Towards anchoring evolutionary fitness for protein stability with virtual chemical environment recovery

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

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China.
[2]Peng Cheng Laboratory, Shenzhen, China.
[3]School of Computer Science, Peking University, 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]Hong Kong University of Science and Technology, Hong Kong, China.

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

Abstract

Protein stability offers valuable insights into protein folding and functionality, making it an integral component of evolutionary fitness. Previous computational methods possess both strengths and weaknesses, leading to practical and interpretational limitations. Here, we propose an interpretable protein stability change

prediction method, S3C, to anchor evolutionary fitness for protein stability with virtual chemical environment recovery. S3C first gets rid of the shackles of high-resolution protein structure data and restores the local chemical environments of the mutations at the sequence level. Subsequently, S3C promotes the evolutionary fitness of protein stability to dominate the fitness landscape under the selective pressure. Naturally, S3C comprehensively outperforms state-of-the-art methods on benchmark datasets while showing ideal generalization when migrated to unseen protein families. More importantly, S3C is demonstrated to be interpretable at multiple scales, including high-fidelity recovery of local structure micro-environment, perception of intricate interaction reconstruction, and accurate mining of rare beneficial mutations. S3C expands the boundaries of protein evolution prediction and provides an ideal candidate for large-scale optimization of protein engineering.

1 Introduction

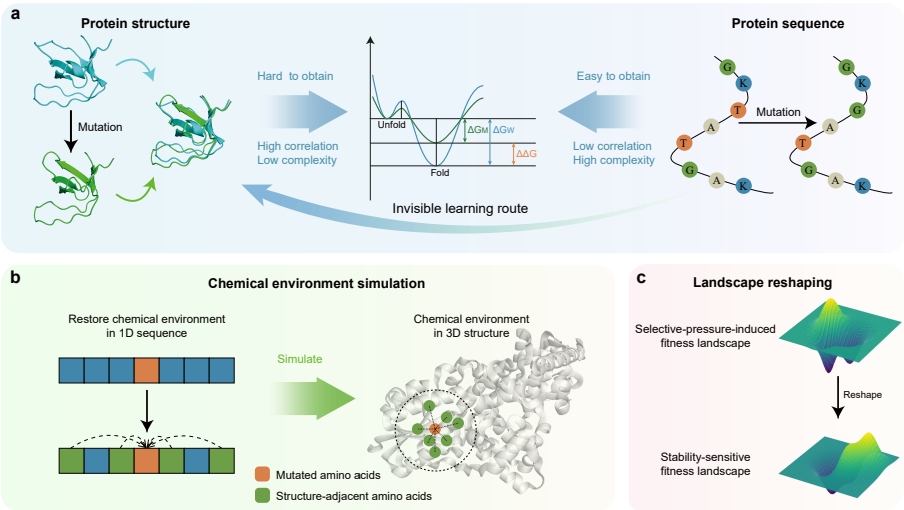

Fig. 1 The motivation and methodology of S3C. **a**, Existing methods for predicting protein stability changes are divided into structure-dependent and sequence-only methods, both of which have corresponding advantages and disadvantages. This drives us to develop an interpretable structure-free $\Delta\Delta G$ prediction method. **b**, Chemical environment simulation module, in which the local chemical environment in the 3D structure is simulated in the 1D sequence, thereby implicitly integrating structural patterns to establish an invisible learning route of 'sequence-to-function' for deep-learning framework. **c**, Landscape reshaping module, in which the fitness landscape under selective pressure is transformed into a state that is sensitive to protein stability, thereby anchoring the evolutionary fitness of protein stability.

Protein is the material basis of life, and the prerequisite for the completion of its specific function is that the corresponding tertiary structure is correctly folded [1, 2]. As a common phenomenon in life activities, mutations may change protein structure to a certain extent, thereby affecting its intended function [3, 4]. Quantifying the mutation-induced effects is fundamental to studying the evolutionary fitness of proteins [5, 6], with protein thermodynamic stability being of particular interest as it provides valuable insights into protein folding and functions [7, 8]. Generally speaking, we adopt the difference in folding free energies between wild-type and mutated proteins ($\Delta\Delta G$) to measure the protein stability change [9, 10]. However, traditional wet-lab methods of estimating protein stability changes are costly in both time and money . Therefore, efficient and low-cost computing methods have always been the focus of continued attention and development.

As shown in Fig.1a, existing computational methods for $\Delta\Delta G$ estimation, including deep-learning approaches, can be roughly divided into two categories: structure-dependent and sequence-only methods. Structure-dependent methods adopt protein structures as input or mix the features of structure and sequence for $\Delta\Delta G$ prediction. Among them, force field-based methods [11–13] require high-resolution 3D protein structures as input and are often computationally intensive. Some methods only extract latent features of the protein structure [14, 15] for $\Delta\Delta G$ prediction, while some other methods utilize both structural and sequence information by mixing input [16–19] or integrating protein structure prediction module [20, 21]. In contrast, sequence-only methods rely solely on sequences [22–24] for $\Delta\Delta G$ prediction. However, data availability and task complexity are contradictory in the above two types of methods. First of all, the high correlation between structure and stability makes the complexity of predicting $\Delta\Delta G$ based on structure lower, but the rare or missing high-resolution protein structure is a devastating disaster faced by this manner. Even though protein structure prediction approaches [25, 26] based on deep learning have made breakthrough progress, the inherent errors of them [27–29] bring catastrophic perturbations to the prediction of mutation-induced stability changes. Secondly, the significant advantage of sequence-only methods is that protein sequences can be easily obtained and suitable for large-scale optimization of protein engineering, but the distance between sequence and stability alone the 'sequence-structure-function' route brings higher task complexity.

How to develop protein stability prediction methods that rely only on sequences but implement virtual structure recovery arouses our interest. Two difficulties impede progress toward this goal, namely the restoration of structural patterns and capturing protein stability patterns in the vast evolutionary fitness space. Firstly, the local chemical environment, that is, the structural context, is strongly correlated with the stability changes caused by amino-acid mutations here, while the protein sequence can only provide the sequence context of a certain mutation, which cannot be directly linked to protein stability. Secondly, protein language models (PLMs) have demonstrated excellent performance in zero-shot mutation prediction [30] and directed evolution of protein engineering [31], which reflects its delineation of the fitness landscape on an evolutionary scale. However, natural protein sequences are the product of selective pressure, which prevents current protein language models trained on them from

capturing stability-sensitive fitness landscapes for better protein stability prediction. An approach to anchor the evolutionary fitness of protein stability from the existing fitness landscape under selective pressure is urgently needed.

To overcome the above challenges, we propose S3C (**S**tructure-**S**imulated **S**tability **C**hange prediction) to interpretably predict protein stability changes caused by mutations through landscape reshaping along invisible learning route. As shown in Fig.1b, the first core module of S3C is 'chemical environment simulation', which establishes an invisible learning route of 'sequence-to-function' for deep-learning framework. It mines structural-adjacent residues of mutations through the attention-based selection strategy to restore the local chemical environments at sequence level. The second core module of S3C is 'landscape reshaping' in Fig.1c, which anchors the evolutionary fitness of protein stability in a comprehensive evolutionary space. With the support of the above two modules, S3C outperforms the state-of-the-art models across the board, stamping a new watermark for protein stability change prediction. More importantly, S3C perceives the intricate interactions at multiple scales. First, S3C not only restores the local chemical environment of the representative protein structure domain with high fidelity but also hints at the interaction system within it. Second, in several enzyme-engineering cases, S3C assigns higher attention to beneficial mutation combinations, indicating that it is sensitive to the reconstruction of intricate interactions behind the improvement of thermal stability. Last but not least, S3C is proven to be able to accurately mine rare beneficial mutations in vast evolutionary space, showing great potential for directed evolution of protein engineering. Overall, S3C captures protein stability patterns in evolutionary fitness space through virtual chemical environment recovery, which upgrades the ability to predict protein evolution and large-scale optimization of protein engineering.

2 Results

2.1 Architecture of S3C

Considering that mutation-induced protein stability changes are significantly affected by the local chemical environments of mutations, we pursue the recovery of structural micro-environment at the sequence level, thereby achieving interpretable structure-free predictions. The PLM has been proven to achieve protein atomic-level structure prediction [26], in which the attention mechanism in the transformer block captures the residue contact map [32, 33], indicating its ability to perceive structural patterns. However, the local chemical environment is the joint participation of some specific residues, therefore cannot be directly represented by the residue-level or sequence-level embeddings extracted by the PLM. To solve this problem, we design the 'chemical environment simulation' module to realize the local chemical environment recovery at the sequence level based on PLM, which brings the distance between sequence and function closer (Fig.2a). Specifically, for a protein sequence, we select structurally adjacent residues through the attention pattern of PLM (ESM here). Several residues (*i.e.*, tokens, here set to 32) with the highest attention scores for the mutated residues are selected, which are considered to be structurally close to the mutations, while the remaining residues are filtered out. It is worth noting that the mutation tokens (their

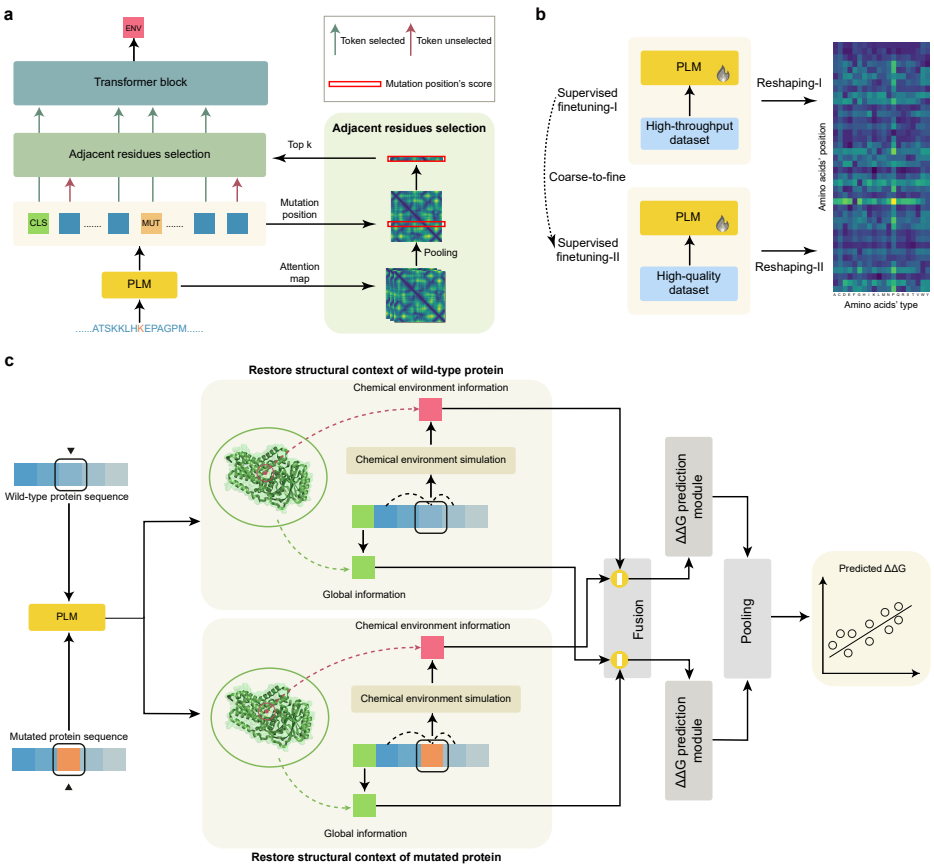

Fig. 2 **a**, Implementation of chemical environment simulation module. An attention-based residue selection strategy mines residues that are structurally adjacent to the mutations at the sequence level, and they are subsequently fused into an ENV embedding that contains the local chemical environment information. **b**, Implementation of landscape reshaping module. The high-throughput dataset and the high-quality dataset are adopted for two-stage supervised fine-tuning respectively, which enables the protein language model to capture the stability-sensitive fitness landscape through the adjustment of the amino acid probability distribution from coarse to fine. **c**, Architecture overview of S3C. The wild-type and mutated protein sequences are input in parallel to extract global and chemical environment information. The above two types of features are respectively fused and fed to two parallel $\Delta\Delta G$ prediction modules, and the corresponding outputs are pooled to obtain the final predicted $\Delta\Delta G$ value.

embeddings are referred to as 'MUT embedding' for short) and classification tokens representing global information (their embeddings are referred to as 'CLS embedding' for short) are included in the selected residues. The embeddings of all tokens selected above are then fused through a learnable transformer block into a novel embedding that contains simulated local chemical environment information, which we call 'ENV

embedding' for short. Obviously, the ENV embedding restores the structural context based on the protein sequence context, thereby building a 'sequence-to-function' invisible learning route to guide the neural network's learning direction.

PLMs are pre-trained on massive natural protein sequences in order to capture sequence patterns on the evolutionary scale. Therefore, these language models are thought to outline the fitness landscape under natural selective pressure [30, 31], which indicates that they are not suitable for zero-shot prediction of protein stability. In order to promote the reshaping of the fitness landscape derived from natural protein sequences in a stability-sensitive direction, we propose 'landscape reshaping' module as shown in Fig.2b. With the advancement of high-throughput wet-lab methods, mutation-induced $\Delta\Delta G$ data can be estimated on a mega scale [34]. Although thermodynamic stabilities inferred by the above method are only accurate under specific conditions [34], the unprecedented scale of assay enables neural networks to understand rough patterns of protein stability changes. Furthermore, some manually curated high-quality datasets [18, 21, 35–39] provide more precise links between mutations and protein stability changes. Considering the above two types of data with different characteristics, a two-stage supervised fine-tuning (SFT) pipeline is adopted to achieve 'coarse-to-fine' landscape reshaping. Specifically, in the first SFT stage, high-throughput single-mutation-induced $\Delta\Delta G$ data are used, and in the second SFT stage, the new dataset S6070 we compiled is adopted. It is worth noting that the parameters of the PLM (ESM here) are not frozen so the evolutionary fitness of protein stability is captured in the parameters through fine-tuning.

On top of the above two modules, the overall architecture of S3C is shown in Fig.2c. S3C takes the form of two parallel inputs, one is a wild-type protein sequence, and the other is a mutated protein sequence. For the above two types of protein sequences, we adopt the same feature extraction process to obtain the corresponding CLS embeddings and ENV embeddings. The CLS embeddings represent global information of a protein sequence to a certain extent and are therefore adopted to avoid missing long-range interactions that may exist outside the local chemical environment. The ENV embeddings we design are tangible simulations of the local chemical environments in which the mutations are located, emphasizing the dominant influence of the structural micro-environment on mutation-induced stability changes. Subsequently, the CLS embeddings of wild-type and mutated protein sequences are subtracted and sent to the $\Delta\Delta G$ prediction module, as is the ENV embeddings. Finally, the outputs of two parallel prediction modules are pooled to obtain the predicted $\Delta\Delta G$ value. Under this framework, S3C performs two-stage supervised fine-tuning for landscape reshaping along an invisible learning route.

2.2 Prediction performance of S3C

In order to comprehensively evaluate the prediction performance of S3C, the widely recognized benchmark dataset, S669 [36], is adopted. It is worth noting that we eliminate protein sequences homologous to S669 in S6070 (the new dataset we compiled for the second supervised fine-tuning), setting 30% sequence identity as the threshold. $\Delta\Delta G$ is determined by the two endpoints of wild-type and mutated protein, i.e., does

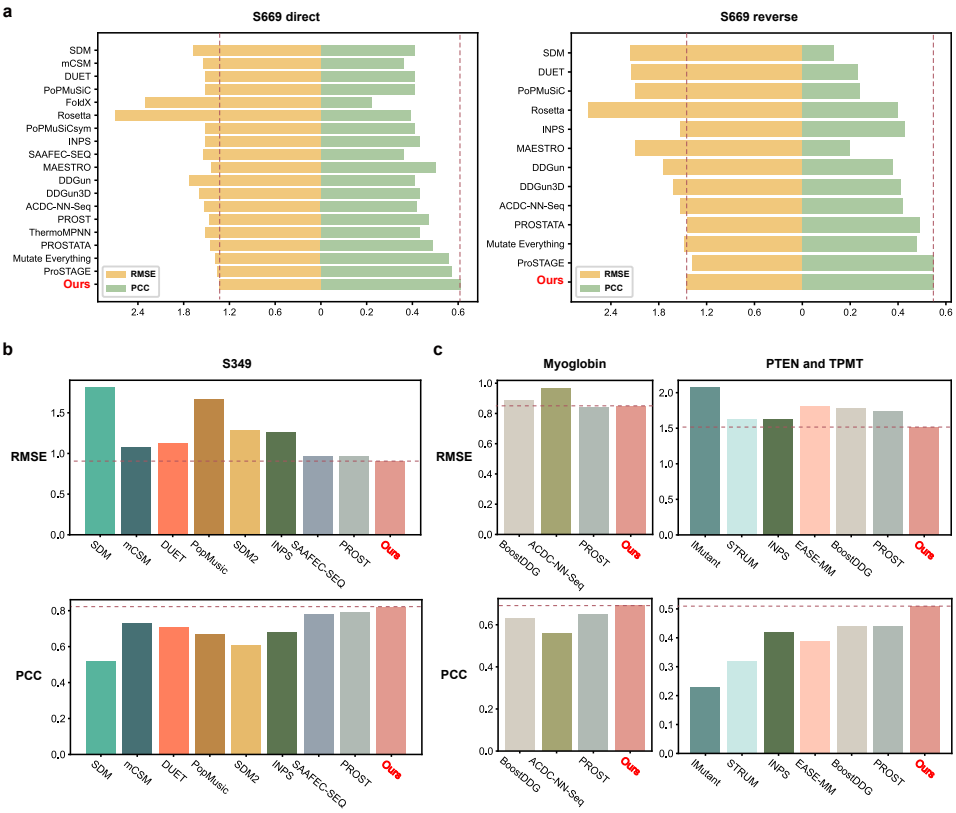

Fig. 3 **a**, Test performance on benchmark dataset S669. S3C outperforms state-of-the-art methods while maintaining a certain degree of predictive symmetry. **b**, A commonly used blind test set, S349, is adopted to further evaluate the prediction performance of S3C. **c**, Two datasets from unseen protein families are employed to evaluate the generalization performance of S3C, including Myoglobin, PTEN and TPMT.

not rely on direct or reverse paths, because free energy is a thermodynamic state function [9]. Therefore, an ideal model is one that is able to achieve the opposite sign of direct mutations when predicting reverse mutations. To evaluate the prediction performance of different models for hypothetical reverse mutations, we set up another benchmark dataset with reverse mutations of S669 (named S669 reverse) to evaluate the symmetry of predictions, which is a reflection of model reliability and robustness. Specifically, unlike the original S669 dataset (named S669 direct), in the S669 reverse dataset, the mutated protein sequence is considered to be the wild-type, and vice versa. Considering that the experimental $\Delta\Delta G$ measurements themselves may contain errors and that Pearson correlation coefficient (PCC), which evaluates prediction trends, can better reflect the model's fitting effect on the overall evolutionary fitness trend, we adopt PCC as the core indicator for comparing prediction performance. As shown in Fig.3a, S3C comprehensively surpasses state-of-the-art methods

in the direct mutation prediction task. Compared to the sequence-only state-of-the-art model PROSTATA [23], S3C improves it over 24.4%. Compared to the state-of-the-art structure-dependent model ProSTAGE [18], S3C achieves an improvement of 7%. In the reverse mutation prediction task, most methods, including S3C, show performance degradation. Even so, the prediction performance of S3C is on par with the state-of-the-art model (*i.e.*, the state-of-the-art structure-dependent model ProSTAGE). Notably, S3C improves the state-of-the-art sequence-only model PROSTATA over 12%. In addition, we also select a commonly used blind test set, S349 (a subset of S2648) [21], to further test the prediction performance of S3C (Fig.3b). As expected, S3C achieves the best prediction performance among competitors. Overall, S3C outperforms state-of-the-art methods across the board while maintaining a certain degree of predictive symmetry.

A protein family [40, 41] is composed of proteins with similar sequence, structure, and function, where the protein sequences are evolutionarily related. A deep learning framework that accurately captures patterns of protein stability changes is expected to generalize to unseen protein families. To verify the generalization ability of S3C, we test it on two datasets (Myoglobin [37, 38], PTEN and TPMT [18, 39]) of mutation-induced stability changes of specific proteins from different protein families. When transferred to unseen protein families, S3C still comprehensively surpasses other methods, showing ideal generalization ability (Fig.3c).

2.3 S3C anchors evolutionary fitness for protein stability

'Landscape reshaping' and 'chemical environment simulation' are core components of S3C, so their contributions to prediction performance require further analysis. Comprehensive ablation studies are conducted to present the impact of each module on prediction performance, involving the ENV embedding (represents chemical environment simulation module) and the first supervised fine-tuning, here referred to as 'SFT-I' (represents landscape reshaping module). As shown in Fig.4a, four groups of experiments are set up, including 'CLS', 'CLS+ENV', 'SFT-I+CLS', and 'SFT-I+CLS+ENV'. Additionally, state-of-the-art sequence-only and structure-dependent methods (PROSTATA [23] and ProSTAGE [18]) are adopted for comparison to demonstrate the effectiveness of the two proposed modules. It is worth noting that all compared methods are trained on the widely recognized training dataset S2648, and then tested on S669 direct and S669 reverse respectively. Since S669 has already eliminated homologous proteins to S2648 based on 25% sequence identity, we do not need to perform additional processing. We find that the separate use of ENV embedding and SFT-I promotes the improvement of prediction performance, and the combined use of the two achieves further improvement. The introduction of ENV embedding brings prediction performance on par with the state-of-the-art sequence-only method PROSTATA that relies on model ensembles, demonstrating the effectiveness of the priors-guided chemical environment simulation module. We can infer that the priors-guided design provides the neural network with an invisible learning route toward evolutionary fitness of protein stability, thereby ensuring its prediction performance beyond blind feature fusion. In addition, compared with the state-of-the-art structure-dependent method ProSTAGE, the introduction of SFT-I enables

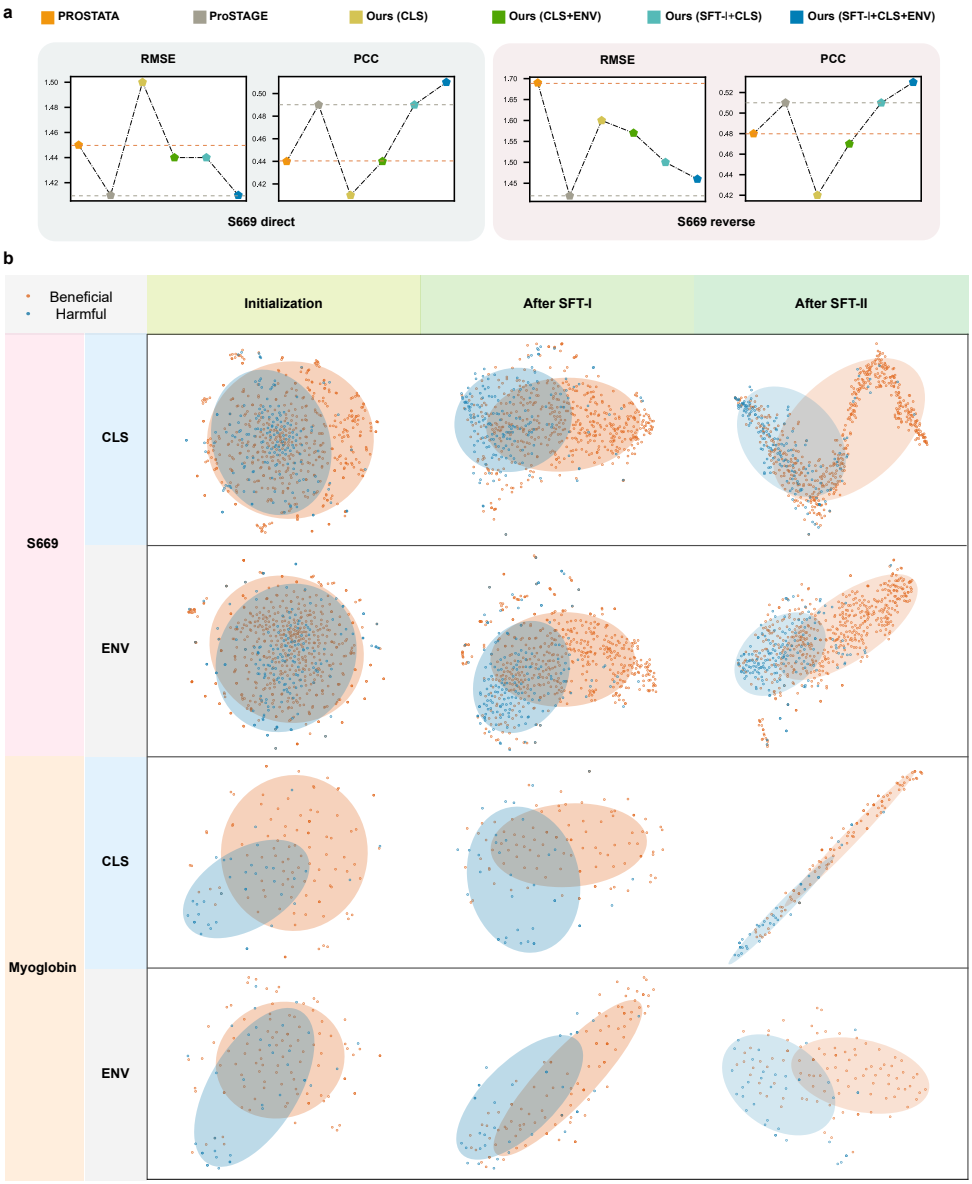

Fig. 4 **a**, Module ablation study on benchmark dataset S669. Based on CLS embedding, the separate use of ENV embedding and SFT-I promotes the improvement of prediction performance, and their combined use leads to better prediction performance. A state-of-the-art structure-dependent and a sequence-only method are additionally included in the comparison. **b**, Dimensionality reduction visualization of CLS embedding and ENV embedding on S669 and Myoglobin datasets. 'Beneficial' refers to mutations that are beneficial to protein stability and 'Negative' refers to those that are harmful to protein stability. 'Initialization' refers to the initial state, in which PLM does not undergo any supervised fine-tuning. 'After SFT-I' refers to the state after the first supervised fine-tuning using the high-throughput dataset, and 'After SFT-II' refers to the state after the second supervised fine-tuning using the high-quality dataset.

prediction performance comparable to the hybrid use of 3D structure information and sequence embeddings. This result implies that landscape reshaping promotes the anchoring of evolutionary fitness of protein stability, thereby compensating for the lack of high-resolution protein structure input in S3C.

In order to further demonstrate the dynamic process of anchoring evolutionary fitness, we perform visual analysis on two key features that affect S3C's prediction performance as shown in Fig.4b, including CLS embedding and ENV embedding. CLS embedding represents the global information of a mutated protein sequence, while ENV embedding represents the local chemical environment information of the mutation, so their joint analysis can reflect the state of the fitness landscape at different molecular-interaction scales. According to the step-by-step progression of landscape reshaping, we analyze the fitness landscape in three states, represented by 'Initialization', 'After SFT-I', and 'After SFT-II' respectively. 'Initialization' represents the initial state, in which PLM (ESM here) is not fine-tuned in a supervised manner. 'After SFT-I' represents the state after the first supervised fine-tuning using the high-throughput dataset, in which the fitness landscape has been transformed in a coarse-grained manner. 'After SFT-II' represents the stage after the second supervised fine-tuning (referred to as 'SFT-II') using the high-quality dataset, which means that the fitness landscape is further transformed in a fine-grained manner.

Changes in protein stability caused by mutations can be divided into three categories: beneficial, harmful, and neutral, which constitute branching patterns of protein stability changes. Naturally, a fitness landscape that is more sensitive to protein stability should be able to screen for different types of mutations. Therefore, we adopt the ability of S3C to distinguish different types of mutations to reflect the sensitivity of the fitness landscape to protein stability. Considering that the proportion of positive and negative samples in the PTEN and TPMT datasets is extremely imbalanced, we select S669 and Myoglobin for analysis. Since there is no $\Delta\Delta G$ value caused by mutation that is strictly equal to 0 in the above datasets, we divide all mutations into two categories: 'Beneficial' ($\Delta\Delta G<0$, beneficial to protein stability) and 'Harmful' ($\Delta\Delta G>0$, harmful to protein stability). As shown in Fig.4b, the CLS embeddings and ENV embeddings of different states on the two datasets are visualized using t-SNE for dimensionality reduction. In the 'initial' state, the above two types of embeddings of harmful mutations and beneficial mutations cannot be distinguished. This supports our hypothesis that PLM pre-trained on natural protein sequences captures fitness landscapes under selective pressure that are insensitive to changes in protein stability. As supervised fine-tuning is performed, the clustering distinction between harmful and beneficial mutations becomes increasingly clear, indicating that the fitness landscape gradually becomes sensitive to protein stability. The above results prove that S3C effectively and significantly dynamically reshapes the state of the fitness landscape, thereby tightly anchoring the evolutionary fitness of protein stability.

2.4 S3C perceives intricate interactions

The proposed chemical environment simulation module aims to rely solely on sequences to restore the structural micro-environment of mutations, which is a key driving force for S3C to achieve excellent prediction performance. In order to

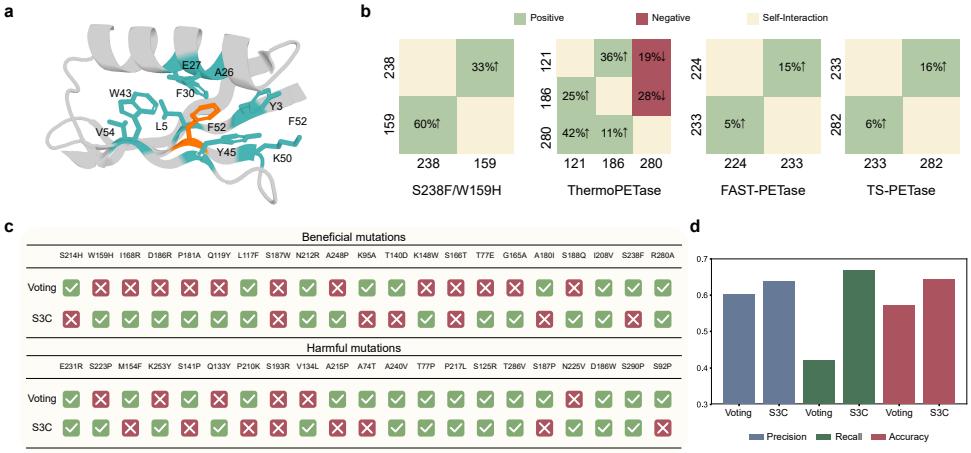

Fig. 5 **a**, S3C restores the local structural micro-environment centered on F52 with high fidelity in the Protein G-B1 domain. The $\pi - \pi$ interactions formed by F52 with some of the identified surrounding residues appear to stabilize the hydrophobic core to some extent, suggesting the perception of intricate molecular interactions by S3C. **b**, Four engineering cases of *Is*PETase with improved thermal stability, in which attention scores between pairwise mutations are calculated. Pairwise mutations that are assigned higher attention scores relative to wild-type pairwise residues are considered 'positive', and the proportion of score increase is marked. On the contrary, pairwise mutations that are assigned lower attention scores relative to wild-type pairwise residues are considered 'negative', and the proportion of score decrease is marked. Overall, more than 83% of the pairwise mutations are assigned higher attention by S3C, which proves that it is sensitive to the reconstruction of intricate interactions behind thermal stability improvement. **c**, A specific engineering case of *Is*PETase with improved thermal stability, where 21 beneficial mutations and 21 harmful mutations proven by wet-lab experiments are selected to form a balanced test set to evaluate the discriminating ability of different models for rare beneficial mutations. Correctly predicted mutations are represented by '✓', while incorrectly predicted mutations are represented by '×'. **d**, S3C comprehensively surpasses the directed evolution method through PLMs voting for efficiently suggesting high-fitness mutations, especially showing an excellent ability to mine rare beneficial mutations, that is, the Recall is increased by 24% relative to its competitor.

demonstrate the recurrence degree of S3C for structural context, we select the structure domain with significant characteristics for case analysis. Protein G [42] is an immunoglobulin-binding protein in which the B1 domain is supported by a hydrophobic core. Previous studies [43, 44] have shown that with F52 as the center, Y3, L5, F30, W43, and Y45 are involved in the nucleation event of its hydrophobic core residues. As shown in Fig.5a, with F52 as the central residue, we highlight structural-adjacent residues (top 10 here) identified by the attention map of S3C to observe the recovery of local chemical environment. Satisfactorily, the residues surrounding F52 are assigned higher attention scores and thus are identified. Interestingly, F52 forms $\pi - \pi$ interactions with several surrounding residues, which likely stabilizes the hydrophobic core to a certain extent. The above results not only show that S3C can restore the local structural micro-environment with high fidelity based on the sequence only, but also imply that it perceives intricate molecular interactions.

Mutations may bring about perturbations in intricate interactions, which ultimately result in changes in protein stability. In order to further confirm that S3C is sensitive to the reconstruction of interactions, we select several enzyme-engineering cases with improved thermal stability for verification. *Is*PETase [45] is a plastic-degrading enzyme that exhibits high hydrolyzing activity toward Polyethylene terephthalate (PET) [46] at ambient temperature and therefore has attracted much attention in mitigating plastic pollution. Previous studies [47–51] have improved its thermal stability to ensure that it still performs well at higher temperatures. Four representative cases, including S238F/W159H [48], ThermoPETase [49], FAST-PETase [50] and TS-PETase [51], are adopted by us. From a macro level, the improvement of thermal stability means that these mutations lead to the reconstruction of intricate interactions that benefit protein stability, so S3C is expected to assign higher attention scores between mutations. As shown in Fig.5b, we calculate and visualize the attention scores between the pairwise mutations of the four enzyme-engineering cases. If the calculated score has increased relative to wild-type pairwise residues, we mark it as 'positive' with an improvement ratio, otherwise, it is marked as 'negative' with a decay ratio. More than 83% of pairwise mutations exhibit increasing attention scores and are identified as 'positive'. In contrast, only 2 pairs of mutations have decreased attention scores and are identified as 'negative'. The above results demonstrate that S3C is sensitive to reconstructions of intricate interactions behind the improvement of thermal stability. It is worth noting that the direct and reverse attention scores between pairwise mutations (each mutation in the pair takes turns doing the query) are not equal, perhaps caused by differences in their intrinsic properties or local chemical environments.

Traditional wet-lab directed evolution of proteins is limited by efficiency, which is a dilemma caused by randomness and the huge evolution space [52]. Generally speaking, in the fitness landscape for protein evolution, mutations that are beneficial to fitness improvement constitute a rare subset [31]. Therefore, accurately mining the rare beneficial mutations is the core of improving the efficiency of directed evolution for protein engineering. By virtue of perceiving mutation-induced intricate interactions, S3C is expected to identify rare beneficial mutations. In order to test the discriminating ability of S3C for rare beneficial mutations, we conduct a verification experiment on a specific case of *Is*PETase engineering with improved thermal stability [53]. In the above enzyme-engineering case, 21 single-site beneficial mutations for thermal stability are screened out, along with the generation of more harmful mutations. For the sake of comprehensive observation, we select all 21 beneficial mutations and randomly sampled 21 harmful mutations to form a balanced test set. A previous study proposed a PLMs-voting (abbreviated as Voting here) method to guide antibody affinity maturation [31], demonstrating a methodology for efficiently suggesting high-fitness mutations. Therefore, we consider it as a competitor in the task of mining rare beneficial mutations.

As shown in Fig.5c, we show the prediction performance of S3C and the Voting method on beneficial mutations and harmful mutations separately. For 21 beneficial mutations, S3C successfully predicts 14 of them (about 67%), while the Voting

method only correctly predicts 9 (about 43%), which proves that S3C captures beneficial mutations more accurately. For 21 randomly sampled harmful mutations, S3C correctly predicts 13 of them, achieving almost the same prediction performance as on the beneficial mutations, which demonstrated its reliability. As expected, the Voting method correctly predicts 15, significantly exceeding the prediction performance on the beneficial mutations. This is because the PLMs that this method relies on captures the evolutionary pattern of natural proteins, in which wild-type residues are often assigned higher distribution probabilities [30, 31], which leads to its tendency to identify mutations as harmful mutations. Considering both harmful mutations and beneficial mutations, S3C comprehensively surpasses the Voting method in three core indicators, including Precision, Recall, and Accuracy (Fig.5d). Among them, the Recall reflects the model's ability to mine rare beneficial mutations. S3C increase this indicator to 67% with a 24% improvement, indicating its great potential for directed evolution of proteins.

3 Discussion

The realization of protein function requires the correct folding of its tertiary structure, and mutations have a certain probability of destroying the overall conformation, ultimately affecting the function carried by the protein. As a basic sub-topic of evolutionary fitness, protein stability is of great value for functional mechanism research and protein engineering. The computational methods to estimate protein stability changes have received continued attention and development, resulting in two major categories of methods: structure-dependent and sequence-only. The advantages and disadvantages of the above two types of methods drive us to develop methods that do not require structure but implicitly incorporate structural patterns. Two difficulties hinder the realization of this goal: how to restore structural patterns and how to anchor the evolutionary fitness of protein stability in a comprehensive evolutionary space.

To address the above issues, we propose S3C, an interpretable prediction method for protein stability changes. First, we design the novel chemical environment simulation module to restore the structural micro-environment of the mutation at the sequence level, thus establishing an invisible learning path of 'sequence-to-function'. Secondly, a landscape reshaping module is proposed to dynamically transform the fitness landscape captured by the protein language model, thereby tightly anchors the evolutionary fitness of protein stability in the vast evolutionary space. As expected, S3C comprehensively outperforms state-of-the-art models on benchmark datasets and additionally exhibits ideal generalization when faced with unseen protein families.

The fitness landscape captured by the protein language model from natural protein sequences at the evolutionary scale is demonstrated to be insensitive to protein stability changes by dimensionality reduction visualization analysis. As the two-stage supervised fine-tuning continues to advance, the fitness landscape captured by S3C becomes increasingly sensitive to protein stability, allowing for an increasingly clear distinction between harmful and beneficial mutations. We can conclude that S3C

dynamically reshapes the fitness landscape of protein evolution, leading to a better understanding of the branching patterns of harmful and beneficial mutations for protein stability.

The interpretability of S3C comes from its perception of intricate molecular interactions. First, S3C restores the local chemical environment of a representative protein structure domain with high fidelity based solely on protein sequence, implying the interaction system that underpins the domain. Second, S3C assigns higher attention to mutation combinations that lead to improved thermal stability, indicating that it is sensitive to the reconstructions of intricate interactions. In addition, S3C shows an excellent ability to mine rare beneficial mutations in the huge evolutionary space, showing its application prospects in the directed evolution of protein engineering. With no need for structural input, S3C is suitable for large-scale optimization and screening of proteins, thereby efficiently engineering multiple protein families such as antibodies and enzymes.

As a next step, we consider self-supervised pre-training using simulated local chemical environments on massive natural protein sequences to prompt the model to understand more general structural patterns. At the same time, we also plan to explore whether the landscape reshaping module promotes a fitness landscape that is simultaneously sensitive to protein stability and other properties. Considering that the fitness landscape under selective pressure is the product of the combined effect of multiple factors, stability-oriented landscape reshaping may promote the protein language model to capture the intersection of protein stability and other properties, such as enzyme activity and antibody affinity. This will contribute to the joint improvement of multiple properties in protein engineering and significantly improve the efficiency of directed evolution. This work still has limitations, that is, the current version of S3C does not support the prediction of stability changes caused by unequal-length mutations (insertion or deletion of amino acids) or multi-site mutations. Looking to the future, we can improve in two aspects: wet-lab data and model architecture. As high-throughput protein folding stability assays are further developed and implemented, more data will facilitate deep-learning methods to understand the epistasis present in multi-site mutations. In addition, the architecture of S3C can be adaptively modified in the face of unequal-length mutations and multi-site mutations. In order to adapt to the prediction of unequal-length mutations, we can set special tokens to represent inserted or deleted amino acids to extract the required features or approximate them by the average of the features of the two adjacent amino acids. For the prediction of protein stability changes caused by multi-site mutations, it is feasible to design a module that integrates the extracted features of multiple mutations.

4 Methods

4.1 Hypothetical reverse mutation

Protein stability changes can be measured by the difference in folding free energies ($\Delta\Delta G$) between wild-type protein and their mutants. Considering that the free energy is a thermodynamic state function, $\Delta\Delta G$ is determined by the two endpoints of wild-type protein and mutant, i.e., path independent. Thus, the change of folding free

energies from the mutant to the wild-type protein (M → W) $\Delta\Delta G_{M\to W}$ is the opposite of $\Delta\Delta G_{W\to M}$, the change of folding free energies from the wild-type to the mutant (W → M):

$$\Delta\Delta G_{M\to W} = -\Delta\Delta G_{W\to M} \tag{1}$$

In this work, we create reverse datasets consisting of hypothetical reverse mutations of the screened high-throughput dataset and the benchmark dataset S669, respectively. Creating a reverse dataset composed of hypothetical reverse mutations for the screened high-throughput dataset is a data augmentation method to expand the size of the training data. In contrast, a reverse dataset composed of hypothetical reverse mutations is created for the benchmark dataset S669 (named S669 reverse) to test the robustness of different methods. The underlying hypothesis is that an ideal model should exhibit a degree of predictive symmetry, meaning it should predict the opposite sign for reverse mutations compared to direct mutations.

4.2 Data

4.2.1 Dataset of landscape reshaping

Dataset for supervised fine-tuning I

High-throughput dataset [34] 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 first supervised fine-tuning (a total of 778,136).

Dataset for supervised fine-tuning II

The S6070 dataset is adopted in the prediction performance comparisons in Fig.3a and generalization evaluation in Fig.3c. When conducting an additional prediction performance comparison in Fig.3b, the S2298 dataset is used. In module ablation experiments of Fig.4a, the S2648 dataset is used. Details of the above high-quality datasets are as follows:

S6070 is a new dataset compiled by us, screened from ProThermDB [54], ThermoMutDB [55], and FireProtDB [56]. We only retain single mutations with experimental stability change values and ensure that the data are non-redundant, resulting in 6,070 single amino acid mutants with known $\Delta\Delta G$. Facing different test datasets, we eliminate homologous sequences in the training dataset in advance, setting 30% sequence identity as the threshold using MMseqs2 easy-cluster tool [57]. Specifically, when testing on S669, PTEN and TPMT, and Myoglobin, we eliminate protein sequences in S6070 that are homologous to those of the three test datasets.

S2648 [35] contains 2,648 non-redundant single amino acid mutants with corresponding $\Delta\Delta G$ obtained from 131 proteins, and is the most used training dataset for predicting protein stability changes.

S2298 [21] is the remaining dataset after excluding a commonly used blind test set S349 and a mutation (PDB ID: 2A01) deleted from the protein database from the above S2648 dataset.

4.2.2 Test dataset

The four test datasets adopted in this work are described as follows:

S669 [36] is a manually curated dataset containing 669 mutants and their $\Delta\Delta G$ measurements, with less than 25% sequence identity to the most used training dataset S2648. Given its comprehensiveness and high quality, the most recently published methods [18, 19, 23] for predicting protein stability are benchmarked on this dataset for a fair comparison.

Myoglobin is a dataset consisting of 134 mutations in cytoplasmic globular proteins that regulate cellular oxygen concentration [37, 38].

PTEN and TPMT [18, 39] is a deep mutational scanning dataset of the phosphatase and tensin homolog (PTEN) and thiopurine S-methyl transferase (TPMT) proteins, with the size of 7,363 mutations.

S349 [21] a commonly used blind test set selected from S2648, including 349 mutations.

4.3 Task definition

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

4.4 Architecture overview

S3C is built on a protein language model that accepts inputs of both wild-type and mutated protein sequences. The protein language model employed in S3C is ESM, which functions as a protein feature extractor. Both the wild-type and mutated protein sequences, denoted as S_{wild} and S_{mut}, are input into a shared ESM model to generate the corresponding sequence features:

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

$$\mathbf{Z}_{mut}^{seq} = \text{ESM}(\text{S}_{mut}) \tag{3}$$

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

Subsequently, the sequence features are fed into the Chemical Environment Simulation (CES) module to obtain the chemical environment features as follows:

$$\mathbf{Z}_{wild}^{env} = \mathrm{CES}(\mathbf{Z}_{wild}^{seq}, \mathrm{pos}_{mut}) \tag{4}$$
$$\mathbf{Z}_{mut}^{env} = \mathrm{CES}(\mathbf{Z}_{mut}^{seq}, \mathrm{pos}_{mut}) \tag{5}$$

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

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

$$\Delta\Delta G_{CLS} = \mathrm{FC}_1(\mathbf{Z}_{wild}^{CLS} - \mathbf{Z}_{mut}^{CLS}) \tag{6}$$
$$\Delta\Delta G_{ENV} = \mathrm{FC}_2(\mathbf{Z}_{wild}^{env} - \mathbf{Z}_{mut}^{env}) \tag{7}$$
$$\Delta\Delta G = (\Delta\Delta G_{CLS} + \Delta\Delta G_{ENV})/2 \tag{8}$$

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

4.5 Chemical environment simulation

In this work, we propose a module called 'chemical environment simulation' to restore the local structural micro-environment of mutations at the sequence level. Given a protein sequence S, we obtain its sequence feature $\mathbf{Z}^{seq}$ by feeding S into the ESM model. Moreover, ESM outputs an attention map every transformer layer. The attention scores in the attention map reflect the importance of the relationship between the paired tokens. Therefore, we design a structural-adjacent residues selection strategy based on the attention map $A^{last} \in \mathbb{R}^{H\times(L+2)\times(L+2)}$ of the last transformer layer to simulate the local chemical environment at the mutation position, in which H is the number of attention heads of a transformer.

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

$$\bar{A}^{last} = \frac{1}{H}\sum_{i}^{H} A_i^{last} \tag{9}$$

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}:

$$A_{mut} = \bar{A}_{\{\mathrm{pos}_{mut}\}}^{last} \tag{10}$$

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

$$(\mathrm{pos}_1, \mathrm{pos}_2, ..., \mathrm{pos}_k) = \mathrm{argmax}_k(A_{mut}) \tag{11}$$
$$\mathbf{Z}^{top_k} = \mathbf{Z}_{\{pos_1, pos_2, ..., pos_k\}}^{seq} \tag{12}$$

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

Finally, we feed the $\mathbf{Z}^{top_k}$ with a learnable embedding vector $\mathbf{Z}^{CLS} \in \mathbb{R}^D$ into a one-layer transformer block to integrate them into the final chemical environment feature carrying chemical environment information. It is worth noting that the output of $\mathbf{Z}^{CLS}$ becomes the final chemical environment feature:

$$\mathbf{Z}^{env} = \text{Extract-CLS}(\phi_\theta(\mathbf{Z}^{CLS} \oplus \mathbf{Z}^{top_k})) \tag{13}$$

where $\mathbf{Z}^{env} \in \mathbb{R}^D$ represents the chemical environment feature, 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.

4.6 Supervised fine-tuning

Unless otherwise stated, S3C is always deployed using a two-stage supervised fine-tuning approach. In the first stage, the high-throughput dataset is employed, while in the second stage, different high-quality datasets are employed. The ESM-series models include multiple versions with different parameter amounts. In this work, we use the ESM-2 model with 650 million parameters, which produces embeddings of 1280 dimensions for every residue. Additionally, the model employs a special token named CLS token for extracting global information, and an END token appended to each sequence. Therefore, the output of ESM-2 for each protein sequence of length L is a vector of size $(L + 2) \times 1280$. During fine-tuning, we employ the Adam optimizer with a weight decay of 0.01 and use mean squared error (MSE) loss. In the first stage, we use a learning rate of 5e-6 and a batch size of 3 for 15 epochs. In the second stage, we adjust to a learning rate of 6e-6 and a batch size of 1 for 20 epochs. In both stages, all parameters of the ESM-2 model are fine-tuned.

4.7 Ablation studies

In order to verify the effectiveness of the different modules, we conduct ablation studies on the SFT-I and ENV embedding. Specifically, we set up four groups of experiments:

- **CLS embedding only**: To assess the effectiveness of other modules, we design a basic model that only performs supervised fine-tuning II and doesn't include the chemical environment simulation module, which means we directly follow the Equation.6 to predict $\Delta\Delta G$.
- **CLS embedding + ENV embedding**: This configuration is designed to test the effectiveness of the chemical environment simulation module. In this setting, we do not perform Supervised Fine-tuning I, focusing solely on the impact of adding the ENV embedding to the CLS embedding.
- **CLS embedding + SFT-I**: To verify the effectiveness of supervised fine-tuning I, we perform supervised fine-tuning I (SFT-I) on the basis of **CLS embedding only** configuration.
- **CLS embedding + ENV embedding + SFT-I**: In this configuration, the ENV embedding and SFT-I using high-throughput datasets are adopted simultaneously.

For a fair comparison, all models with different configurations are trained on the S2648 dataset and tested separately on S669 direct and S669 reverse.

4.8 Calculation and visualization of attention scores

4.8.1 Chemical environment restoration of Protein G-B1 domain

Given the mutation position pos_{mut}, we can obtain the attention score vector A_{mut} corresponding to this position. In order to observe the recovery degree of S3C to the local chemical environment, we select Protein G-B1 domain for case analysis. We consider F52 as the residue at the mutated position to obtain the corresponding A_{mut}, where the top 10 most relevant residues are identified (Equation 11), excluding itself and any special tokens, for visualization purposes.

4.8.2 Pairwise mutations of four enzyme-engineering cases

To analyze S3C's perception of interaction reconstruction, four cases of enzyme engineering 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 wild-type protein and mutated protein:

$$\text{score}_w(i, j) = A_w(i, j) \tag{14}$$

$$\text{score}_m(i, j) = A_m(i, j) \tag{15}$$

$$\Delta\text{score(i, j)} = \text{score}_m(i, j) - \text{score}_w(i, j) \tag{16}$$

where A_w is the pooled attention map outputted by the last transformer block of ESM-2 of wild-type protein and A_m is that of mutated protein.

4.9 Rare beneficial mutations mining for directed evolution

Accurately predicting rare beneficial mutations in the huge evolutionary space can significantly improve the efficiency of directed evolution for protein engineering. In order to evaluate the ability of S3C to mine rare beneficial mutations, we select the same number of beneficial mutations and harmful mutations to form a balanced test set from a specific case of *Is*PETase engineering with improved thermal stability. The voting method proposed by Hie et al. [31] is adopted to compare with S3C. This voting method relies on voting from multiple protein language models, employing six ESM models (esm1b_t33_650M_UR50S, esm1v_t33_650M_UR90S_1, esm1v_t33_650M_UR90S_2, esm1v_t33_650M_UR90S_3, esm1v_t33_650M_UR90S_4, esm1v_t33_650M_UR90S_5). The above six ESM models are represented by $P = \{p_1, p_2, ..., p_6\}$ and the conditional likelihoods produced by them are represented by $p_j(s^i|S)$. Specifically, we first calculate whether an amino acid substitution results in a higher language model likelihood than the wild-type residue at the mutation position, as follows:

$$\mathcal{M}(p_j, s_{mut}^{\text{pos}_{mut}}) = \mathbb{I}\{\frac{p_j(s_{mut}^{\text{pos}_{mut}}|S)}{p_j(s_{wild}^{\text{pos}_{mut}}|S)} > 1\},\ j = 1, 2, ..., 6 \tag{17}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

For the mutated residue $s_{mut}^{\text{pos}_{mut}}$, we calculate:

$$f(s_{mut}^{\text{pos}_{mut}}) = \sum_{p_j \in P} \mathcal{M}(p_j, s_{mut}^{\text{pos}_{mut}}) \tag{18}$$

Subsequently, if more than three models agree that a mutation is beneficial ($f(s_{mut}^{\text{pos}_{mut}}) > 3$), the voting method deems it to be a beneficial mutation. If not, the mutation is considered harmful.

For S3C, we judge the category of a mutation through the predicted $\Delta\Delta G$, where a negative $\Delta\Delta G$ indicates that the corresponding mutation is beneficial to protein stability, while a positive $\Delta\Delta G$ indicates the opppsite.

4.10 Evaluation metrics

4.10.1 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:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{19}$$

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

4.10.2 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:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{20}$$

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

4.10.3 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:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

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.

Data availability

S2648, S2298, S669, Myoglobin, PTEN and TPMT, and S349 are freely available from previous studies [18, 21, 35–39]. The processed high-throughput dataset and S6070 can be downloaded from https://github.com/ZhiweiNiepku/S3C.

Code availability

Relevant code and models are available at https://github.com/ZhiweiNiepku/S3C.

References

[1] Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Current opinion in structural biology* **14**, 70–75 (2004).

[2] Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).

[3] Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* **11**, 572–582 (2010).

[4] Studer, R. A., Dessailly, B. H. & Orengo, C. A. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical journal* **449**, 581–594 (2013).

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

[6] 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).

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

[8] 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).

[9] Fang, J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings in bioinformatics* **21**, 1285–1292 (2020).

[10] Parthiban, V., Gromiha, M. M. & Schomburg, D. Cupsat: prediction of protein stability upon point mutations. *Nucleic acids research* **34**, W239–W242 (2006).

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

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

[13] 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).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

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

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

[24] 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).

[25] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

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

[27] Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Science* **31**, e4379 (2022).

[28] Stevens, A. O. & He, Y. Benchmarking the accuracy of alphafold 2 in loop structure prediction. *Biomolecules* **12**, 985 (2022).

[29] McDonald, E. F., Jones, T., Plate, L., Meiler, J. & Gulsevin, A. Benchmarking alphafold2 on peptide structure prediction. *Structure* **31**, 111–119 (2023).

[30] Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems* **34**, 29287–29303 (2021).

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

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

[33] 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).

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

[35] 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).

[36] 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).

[37] Ordway, G. A. & Garry, D. J. Myoglobin: an essential hemoprotein in striated muscle. *Journal of Experimental Biology* **207**, 3441–3446 (2004).

[38] Wang, S., Tang, H., Shan, P., Wu, Z. & Zuo, L. Pros-gnn: predicting effects of mutations on protein stability using graph neural networks. *Computational Biology and Chemistry* **107**, 107952 (2023).

[39] Lv, X. *et al.* Accurately predicting mutation-caused stability changes from protein sequences using extreme gradient boosting. *Journal of chemical information and modeling* **60**, 2388–2395 (2020).

[40] Punta, M. *et al.* The pfam protein families database. *Nucleic acids research* **40**, D290–D301 (2012).

[41] Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42**, D222–D230 (2014).

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

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

[44] 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).

[45] Kan, Y., He, L., Luo, Y. & Bao, R. Ispetase is a novel biocatalyst for poly (ethylene terephthalate)(pet) hydrolysis. *ChemBioChem* **22**, 1706–1716 (2021).

[46] Nisticò, R. Polyethylene terephthalate (pet) in the packaging industry. *Polymer Testing* **90**, 106707 (2020).

[47] Bell, E. L. *et al.* Directed evolution of an efficient and thermostable pet depolymerase. *Nature Catalysis* **5**, 673–681 (2022).

[48] 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).

[49] 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).

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

[51] 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).

[52] Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nature Reviews Genetics* **16**, 379–394 (2015).

[53] 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).

[54] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic acids research* **32**, D120–D121 (2004).

[55] Xavier, J. S. *et al.* Thermomutdb: a thermodynamic database for missense mutations. *Nucleic acids research* **49**, D475–D479 (2021).

[56] Stourac, J. *et al.* Fireprotdb: database of manually curated protein stability data. *Nucleic acids research* **49**, D319–D324 (2021).

[57] Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

Acknowledgements

This work is financially supported by the National Key R&D Program of China (No. 2022ZD0118201, 2020YFA0908100, 2023YFF1204401), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465, 61825101, 62088102, 21991132, 21925102, 92056118, 22331003, 22301010), Shenzhen Medical Research Fund (No. B2302037), Beijing National Laboratory for Molecular Sciences (BNLMS-CXXM-202006), and the Major Key Project of PCL (PCL2021A13). The work is also supported by the National Center for Protein Sciences at Peking University. W.-B.Z. acknowledges Bayer Pharma for the Bayer Investigator Award. We appreciate the support of Peng Cheng Cloud-Brain and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China.

Competing interests

The authors declare no competing interests.

Supplementary information

S1 Test performance on benchmark dataset S669

Model	S669 direct		S669 reverse	
	RMSE	**PCC**	**RMSE**	**PCC**
SDM	1.67	0.41	2.16	0.13
mCSM	1.54	0.36	-	-
DUET	1.52	0.41	2.14	0.23
PoPMuSiC	1.51	0.41	2.09	0.24
FoldX	2.30	0.22	-	-
Rosetta	2.70	0.39	2.68	0.40
PopMusicSym	1.51	0.41	-	-
INPS	1.52	0.43	1.53	0.43
SAAFEC-SEQ	1.54	0.36	-	-
MAESTRO	1.44	0.50	2.10	0.20
DDGun	1.72	0.41	1.75	0.38
DDGun3D	1.60	0.43	1.62	0.41
ACDC-NN-Seq	1.53	0.42	1.53	0.42
PROST	1.46	0.47	-	-
PROSTATA	1.45	0.49	1.45	0.49
ThermoMPNN	1.52	0.43	-	-
Mutate Everything	1.38	0.56	1.48	0.48
ProSTAGE	1.36	0.57	**1.38**	**0.55**
S3C	**1.33**	**0.61**	1.46	**0.55**

Table S1 Test performance on benchmark dataset S669.

S2 Test performance on S349, Myoglobin, and PTEN and TPMT

Model	Myoglobin		PTEN and TPMT		S349	
	RMSE	**PCC**	**RMSE**	**PCC**	**RMSE**	**PCC**
SDM	-	-	-	-	1.81	0.52
mCSM	-	-	-	-	1.08	0.73
IMutant	-	-	2.07	0.23	-	-
STRUM	-	-	1.62	0.32	-	-
DUET	-	-	-	-	1.13	0.71
PoPMuSiC	-	-	-	-	1.67	0.67
SDM2	-	-	-	-	1.29	0.61
INPS	-	-	1.63	0.42	1.26	0.68
EASE-MM	-	-	1.81	0.39	-	0.56
BoostDDG	0.89	0.63	1.78	0.44	-	-
SAAFEC-SEQ	-	-	-	-	0.96	0.78
ACDC-NN-Seq	0.97	0.56			-	-
PROST	**0.84**	0.65	1.74	0.44	0.96	0.79
PROSTATA	0.85	0.64	-	-	-	-
S3C	0.85	**0.69**	**1.51**	**0.51**	**0.92**	**0.81**

Table S2 Test performance on S349, Myoglobin, and PTEN and TPMT.

S3 Module ablation study on benchmark dataset S669

Model	S669 direct		S669 reverse	
	RMSE	**PCC**	**RMSE**	**PCC**
PROSTATA (S2648)	1.45	0.48	1.69	0.44
ProSTAGE (S2648)	**1.41**	0.51	**1.42**	0.49
Ours (CLS)	1.50	0.42	1.60	0.41
Ours (CLS+ENV)	1.44	0.47	1.57	0.44
Ours (SFT-I+CLS)	1.44	0.51	1.50	0.49
Ours (SFT-I+CLS+ENV)	**1.41**	**0.53**	1.46	**0.51**

Table S3 Module ablation study on benchmark dataset S669.

S4 Attention scores of pairwise mutations

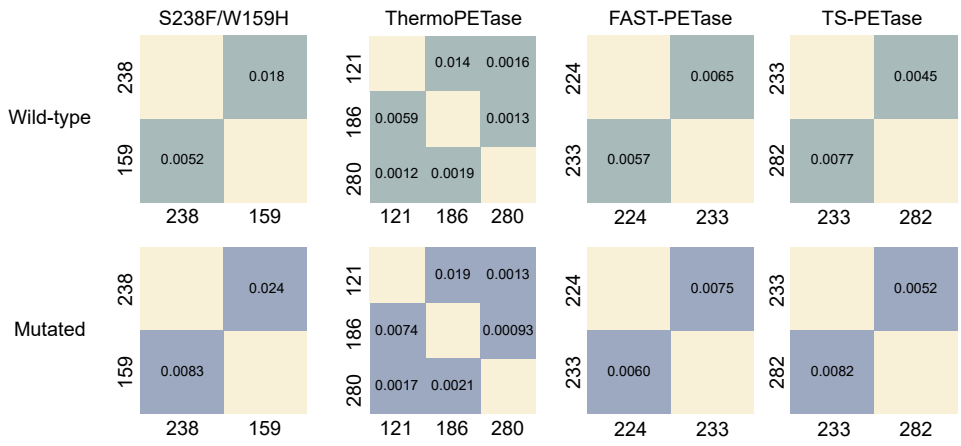

Fig. S1 Attention scores of pairwise mutations in four engineering cases of *Is*PETase with improved thermal stability.

S5 Test performance for directed evolution

Model	Precision	Recall	Accuracy
Voting	0.60	0.43	0.57
S3C	**0.64**	**0.67**	**0.64**

Table S4 Test performance for directed evolution in a specific engineering case of *Is*PETase with improved thermal stability.

S6 Hyperparameters

Hyperparameter	Value
Learning rate in SFT-I	5e-6
Learning rate in SFT-II	6e-6
Batch size in SFT-I	3
Batch size in SFT-II	1
Number of epochs in SFT-I	15
Number of epochs in SFT-II	20
Weight decay	0.01
Parameters of ESM-2	650M

Table S5 Hyperparameters of the model.