M, Nakazawa H, Niide T, Kameda T, Tsuda K, et al. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. ACS Synth Biol. 2018;7(9):2014-22.
4. Freschlin CR, Fahlberg SA, Romero PA. Machine learning to navigate fitness landscapes for protein engineering. Curr Opin Biotechnol. 2022;75:102713.
5. Needham CJ, Bradford JR, Bulpitt AJ, Care MA, Westhead DR. Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. BMC Bioinformatics. 2006;7:405.
6. Yang Y, Chen B, Tan G, Vihinen M, Shen B. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids. 2013;44(3):847-55.
7. David A, Sternberg MJE. Protein structure-based evaluation of missense variants: Resources, challenges and future directions. Curr Opin Struct Biol. 2023;80:102600.
8. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nature methods. 2019;16(8):687-94.
9. Kouba P, Kohout P, Haddadi F, Bushuiev A, Samusevich R, Sedlar J, et al. Machine learning-guided protein engineering. ACS catalysis. 2023;13(21):13863-95.
10. Mazurenko S, Prokop Z, Damborsky J. Machine learning in enzyme engineering. ACS Catalysis. 2019;10(2):1210-23.
11. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented

software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545-74.

12. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123-30.
13. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in neural information processing systems. 2021;34:29287-303.
14. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604(7904):175-83.
15. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 2014;42(14):e112.
16. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. Am J Hum Genet. 2021;108(9):1735-51.
17. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
18. Sun J, Zhu T, Cui Y, Wu B. Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation. Innovation (Camb). 2025;6(1):100750.
19. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic acids research. 2005;33(suppl_2):W382-W8.
20. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert opinion on drug discovery. 2015;10(5):449-61.
21. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79(3):830-8.
22. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. The Journal of Physical Chemistry B. 2018;122(21):5389-99.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic acids research. 2000;28(1):235-42.
24. Harkiolaki M, Tsirka T, Lewitzky M, Simister PC, Joshi D, Bird LE, et al. Distinct binding modes of two epitopes in Gab2 that interact with the SH3C domain of Grb2. Structure. 2009;17(6):809-22.
25. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. Crystal structures of a complexed and peptide-free membrane protein–binding domain: molecular basis of peptide recognition by PDZ. Cell. 1996;85(7):1067-76.
26. Sauer-Eriksson AE, Kleywegt GJ, Uhlén M, Jones TA. Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. Structure. 1995;3(3):265-78.
27. Nurizzo D, Shewry SC, Perlin MH, Brown SA, Dholakia JN, Fuchs RL, et al. The crystal structure of aminoglycoside-3′-phosphotransferase-IIa, an enzyme responsible for antibiotic resistance. Journal of molecular biology. 2003;327(2):491-506.
28. Wester MR, Yano JK, Schoch GA, Yang C, Griffin KJ, Stout CD, et al. The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-Å resolution. Journal of Biological Chemistry. 2004;279(34):35630-7.
29. Georgiev AG. Interpretable numerical descriptors of amino acid space. J Comput Biol. 2009;16(5):703-23.

30. Liang G, Chen G, Niu W, Li Z. Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. Chem Biol Drug Des. 2008;71(4):345-51.
31. Zaliani A, Gancia E. MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. Journal of chemical information and computer sciences. 1999;39(3):525-33.
32. Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. Journal of molecular structure. 2007;830(1-3):106-15.
33. Yang L, Shu M, Ma K, Mei H, Jiang Y, Li Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino acids. 2010;38:805-16.
34. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. Journal of medicinal chemistry. 1998;41(14):2481-91.
35. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Peptide Science: Original Research on Biomolecules. 2005;80(6):775-86.
36. van Westen GJ, Swier RF, Cortes-Ciriano I, Wegner JK, Overington JP, IJzerman AP, et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. Journal of cheminformatics. 2013;5:1-20.
37. Breiman L. Random forests. Machine learning. 2001;45:5-32.
38. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14:199-222.
39. Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of mathematical psychology. 2018;85:1-16.
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.