Protein stability prediction with dual-view ensemble learning from single sequence

Zhiwei Nie[1,2†], Yiming Ma[1†], Yutian Liu[3], Xiansong Huang[2], Zhihong Liu[4], Peng Yang[5], Fan Xu[2], Feng Yin[4], Zigang Li[6,4], Jie Fu[7*], Zhixiang Ren[2*], Wen-Bin Zhang[5*], Jie Chen[1,2*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China.
[2]Pengcheng Laboratory, Shenzhen, China.
[3]School of Computer Science, Peking University, Beijing, China.
[4]Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen, China.
[5]Beijing National Laboratory for Molecular Sciences, Key Laboratory of Polymer Chemistry & Physics of Ministry of Education, Center for Soft Matter Science and Engineering, College of Chemistry and Molecular Engineering, Peking University, China.
[6]State Key Laboratory of Chemical Oncogenomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen, China.
[7]Shanghai AI Laboratory, Shanghai, China.

*Corresponding author(s). E-mail(s): jiefu@ust.hk; renzhx@pcl.ac.cn; wenbin@pku.edu.cn; jiechen2019@pku.edu.cn;
†These authors contributed equally to this work.

Abstract

Predicting the protein stability changes upon mutations is one of the effective ways to improve the efficiency of protein engineering. Here, we propose a dual-view ensemble learning-based framework, DVE-stability, for protein stability prediction from single sequence. DVE-stability integrates the global and local dependencies of mutations to capture the intramolecular interactions from two views through ensemble learning, in which a structural microenvironment simulation module is designed to indirectly introduce the information of structural

microenvironment at the sequence level. DVE-stability not only achieves comparable prediction performance to state-of-the-art methods, but also achieves superior performance in predicting rare beneficial mutations that are critical for directed evolution of proteins. More importantly, DVE-stability identifies important intramolecular interactions via attention scores, demonstrating interpretable. Overall, DVE-stability provides a flexible and efficient tool for protein stability prediction in an interpretable ensemble learning manner.

Keywords: protein stability, dual-view, ensemble learning, microenvironment simulation

1 Introduction

Quantifying the mutation-induced effects is fundamental for exploring the evolutionary fitness landscape of proteins [1, 2]. Protein thermodynamic stability is of particular interest as it provides valuable insights into protein folding and functions [3, 4], thereby promoting the protein engineering [5–8] for pharmaceutical biologics and industrial biocatalysts. Considering that traditional experimental measurements of protein stability are costly in both time and money, computational methods for predicting protein stability changes have emerged in an endless stream [9–12].

Existing computational methods for protein stability prediction, especially deep-learning approaches, can be divided into two categories: structure-based and sequence-based methods. The structure-based methods directly encoded the structural microenvironment of mutations in protein 3D structures for pretraining or end-to-end training [13–20]. In contrast, sequence-based methods exploited only sequence information [21, 21–25] or employed structure prediction modules to extract structure-related primary features [26–28] for protein stability prediction. Sequence-based methods are often easier to fine-tune for customization because sequences are less difficult to obtain than high-quality structures. Therefore, protein stability prediction methods from single sequence are worth exploring, which is beneficial for efficient protein engineering.

In this work, we propose a **D**ual-**V**iew **E**nsemble learning-based framework, DVE-stability, for protein stability prediction from single sequence. DVE-stability integrates the global dependencies of mutations dominated by long-range interactions and local dependencies of mutations determined by the structural microenvironments, enabling the neural network to capture the intramolecular interaction network in which the mutation is located from two views through ensemble learning. Considering the importance of the structural microenvironments for stability changes caused by mutations, we argue that it is beneficial to introduce information about the structural microenvironment into sequence-based methods in an indirect way. To this end, we propose the structural microenvironment simulation (SMS) module to embed in the DVE-stability framework to indirectly introduce the information of structural microenvironment at the sequence level. Finally, DVE-stability achieves comparable performance to state-of-the-art methods and superior performance in predicting rare beneficial mutations that are critical for directed evolution of proteins. In addition, DVE-stability is

demonstrated to be interpretable, identifying important intramolecular interactions via attention scores. Overall, DVE-stability represents an interpretable ensemble learning approach for protein stability prediction from single sequence, which is expected to provide a flexible and efficient tool for protein engineering.

2 Results

2.1 The dual-view ensemble learning architecture of DVE-stability

The dual-view ensemble learning architecture of DVE-stability (Fig.1a) is designed to simultaneously internalize the global dependencies of mutations dominated by long-range interactions and local dependencies of mutations determined by the structural microenvironments. First, a mutated protein sequence and its wild-type sequence are encoded in parallel by the pretrained protein language model (PLM), ESM-2-650M [29], to extract corresponding amino-acid-level embeddings. Second, the dual-view interaction dependencies of mutations are extracted in the mutated protein, as well as in the wild-type protein. The CLS token of PLM, which internalizes global information of an entire sequence to some extent [30], is employed to represent the global dependencies of mutations on the entire protein sequence, while the local dependencies of mutations on its structurally adjacent residues is built through a structural microenvironment simulation (SMS) module. As shown in Fig.1b, considering that fitness effects upon mutations are largely determined by the structural microenvironment of mutations [31–33], we pursue the simulation of structural microenvironment at the sequence level, thereby indirectly constructing and encoding the local dependencies of mutations in 3D structures. To this end, the SMS module is designed to identify the related residues that may constitute the structural microenvironment of mutations and perform feature fusion. Specifically, considering that the attention map of PLM has been proven to capture residue contact relationships to a certain extent [34, 35], ESM-2-650M with frozen parameters is adopted for the identification of structurally adjacent residues based on the assigned attention scores. The top 3 residues with the highest attention scores are selected, while the remaining residues are filtered out. Subsequently, the embeddings of selected residues are fused through a transformer block into an "ENV" embedding that represents local dependencies of mutations (Fig.1c).

Third, the global dependencies of the mutation in a mutated protein sequence is subtracted by that in its wild-type sequence to highlight the mutational effects on the intramolecular interaction network of proteins, and the local dependencies of the mutation are also processed in the same way. Fourth, the difference between global dependencies and local dependencies is used for the prediction of protein stability changes upon mutations (i.e. $\Delta\Delta G$) separately, and the final prediction is the average of the two branch predictions.

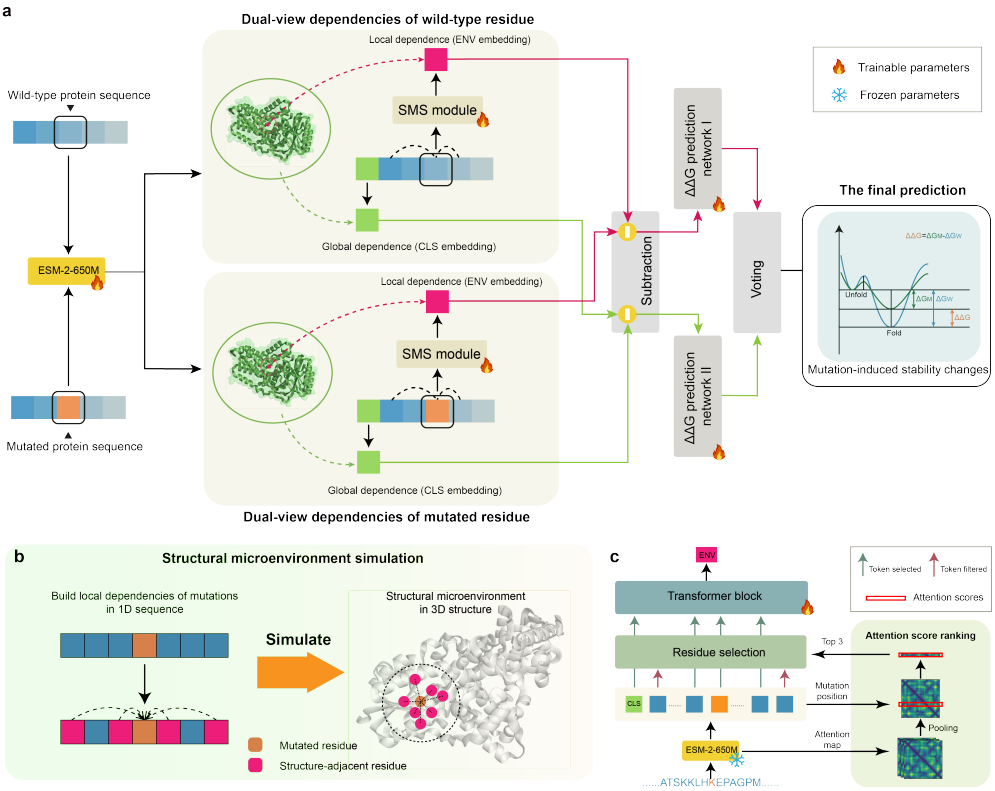

Fig. 1 The model architecture of DVE-stability. **a**, The architecture overview of DVE-stability. The wild-type and mutated protein sequences are input in parallel to extract dual-view interaction dependencies of mutations (CLS embedding and ENV embedding) for subsequent ensemble learning. **b**, Illustration of structural microenvironment simulation (SMS) module, where residues structurally adjacent to the mutated residue are identified in the protein sequence to build the local dependencies of the mutation, thereby simulating the structural microenvironment in 3D structure. **c**, Internal architecture of the SMS module. The off-the-shelf protein language model ESM-2-650M is employed to obtain the ranking of the attention scores of the residue types at the mutation position, where the residues with the top 3 largest attention scores are selected. A transformer block is adopted to fuse the above selected residues to obtain an "ENV" embedding that represents local dependencies of mutations.

2.2 Prediction performance for protein stability changes upon mutations

To evaluate the prediction performance of DVE-stability for protein stability changes upon mutations, the widely recognized benchmark dataset S669 [9] is adopted for the blind test after training on the most used training dataset S2648 [36]. Considering that an ideal predictive model is one that is able to achieve the opposite sign of direct mutations when predicting reverse mutations, we set up the dataset with hypothetical reverse mutations of S669 (named S669 reverse dataset) to evaluate the symmetry

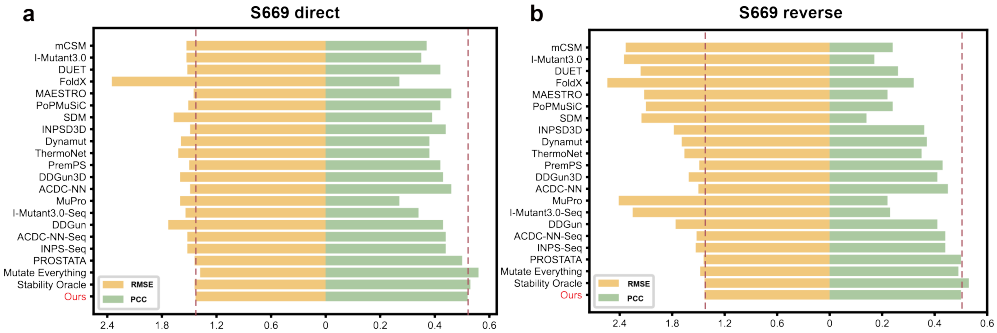

Fig. 2 Prediction performance on the widely recognized benchmark blind test dataset S669 [9], including S669 direct dataset (a) and S669 reverse dataset (b). The results of other methods are obtained from the reported values of previous work Mutate Everything [26].

of predictions, which is a reflection of model robustness. As shown in Fig.2, DVE-stability is comparable in prediction performance to two state-of-the-art sequence-based methods (PROSTATA [22] and Mutate Everything [26]) and a state-of-the-art structure-based method (Stability Oracle [37]), while maintaining the ideal prediction symmetry on direct and reverse datasets.

2.3 Prediction performance for rare beneficial mutations

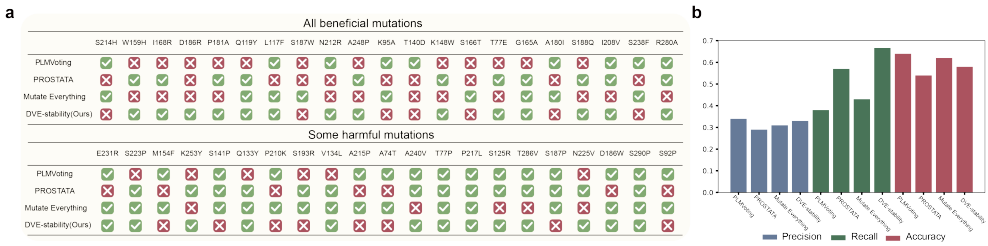

Fig. 3 Prediction performance of different methods on beneficial and harmful mutations of a representative enzyme engineering case [38]. **a**, Illustration of prediction results for beneficial and harmful mutations, where correctly predicted mutations are marked by '✓', and incorrectly predicted ones are marked by '×'. **b**, Prediction performance on the test set containing all mutations.

Generally speaking, mutations that are beneficial to fitness improvement constitute a rare subset of the fitness landscape for protein evolution [39]. Therefore, accurately mining rare beneficial mutations is one of the potential means to improve the efficiency of directed evolution of proteins. To test the predictive ability of DVE-stability for rare beneficial mutations, we select a representative enzyme engineering case (*Is*PETase engineering with improved thermal stability [38]) to set up a validation

experiment. Specifically, we use different methods to predict whether a single mutation is a beneficial mutation or a harmful mutation for improving enzyme thermal stability. The method of ensemble voting of multiple existing protein language models (referred to as PLMVoting) [39], two state-of-the-art protein stability prediction methods (PROSTATA [22] and Mutate Everything [26]) are adopted to compare with our DVE-stability model. As shown in Fig.3a, the prediction results of all 21 beneficial mutations and the same number of harmful mutations (see the prediction results of all 64 harmful mutations in Supplementary information S2) from the original literature of *Is*PETase engineering case [38] are shown. For all 21 beneficial mutations, DVE-stability successfully predicts 14 of them, exceeding 9 of PLMVoting, 11 of PROSTATA and 9 of Mutate Everything, which corresponds to the best recall metric of DVE-stability in Fig.3b. 21 harmful mutations are randomly selected to demonstrate the prediction tendency of different methods for harmful mutations, where the four methods achieve similar prediction success rates (15 of PLMVoting, 14 of PROSTATA, 16 of Mutate Everything and 13 of DVE-stability). As shown in Fig.3b, looking at all mutation samples, the precision metrics of the four methods are close. Even though PLMVoting method shows a certain advantage in the accuracy metric, its higher accuracy value comes from the successful prediction of harmful mutation samples that occupy most of the test set, which cannot reflect its real predictive ability. Overall, DVE-stability has superior performance in predicting rare beneficial mutations that are critical for protein engineering.

2.4 Interpretability of DVE-stability

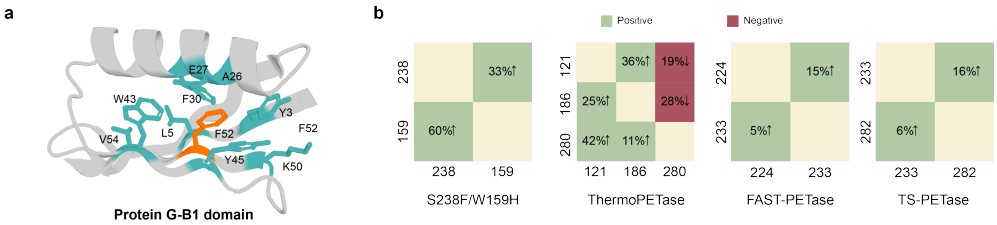

Fig. 4 Interpretability of DVE-stability. **a**, The case analysis of representative protein G-B1 domain, where the top 10 residues with the highest attention scores with F52 as the central residue are highlighted. **b**, The attention scores between pairwise mutations from four representative enzyme-engineering cases, where the increased attention scores relative to pairwise wild-type residues are marked as "positive" with corresponding improvement ratio, and the decreased scores are marked as "negative" with corresponding decay ratio.

To demonstrate the interpretability of DVE-stability for intramolecular interactions, representative protein G-B1 domain [40] is adopted for case analysis. As shown in Fig.4a, with F52 as the central residue, the top 10 residues with the highest attention scores assigned by DVE-stability model are highlighted. We can find that DVE-stability assigns higher attention scores to some residues (Y3, L5, F30, W43,

and Y45) that form $\pi - \pi$ interactions with F52, which is consistent with the involvement of F52 and these residues in the nucleation event of hydrophobic core [41, 42]. In addition, several enzyme-engineering cases with improved thermal stability are adopted to demonstrate the sensitivity of DVE-stability to fitness effects caused by pairwise mutations. Specifically, the attention scores between pairwise mutations from four representative cases are calculated and visualized, including S238F/W159H on top of *Is*PETase [43], ThermoPETase [44], FAST-PETase [45] and TS-PETase [46]. If the calculated attention score has increased relative to pairwise wild-type residues, we mark it as "positive" with corresponding improvement ratio, otherwise, it is marked as "negative" with corresponding decay ratio. As illustrated in Fig.4b, more than 83% of pairwise mutations exhibit increasing attention scores and only 2 pairs of mutations have decreased attention scores, demonstrating that DVE-stability is sensitive to mutation-induced fitness effects behind the improvement of eznyme thermal stability.

2.5 Pretraining on larger datasets improves the prediction performance

With the advancement of high-throughput wet-lab methods, mutation-induced $\Delta\Delta G$ can be estimated on a mega scale [47], which is suitable for pretraining to further improve the prediction performance of DVE-stability. First, the CLS embedding and ENV embedding of DVE-stability are visualized using t-SNE for dimensionality reduction to intuitively analyze the effectiveness of pretraining (Fig.5a). Compared with the original ESM-2-650M, the CLS embedding and ENV embedding after pretraining can distinguish between harmful (reduced stability) and beneficial (increased stability) mutations on the S669 dataset with boundaries, which qualitatively proves that pretraining enables DVE-stability to capture the mapping between mutations and stability changes. Second, as shown in Fig.5b, we perform ablation experiments to quantify the performance improvement by pretraining. In addition, the impact of the SMS module on the prediction ability of DVE-stability is also explored. We can find that pretraining is always beneficial to DVE-stability's performance improvement, regardless of whether the SMS module (i.e., the use of ENV embedding) is introduced. Similarly, the introduction of SMS module is always beneficial to DVE-stability's performance improvement, regardless of whether pretraining is performed. The above ablation results show that SMS module, the core component of DVE-stability, is effective, and pretraining on a larger dataset significantly improves the prediction performance. Overall, pretraining on larger datasets can help DVE-stability further improve its predictive power, even if the dataset involves proteins from different families.

3 Discussion

With dual-view ensemble learning, DVE-stability fuses the global interaction dependencies and local interaction dependencies of mutations to predict protein stability changes upon mutations from single sequence. DVE-stability is comparable to the state-of-the-art methods on benchmark blind test sets and is superior in predicting rare

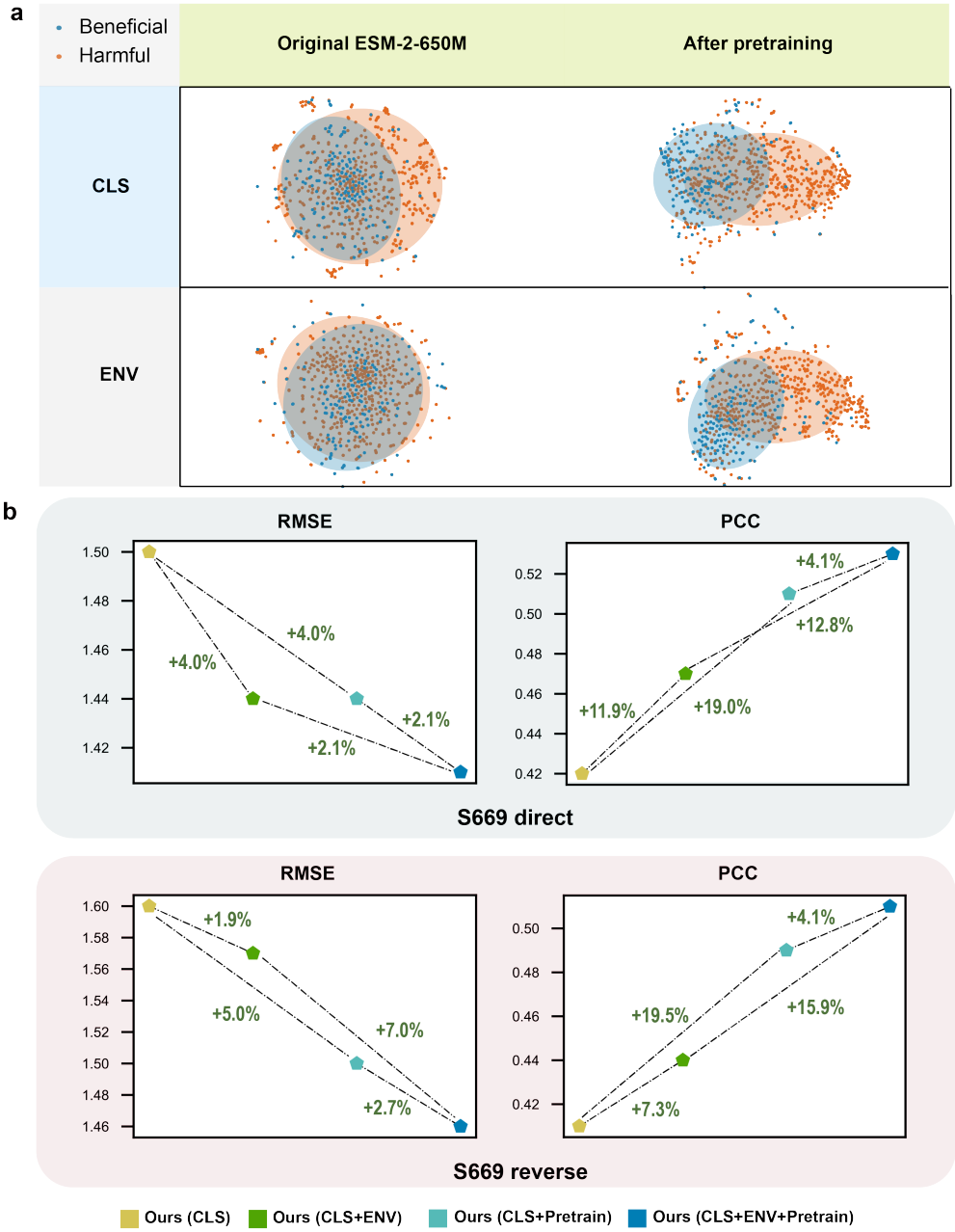

Fig. 5 Ablation study for pretraining with larger datasets. **a**, The dimensionality reduction of CLS embedding and ENV embedding of harmful (reduced stability) and beneficial (increased stability) mutations on the S669 dataset. **b**, The ablation results for pretraining and the SMS module (i.e., the use of ENV embedding).

beneficial mutations that are critical for protein engineering. More importantly, DVE-stability is interpretable in identifying residue interactions. However, DVE-stability still has clear limitations, including not supporting the prediction of unequal-length mutations (insertion or deletion of amino acids) or multi-site mutations. Next, for the prediction of unequal-length mutations, we plan to set special tokens to represent inserted or deleted residues, and for multi-site mutations, it is expected that a fusion module combining the effects of multiple mutations will be effective. Looking to the future, as the available data in the community gradually grows, DVE-stability is expected to further improve its prediction performance by pretraining on larger datasets.

4 Methods

4.1 Dataset

4.1.1 Training and test dataset

S669 [9] is a manually curated blind test dataset containing 669 mutants and their $\Delta\Delta G$ measurements, with less than 25% sequence identity to our training dataset S2648 [36], which contains 2,648 non-redundant single amino acid mutants with corresponding $\Delta\Delta G$ obtained from 131 proteins.

4.1.2 High-throughput dataset for pretraining

High-throughput dataset [47] is a mega-scale folding stability measurement dataset produced by cDNA display proteolysis. This dataset consists of approximately 776,000 high-throughput fold stability measurements, of which 448,788 single amino acid mutants are selected by us. After eliminating the two types of mutations, insertions and deletions of amino acids, 389,068 variants remain to form the screened high-throughput dataset. We create a reverse dataset consisting of hypothesized reverse mutations of the above-screened high-throughput dataset, which is used together with its direct dataset for the pretraining (a total of 778,136).

4.2 Training settings

The training settings of DVE-stability on the S2648 dataset are as follows: a learning rate of 6e-6 and a batch size of 1 for 20 epochs. The pretraining settings of DVE-stability on high-throughput dataset are as follows: a learning rate of 5e-6 and a batch size of 3 for 15 epochs. For the above training and pretraining processes, we employ the Adam optimizer with a weight decay of 0.01 and use mean squared error (MSE) loss.

4.3 Architecture overview

Amino acid has 20 different types, represented by a set $AA = \{A, C, ..., Y\}$. A protein of length L can be defined as $S = (s^1, s^2, ..., s^L)$ where $s^l \in AA$. Given the position of mutation pos_{mut} and a mutated residue $s_{mut}^{\text{pos}_{mut}} \in AA$ substituting the wild-type residue $s_{wild}^{\text{pos}_{mut}} \in AA$, we can get the wild-type protein

$S_{wild} = (s^1_{wild}, s^2_{wild}, ..., s^{\text{pos}_{mut}}_{wild}, ..., s^L_{wild}) \in AA^L$ and mutated protein $S_{mut} = (s^1_{mut}, s^2_{mut}, ..., s^{\text{pos}_{mut}}_{mut}, ..., s^L_{mut}) \in AA^L$. The target of predicting protein stability changes upon point mutation is to calculate the mutation-induced stability change $\Delta\Delta G$.

DVE-stability is built on the protein language model, ESM-2-650M, that accepts inputs of both wild-type and mutated protein sequences. Both the wild-type and mutated protein sequences, denoted as S_{wild} and S_{mut}, are input into a shared ESM-2-650M model to generate the corresponding sequence features:

$$\mathbf{Z}^{seq}_{wild} = \text{ESM}(\text{S}_{wild}) \tag{1}$$

$$\mathbf{Z}^{seq}_{mut} = \text{ESM}(\text{S}_{mut}) \tag{2}$$

where $\mathbf{Z}^{seq}_{wild} = (\mathbf{Z}^{CLS}_{wild}, \mathbf{Z}^1_{wild}, \mathbf{Z}^2_{wild}, ..., \mathbf{Z}^L_{wild}, \mathbf{Z}^{END}_{wild}) \in \mathbb{R}^{(L+2)\times D}$ and $\mathbf{Z}^{seq}_{mut} = (\mathbf{Z}^{CLS}_{mut}, \mathbf{Z}^1_{mut}, \mathbf{Z}^2_{mut}, ..., \mathbf{Z}^L_{mut}, \mathbf{Z}^{END}_{mut}) \in \mathbb{R}^{(L+2)\times D}$ represent the features of wild-type and mutated protein sequences output by ESM-2-650M model, in which D represents the dimension of the features.

Subsequently, the sequence features are fed into the structural microenvironment simulation (SMS) module to obtain the local dependencies of mutations as follows:

$$\mathbf{Z}^{env}_{wild} = \text{CES}(\mathbf{Z}^{seq}_{wild}, \text{pos}_{mut}) \tag{3}$$

$$\mathbf{Z}^{env}_{mut} = \text{CES}(\mathbf{Z}^{seq}_{mut}, \text{pos}_{mut}) \tag{4}$$

where $\mathbf{Z}^{env}_{wild} \in \mathbb{R}^D$ and $\mathbf{Z}^{env}_{mut} \in \mathbb{R}^D$ represent the local dependencies of wild-type residue and mutated residue.

Finally, $\mathbf{Z}^{CLS}_{wild}$ and $\mathbf{Z}^{CLS}_{mut}$ are fed into a $\Delta\Delta G$ prediction neural network, while $\mathbf{Z}^{env}_{wild}$ and $\mathbf{Z}^{env}_{mut}$ are fed into another parallel $\Delta\Delta G$ prediction neural network. The final predicted $\Delta\Delta G$ is the average of the above two predictions.

$$\Delta\Delta G_{CLS} = \text{FC}_1(\mathbf{Z}^{CLS}_{wild} - \mathbf{Z}^{CLS}_{mut}) \tag{5}$$

$$\Delta\Delta G_{ENV} = \text{FC}_2(\mathbf{Z}^{env}_{wild} - \mathbf{Z}^{env}_{mut}) \tag{6}$$

$$\Delta\Delta G = (\Delta\Delta G_{CLS} + \Delta\Delta G_{ENV})/2 \tag{7}$$

where FC_1 and FC_2 adopt the same configuration, that is, a single fully connected layer, to predict the $\Delta\Delta G$.

4.4 SMS module implementation

Given a protein sequence S, we obtain its sequence feature $\mathbf{Z}^{seq}$ by feeding S into the ESM-2-650M model. Moreover, ESM-2-650M outputs an attention map every transformer layer, where the attention scores reflect the importance of the relationship between the paired tokens. Therefore, we select the structurally adjacent residues based on the attention map $A^{last} \in \mathbb{R}^{H\times(L+2)\times(L+2)}$ of the last transformer layer, in which H is the number of attention heads of a transformer.

First, A^{last} is pooled as follows:

$$\bar{A}^{last} = \frac{1}{H}\sum_{i}^{H} A^{last}_i \tag{8}$$

where A_i^{last} denotes the i-th head's attention map of the A^{last}.

Subsequently, given the mutation position pos_{mut}, we obtain the attention score vector $A_{mut} \in \mathbb{R}^{L+2}$ that represents the attention scores to the mutation token (abbreviated as MUT token) by indexing $A_{average}^{last}$ with mutation position pos_{mut}:

$$A_{mut} = \bar{A}_{\{\text{pos}_{mut}\}}^{last} \tag{9}$$

Based on A_{mut}, we select 3 tokens related to MUT token, which rank among the top 3 in attention scores:

$$(\text{pos}_1, \text{pos}_2, \text{pos}_3) = \text{argmax}_3(A_{mut}) \tag{10}$$

$$\mathbf{Z}^{top_3} = \mathbf{Z}_{\{pos_1, pos_2, pos_3\}}^{seq} \tag{11}$$

where $\text{argmax}_3(\cdot)$ represents the function to get the 3 most largest elements' positions of the given vector, and $\mathbf{Z}^{top_3} \in \mathbb{R}^{3\times D}$ represents the 3 most important token related to MUT token.

Finally, we feed the $\mathbf{Z}^{top_3}$ with a learnable embedding vector $\mathbf{Z}^{CLS} \in \mathbb{R}^D$ into a one-layer transformer block to integrate them into the final "ENV" embedding of local dependencies.

$$\mathbf{Z}^{env} = \text{Extract-CLS}(\phi_\theta(\mathbf{Z}^{CLS} \oplus \mathbf{Z}^{top_3})) \tag{12}$$

where $\mathbf{Z}^{env} \in \mathbb{R}^D$ represents the feature of local dependencies, Extract-CLS denotes extracting the CLS token's embedding, ϕ_θ represents a one-layer transformer block with trainable parameter θ, and $\oplus$ denotes concatenating the two features.

4.5 Attention score calculation of pairwise mutations

In this work, four enzyme-engineering cases are employed to calculate attention scores between pairwise mutations. Specifically, if there are beneficial mutations with mutation positions $P = \{\text{pos}_1, \text{pos}_2, ..., \text{pos}_m\}$, we will calculate the difference of attention scores of any two positions $(i, j) \in P \times P$ between pairwise mutations and pairwise wild-type residues:

$$\text{score}_w(i, j) = A_w(i, j) \tag{13}$$

$$\text{score}_m(i, j) = A_m(i, j) \tag{14}$$

$$\Delta\text{score(i, j)} = \text{score}_m(i, j) - \text{score}_w(i, j) \tag{15}$$

where A_w is the pooled attention map outputted by the last transformer block of ESM-2-650M of pairwise wild-type residues and A_m is that of pairwise mutations.

4.6 Evaluation metrics

4.6.1 Root mean square error (RMSE)

RMSE is a standard way to measure how closely the predicted measurements align with the true measurements. The formula for RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{16}$$

where N is the number of obeservations, y_i is the actual value of an observation, and $\hat{y}_i$ is the predicted value.

4.6.2 Pearson correlation coefficient (PCC)

Pearson correlation coefficient is a usual way to measure the linear correlation between two variables, offering a comprehensive analysis when the data distribution permits:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{17}$$

where x_i and y_i are the individual sample points indexed with i, $\bar{x}$ and $\bar{y}$ are the sample means.

4.6.3 Accuracy, Precision, and Recall

Accuracy, precision, and recall are three measures of observational error and are commonly used in classification task. The formula for accuracy, recall, and precision is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Key Points

- We propose a dual-view ensemble learning-based framework, DVE-stability, for protein stability prediction from single sequence.
- DVE-stability integrates the global dependencies of mutations dominated by long-range interactions and local dependencies of mutations determined by the structural microenvironments, enabling the neural network to capture the intramolecular interaction network in which the mutation is located from two views through ensemble learning.
- A structural microenvironment simulation (SMS) module is designed to indirectly introduce the information of structural microenvironment at the sequence level.
- DVE-stability achieves comparable performance to state-of-the-art methods and superior performance in predicting rare beneficial mutations that are critical for directed evolution of proteins, and identifies important intramolecular interactions via attention scores.

- DVE-stability represents an interpretable ensemble learning approach for protein stability prediction from single sequence, which is expected to provide a flexible and efficient tool for protein engineering.

Data availability

S2648, S669 and the original high-throughput dataset are freely available from previous studies [9, 36, 47]. The processed high-throughput dataset is publicly available via GitHub at https://github.com/ZhiweiNiepku/DVE-stability.

Code availability

Relevant code and models are available via GitHub at https://github.com/ZhiweiNiepku/DVE-stability.

References

[1] Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118–e118 (2011).

[2] Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences* **108**, 9916–9921 (2011).

[3] Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current opinion in structural biology* **19**, 596–604 (2009).

[4] Pucci, F., Bernaerts, K. V., Kwasigroch, J. M. & Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **34**, 3659–3665 (2018).

[5] Wu, S., Snajdrova, R., Moore, J. C., Baldenius, K. & Bornscheuer, U. T. Biocatalysis: enzymatic synthesis for industrial applications. *Angewandte Chemie International Edition* **60**, 88–119 (2021).

[6] Bell, E. L. *et al.* Biocatalysis. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

[7] Jay, S. M. & Lee, R. T. Protein engineering for cardiovascular therapeutics: untapped potential for cardiac repair. *Circulation research* **113**, 933–943 (2013).

[8] Meghwanshi, G. K. *et al.* Enzymes for pharmaceutical and therapeutic applications. *Biotechnology and applied biochemistry* **67**, 586–601 (2020).

[9] Pancotti, C. *et al.* Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics* **23**, bbab555 (2022).

[10] Pucci, F., Schwersensky, M. & Rooman, M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Current opinion in structural biology* **72**, 161–168 (2022).

[11] Caldararu, O., Mehra, R., Blundell, T. L. & Kepp, K. P. Systematic investigation of the data set dependency of protein stability predictors. *Journal of Chemical Information and Modeling* **60**, 4772–4784 (2020).

[12] Qiu, Y., Huang, T. & Cai, Y.-D. Review of predicting protein stability changes upon variations. *Proteomics* 2300371 (2024).

[13] Schymkowitz, J. *et al.* The foldx web server: an online force field. *Nucleic acids research* **33**, W382–W388 (2005).

[14] Sheffler, W. & Baker, D. Rosettaholes: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Science* **18**, 229–239 (2009).

[15] Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science* **11**, 2714–2726 (2002).

[16] Pires, D. E., Ascher, D. B. & Blundell, T. L. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335–342 (2014).

[17] Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks. *PLoS computational biology* **16**, e1008291 (2020).

[18] Savojardo, C., Fariselli, P., Martelli, P. L. & Casadio, R. Inps-md: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* **32**, 2542–2544 (2016).

[19] Li, G., Yao, S. & Fan, L. Prostage: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks. *Journal of Chemical Information and Modeling* **64**, 340–347 (2024).

[20] Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences* **121**, e2314853121 (2024).

[21] Pancotti, C. *et al.* A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes* **12**, 911 (2021).

[22] Umerenkov, D. *et al.* Prostata: a framework for protein stability assessment using transformers. *Bioinformatics* **39**, btad671 (2023).

[23] Folkman, L., Stantic, B., Sattar, A. & Zhou, Y. Ease-mm: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *Journal of molecular biology* **428**, 1394–1405 (2016).

[24] Montanucci, L., Capriotti, E., Frank, Y., Ben-Tal, N. & Fariselli, P. Ddgun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC bioinformatics* **20**, 1–10 (2019).

[25] Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* **62**, 1125–1132 (2006).

[26] Ouyang-Zhang, J., Diaz, D., Klivans, A. & Krähenbühl, P. Predicting a protein's stability under a million mutations. *Advances in Neural Information Processing Systems* **36** (2024).

[27] Iqbal, S. *et al.* Prost: Alphafold2-aware sequence-based predictor to estimate protein stability changes upon missense mutations. *Journal of chemical information and modeling* **62**, 4270–4282 (2022).

[28] Chen, Y., Xu, Y., Liu, D., Xing, Y. & Gong, H. An end-to-end framework for the prediction of protein structure and fitness from single sequence. *Nature Communications* **15**, 7400 (2024).

[29] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[30] Kovaleva, O., Romanov, A., Rogers, A. & Rumshisky, A. in *Revealing the dark secrets of bert* 4356–4365 (Association for Computational Linguistics, 2019).

[31] Shroff, R. *et al.* Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS synthetic biology* **9**, 2927–2935 (2020).

[32] Lu, H. *et al.* Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature* **604**, 662–667 (2022).

[33] d'Oelsnitz, S. *et al.* Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme. *Nature Communications* **15**, 2084 (2024).

[34] Vig, J. *et al. Bertology meets biology: Interpreting attention in protein language models* (2021).

[35] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

[36] Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0.

Bioinformatics **25**, 2537–2543 (2009).

[37] Diaz, D. J. *et al.* Stability oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nature Communications* **15**, 6170 (2024).

[38] Cui, Y. *et al.* Computational redesign of a petase for plastic biodegradation under ambient condition by the grape strategy. *Acs Catalysis* **11**, 1340–1350 (2021).

[39] Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology* **42**, 275–283 (2024).

[40] Sjöbring, U., Björck, L. & Kastern, W. Streptococcal protein g. gene structure and protein binding properties. *Journal of biological chemistry* **266**, 399–405 (1991).

[41] Kmiecik, S. & Kolinski, A. Folding pathway of the b1 domain of protein g explored by multiscale modeling. *Biophysical journal* **94**, 726–736 (2008).

[42] Blanco, F. J. & Serrano, L. Folding of protein g b1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *European journal of biochemistry* **230**, 634–649 (1995).

[43] Austin, H. P. *et al.* Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences* **115**, E4350–E4357 (2018).

[44] Son, H. F. *et al.* Rational protein engineering of thermo-stable petase from ideonella sakaiensis for highly efficient pet degradation. *Acs Catalysis* **9**, 3519–3526 (2019).

[45] Lu, H. *et al.* Deep learning redesign of petase for practical pet degrading applications. *Biorxiv* 2021–10 (2021).

[46] Zhong-Johnson, E. Z. L., Voigt, C. A. & Sinskey, A. J. An absorbance method for analysis of enzymatic degradation kinetics of poly (ethylene terephthalate) films. *Scientific Reports* **11**, 928 (2021).

[47] Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).

Supplementary information

S1 Prediction performance on blind test dataset S669

Method	S669 direct		S669 reverse	
	RMSE	**PCC**	**RMSE**	**PCC**
mCSM	1.53	0.37	2.33	0.24
I-Mutant3.0	1.53	0.35	2.35	0.17
DUET	1.52	0.42	2.16	0.26
FoldX	2.35	0.27	2.54	0.32
MAESTRO	1.45	0.46	2.12	0.22
PopMusic	1.51	0.42	2.10	0.24
SDM	1.67	0.39	2.15	0.14
INPS3D	1.49	0.44	1.78	0.36
Dynamut	1.59	0.38	1.69	0.37
ThermoNet	1.62	0.38	1.66	0.35
PremPS	1.50	0.42	1.49	0.43
DDGun3D	1.60	0.43	1.61	0.41
ACDC-NN	1.49	0.46	1.50	0.45
MuPro	1.60	0.27	2.41	0.22
I-Mutant3.0-Seq	1.54	0.34	2.25	0.23
DDGun	1.73	0.43	1.76	0.41
ACDC-NN-Seq	1.52	0.44	1.52	0.44
INPS-Seq	1.52	0.44	1.53	0.44
PROSTATA	1.44	0.50	1.44	0.50
Mutate Everything	**1.38**	**0.56**	1.48	0.49
Stability Oracle	1.44	0.53	1.43	**0.53**
DVE-stability(Ours)	1.43	0.52	**1.43**	0.50

Table S1 Prediction performance on benchmark dataset S669. Bold values represent the best performance, and underlined values represent the second best performance.

S2 Prediction results of all harmful mutations

	N37P	A40P	A47W	T51P	S54W	G66A	A74T	T77P	A82V	R90D	Q91P	S92P	R100F	H104F	D112R	T113P
PLM Voting	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
PROSTATA	✓	✓	✓	✓	×	✓	×	✓	✓	✓	✓	×	×	×	✓	✓
Mutate Everything	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	×	✓	✓
DVE-stability (Ours)	✓	✓	×	×	✓	✓	×	✓	✓	✓	×	×	✓	✓	×	✓
	T116P	**S121D**	**R123V**	**S125R**	**Q133Y**	**V134L**	**S136Q**	**G139N**	**S141P**	**G147T**	**T151P**	**A152S**	**M154F**	**G155A**	**S175W**	**Q182L**
PLM Voting	×	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓	✓	✓	×
PROSTATA	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	×	×	✓	×	×
Mutate Everything	✓	✓	✓	×	✓	✓	✓	×	✓	✓	✓	×	✓	✓	×	×
DVE-stability (Ours)	✓	×	×	✓	✓	✓	×	×	×	✓	✓	×	×	✓	✓	✓
	A183T	**D186W**	**S187P**	**T189K**	**S193R**	**T198V**	**E204K**	**P210K**	**A215P**	**P217L**	**I218F**	**D220E**	**S223P**	**N225V**	**A226P**	**Q228V**
PLM Voting	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓	×	×	×	×
PROSTATA	✓	×	✓	✓	✓	✓	✓	×	×	✓	×	×	✓	✓	×	✓
Mutate Everything	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	✓	×	×	✓
DVE-stability (Ours)	×	✓	×	×	×	✓	✓	×	×	✓	✓	×	✓	✓	✓	✓
	E231M	**E231R**	**A240V**	**A240P**	**N241L**	**N246W**	**K253Y**	**G254V**	**K259F**	**T266P**	**T270Q**	**A272L**	**E274P**	**T279P**	**T286V**	**S290P**
PLM Voting	✓	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	×	✓	✓	✓
PROSTATA	✓	×	✓	×	✓	×	✓	×	✓	✓	×	×	✓	×	✓	✓
Mutate Everything	✓	✓	×	×	✓	×	×	×	✓	✓	×	×	✓	✓	×	✓
DVE-stability (Ours)	✓	✓	✓	✓	×	✓	✓	✓	×	✓	✓	✓	×	✓	✓	✓

Table S2 Prediction results of all harmful mutations of different methods.

S3 Attention scores of pairwise mutations

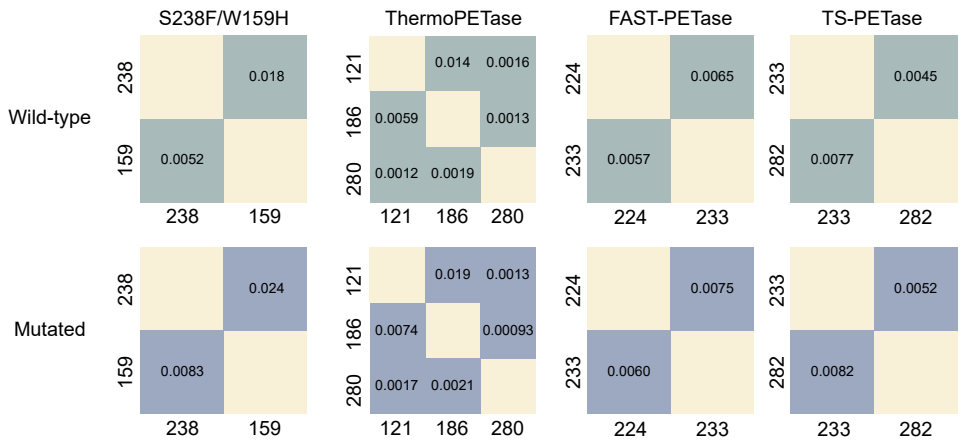

Fig. S1 Attention scores of pairwise mutations in four engineering cases of *Is*PETase with improved thermal stability.

S4 Ablation study on blind test dataset S669

Model	S669 direct		S669 reverse	
	RMSE	PCC	RMSE	PCC
Ours (CLS)	1.50	0.42	1.60	0.41
Ours (CLS+ENV)	1.44	0.47	1.57	0.44
Ours (CLS+Pretrain)	1.44	0.51	1.50	0.49
Ours (CLS+ENV+Pretrain)	1.41	0.53	1.46	0.51

Table S3 Ablation results on S669 dataset.