JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

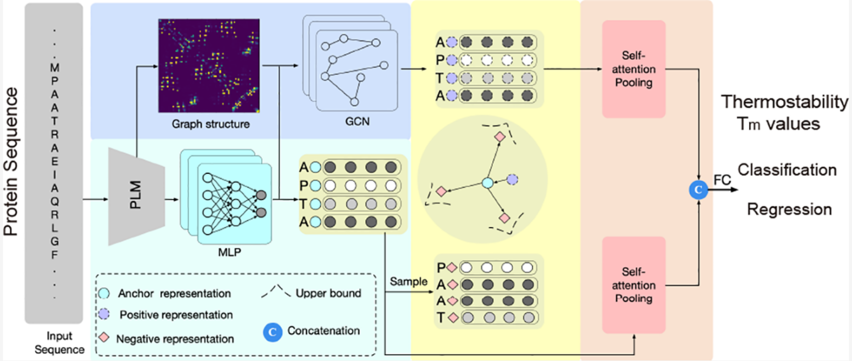
ProCeSa: Contrast-Enhanced Structure-Aware Network for Thermostability Prediction with Protein Language Models

Feixiang Zhou, Shuo Zhang, Huifeng Zhang,* and Jian K. Liu*

Cite This: *J. Chem. Inf. Model.* 2025, 65, 2304−2313

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Proteins play a fundamental role in biology, and their thermostability is essential for their proper functionality. The precise measurement of thermostability is crucial, traditionally relying on resource-intensive experiments. Recent advances in deep learning, particularly in protein language models (PLMs), have significantly accelerated the progress in protein thermostability prediction. These models utilize various biological characteristics or deep representations generated by PLMs to represent the protein sequences. However, effectively incorporating structural information, based on the PLM embeddings, while not considering atomic protein structures, remains an open and formidable challenge. Here, we propose a novel Protein Contrast-enhanced Structure-Aware (ProCeSa) model that seamlessly integrates both sequence and structural information extracted from PLMs to enhance thermostability prediction. Our model employs a contrastive learning scheme guided by the categories of amino acid residues, allowing it to discern intricate patterns within protein sequences. Rigorous experiments conducted on publicly available data sets establish the superiority of our method over state-of-the-art approaches, excelling in both classification and regression tasks. Our results demonstrate that ProCeSa addresses the complex challenge of predicting protein thermostability by utilizing PLM-derived sequence embeddings, without requiring access to atomic structural data.

1. INTRODUCTION

Protein stability is a fundamental aspect that affects the structure, function, and activity of proteins, playing a crucial role in various biological processes.[1] The ability of proteins to maintain their native conformation and activity under high-temperature conditions is of great importance in various fields, including biotechnology,[2] enzyme engineering,[3] and drug design.[4] Thermostable proteins have attracted significant attention due to their potential applications in industrial production and as biocatalysts for high-temperature reactions. Consequently, the accurate prediction of protein thermostability has emerged as a crucial effort to leverage the full potential of proteins in various biotechnological applications.

The stability of proteins is influenced by a multitude of factors, including intrinsic and extrinsic components. Intrinsic factors involve biological characteristics, such as amino acid distribution,[5] dipeptide composition,[6,7] salt bridges,[8,9] hydrogen bonding,[10] and hydrophobic interactions,[11] all of which contribute to the overall stability of the three-dimensional structures of proteins. Additionally, extrinsic factors such as pH, ionic strength, and temperature play a crucial role in modulating protein stability.[12,13] Among these factors, temperature holds particular significance because of its direct impact on the conformational dynamics of proteins, making it a key determinant of protein thermostability.

The protein thermostability can be measured by biological experiments, which are expensive, time-consuming, and labor-intensive.[14] However, in recent years, computational methods[12,15−17] have gained prominence in the prediction of protein thermostability with higher throughput and lower

Received: October 4, 2024
Revised: February 14, 2025
Accepted: February 17, 2025
Published: February 24, 2025

https://doi.org/10.1021/acs.jcim.4c01752

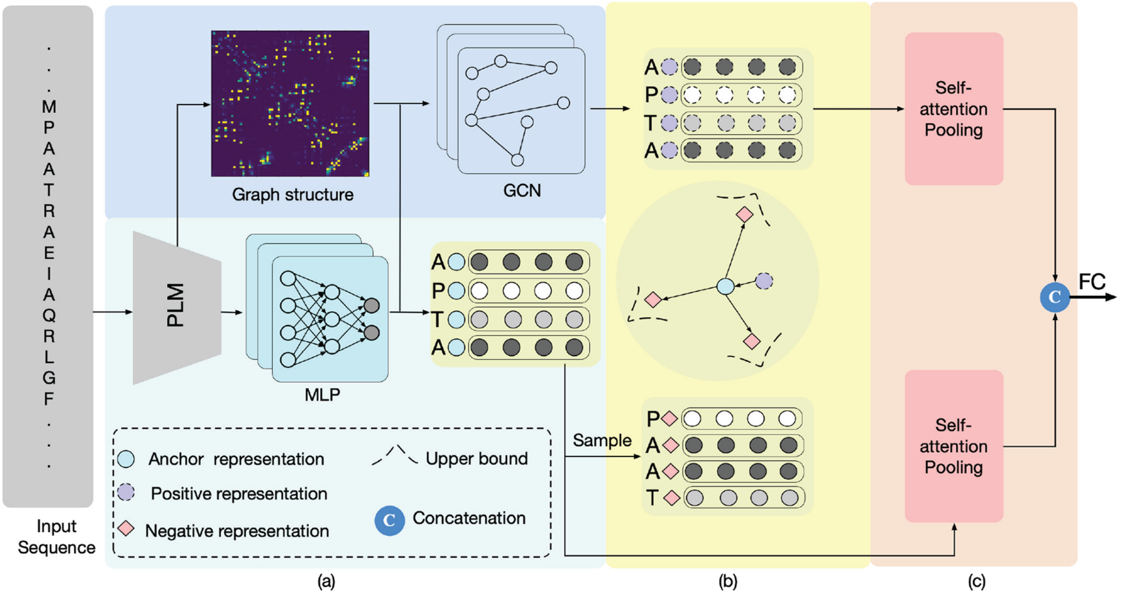


Figure 1. Overview of the proposed ProCeSa for protein thermostability prediction. (a) Sequence and Structure information encoding. A pretrained PLM is first employed to extract sequence features and generate a contact map, representing the graph structure of the protein sequence. These sequence features are further processed through an MLP to produce high-level sequence representations, which are then integrated with the graph structure using a GCN to effectively capture the intrinsic structural relationships among amino acid residues. (b) Contrast-enhanced representation learning. Building on the structure and sequence representations, a contrastive learning scheme is developed, guided by amino acid residue types, to enhance the overall representation quality. (c) Structure-sequence feature aggregation. The enhanced representations are then aggregated using self-attention pooling and subsequently fused to predict protein thermostability.

resource costs compared to traditional experimental approaches. Most current state-of-the-art computational approaches rely on different types of features, such as physicochemical properties,[17,18] amino acid categories,[17] and sequence-based biological features,[16,19] which can then be fed into traditional machine learning models or deep learning networks to predict protein thermostability. The amino acid composition and the dipeptide composition were adopted as inputs to the support vector machine (SVM) to predict protein classes (thermophilic or mesophilic) in a small data set.[20] By combining the biological characteristics of seven groups of protein sequences, a multilayer perceptron (MLP) was used to distinguish thermophilic proteins from mesophilic proteins.[19] In addition to these approaches, which focus on the classification of the thermostability of proteins, recent work has emerged on thermostability regression. ProTstab2 used gradient boosting to predict protein melting temperatures.[21] Similarly, DeepET was used to estimate optimal growth temperatures of source organisms using a large data set of 3 million enzymes in a wide range of organisms, where biological characteristics were extracted as input to the model.[16] Although these methods have achieved promising results in thermostability prediction, they mainly rely on customized biological or physicochemical features, where some irrelevant and redundant features would affect the prediction performance.

Traditional approaches to protein thermostability prediction rely heavily on detailed structural information, particularly when calculating the relative stability changes ($\Delta\Delta G$) between wild-type and mutated states.[22,23] These methods require atomic-level structural data to evaluate the conformational energy differences. However, a fundamental challenge exists: while protein sequence databases are vast, high-quality structural data remain scarce for most proteins. Recent breakthroughs in protein structure prediction, particularly AlphaFold2,[24] have inspired a new approach to address this limitation. Since these models have been trained on large-scale data sets to extract structural information from sequences, researchers have attempted to leverage the powerful feature extraction capabilities. By using the high-dimensional extracted features as input to downstream neural networks, they aim to predict protein thermostability.[25,26] However, studies have shown that thermostability predictions are highly sensitive to structural accuracy, and even cutting-edge structure prediction models cannot consistently provide the precision needed for reliable thermostability prediction.[27] Therefore, it is necessary to develop models that directly predict the melting temperature (T_m) from sequence information without relying on detailed structural features. Such models are highly demanding for facilitating protein engineering without expensive wet-lab experiments.[28]

With advances in protein language models (PLMs) that were trained on hundreds of millions of natural protein sequences,[29−32] transfer learning has been used to extract protein representations generated by language model encoders. Such informative representations have already been shown to be suitable inputs for various predictive tasks.[33,34] A large-scale protein data set HotProtein includes temperature annotations related to thermostability was proposed.[15] Alongside this, a novel learning model was introduced, together with the HotProtein data set. By combining the ESM-1B[29] with the proposed feature augmentation and pretraining schemes, this model improves the prediction performance. However, the pretraining approach can be complex because it requires protein language model ESM-1B and ESM-IF1[35] simultaneously to process sequence and 3D coordinate information, respectively. Additionally, the existing methods based on PLMs have limited ability to capture both sequence and spatial structure information on protein sequences.

In this paper, we propose a novel protein contrast-enhanced structure-aware (ProCeSa) model for thermostability prediction from the generated sequence representations and contact maps by PLMs. The contact maps represent pairwise interactions or contacts between amino acid residues within a protein's 3D structure. Specifically, we first leverage an MLP to extract high-level sequence information based on pretrained sequence representations, followed by designing a graph convolutional network (GCN)[36] to learn local and global structural features. Afterward, we propose a contrastive learning scheme to enhance the learned sequence and structure representations by discriminating the similarity and dissimilarity between different amino acid residues. As shown in Figure 1, we consider the structure and sequence representations as positive and anchor representations, respectively, followed by constructing the negative representation by sampling the specific amino acid representation from the anchor representation. For positive pairs, we sample the corresponding amino acids from both sequence representation and structure representation, enabling the model to learn the inherent relationships between sequence and structural features that influence protein stability. For negative pairs, we strategically sample amino acids with labels different from the sequence representation, which helps the model distinguish between sequence characteristics associated with varying degrees of thermostability. This sampling and learning process enhances the model's ability to capture thermostability-related features through joint sequence-structure encoding. The enhanced sequence and structure representations are then aggregated by self-attention pooling to generate feature embeddings associated with protein thermostability.

2. PROPOSED METHOD

Given a protein sequence consisting of L amino acids, its spatial structure can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$. $\mathcal{V} = \{v_1, v_2, \cdots, v_L\}$ is the set of all nodes in the protein sequence, $\mathcal{E}$ represents the edge set $\mathcal{E}_S = \{(v_n, v_m)|1 \leq n, m \leq L\}$ that describes the relationship between any pair of amino acids (v_n, v_m). $\mathcal{X} = \{x_1, x_2, \cdots, x_L\}$ is a set of node features, which is represented as a matrix $\mathbf{X} \in \mathbb{R}^{L\times D}$. $\mathcal{E}$ of the protein sequence can be formulated as an adjacency matrix $\mathbf{A} \in \mathbb{R}^{L\times L}$ where the element $a_{n,m}$ reflects the correlation strength between v_n and v_m. Here, we explore two tasks related to protein thermostability prediction, i.e., thermostability classification and regression. For the classification task, each graph (Protein sequence) is associated with a class label $y \in \{1, \cdots, \text{and } C\}$ where C denotes the number of classes. For the regression task, each graph has a ground-truth temperature value (e.g., growth temperature and melting temperature). As shown in Figure 1, our proposed ProCeSa jointly utilizes sequence and structure information to extract rich representations related to thermostability and then incorporates a novel contrastive learning scheme into our model to learn more discriminative representations.

2.1. Sequence and Structure Information Encoding.

Similar to most existing work,[37,34] we adopt an MLP to extract a high-level representation with sequence information from per-trained PLMs. Formally, given the pretrained protein feature $\mathbf{P} \in \mathbb{R}^{L\times F}$, sequence representation $\mathbf{X} \in \mathbb{R}^{L\times D}$ can be formulated as

$$\begin{aligned} \mathbf{P}^{(l+1)} &= \sigma(\mathbf{P}^{(l)}\mathbf{W}^{(l)} + b^{(l)}) \\ \mathbf{X} &= \mathbf{P}^{(l+1)}\mathbf{W}^{(l+1)} + b^{(l+1)} \end{aligned} \tag{1}$$

where $\mathbf{P}^{(l)} \in \mathbb{R}^{L\times D^l}$ denotes the input hidden representation from the previous layer ($\mathbf{P}^{(0)} = \mathbf{P}$) and $\mathbf{W}^{(l)} \in \mathbb{R}^{D^l\times D^{l+1}}$ and $b^{(l)}$ are trainable parameters. σ is an activation function, such as ReLU(·). F is the dimension of pretrained features. The MLP maps the feature dimensions from F to D, where D is the dimension of the learned sequence representations.

To encode structure information within the protein sequence, we apply the widely used GCN[36] to model structural relations among amino acids, which can be defined as

$$\begin{aligned} \mathbf{H}^{(l+1)} &= \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \\ \mathbf{G} &= \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l+1)}\mathbf{W}^{(l+1)}) \end{aligned} \tag{2}$$

where $\mathbf{H}^{(0)} = \mathbf{X}$ and $\mathbf{G} \in \mathbb{R}^{L\times D}$ is the structure representation. $\hat{\mathbf{A}} = \mathbf{D}^{-1/2}\tilde{\mathbf{A}}\mathbf{D}^{-1/2} \in \mathbb{R}^{L\times L}$ represents a symmetrically normalized adjacency matrix. $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the diagonal matrix, where $\mathbf{D}_{ii} = \sum_j(\tilde{\mathbf{A}}_{i\,j}) + c$, and c is a small constant avoiding empty rows. $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_L$ where $\mathbf{I}_L$ is the identity matrix.

2.2. Contrast-Enhanced Representation Learning.

Recently, contrastive learning has become increasingly popular across a range of computer vision domains, including image representation learning,[38,39] video representation learning,[40,41] time series representation learning.[42,43] Among these approaches, contrastive loss[15,38,39] and triplet loss[44,43] have emerged as two prevalent loss functions for conducting contrasting tasks. The foundational concept underlying the former is to bring together representations of augmented samples from the same image or video clips (i.e., positive pairs), while simultaneously pushing apart those originating from different instances (i.e., negative pairs). The latter shares the same objective but involves defining a triplet consisting of anchor, positive, and negative pairs. The positive pairs composed of anchor and positive representations are encouraged to be close, whereas the negative pairs comprising anchor and negative representations should be far away.

Different from,[15] our contrastive learning strategy is built on the triplet loss where we explicitly construct the anchor, positive and negative representations based on the sequence representation $\mathbf{X}$ and structure representation $\mathbf{G}$ for enhancing representation learning. In particular, we utilize prior information on the amino acid labels to guide the construction of positive and negative pairs. Specifically, we first sample a fixed number of amino acids from $\mathbf{X}$ and $\mathbf{G}$ to form $\mathbf{X}_s$ and $\mathbf{G}_s$ (In experiments, we sample L_s amino acids in each mini-batch). This is because it is too computationally expensive to measure each amino acid in a mini-batch. Then we identify $\mathbf{G}_s$ and $\mathbf{X}_s$ as positive and anchor representations, respectively, and our goal is to pull them (i.e., positive pairs) together. Inspired by,[43] given $\mathbf{G}_s$ and $\mathbf{X}_s$, we formalize the objective function of similarity learning between amino acids as

$$\begin{aligned} \mathbf{\Psi}_p &= -[\log(\sigma(\mathbf{X}_s^T\mathbf{G}_s/\omega))] \odot \mathbf{I} \\ &= \begin{bmatrix} \mathbf{\Psi}_p[1,1] & & \\ & \ddots & \\ & & \mathbf{\Psi}_p[L_s, L_s] \end{bmatrix} \end{aligned} \tag{3}$$

where $\odot$ represents the element-wise product. $\mathbf{I} \in \mathbb{R}^{L_s \times L_s}$ represents the identity matrix used to select pairwise amino acids with the same index, and $0 < \omega < 1$ is a scale factor to adjust the correlation between vectors. $\mathbf{\Psi}[,]$ is the element of matrix $\mathbf{\Psi}$.

Afterward, we can obtain the overall loss $\mathcal{L}_p$ between positive pairs below:

$$\mathcal{L}_p = \frac{1}{L_s} \sum_{i=1}^{L_s} (\Psi_p[i, i]) \tag{4}$$

In eq 3, we construct positive pairs based on the inherent similarity between sequence and structure representations for direct contrast. The representations of pairwise amino acids with the same index in $\mathbf{G}_s$ and $\mathbf{X}_s$ indicate the same amino acid. In contrast, the construction of negative pairs relies on prior amino acid labels. In detail, for each amino acid in anchor representation $\mathbf{X}_s$, we sample K amino acids that have different labels to it from $\mathbf{X}_s$ to form the set of negative representations, i.e., $\{\mathbf{Z}_s^1, \mathbf{Z}_s^2, \cdots, \mathbf{Z}_s^K\}$. The element $\mathbf{Z}_s^k$ can be represented as $\mathbf{Z}_s^k[i] \in \{\mathbf{X}_s[j] | l[j] \neq l[i], 1 \leq j \leq L_s\}$ where $l[j]$ is the class label of the j-th amino acid residue. Therefore, the negative representation $\mathbf{Z}_s^k$ and the anchor representation $\mathbf{X}_s$ should be separate. The function $\mathbf{\Psi}_n^k$ to be minimized with the corresponding loss $\mathcal{L}_n$ is expressed as

$$\mathbf{\Psi}_n^k = -[\log(\sigma(-\mathbf{X}_s^T \mathbf{Z}_s^k / \omega))] \odot \mathbf{I} \tag{5}$$

$$\mathcal{L}_n = \frac{1}{L_s} \sum_{k=1}^{K} \sum_{i=1}^{L_s} (\Psi_n^k[i, i]) \tag{6}$$

Afterward, we can obtain the contrastive loss $\mathcal{L}_{\text{con}}$ as follows:

$$\mathcal{L}_{\text{con}} = \mathcal{L}_p + \mathcal{L}_n \tag{7}$$

2.3. Structure-Sequence Feature Aggregation.

In the field of protein analysis, it is crucial to acknowledge the inherent diversity of protein lengths. Proteins can range from relatively short sequences to incredibly long ones, and this variation poses a challenge when aiming to capture their overall characteristics. To effectively characterize a protein, we need to aggregate the distinctive features of all its constituent amino acids into a unified protein-level representation.[45] This aggregation process allows us to encapsulate the essential information contained within the entire protein sequence, enabling a more comprehensive and meaningful analysis of its structural and functional attributes. Similar to previous studies,[46,47] we employ self-attention pooling to aggregate all amino acid features. Given the structure representation $\mathbf{G}$, the attention matrix $\mathbf{A}_g \in \mathbb{R}^{D'' \times L}$ can be computed as

$$\mathbf{A}_g = \text{softmax}(\mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{G}^T)) \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D' \times D}$ and $\mathbf{W}_2 \in \mathbb{R}^{D'' \times D'}$ are two learnable parameters. D'' is the number of groups of attention vectors, each of which evaluates the association of each residue with thermostability from a specific perspective. The softmax$(\cdot)$ is performed along the second dimension of its input. We then compute the aggregated protein embedding $\tilde{\mathbf{G}}$ by multiplying the attention matrix $\mathbf{A}_g$ and the encoded structure representation $\mathbf{G}$, and average all groups of attention vectors as follows:

$$\begin{aligned} \tilde{\mathbf{G}} &= \mathbf{A}_g \mathbf{G} \\ \bar{\mathbf{G}} &= \frac{1}{H} \sum_{i=1}^{H} \tilde{\mathbf{G}}[i, :] \end{aligned} \tag{9}$$

where $\tilde{\mathbf{G}}[i,:]$ denotes the i-th row of the matrix $\tilde{\mathbf{G}}$. $\bar{\mathbf{G}} \in \mathbb{R}^{1 \times D}$ is the final graph-level representation. Similarly, we apply the same self-attention pooling to sequence representation $\mathbf{X}$ to obtain the final representation $\bar{\mathbf{X}} \in \mathbb{R}^{1 \times D}$. Finally, we combine the two types of representations (i.e., $\bar{\mathbf{G}}$ and $\bar{\mathbf{X}}$) by the concatenation operation, followed by adopting a fully connected layer to generate prediction results. By combining the traditional prediction loss (cross-entropy loss for classification and root-mean-square error (RMSE) for regression) with our proposed contrastive loss, we are able to produce more accurate results. The overall loss is defined as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{pre}} \tag{10}$$

where α is the weight of $\mathcal{L}_{\text{con}}$.

3. EXPERIMENTAL SETUP

3.1. Data Sets.

HotProtein data set is a large-scale protein data set with organism-level temperature annotations for both classification and regression tasks.[15] The detailed experimental protocol for obtaining the values of thermostability can be found in ref 14. Briefly, the thermostability is measured as a lower bound of the protein's melting temperature, which can be used for regression prediction. It contains 182 K amino acid sequences of proteins from 230 different species, covering five thermostability types, e.g., Cryophilic (−20−5 Celsius), Psychrophilic (5−25 Celsius), Mesophilic (25−45 Celsius), Thermophilic (45−75 Celsius), and Hyperthermophilic (>75 Celsius). HotProtein consists of 4 distinct subsets with different scales, including HP-S^2C2 (2 classes), HP-S^2C5 (5 classes), HP-S (5 classes) and HP-SC2 (2 classes) as developed in ref 15. Briefly, four distinct testbeds are derived from the HotProtein data set that differed in scale: (1) HP-S^2C2 has 1026 "hot" (≥45 °C) and 939 "cold" (<45 °C) proteins from 61 and 4 species, respectively. (2) HP-S^2C5 consists {73, 387, 195, 196, 189} proteins sampled from the five categories, from Cryophilic to Hyperthermophilic. (3) HP-S is the entire sequence HotProtein data set with {6390, 34946, 30333, 79087, 31549} sequences from {3, 32, 31, 116, 48} different species, of five classes ordered from Cryophilic to Hyperthermophilic. (4) HP-SC2 is a 2-class variant created by merging Hyperthermophilic and Thermophilic as "hot" class and the other three as "cold" class. Given their sample size, HP-S^2C2/C5 and HP-S/SC2 are regarded as small- and large-scale data sets.

The second data set used in this work is another large-scale protein thermostability data set developed in DeepStabP,[48] derived from the Meltome Atlas.[14] It contains 35,112 protein sequences retrieved from UniProt[49] with annotated melting temperatures from diverse species.

3.2. Evaluation Metrics.

The performance of the methods in classification is evaluated with two measures as fellows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

Table 1. Performance of Predicting Thermostability with Classification Tasks Using the HotProtein Dataset; Accuracy (%) Is Reported for All Three Datasets, and Precision (%) Is Calculated for the Two-Class Classification on HP-S^2C2 and HP-SC2

	HP-S^2C2		HP-S^2C5	HP-S	HP-SC2	
method	accuracy	precision	accuracy	accuracy	accuracy	precision
3D GCN[50]	78.88 ± 1.57	73.39 ± 2.76	67.40 ± 2.11	N/A	N/A	N/A
TAPE[51]	83.31 ± 1.10	76.42 ± 3.06	66.44 ± 2.30	64.75 ± 0.23	76.37 ± 0.25	80.64 ± 0.50
ESM-IF1[35]	79.08 ± 0.85	76.49 ± 3.96	58.75 ± 2.46	N/A	N/A	N/A
ESM-1B[29]	91.19 ± 0.47	84.18 ± 1.71	83.26 ± 1.54	69.50 ± 0.16	86.24 ± 0.22	88.14 ± 1.62
HotProtein[15]	92.36 ± 0.58	86.51 ± 1.67	86.25 ± 1.03	73.21 ± 0.13	87.57 ± 0.10	89.07 ± 1.29
ProCeSa-ESM-1B	92.94 ± 0.53	89.30 ± 2.83	87.02 ± 1.88	74.54 ± 0.02	87.87 ± 0.15	91.20 ± 0.48
ProCeSa-ESM-1B (+Cont)	93.39 ± 1.07	**91.38 ± 2.68**	87.50 ± 2.31	75.52 ± 0.10	88.13 ± 0.28	91.38 ± 0.43
ProCeSa-ESM-C	93.96 ± 0.34	89.57 ± 0.91	92.85 ± 0.06	80.37 ± 0.06	91.15 ± 0.16	**92.97 ± 1.38**
ProCeSa-ESM-C (+Cont)	**94.17 ± 0.06**	89.52 ± 0.40	**93.36 ± 0.45**	**80.53 ± 0.31**	**91.15 ± 0.14**	92.48 ± 1.72

Accuracy measures the overall correctness of the model across the entire data set. Precision evaluates a model's classification performance on a specific class, which represents the proportion of samples predicted as the positive class that are true positives. TP represents True Positives (correctly predicted hot proteins), TN represents True Negatives (correctly predicted cold proteins), FP represents False Positives (cold proteins incorrectly predicted as hot), and FN represents False Negatives (hot proteins incorrectly predicted as cold).

The Area under the receiver operating characteristic curve (AUC-ROC or AUC) evaluates the model's ability to distinguish between classes across various classification thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values, where

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{13}$$

The AUC ranges from 0 to 1, with 1 representing perfect classification and 0.5 indicating a random guess. A higher AUC reflects better model discrimination between classes of melting temperatures across different decision thresholds.

We evaluate the regression performance using three metrics: Pearson correlation coefficient, Spearman correlation coefficient, and R^2. The Pearson correlation coefficient measures the strength and direction of the linear relationship between two continuous variables. It quantifies this relationship by computing the covariance between the two variables divided by the product of their respective standard deviations, where its values range from −1 (perfect negative correlation) to 1 (perfect positive correlation). The formula for the Pearson correlation coefficient is defined as

$$\mathrm{Pearson} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[X^2] - \mu_X^2}\sqrt{E[Y^2] - \mu_Y^2}} \tag{14}$$

where $\mathrm{cov}(\cdot)$ is the covariance. σ_X and σ_Y are the standard deviations of X and Y (ground-truth and predicted values), respectively. μ_X and μ_Y are the mean values of X and Y. E is the expectation.

The Spearman correlation coefficient assesses the strength and direction of monotonic relationships between two variables. It achieves this by first converting the observed values of each variable into ranks and then calculating the Pearson correlation coefficient between these ranks. The formula for the Spearman correlation coefficient is as follows:

$$\mathrm{Spearman} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{15}$$

where d_i represents the differences between ranks for each pair of observations. n is the sample size.

R^2 measures how well the model fits the data. It represents the proportion of variance in the dependent variable explained by the independent variables. R^2 ranges from 0 to 1, where 1 indicates a perfect prediction. The formula is

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{16}$$

where y_i are the actual values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of actual values.

3.3. Implementation Details.

All models were trained on a single NVIDIA A100 40GB GPU using PyTorch. We employed the Adam optimizer with fixed learning rates throughout the training. All models were trained for 10 epochs, and we report the test set performance of the best-performing model selected across epochs. For model training, we used a batch size of 4 for HP-S^2C2 and HP-S^2C5, 8 for DeepStabP, and 512 for HP-S. The learning rates were set to 8e-5 for HP-S^2C2, 8e-4 for HP-S^2C5, and 1e-4 for HP-S and DeepStabP.

4. RESULTS

Using the large-scale HotProtein data set, we evaluated the performance of two configurations of models: ProCeSa, which purely integrates structural and sequential information, and ProCeSa(+Cont), which additionally performs contrastive learning of the two protein representations. The results show that our proposed ProCeSa has a better performance compared to those of other SOTA models.

We compare ProCeSa to five outstanding models: 3D GCN,[50] TAPE,[51] ESM-IF1,[35] ESM-1B,[29] and HotProtein.[15] For a fair comparison, we use the same data sets as in HotProtein.[15] The results of the 3D GCN, TAPE, ESM-IF1, ESM-1B, and HotProtein model are cited from the original HotProtein study.[15] We also compare the performance of ProCeSa installed with different encoders of ESM-1B and ESM-C.[52]

For classification tasks, accuracy is utilized as the primary evaluation metric for both two-class and five-class classification tasks, while binary precision is specifically calculated for the two-class classification scenario (Table 1). Across all subdata sets of the HotProtein benchmark, ProSeCa consistently outperforms all baseline models. Furthermore, by leveraging

Table 2. Performance of Predicting Thermostability with Regression Tasks Using the HotProtein Dataset[a]

	HP-S^2C2		HP-S^2C5		HP-S	
method	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
3D GCN[50]	0.490 ± 0.019	0.469 ± 0.019	0.291 ± 0.053	0.301 ± 0.074	N/A	N/A
TAPE[51]	0.432 ± 0.061	0.386 ± 0.065	0.367 ± 0.063	0.364 ± 0.047	0.504 ± 0.013	0.453 ± 0.031
ESM-IF1[35]	0.589 ± 0.040	0.547 ± 0.036	0.373 ± 0.036	0.377 ± 0.035	N/A	N/A
ESM-1B[29]	0.890 ± 0.018	0.893 ± 0.024	0.712 ± 0.043	0.804 ± 0.023	0.807 ± 0.001	0.809 ± 0.001
HotProtein[15]	0.906 ± 0.010	0.923 ± 0.012	0.754 ± 0.035	0.837 ± 0.019	0.823 ± 0.001	0.827 ± 0.003
ProCeSa-ESM-1B	0.930 ± 0.012	0.958 ± 0.013	0.856 ± 0.053	0.900 ± 0.039	0.852 ± 0	0.858 ± 0
ProCeSa-ESM-1B (+Cont)	0.933 ± 0.014	0.961 ± 0.015	0.861 ± 0.050	0.904 ± 0.045	0.855 ± 0.001	0.861 ± 0.001
ProCeSa-ESM-C	0.942 ± 0.001	0.969 ± 0.000	0.919 ± 0.000	0.949 ± 0.002	0.892 ± 0.000	0.897 ± 0.001
ProCeSa-ESM-C (+Cont)	**0.943 ± 0.001**	**0.969 ± 0.000**	**0.921 ± 0.002**	**0.954 ± 0.005**	**0.894 ± 0.001**	**0.900 ± 0.001**

[a]Correlation coefficients of Spearman and Pearson are reported for all three datasets; 95(%) confidence intervals are computed via the 10-fold evaluation on HP-S^2C2/C5 and 3 replicates on HP-S.

Table 3. Performance of Predicting Thermostability with Regression Tasks Using the Dataset in DeepStabP[a]

method	R^2	Spearman	Pearson	model parameter
DeepStabP (ProtT5-XL)[48] (reproduced)	0.792 ± 0.021	0.732 ± 0.005	0.897 ± 0.003	3B
ProCeSa-ESM-C (+Cont)	0.788 ± 0.004	0.717 ± 0.004	0.889 ± 0.004	600M

[a]R^2, Spearman, and Pearson correlation coefficients are reported. The approximate number of model parameters is listed for both models.

the advanced ESM-C encoder in conjunction with contrastive learning, ProSeCa demonstrates superior performance in the majority of cases.

For regression tasks, the model performance is evaluated using Spearman and Pearson correlation coefficients. Analyzing different subsets of the HotProtein data set (Table 2), ProSeCa equipped with the ESM-C encoder and contrastive learning achieves the best performance. These results highlight the effectiveness of the ESM-C encoder in enhancing the representation quality. Additionally, contrastive learning contributes to the improved modeling of both sequence and structural features, although the observed performance gains are relatively modest.

To evaluate the generalizability of the ProCeSa model, we tested its performance on another thermostability data set, DeepStabP[48] for the regression task. Spearman correlation, Pearson correlation, and R^2 values were calculated to ensure a fair comparison with the results reported in the original DeepStabP study[48] (Table 3). The ProCeSa-ESM-C model shows comparable performance compared to the DeepStabP model based on the ProtT5-XL PLM. Notably, ProCeSa achieves this level of performance with significantly fewer parameters (600 M vs 3B), highlighting its efficiency in capturing capabilities comparable to those of larger models.

HotProtein makes use of AlphaFold2 to predict protein 3D coordinates and feeds this 3D information to their model. Considering taking advantage of interactions between amino acids, we extract the graph information using ESM-series models following HybridGCN[53] and feed the extracted contact map into the GCN network. This allows our model to fully learn the interactions between amino acids, thereby improving the accuracy of predictions and the generalization ability of the model. On the other hand, contrastive learning can bring an increase in performance. These stable improvements demonstrate that our specially designed triplet-loss contrastive learning mode for two different modalities enables the model to explore the similarities and differences between the protein structure and sequence.

We compare the macro ROC curves of models on the HP-S^2C5 subdata set (Figure 2). Macro ROC is selected over micro ROC because it evaluates the performance of each class equally, which is important given the class imbalance in this data set. Our ProCeSa(+Cont) demonstrates superior performance compared with the other models. The performance difference between ProCeSa-ESM-C and ProCeSa-ESM-1B indicates that ESM-C extracts features more effectively than ESM-1B, leading to better prediction accuracy.

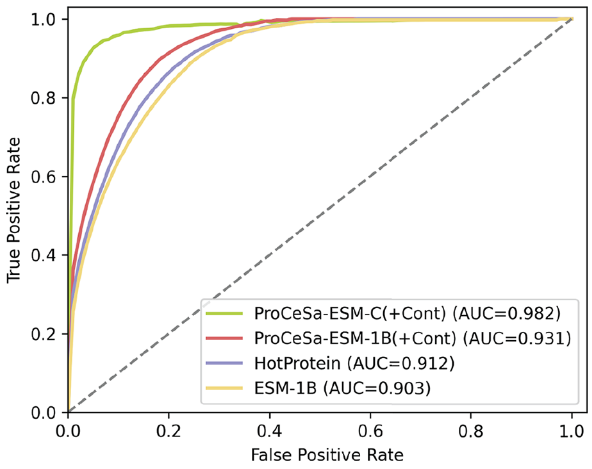


Figure 2. ROC curves performance comparison of ESM-1B, HotProtein, ProCeSa-ESM-1B, and ProCeSa-ESM-C models on HP-S^2C5.

Using macro ROC curves to evaluate performance across the five protein categories in HP-S^2C5 (Figure 3), we observe distinct trends. Most models achieve their highest AUC in the Cryophilic category with scores ranging from 0.997 to 0.998. However, ProCeSa-ESM-C achieves its highest AUC in the thermophilic category (0.994). This difference may stem from ProCeSa-ESM-C being pretrained on data that better captures features relevant to high-temperature proteins, allowing it to perform particularly well in the thermophilic category.

In the mesophilic category, ESM-1B and ProCeSa-ESM-C show the lowest AUC values (0.800 and 0.974, respectively). HotProtein and ProCeSa-ESM-1B perform the worst in the Thermophilic category, with AUC scores of 0.814 and 0.846,

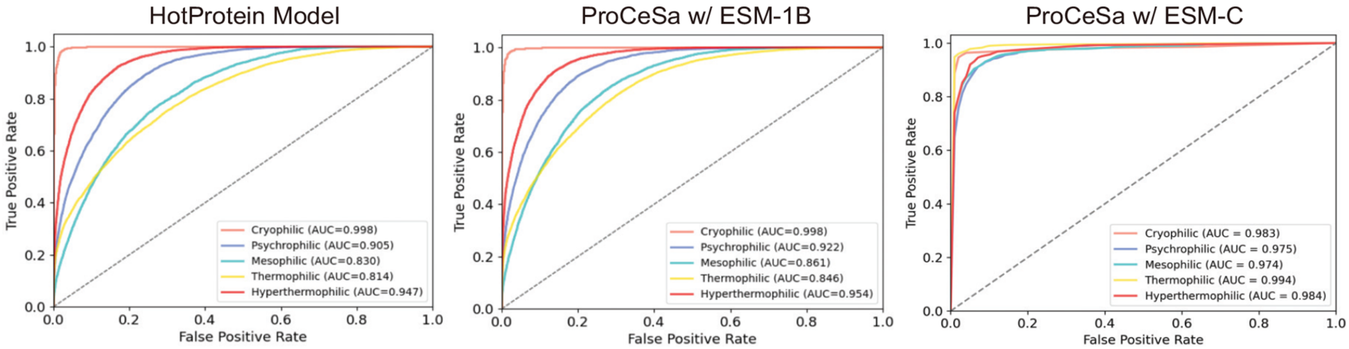


Figure 3. Individual ROC curves comparison of HotProtein, Procesa-ESM-1B, and Procesa-ESM-C on 5 classes of the HP-S^2C5 Data set.

respectively. Notably, ProCeSa-ESM-C demonstrates a balanced performance across all categories, suggesting its robustness in handling diverse protein types.

These findings highlight the importance of considering the specific protein category when training thermostability prediction models. Future work could focus on improving performance in individual categories.

5. DISCUSSION

In this work, we introduce ProCeSa, a novel deep-learning network model tailored specifically for predicting protein thermostability. ProCeSa uniquely leverages structural and sequential features extracted by large language models, pioneering a departure from conventional approaches that rely solely on one type of feature or on conventional biophysical features. By the integration of both types of features, our model capitalizes on the complementary nature of structural and sequential information, thus enriching the predictive capacity for protein thermostability.

Central to ProCeSa is the utilization of a contrastive learning scheme, which enables the exploration of implicit interactions between structural and sequence features. This approach not only enhances the model's ability to capture meaningful relationships within protein sequences and structures but also facilitates robust performance across diverse data sets.

Contrastive learning has been recently used for various tasks in protein engineering.[54−56] The primary benefit of this learning scheme is to reduce dependence on labeled data. In protein sequence analysis, labeled data can be scarce. Contrastive learning effectively utilizes unlabeled data, mitigating the need for extensive labeled data sets and facilitating the development of accurate models even in data-constrained scenarios, particularly for enzyme functional annotation, such as enzyme commission (EC) number.[57] Beyond data efficiency, contrastive learning enhances model robustness through controlled noise during training, helping models learn stable protein representations and reduce overfitting.[56] This approach also improves the discrimination ability between similar proteins, which is crucial for tasks like identifying true binding partners in drug-target interactions.[55] Models trained with contrastive objectives show better generalization across diverse protein families and maintain consistent performance on previously unseen sequences.[54]

Compared to the abundant protein sequence databases, high-quality experimentally determined structural data remain scarce for most proteins. Recent advancements in structure prediction from sequences, such as RoseTTAFold,[58,59] AlphaFold,[24,60] ESMfold,[31] and ESM Cambrian,[52] have made it possible to generate 3D atomic structures from virtually any protein sequence. Incorporating these predicted structures into computational models has been shown to enhance performance in general-purpose protein engineering tasks, as demonstrated by models such as S-PLM[61] and ProTrek,[62] as well as in specific applications like thermostability prediction, exemplified by SPIRED[25] and ProSTAGE.[26]

S-PLM, for instance, utilizes contact maps derived from physical distances in predicted 3D structures, while attention-based contact maps generated by ESM language models capture both physical contacts and functional relationships between residues, independent of the structure prediction accuracy. Predicted 3D structures have proven useful for estimating relative stability changes ($\Delta\Delta G$ or ΔT_m) between wild-type and mutated proteins.[22,23] However, recent studies have highlighted that thermostability predictions are highly sensitive to structural precision, and even state-of-the-art structure prediction models cannot consistently deliver the level of precision required for reliable $\Delta\Delta G$ calculations.[27] This highlights the need to develop models capable of directly predicting melting temperatures (T_m) from sequence information without relying on detailed structural features. Such models would significantly advance protein engineering by reducing dependence on expensive and time-consuming wet-lab experiments.[28]

Future advancements in thermostability prediction will depend on the integration of diverse methodologies, combining data from multiple levels of protein characterization.[63] Both the HotProtein and DeepStabP data sets are derived from the experimental Meltome Atlas,[14] which primarily reflects organismal optimal growth temperatures. While these temperatures show some correlation with protein thermostability, they are not always reliable proxies. Instead, they can be considered a lower bound for the melting temperature (T_m), a parameter that is relatively easy to measure through wet-lab experiments. In practical applications, however, the most useful predictor of thermostability is the change in the free energy ($\Delta\Delta G$). Unfortunately, all deep learning models face limitations due to how databases are annotated. For example, critical factors such as pH and salinity, which significantly influence thermostability, are poorly represented in existing open-access data sets. Additionally, the sequences in HotProtein often lack full structural information, particularly in regions with large unstructured segments—a common issue across data sets used for training deep learning models. This poses a significant challenge when encoding detailed structural features, especially when relying on limited predictability accuracy from structure prediction models such as AlphaFold or other methods.[27] Addressing these gaps will be essential for improving the accuracy and applicability of thermostability prediction models.

Existing classical thermostability prediction models have advanced by incorporating key structural- and sequence-based factors. PROSS[64,65] is a computational framework designed to engineer stabilized protein variants without compromising their functional activity. The method begins by analyzing a high-resolution protein structure and multiple sequence alignments of homologous sequences. It then employs a series of steps of optimization to enhance protein stability by phylogenetic filtering for evolutionarily conserved amino acids and eliminating destabilized mutations to select the most stabilizing mutations while ensuring the protein's functional integrity is preserved. PROSS achieves its goals by targeting specific regions of the protein: core residues are modified to improve packing efficiency, surface residues are adjusted to reduce aggregation and backbone interactions are optimized to enhance rigidity. As a result, PROSS-designed proteins often exhibit significantly increased thermostability and higher expression levels in bacterial systems. Additionally, these stabilized variants frequently fold correctly without the need for chaperone assistance, making them more suitable for industrial and biomedical applications. This approach demonstrates how computational design can effectively bridge the gap between protein stability and functionality.

With the recent advances in PLMs, a two-stage prediction framework could effectively address existing limitations by integrating the strengths of PLM-based models and classical models such as PROSS. In the first stage, a foundation PLM, such as the one proposed in this study, could provide an initial estimation of protein thermostability directly from sequence information. This would enable rapid screening of new protein sequences for potential thermostability, leveraging the metric of melting temperature. By predicting T_m values, the PLM could identify sequences with favorable stability profiles, serving as a high-throughput filter for further analysis. In the second stage, more detailed structural information could be integrated to refine predictions.[27] This stage would involve the selection and elimination of single or multiple point mutations,[66] capturing the nuanced sequence-structure−function relationships. PROSS could then be applied to optimize these mutations, focusing on core packing, surface residue adjustments, and backbone interactions to further enhance the stability and expression levels. This hierarchical approach would combine the speed and scalability of PLMs with the precision of structure-based methods such as PROSS, significantly improving predictive accuracy and applicability. By bridging sequence-based predictions with experimentally validated structural insights, this framework would not only enhance thermostability predictions but also accelerate advancements in protein engineering. It would enable researchers to design proteins with improved coarse-grained thermostability (T_m) and fine-grained energetically favorable folding ($\Delta\Delta G$), reducing reliance on experimental screening and facilitating the development of proteins for industrial, therapeutic, and biotechnological applications.

■ ASSOCIATED CONTENT

Data Availability Statement

The code and data used to generate the results in this work are available on the project page: https://github.com/notabigfish/procesa.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c01752.

Replies to the comments from the editor and referee (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Huifeng Zhang − *Readline Intelligence, Birmingham B29 6SQ, U.K.*; Email: ben.zhang@readlineint.com

Jian K. Liu − *School of Computer Science, University of Birmingham, Birmingham B15 2TT, U.K.*; orcid.org/0000-0002-5391-7213; Email: j.liu.22@bham.ac.uk

Authors

Feixiang Zhou − *Readline Intelligence, Birmingham B29 6SQ, U.K.*

Shuo Zhang − *School of Computer Science, University of Birmingham, Birmingham B15 2TT, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c01752

Author Contributions

F.Z. and S.Z. developed the workflow, performed the data gathering and analyses, the model development and training. H.Z. and J.K.L. conceived the project. All authors wrote, edited, read and approved the manuscript. F.Z. and S.Z. contributed equally to this work.

Funding

None.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Rongqi Hong for helpful discussions.

■ REFERENCES

(1) Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (2), 452−456.

(2) Chen, Y.-C.; Smith, T.; Hicks, R. H.; Doekhie, A.; Koumanov, F.; Wells, S. A.; Edler, K. J.; van den Elsen, J.; Holman, G. D.; Marchbank, K. J.; et al. Thermal stability, storage and release of proteins with tailored fit in silica. *Sci. Rep.* **2017**, *7* (1), 46568.

(3) Chandler, P. G.; Broendum, S. S.; Riley, B. T.; Spence, M. A.; Jackson, C. J.; McGowan, S.; Buckle, A. M. Strategies for increasing protein stability. *Protein Nanotechnology: Protocols, Instrumentation, and Applications* **2020**, *2073*, 163−181.

(4) De Carvalho, C. C. Enzymatic and whole cell catalysis: finding new strategies for old processes. *Biotechnology advances* **2011**, *29* (1), 75−83.

(5) Zhou, X.-X.; Wang, Y.-B.; Pan, Y.-J.; Li, W.-F. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino acids* **2008**, *34*, 25−33.

(6) Nakariyakul, S.; Liu, Z.-P.; Chen, L. Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. *Amino Acids* **2012**, *42*, 1947−1953.

(7) Zhang, G.; Fang, B. Logitboost classifier for discriminating thermophilic and mesophilic proteins. *Journal of biotechnology* **2007**, *127* (3), 417−424.

(8) Sadeghi, M.; Naderi-Manesh, H.; Zarrabi, M.; Ranjbar, B. Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.* **2006**, *119* (3), 256−270.

(9) Ge, M.; Xia, X.-Y.; Pan, X.-M. Salt bridges in the hyperthermophilic protein ssh10b are resilient to temperature increases. *J. Biol. Chem.* **2008**, *283* (46), 31690−31696.

(10) Bleicher, L.; Prates, E. T.; Gomes, T. C.; Silveira, R. L.; Nascimento, A. S.; Rojas, A. L.; Golubev, A.; Martínez, L.; Skaf, M. S.; Polikarpov, I. Molecular basis of the thermostability and thermophilicity of laminarinases: X-ray structure of the hyperthermostable laminarinase from rhodothermus marinus and molecular dynamics simulations. *J. Phys. Chem. B* **2011**, *115* (24), 7940−7949.

(11) Gromiha, M. M.; Pathak, M. C.; Saraboji, K.; Ortlund, E. A.; Gaucher, E. A. Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins: Struct., Funct., Bioinf.* **2013**, *81* (4), 715−721.

(12) Cao, H.; Wang, J.; He, L.; Qi, Y.; Zhang, J. Z. Deepddg: predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model.* **2019**, *59* (4), 1508−1514.

(13) Pucci, F.; Kwasigroch, J. M.; Rooman, M. Scoop: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics* **2017**, *33* (21), 3415−3422.

(14) Jarzab, A.; Kurzawa, N.; Hopf, T.; Moerch, M.; Zecha, J.; Leijten, N.; Bian, Y.; Musiol, E.; Maschberger, M.; Stoehr, G.; et al. Meltome atlas—thermal proteome stability across the tree of life. *Nat. Methods* **2020**, *17* (5), 495−503.

(15) Chen, T.; Gong, C.; Diaz, D. J.; Chen, X.; Wells, J. T.; Liu, G.; Wang, Z.; Ellington, A.; Dimakis, A.; Klivans, A. In Hotprotein: A novel framework for protein thermostability prediction and editing, The Eleventh International Conference on Learning Representations, 2023.

(16) Li, G.; Buric, F.; Zrimec, J.; Viknander, S.; Nielsen, J.; Zelezniak, A.; Engqvist, M. K. Learning deep representations of enzyme thermal adaptation. *Protein Sci.* **2022**, *31* (12), No. e4480.

(17) Zhao, J.; Yan, W.; Yang, Y. Deeptp: A deep learning model for thermophilic protein prediction. *International Journal of Molecular Sciences* **2023**, *24* (3), 2217.

(18) Shen, B.; Vihinen, M. Conservation and covariance in ph domain sequences: physicochemical profile and information theoretical analysis of xla-causing mutations in the btk ph domain. *Protein Engineering Design and Selection* **2004**, *17* (3), 267−276.

(19) Ahmed, Z.; Zulfiqar, H.; Khan, A. A.; Gul, I.; Dao, F.-Y.; Zhang, Z.-Y.; Yu, X.-L.; Tang, L. ithermo: a sequence-based model for identifying thermophilic proteins using a multi-feature fusion strategy. *Frontiers in Microbiology* **2022**, *13*, No. 790063.

(20) Lin, H.; Chen, W. Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods* **2011**, *84* (1), 67−70.

(21) Yang, Y.; Zhao, J.; Zeng, L.; Vihinen, M. Protstab2 for prediction of protein thermal stabilities. *International Journal of Molecular Sciences* **2022**, *23* (18), 10798.

(22) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **2002**, *320* (2), 369−387.

(23) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (3), 830−838.

(24) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with alphafold. *Nature* **2021**, *596* (7873), 583−589.

(25) 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. *Nat. Commun.* **2024**, *15* (1), 7400.

(26) Li, G.; Yao, S.; Fan, L. Prostage: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks. *J. Chem. Inf. Model.* **2024**, *64* (2), 340−347.

(27) Zhang, D.; Zeng, Y.; Hong, X.; Xu, J. Leveraging multimodal protein representations to predict protein melting temperatures. *arXiv* **2024**.

(28) Cheng, P.; Mao, C.; Tang, J.; Yang, S.; Cheng, Y.; Wang, W.; Gu, Q.; Han, W.; Chen, H.; Li, S.; Chen, Y.; Zhou, J.; Li, W.; Pan, A.; Zhao, S.; Huang, X.; Zhu, S.; Zhang, J.; Shu, W.; Wang, S. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell Research* **2024**, *34* (9), 630−647.

(29) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), No. e2016239118.

(30) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29287−29303.

(31) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123−1130.

(32) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **2022**, *44* (10), 7112−7127.

(33) Fenoy, E.; Edera, A. A.; Stegmayer, G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Briefings Bioinf.* **2022**, *23* (4), bbac232.

(34) Dallago, C.; Mou, J.; Johnston, K. E.; Wittmann, B. J.; Bhattacharya, N.; Goldman, S.; Madani, A.; Yang, K. K. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv* **2021**.

(35) Hsu, C.; Verkuil, R.; Liu, J.; Lin, Z.; Hie, B.; Sercu, T.; Lerer, A.; Rives, A. In Learning inverse folding from millions of predicted structures, International Conference on Machine Learning. PMLR, 2022; pp 8946−8970.

(36) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**.

(37) Huang, H.-L.; Charoenkwan, P.; Kao, T.-F.; Lee, H.-C.; Chang, F.-L.; Huang, W.-L.; Ho, S.-J.; Shu, L.-S.; Chen, W.-L.; Ho, S.-Y. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinf.* **2012**, *13*, S3.

(38) He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. In Momentum contrast for unsupervised visual representation learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; pp 9729−9738.

(39) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. In A simple framework for contrastive learning of visual representations, International Conference on Machine Learning. PMLR, 2020; pp 1597−1607.

(40) Dorkenwald, M.; Xiao, F.; Brattoli, B.; Tighe, J.; Modolo, D. In Scvrl: Shuffled contrastive video representation learning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; pp 4132−4141.

(41) Hu, K.; Shao, J.; Liu, Y.; Raj, B.; Savvides, M.; Shen, Z. In Contrast and order representations for video self-supervised learning, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp 7939−7949.

(42) Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; Xu, B. In Ts2vec: Towards universal representation of time series, Proceedings of the AAAI Conference on Artificial Intelligence, 2022; Vol. *36*, pp 8980−8987.

(43) Franceschi, J.-Y.; Dieuleveut, A.; Jaggi, M. In Unsupervised scalable representation learning for multivariate time series, Advances in Neural Information Processing Systems, **2019**; Vol. *32*.

(44) Schroff, F.; Kalenichenko, D.; Philbin, J. In Facenet: A unified embedding for face recognition and clustering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015; pp 815−823.

(45) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying structure–property relationships through smiles syntax analysis with self-attention mechanism. *J. Chem. Inf. Model.* **2019**, *59* (2), 914–923.

(46) Lin, Z.; Feng, M.; dos Santos, C.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. In A structured self-attentive sentence embedding, International Conference on Learning Representations, ICLR, 2017.

(47) Chen, J.; Zheng, S.; Zhao, H.; Yang, Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J. Cheminf.* **2021**, *13* (1), 7.

(48) Jung, F.; Frey, K.; Zimmer, D.; Mühlhaus, T. Deepstabp: A deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences* **2023**, *24* (8), 7444.

(49) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Garmiri, P.; da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Neto, T. M. B.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; de Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **2022**, *51* (D1), D523–D531.

(50) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; Xavier, R. J.; Knight, B.; Cho, K.; Bonneau, R. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12* (1), 3168.

(51) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; Song, Y. Evaluating protein transfer learning with tape. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9689–9701.

(52) ESM Team *ESM cambrian: Revealing the mysteries of proteins with unsupervised learning*, 2024. https://evolutionaryscale.ai/blog/esm-cambrian.

(53) Chen, L.; Wu, R.; Zhou, F.; Zhang, H.; Liu, J. K. HybridGCN for protein solubility prediction with adaptive weighting of multiple features. *J. Cheminf.* **2023**, *15* (1), 118.

(54) Yang, Y.; Jerger, A.; Feng, S.; Wang, Z.; Brasfield, C.; Cheung, M. S.; Zucker, J.; Guan, Q. Improved enzyme functional annotation prediction using contrastive learning with structural inference. *Commun. Biol.* **2024**, *7* (1), 1690.

(55) Singh, R.; Sledzieski, S.; Bryson, B.; Cowen, L.; Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120* (24), No. e2220778120.

(56) Gu, Z.; Luo, X.; Chen, J.; Deng, M.; Lai, L. Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics* **2023**, *39* (7), btad410.

(57) Yu, T.; Cui, H.; Li, J. C.; Luo, Y.; Jiang, G.; Zhao, H. Enzyme function prediction using contrastive learning. *Science* **2023**, *379* (6639), 1358–1363.

(58) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373* (6557), 871–876.

(59) Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; McHugh, R.; Vafeados, D.; Li, X.; Sutherland, G. A.; Hitchcock, A.; Hunter, C. N.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Baek, M.; DiMaio, F.; Baker, D. Generalized biomolecular modeling and design with rosettafold all-atom. *Science* **2024**, *384* (6693), No. eadl2528.

(60) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **2024**, *630* (8016), 493–500.

(61) Wang, D.; Pourmirzaei, M.; Abbas, U. L.; Zeng, S.; Manshour, N.; Esmaili, F.; Poudel, B.; Jiang, Y.; Shao, Q.; Chen, J.; Xu, D. S-PLM: Structure-aware protein language model via contrastive learning between sequence and structure. *bioRxiv* **2024**.

(62) Su, J.; Zhou, X.; Zhang, X.; Yuan, F. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv* **2024**.

(63) Jun. Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **2018**, *87* (1), 105–129.

(64) Goldenzweig, A.; Goldsmith, M.; Hill, S.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R.; Aharoni, A.; Silman, I.; Sussman, J.; Tawfik, D.; Fleishman, S. Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* **2016**, *63* (2), 337–346.

(65) Apr. Weinstein, J. J.; Goldenzweig, A.; Hoch, S.; Fleishman, S. J. PROSS 2: A new server for the design of stable and highly expressed protein variants. *Bioinformatics* **2021**, *37* (1), 123–125.

(66) Zhou, B.; Zheng, L.; Wu, B.; Tan, Y.; Lv, O.; Yi, K.; Fan, G.; Hong, L. Protein engineering with lightweight graph denoising neural networks. *J. Chem. Inf. Model.* **2024**, *64* (9), 3650–3661.