Computational and Structural Biotechnology Journal 21 (2023) 5544–5560

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences

Mengyu Li[a,1], Hongzhao Wang[a,1], Zhenwu Yang[a], Longgui Zhang[b], Yushan Zhu[a,c,*]

[a] College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
[b] SINOPEC Beijing Research Institute of Chemical Industry, Beijing 100013, China
[c] National Energy R&D Center for Biorefinery, Beijing University of Chemical Technology, Beijing 100029, China

ARTICLE INFO

Keywords:
Protein thermal stability
Melting temperature
Deep learning predictor
Graph convolutional neural network
Self-attention network
Sequence features

ABSTRACT

Thermally stable proteins find extensive applications in industrial production, pharmaceutical development, and serve as a highly evolved starting point in protein engineering. The thermal stability of proteins is commonly characterized by their melting temperature (T_m). However, due to the limited availability of experimentally determined T_m data and the insufficient accuracy of existing computational methods in predicting T_m, there is an urgent need for a computational approach to accurately forecast the T_m values of thermophilic proteins. Here, we present a deep learning-based model, called DeepTM, which exclusively utilizes protein sequences as input and accurately predicts the T_m values of target thermophilic proteins on a dataset consisting of 7790 thermophilic protein entries. On a test set of 1550 samples, DeepTM demonstrates excellent performance with a coefficient of determination (R^2) of 0.75, Pearson correlation coefficient (P) of 0.87, and root mean square error (RMSE) of 6.24 ℃. We further analyzed the sequence features that determine the thermal stability of thermophilic proteins and found that dipeptide frequency, optimal growth temperature (OGT) of the host organisms, and the evolutionary information of the protein significantly affect its melting temperature. We compared the performance of DeepTM with recently reported methods, ProTstab2 and DeepSTABp, in predicting the T_m values on two blind test datasets. One dataset comprised 22 PET plastic-degrading enzymes, while the other included 29 thermally stable proteins of broader classification. In the PET plastic-degrading enzyme dataset, DeepTM achieved RMSE of 8.25 ℃. Compared to ProTstab2 (20.05 ℃) and DeepSTABp (20.97 ℃), DeepTM demonstrated a reduction in RMSE of 58.85% and 60.66%, respectively. In the dataset of thermally stable proteins, DeepTM (RMSE=7.66 ℃) demonstrated a 51.73% reduction in RMSE compared to ProTstab2 (RMSE=15.87 ℃). DeepTM, with the sole requirement of protein sequence information, accurately predicts the melting temperature and achieves a fully end-to-end prediction process, thus providing enhanced convenience and expediency for further protein engineering.

1. Introduction

Thermally stable proteins possess the characteristic of maintaining their structural integrity and functionality under high-temperature conditions. As a result, they are widely employed in industrial production [1–3], pharmaceutical development [4–6], and serve as highly evolved starting points in protein engineering [7–11]. To assess the thermal stability of proteins, it is commonly characterized by the melting temperature (T_m) of the protein. However, with the exponential growth of protein sequence numbers, the existing protein and mutant

Abbreviations: T_m, melting temperature; OGT, optimal growth temperature of the host organisms; T_{env}, environmental survival temperature; R^2, coefficient of determination; P, Pearson correlation coefficient; RMSE, root mean square error; TPP, Thermal Proteome Profiling; PET, Polyethylene terephthalate; BLOSUM, Blocks Amino Acid Substitution Matrices; PSSM, the Position Specific Scoring Matrix; MSA, Multiple Sequence Alignment; HMM, the Hidden Markov Model; SPD33, a series of protein secondary structure descriptors; ASA, Solvent Accessible Surface Area; HSE, Half Sphere Exposure; PP7, the seven physicochemical properties; AA, Amino Acid Frequency; Dipeptide, Dipeptide Frequency; Tm50 dataset, The dataset consists of proteins from the T_m dataset with $T_m \geq 50$ ℃; LCC, Leaf-branch compost cutinase.

[*] Corresponding author at: College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China.
E-mail address: zhuys@mail.buct.edu.cn (Y. Zhu).

[1] These authors contribute equally.

https://doi.org/10.1016/j.csbj.2023.11.006
Received 23 August 2023; Received in revised form 2 November 2023; Accepted 2 November 2023
Available online 4 November 2023

thermodynamic database, ProThermDB [12], only contains approximately 120,000 records, of which fewer than 10,000 contain T_m information. In contrast, the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (Release 2023_01) [13] databases already include over 200 million protein sequence entries. This poses a significant challenge to the current endeavors in thermal stability assessment. To accurately determine the T_m value of a protein in the laboratory, a series of complex procedures are required, involving steps such as protein expression, purification, and the utilization of specialized instrumentation. Currently, the commonly applied methods for determining T_m values of proteins include Differential Scanning Calorimetry (DSC), Circular Dichroism (CD) spectroscopy, and Differential Scanning Fluorimetry (DSF) [14]. Differential Scanning Calorimetry (DSC) measures the relationship between the power supplied to the test sample and the reference material with respect to temperature, achieved through controlled temperature programming, to obtain the heat capacity. Circular Dichroism (CD) spectroscopy uses left- and right-circularly polarized light passing through optically active media with chiral structures like biomolecules. During a temperature ramp, it measures the circular dichroism spectrum of substances such as proteins, reflecting the trend of structural changes with increasing temperature. Differential Scanning Fluorimetry (DSF) employs the characteristic of exposing hydrophobic groups inside the protein upon heating. There are two approaches: one involves adding dyes that generate a fluorescence signal by binding selectively to the hydrophobic regions of the protein, and the fluorescence signal curve during the temperature ramp is recorded to determine the T_m value. The other approach detects changes in the hydrophobic environment of aromatic amino acids (tryptophan and tyrosine) with fluorescent chromophores during the protein unfolding process to measure the T_m value. In certain scenarios, particularly when screening a large number of protein scaffolds, computational prediction may be more practical compared to experimental methods. Additionally, the challenges associated with protein expression and purification can be limiting factors for experimentally determining T_m values. In summary, for the vast number of proteins with unknown T_m values, it is evidently impractical to measure them using experimental methods, primarily due to the high cost and lengthy experimental duration [15]. Therefore, the utilization of rational computational methods to predict the melting temperature of proteins has become an urgent task and a crucial solution to address the discrepancy between protein sequence information and thermal stability data.

Currently, there are two types of computational methods available for predicting the melting temperature of proteins: statistical-based methods and machine learning-based methods (see Supplementary Table 1). For instance, Zhang et al. [16] employed support vector machines to classify thermophilic and mesophilic proteins by extracting the composition of the 20 amino acids in the primary structure as feature vectors. They achieved an accuracy of 90.1% on a dataset consisting of 152 proteins. Pucci et al. [17] utilized temperature-dependent statistical potentials to predict the thermal stability of homologous proteins in a dataset containing 166 proteins from 11 homologous families. In cross-validation, the RMSE between the experimental and predicted T_m values was found to be 13.6 ℃. By excluding the six proteins with the poorest predictions, the RMSE decreases to 8.3 ℃. Dehouck et al. [18] computed a correlation coefficient of 0.59 between the environmental survival temperature (T_{env}, equivalent to the optimal growth temperature of the host organisms, OGT) and the protein T_m based on a dataset of 127 proteins. The regression equation derived from their analysis is approximately $T_m \approx 42.9℃ + 0.62T_{env}$. Ku et al. [19] categorized proteins into high T_m group ($T_m > 65$ ℃) and low T_m group ($T_m < 55$ ℃) based on their protein T_m values. They employed the composition of dipeptides within protein sequences as features for classification. Remarkably, they achieved 100% accuracy in their classification on a dataset consisting of 35 proteins. Gromiha et al. [20] calculated a correlation coefficient of 0.91 between T_{env} and the protein T_m for each protein family. The corresponding regression equation derived from their analysis is $T_m = 24.4℃ + 0.93T_{env}$. Pucci et al. [21] proposed a method for predicting T_m, which relies on the optimal growth temperature of the protein host organisms, protein properties, standard statistical potentials, and temperature-dependent statistical potentials. In their experiments, this method yielded a Pearson correlation coefficient (P) of 0.72 when applied to a dataset comprising 22 proteins. Gorania et al. [15] characterized proteins by utilizing amino acid composition and pseudo-amino acid composition, and employed an artificial neural network approach to predict the melting temperature of proteins. Li et al. [22] utilized transfer learning methods to predict the melting temperature of proteins based on a prediction model for the optimal growth temperature of the host organisms (OGT). They conducted these predictions on two distinct datasets comprising 2506 and 41,725 protein sequences, respectively. The results revealed that the R^2, attained values of 0.73 and 0.58 on the two datasets, respectively. Yang et al. [23] developed the protein melting temperature prediction algorithm, ProTstab2, utilizing the gradient boosting algorithm. They achieved a determination coefficient (R^2) of 0.58, a Pearson correlation coefficient (P) of 0.753, and a root mean square error (RMSE) of 7.005 ℃ when evaluated on a dataset containing 34,913 protein melting temperatures. Jung et al. [24] developed the DeepSTABp protein melting temperature prediction algorithm based on a Transformer-based language pre-training model. By incorporating experimental conditions used in Thermal Proteome Profiling (TPP), amino acid sequences, and OGT, the algorithm predicted protein melting temperatures on a dataset containing 35,112 protein sequences. The results exhibited a determination coefficient (R^2) of 0.8, a Pearson correlation coefficient (P) of 0.9, and a root mean square error (RMSE) of 4.3 ℃. Although there are numerous tools available for characterizing the thermal stability of proteins, accurately predicting the T_m value of proteins based solely on their sequences remains a significant challenge. Research [18,20] has revealed a strong correlation between the optimal growth temperature of the host organisms and the T_m value of proteins. Furthermore, the utilization of information from homologous protein families allows for a more robust characterization of protein thermal stability [17,18,25–28]. Simultaneously, protein contact maps [29] can capture the underlying relationships between residue-residue pairs in the spatial dimension of proteins [30]. Therefore, utilizing the evolutionary information [31,32] of proteins and protein contact maps to predict their melting temperatures is a promising direction worth exploring.

This study presents a novel protein representation method that combines the host's optimal growth temperature, evolutionary information of the protein, protein contact map, seven physicochemical properties (steric parameters, hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability), protein secondary structure descriptors, as well as amino acid and dipeptide frequencies to characterize proteins. Subsequently, we constructed a deep neural network model (Fig. 1) called DeepTM, using graph convolutional neural networks [33] and self-attention networks, with the aim of directly predicting the T_m values of proteins from their sequence information. Additionally, to validate the practical performance of the model, we collected a dataset (Supplementary Table 2) comprising 22 PET (Polyethylene terephthalate) plastic-degrading enzymes along with their experimentally determined T_m values, as well as a dataset (Supplementary Table 3) containing 29 thermally stable proteins with a broader classification and their corresponding experimental T_m values to serve as two external validation set, demonstrating the applicability of our model in real-world scenarios. In the biodegradation pathway of polyester waste recycling, the search for thermally stable PET plastic-degrading enzymes is a crucial step. Currently, PET plastic-degrading enzymes lack robustness in terms of temperature range and exhibit slow reaction rates [34]. Therefore, precise prediction of protein melting temperatures can offer convenience in this regard and aid researchers in identifying more stable PET plastic-degrading enzymes.

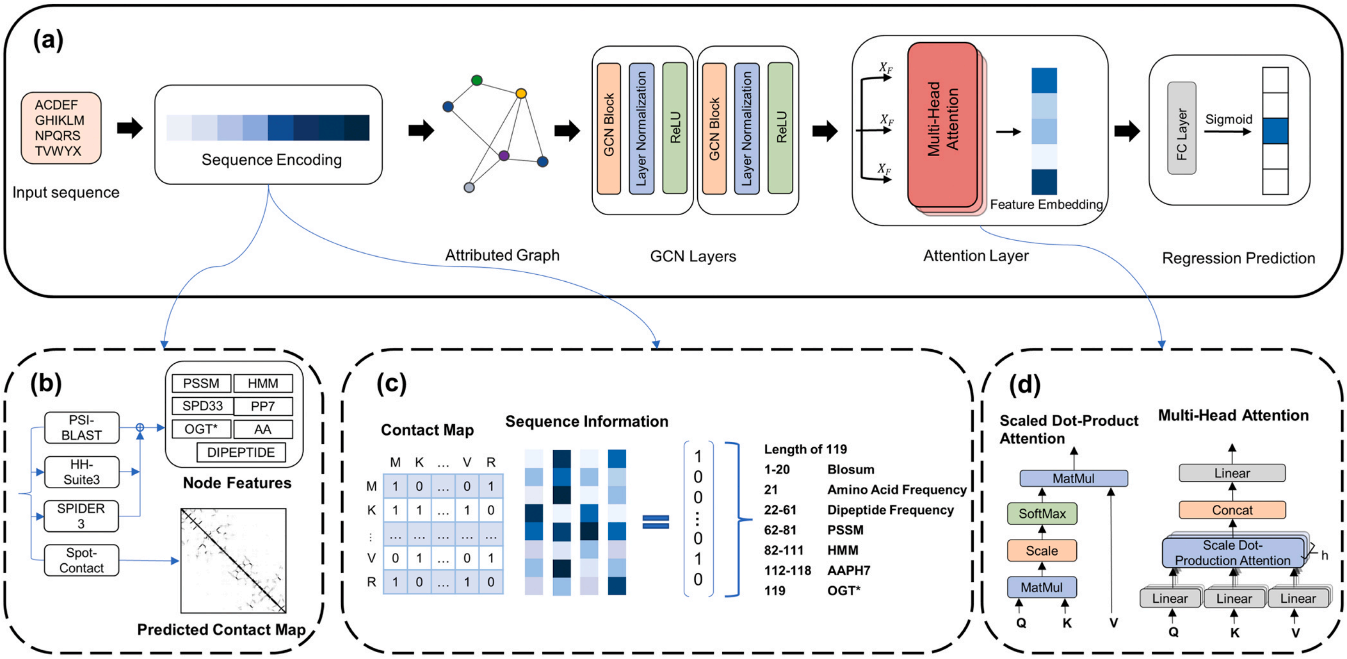


Fig. 1. Architecture of deep learning predictor DeepTM. (a) Four steps of neural network model: Input protein sequence data; Encode protein vector representation and input it to the neural network in the form of a graph; Train the neural network using GCN layers and attention layer; Predict continuous T_m values. GCN Layers consist of two graph convolutional layers. After each graph convolutional layer, layer Normalization is applied to prevent overfitting. The activation function used in the graph convolutional layers is ReLU(·). FC Layer refers to a fully connected layer. (b) Stepwise Feature Extraction. The PSSM, HMM, SPD33, and Contact Map profiles are generated using the PSI-BLAST [44], HH-Suite3 [47], SPIDER3 [50], and Spot-contact [29] programs, respectively. The PP7 was obtained from the literature [51]. The amino acid frequency (AA) and dipeptide frequency (Dipeptide) were calculated. OGT was predicted using the OGT predicting model. The symbol (*) indicates that this feature was obtained using models trained specifically for this experiment. (c) Feature vectors encoding contact maps and sequence information. (d) Multi-head attention mechanism [56].

2. Methods

2.1. Dataset

OGT Dataset Our OGT dataset consists of two parts, with both datasets provided by Li. [35]. Specifically, we downloaded the brenda_sequences_20180109.fasta and enzyme_ogt_topt.tsv files from https://zenodo.org/record/2539114 and subjected them to the following processing steps: (1) Select sequences containing experimental OGT data, retaining only sequence ID, experimental OGT value, and protein sequence; (2) Remove duplicate sequences and apply the CD-HIT [36] algorithm for clustering (-c 0.5 -n 3 -M 16000 -d 0 -T 8, with default values for other parameters); (3) Remove protein sequences containing non-standard amino acids (X, U, O); (4) Remove protein sequences with OGT values less than 0 ℃ or greater than 110 ℃; (5) Remove protein sequences with a length exceeding 1028 (since on a

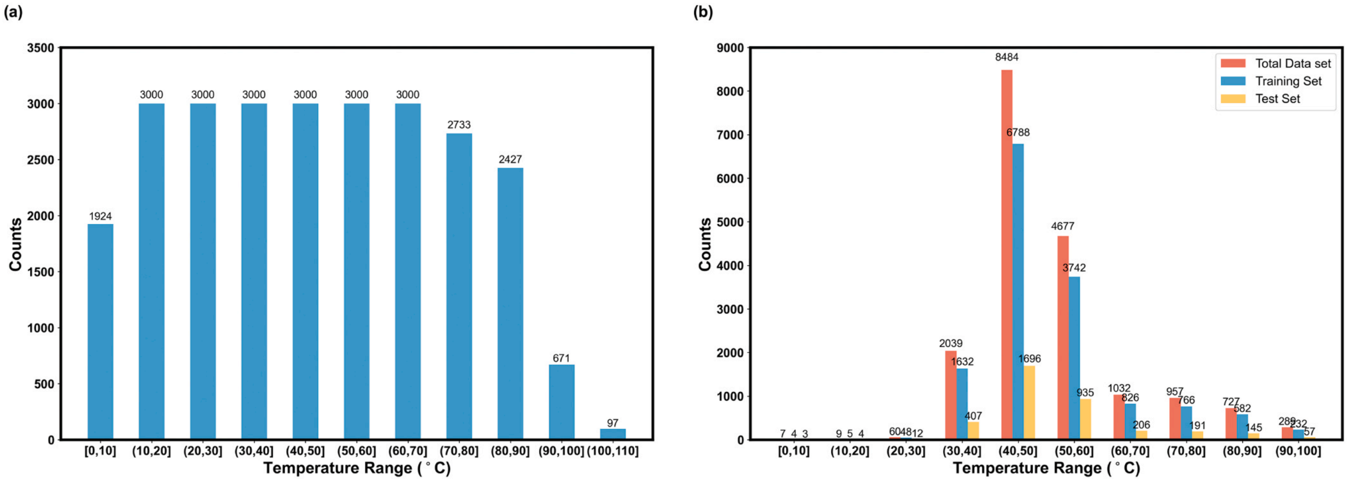


Fig. 2. Distribution of OGT and T_m for all proteins in the OGT dataset and T_m dataset. (a) Distribution of the number of protein sequences in the OGT dataset for each temperature interval of OGT. The abscissa represents the temperature intervals of OGT (starting from 0 ℃ with each interval size of 10 ℃), while the ordinate represents the number of protein sequences within each corresponding temperature interval. T_m: Melting Temperature. OGT: Optimal Growth Temperature of the host organisms. (b) Distribution of the number of protein sequences in the T_m dataset for each temperature interval of T_m. Comparison of sequence distributions among the entire T_m dataset, T_m training set, and T_m test set, represented by orange, blue, and yellow colors, respectively. The abscissa represents the temperature intervals of T_m (starting from 0 ℃ with each interval size of 10 ℃). The data labels on top of each bar indicate the number of sequences within that interval.

server with 12 GB RAM, the CCMPred [37] software can generate mat profiles with a maximum sequence length of 1028); (6) Divide the OGT data into intervals of 10 °C. If the number of sequences within a temperature interval exceeds 3000, select 3000 sequences. If it is less than 3000, select the actual number of sequences. Ensure a balanced distribution of the dataset across temperature levels. Following these preprocessing steps, we filtered out 2822 and 23,030 protein sequences from the respective datasets. Consequently, we constructed a final dataset comprising 25,852 protein sequences (Fig. 2a, Supplementary Fig. 1a, Supplementary Fig. 1c). Furthermore, the dataset also includes the UniprotKB ID and amino acid sequence information for each protein.

T_m Dataset We have gathered data from the Meltome atlas [38] and performed pre-processing. The pre-processing steps for this dataset encompassed the following procedures: (1) Removal of duplicate sequences and clustering using the CD-HIT [36] algorithm with parameters (-c 0.5 -n 3 -M 16000 -d 0 -T 8, default values for other parameters); (2) Exclusion of protein sequences containing non-standard amino acids (X, U, O); (3) Elimination of protein sequences with melting temperatures (T_m) below 0 °C or above 100 °C; (4) Removal of protein sequences exceeding a length of 1028, ensuring consistency in sequence length. Finally, we obtained a dataset comprising 18,281 protein sequences accompanied by their corresponding experimental T_m values.

Dataset of Plastic-Degrading Enzymes The plastic-degrading enzyme dataset was employed as a blind test dataset to evaluate the performance of the model in practical applications. We selected protein sequences from Erickson et al. [39] that included experimentally determined T_m values and OGT values. These sequences were subjected to 50% clustering using the CD-HIT [36] algorithm. After processing, the final dataset comprised 22 protein sequences, all of which exhibited experimentally determined T_m values above 50 °C (Supplementary Table 2).

Dataset of Thermally Stable Proteins The thermally stable protein dataset, serving as the second blind test dataset, encompasses a broader range of protein classes. This dataset comprises two components. Firstly, we obtained protein sequences with experimentally determined T_m values from the ProThermDB [40] and MPTherm [41] databases. Subsequently, we employed the experimental DSC method to determine the T_m values for three esterase sequences. Following this, we applied the CD-HIT [36] algorithm for sequence deduplication (-c 0.5 -n 3 -M 16000 -d 0 -T 8, default values for other parameters). After processing, the final dataset contains 29 protein sequences (Supplementary Table 3), including three esterase sequences, fifteen membrane protein sequences, nine antibody sequences, and two transcription factor sequences. All these proteins have experimentally determined T_m values above 50 °C.

2.2. Protein representation

2.2.1. Node features

Blocks Amino Acid Substitution Matrices We employ BLOSUM62 [42,43] encoding to represent amino acid residues, which is a substitution matrix commonly used in bioinformatics for sequence alignment purposes. The BLOSUM62 encoding is derived from the observation and statistical analysis of highly conserved sequences within protein families in the BLOCKS database, leading to the compilation of amino acid substitution probabilities. The matrix size is 24 × 24, and in the calculations, we utilized the first 20 rows and 20 columns to represent the 20 natural amino acids.

Position Specific Scoring Matrix The Position Specific Scoring Matrix (PSSM) represents the evolutionary relationships among a set of proteins. It is generated through Multiple Sequence Alignment (MSA). During the process of Multiple Sequence Alignment (MSA), multiple homologous sequences are aligned, resulting in a series of amino acids at corresponding positions. Subsequently, the frequency of occurrence of different amino acids at each position is calculated. Specifically, in the PSSM, an element p_{ij} represents the likelihood of an amino acid at position j in the sequence mutating into the i-th amino acid during the process of evolution. If the value is positive, it indicates a higher likelihood, whereas if the value is negative, it indicates a lower likelihood. In the computation, we utilized the PSIBLAST2.6.0 +[44] program to perform three iterations on the Uniref90 [45] database, generating protein PSSM profiles in bulk.

Hidden Markov Matrix The Hidden Markov Model (HMM) encapsulates the evolutionary information [46] of proteins and is commonly used to depict specific patterns and probability distributions that arise within diverse biological sequences. Additionally, it provides information about the relative positions of amino acids and the probabilities of their transitions within these sequences. In this experiment, we employed the HHblits [47] program to perform searches in the Uniclust30_2020_06 [48] database, generating HMM profiles in bulk. These HMM profiles also include seven transition probabilities and three local alignment diversity values, which we consider as features [49].

SPD33 Representing a series of protein secondary structure descriptors. This includes one Solvent Accessible Surface Area (ASA), two Half Sphere Exposure (HSE) based on C_α atoms, eight sine and cosine values of main chain torsion angles (φ, ψ, θ, τ), and three probabilities for predicting secondary structure [50]. In the computation, we utilized the SPIDER3 [50] program to generate spd33 profiles in bulk.

Physicochemical Properties The seven physicochemical properties (PP7) of amino acids are as follows: steric parameters, hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability. The numerical values of these physicochemical properties were obtained from Meiler et al. [51].

Amino Acid Frequency and Dipeptide Frequency The frequency of amino acids and dipeptides in proteins refers to the occurrence rate of each amino acid and the occurrence rate of adjacent dipeptides formed by two amino acids in the protein sequence. We calculate the corresponding frequencies for each protein.

Finally, these features constitute node features $X_{L\times f}$, where f represents the dimension of node features and L represents the length of the protein sequence. Supplementary Table 4 presents all node features and their dimensions. Before inputting into the neural network, all data are standardized to have zero mean and unit variance.

2.2.2. Edge feature

Protein Contact Map The protein contact map is a two-dimensional matrix representation that captures the contact relationships between each amino acid residue and other residues within a protein. Each element in this matrix indicates whether there is a contact between two amino acid residues. If two residues are in close proximity in the native structure, meaning that the Euclidean distance between their C_β atoms is less than 8 Å, they are considered to form a contact [29]. In computing, we utilize the SPOT-Contact [29] software to predict protein contact maps from protein sequences, where each element represents the probability of contact between two amino acid residues.

2.3. Deep Learning Predictor

In this study, we propose a protein T_m value prediction model based on deep neural networks. The neural network framework of this model is illustrated in Fig. 1, combining three distinct modules: Graph Convolutional Neural Network (GCN), Self-attention Network, and Multilayer Perceptron. We refer to this model as DeepTM. The model takes amino acid sequences as input, and after extracting features from the sequences, it passes through a graph convolutional neural network to embed node features and edge features, thereby obtaining protein sequence representations. Subsequently, this vector is fed into a self-attention layer to extract the parts that are relevant to the T_m value. Then, through a fully connected layer, the sequence representation vector is transformed to the desired size and outputs the model's predicted T_m value of proteins.

In this study, we utilized the PyTorch [52] deep learning framework, along with Scikit-learn [53] and Python 3.7, to construct our neural network models. To train this model, we employed the Adam [54] optimizer with a mean squared error loss function. We set the batch size for each training iteration to 32 and trained the model for 100 epochs. The ReLU($\cdot$) [55] activation function was utilized. In addition, to avoid overfitting, we applied L2 regularization. Finally, we selected the best model based on higher coefficients of determination (R^2) and Pearson correlation coefficients (P), as well as lower root mean square error (RMSE), as our final model.

Graph Convolution Network A approach using neural networks to learn graph-structured data, with the primary objective of extracting and uncovering features and patterns within the graph structure to meet the requirements of subsequent tasks. Given an amino acid sequence of length L, the node features are represented as $X_{L\times f}$, where f denotes the dimensionality of the node features. The node features consist of the BLOSUM62 encoding, PSSM, HMM, seven physicochemical properties (PP7), secondary structure features, amino acid frequencies, and dipeptide frequencies of the protein sequence. The edge features are represented as $A_{L\times L}$, which is composed of a protein contact map where each element represents the contact probability between two amino acids. To ensure that the values of node features remain within the range of $-$ 1–1, it is necessary to normalize the node features by

$$X'_{L\times f} = \frac{X_{L\times f} - \overline{X}_{L\times f}}{\sigma_{L\times f}} \tag{1}$$

where $\overline{X}$ represents the mean of node features, σ is the variance, and X' is the normalized node features. To avoid altering the original distribution of features after multiplying them by the feature matrix, we standardize the edge features by:

$$\widetilde{A} = DAD \tag{2}$$

$$D = diag\left(\left(\sum_{k} A_{ik}\right)^{-\frac{1}{2}}\right) \tag{3}$$

where D refers to a diagonal matrix with its diagonal elements being the negative 1/2 power of the row sums of the adjacency matrix A. $\widetilde{A}$ is the normalized adjacency matrix. Ultimately, the formula for graph convolutional neural networks is as follows:

$$G^{(l+1)} = \sigma\left(\widetilde{A}G^{(l)}W^{(l)} + B^{(l)}\right) \tag{4}$$

where the initial state $G^{(0)} = X'$, $G^{(l)} \in \mathbb{R}^{L\times f}$ is the output of the l-th layer after applying the activation function. $W^{(l)} \in \mathbb{R}^{f\times f'}$ represents the weight matrix of a specific layer, which maps the feature vectors from dimension f to dimension f'. $B^{(l)} \in \mathbb{R}^{f\times f'}$ is the bias matrix of a specific layer. σ represents the activation function ReLU($\cdot$). To accelerate convergence and prevent overfitting, we add a Normalization layer after each graph convolutional layer. The final output of the graph convolutional neural network is denoted as follows:

$$U_{L\times p} = (v_1, v_2, v_3, \ldots, v_L) \tag{5}$$

U is a two-dimensional matrix, where v_i represents the embedding of a node, and p represents the dimensionality of the embedding vector.

Self-attention Network The attention network [56] allows the neural network to focus only on the parts that are of interest while ignoring the uninteresting parts. We use the output matrix U of the graph convolutional neural network as the input to the self-attention network.

$$T = Attention(U) = softmax\left(W_2\cdot\tanh\left(W_1U^T\right)\right)U \tag{6}$$

$$H = \frac{1}{r}\sum\nolimits_{k}^{r}(T)_k \tag{7}$$

Here, U^T represents the transpose of the matrix $U \in \mathbb{R}^{L\times p}$, and $W_1 \in \mathbb{R}^{q\times p}$ and $W_2 \in \mathbb{R}^{r\times q}$ are two weight matrices. The parameters p, q, and r are hyperparameters. In self-attention networks, the *softmax* function normalizes the weight matrix along each row, ensuring that the sum of each row equals 1. Next, we use weight matrices W_1 and W_2 to transform U, and then apply the *softmax* function to obtain the attention matrix T. A set of r different coefficients is used to evaluate the correlation between each residue and the T_m value from different perspectives. Finally, we average the r sets of attention coefficients to obtain the final feature representation $H \in \mathbb{R}^{1\times p}$.

Multi-layer Perceptron The output of the self-attention network is passed as input to the multi-layer perceptron (MLP), which performs a series of nonlinear transformations on the data to predict the T_m value. Specifically, it can be expressed by the following equation:

$$O = Sigmoid\left(W_3H^T + B\right) \tag{8}$$

where $W_3 \in 1\times p$ is the weight matrix and $B \in \mathbb{R}$ is the bias term. The *sigmoid* activation function maps the predicted values to the (0, 1) interval, providing a probabilistic interpretation to the output.

In order to enhance memory utilization and model training speed, thus ensuring greater accuracy in the direction of gradient descent, we have implemented measures for parallelization. Specifically, we conducted transformations on the node features and edge features of all proteins to ensure they have the same dimensionality, with the value of N set to 1028. This value was determined based on the maximum length of protein sequences in the T_m dataset. For node features and edge features with sequence lengths smaller than 1028, we employed a padding strategy known as "zero-padding" (Supplementary Fig. 2). After performing the "zero-padding" operation, the node features are represented as $X_{N\times f}$, and the edge features are represented as $A_{N\times N}$. The original node feature matrix and edge feature matrix are located in the top left corner of the new matrix, while the remaining positions are filled with zeros to maintain the integrity of the matrix.

2.4. Training and evaluation

2.4.1. Hyper-parameter optimization

To train our final model, we employed a random search strategy to adjust the hyper-parameters by sampling values from the defined parameter spaces (Supplementary Table 5). Approximately 50–100 sets of parameters were sampled and tested. The optimal hyper-parameters were chosen based on the parameter set with the lowest R^2 value on the validation set. Subsequently, we conducted an analysis of the results obtained from the random search strategy and manually fine-tuned several hyper-parameters. The best model, determined by the lowest R^2 value on the validation set, was selected as the final set of hyper-parameters. For detailed information on the specific hyper-parameters adjustment process, please refer to Supplementary Table 6. To assess the impact of different hardware on model training, we utilized two distinct hardware configurations: GPU (GeForce® RTX 4090) and pure CPU (64-Core AMD® EPYC 7H12 @ 1.5 GHz). It is worth noting that GPU training can be completed in just 3 h, whereas only-CPU training takes a significantly longer time, lasting up to 17 h. The source code of DeepTM can be accessed publicly through the following link: https://github.com/liimy1/DeepTM (accessed on 13 June 2023). It is worth mentioning that, on a single-core CPU, DeepTM has an average execution time of less than 180 min for protein sequences containing 1000 amino acids. Among these, the primary time is spent on the stage of generating PSSM features using PSI-BLAST [44], which accounts for approximately 94% of the total duration. More importantly, given the corresponding hardware support, DeepTM has the ability to parallel process protein sequences during the feature extraction step. This implies that DeepTM is capable of simultaneously processing feature extraction of multiple protein sequences, thereby significantly enhancing the efficiency of the entire model. After the completion of the feature extraction process, the model's prediction time can reach a level measured in seconds.

2.4.2. Evaluation metrics

In order to assess the performance of the model, we employed three distinct sets of metrics. Root Mean Square Error (RMSE) is a metric used to quantify the disparity between predicted values and observed values, serving as a measure of the model's performance. In the experiment, we calculated RMSE using the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}{n}} \tag{9}$$

where n is the number of samples, y_i is the observed values, $\widehat{y}_i$ is the predicted values by the model.

R-squared, also known as the coefficient of determination, is a statistical measure used to quantify the goodness of fit of a model. It reflects the degree of correlation between the predicted values of the model and the actual observed values. The calculation method is as follows:

$$R^2 = 1 - \frac{\sum\left(y - \widehat{y}\right)^2}{\sum\left(y - \overline{y}\right)^2} \tag{10}$$

where $\overline{y}$ is the mean value of the observed values, and $\widehat{y}$ is the predicted values of the model.

The Pearson correlation coefficient (P) is a statistical measure that quantifies the linear relationship between two variables. Its calculation method is as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X - \mu x)(Y - \mu y)]}{\sigma_X\sigma_Y} \tag{11}$$

where $cov(X, Y)$ is the covariance between the two variables and σ_X is the standard deviation.

2.5. DeepTM web application

DeepTM is available for free as a web application on http://deeptm.top/index.html (accessed on August 22, 2023). The program uses protein sequence(s) as input. DeepTM provides prediction results, which are sent to users by email when ready. This website contains datasets for training and testing.

3. Results and discussion

3.1. Data partitioning

We employ two distinct datasets to train deep neural network models, each dedicated to predicting OGT values and T_m values of proteins, respectively. These two datasets are referred to as the OGT dataset and the T_m dataset, respectively, with the OGT experimental values and T_m experimental values serving as the labels. The OGT dataset and T_m dataset encompass experimental values within temperature ranges of [0, 110 ℃] and [0, 100 ℃], respectively (Fig. 2). To train the model, we split these two datasets into separate training and test sets. Specifically, when training the OGT model, the ratio of training set to test set in the OGT dataset is 19:1. Additionally, 20% of the data from the training set is selected as the validation set. To ensure a balanced distribution of the OGT dataset at the OGT level, we use a temperature interval size of 10 ℃ and select a maximum of 3000 sequences within each temperature interval. In cases where the number of protein sequences within a temperature interval exceeds 3000, we randomly sample 3000 sequences. For intervals with fewer than 3000 sequences, we include all the available protein sequences in that interval. Finally, Fig. 2a illustrates the distribution of the number of protein sequences in each temperature interval for the OGT dataset. The number of protein sequences for each temperature interval in the range of 10–90 ℃ remains between 2000 and 3000. However, there are fewer protein sequences in the 90–110 ℃ range, totaling 768 sequences, which accounts for 2.97% of the entire OGT dataset. There are a total of 1924 protein sequences in the OGT dataset with OGT below 10 ℃, accounting for 7.44% of the entire OGT dataset. When training DeepTM, we partitioned the T_m dataset into training and test sets in an 8:2 ratio, and further selected 20% of the data from the training set as the validation set. To maintain a consistent data ratio between the training set and the test set within a small range, we partitioned the T_m dataset based on the experimental T_m values of proteins into intervals of 10 ℃. Subsequently, we selected protein sequences from each interval in a ratio of 8:2 for the training set and test set, respectively (Fig. 2b). Therefore, in the entire T_m dataset, as well as in the training set and test set, the proportion of protein sequences contained within each temperature interval is similar. Within the entire T_m dataset, 83% of protein sequences exhibit T_m values ranging from 40 ℃ to 60 ℃. Protein sequences with T_m values below 30 ℃ account for a mere 0.4% of the entire T_m dataset, totaling 76 sequences. On the other hand, there are 1017 protein sequences with T_m values exceeding 80 ℃, representing 5.6% of the entire T_m dataset. The experimental T_m values of proteins in the T_m dataset predominantly concentrate around 50 ℃, with a decreasing frequency of data as one moves towards the extremes. In order to mitigate the occurrence of random variations in the training results, we employed a 5-fold cross-validation method to train the model. In neural network modeling, using larger-scale datasets can potentially enhance the model's generalization performance. Currently, public datasets containing information about protein's T_m are relatively scarce. If other melting data related to T_m, such as Tonset and distribution width, could be included in public databases, it would likely enrich the dataset and lead to more accurate predictions by neural network models.

3.2. Predicted results for OGT

We trained a deep neural network model using the OGT dataset, as depicted in Supplementary Fig. 3, to predict the OGT values of proteins. The trained model is referred to as the OGT model. The model takes protein sequences as input, where node features are composed of BLOSUM62 encoding, PSSM, HMM, seven physicochemical properties (PP7), amino acid frequencies, and dipeptide frequencies. The edge features are represented by the protein contact map. Through the utilization of graph convolutional neural networks and attention mechanisms, the model is capable of extracting OGT-related information from protein sequences. Subsequently, the extracted representation vectors are processed through a multi-layer perceptron (MLP) to predict the protein's OGT. On the training set, the performance evaluation results of the OGT model are as follows: $R^2 = 0.78$, $P = 0.89$, RMSE $= 11.8$ ℃ (Fig. 3a). On the test set, the model performs as follows: $R^2 = 0.74$, $P = 0.86$, RMSE $= 12.97$ ℃ (Fig. 3b). However, when the OGT of proteins is above 90 ℃, the model tends to underestimate the OGT values. On the other hand, when the OGT of proteins is in the range of 0–10 ℃, the model tends to overestimate the OGT values. This could be attributed to a larger number of protein sequences with OGT below 90 ℃ compared to those with OGT above 90 ℃ during the model training. Additionally, there is a significantly greater quantity of sequences with OGT above 10 ℃ compared to those with OGT below 10 ℃ in the OGT dataset. This could be attributed to a larger number of protein sequences with OGT below 90 ℃ compared to those with OGT above 90 ℃ during the model training. Additionally, there is a significantly greater quantity of sequences with OGT above 10 ℃ compared to those with OGT below 10 ℃ in the OGT dataset. As a result, the model learned more extensively from the OGT features within the range of 10–90 ℃, causing it to exhibit a tendency to predict OGT values closer to this range for proteins with OGT values outside the 10–90 ℃ range. However, on the test set, our model achieved a Pearson correlation coefficient (P) of 0.86 between the predicted OGT values and the experimentally measured OGT values of proteins. Therefore, we can utilize the trained OGT model to predict

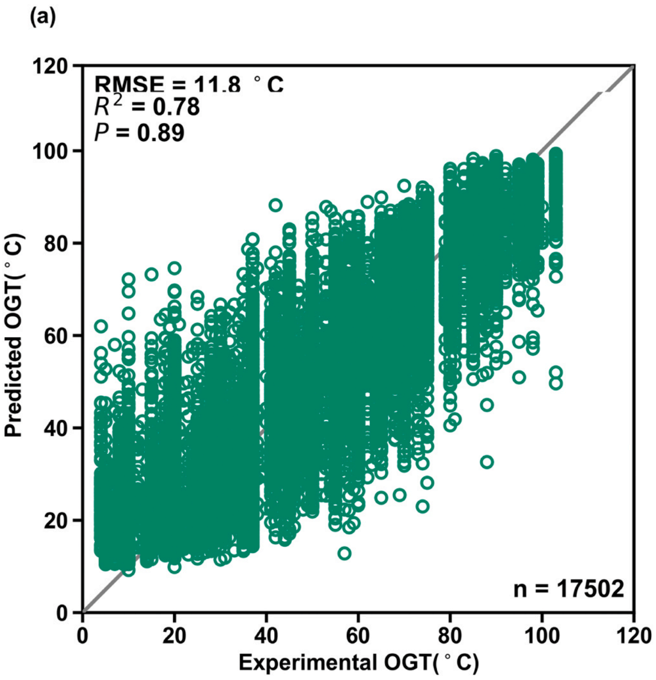


Fig. 3. Prediction results of OGT trained on the OGT dataset. The abscissa represents the experimental values of OGT, while the ordinate represents the predicted values of OGT. The points on the diagonal represent the experimental values of OGT and the corresponding predicted values are identical. RMSE: Root mean square error. R^2: Coefficient of determination. P: Pearson correlation coefficient. n: Number of data points. OGT: Optimal Growth Temperature of the host organisms. (a) Comparison between the experimental values and predicted values of OGT in the training set. (b) Comparison between the experimental values and predicted values of OGT in the test set.

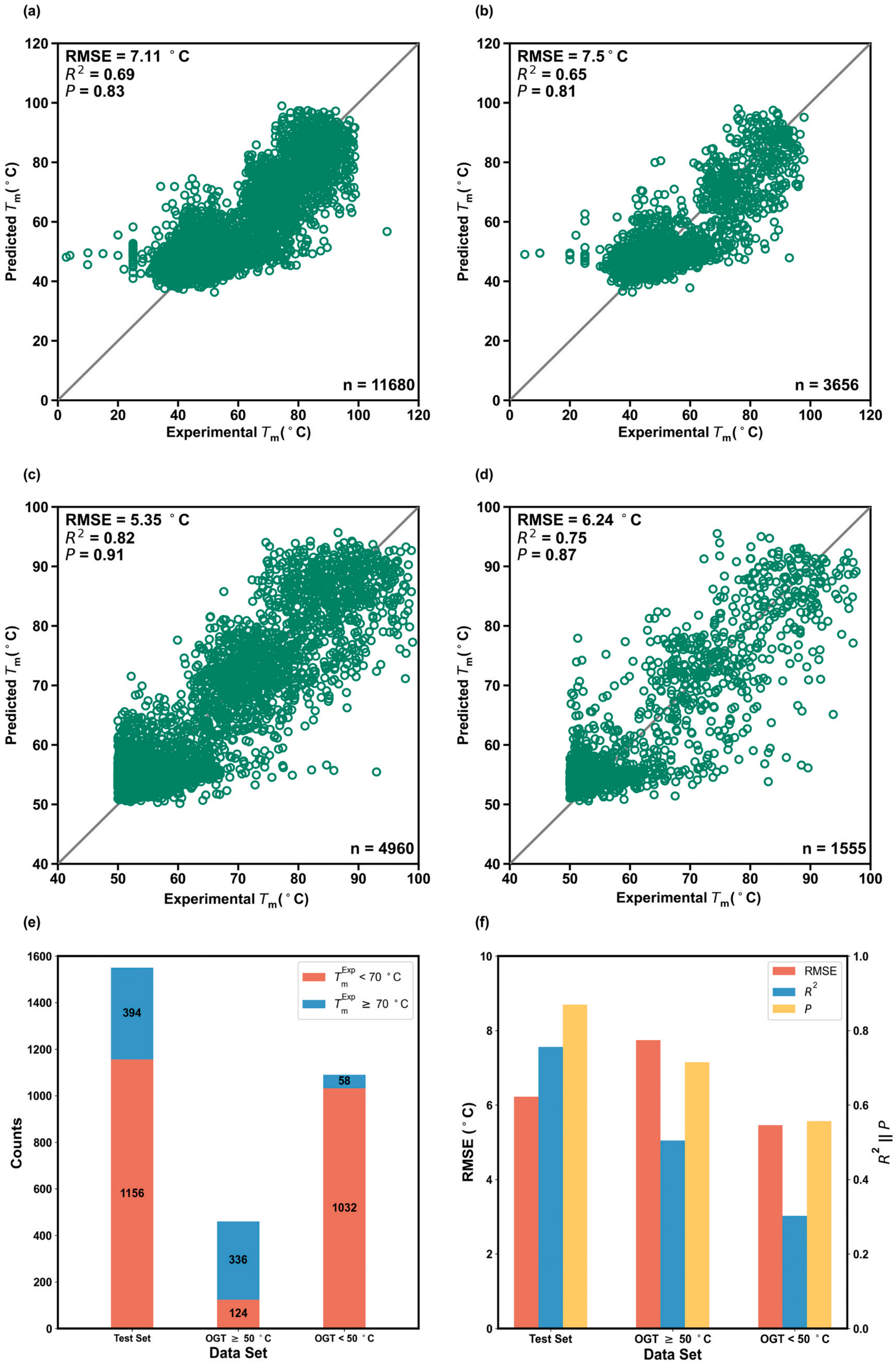
the OGT information of proteins. During the training process, the top-performing OGT model utilized the feature combination from the 11th group in Supplementary Table 7 for training, with the following hyper-parameter settings: a learning rate of 0.0001, 25 epochs for iteration, and a batch size of 1 for inputting into the neural network. In the subsequent section on T_m prediction task, we employ this model to forecast the OGT values of proteins.

3.3. Predicted results for T_m

We trained a deep neural network model called DeepTM, as shown in Fig. 1, using the T_m dataset to predict the protein's T_m values. On the training set, the model achieved the following performance evaluation results: $R^2 = 0.69$, $P = 0.83$, RMSE = 7.11 ℃ (Fig. 4a). As for the test set, the model exhibited the following performance: $R^2 = 0.65$, $P = 0.81$, RMSE = 7.5 ℃ (Fig. 4b). We hypothesize that protein sequences with higher T_m values are more capable of accurately capturing the thermal stability characteristics of proteins. To validate this hypothesis, protein sequences are initially classified into four categories based on their T_m values: hyperthermophilic proteins (> 70 ℃), thermophilic proteins (50–70 ℃), mesophilic proteins (25–50 ℃), and psychrophilic proteins (5–25 ℃)[57]. In this study, we exclusively focus on thermophilic proteins and hyperthermophilic proteins. Next, we select protein sequences with $T_m \geq 50$ ℃ from the training and test sets of the entire T_m dataset to create a new dataset, referred to as the Tm50 dataset in the following sections (Supplementary Figs 1b, 1c), and retrain the model on this new training set. On the training set of the Tm50 dataset, the performance evaluation of the model yields the following results: $R^2 = 0.81$, $P = 0.9$, RMSE = 5.52 ℃ (Fig. 4c). On the test set of the Tm50 dataset, the model demonstrates the following performance: $R^2 = 0.76$, $P = 0.87$, RMSE = 6.17 ℃ (Fig. 4d). It is worth noting that the latter model ($T_m \geq 50$ ℃) shows significant improvements in the evaluation metrics on the test set compared to the former. The coefficient of determination (R^2) has increased by 17.4%, the Pearson correlation coefficient (P) has increased by 7.4%, and the root mean square error (RMSE) has decreased by 22.3%. This result validates our hypothesis that training the model exclusively on protein sequences with $T_m \geq 50$ ℃ leads to a stronger predictive ability for the melting temperatures of thermophilic proteins. Subsequently, we re-adjusted the scale of the training dataset and retrained the model using experimental T_m values within the temperature ranges of [30, 60 ℃], [40, 60 ℃], [40, 100 ℃], and [50, 100 ℃] (Supplementary Table 8) to assess the impact of training set variations on model performance. The computational results indicate that the model performs optimally when using experimental T_m values within the [50, 100 ℃] range. We speculate that the inclusion of proteins with T_m values in the 30–50 ℃ range may have led the model to learn a significant number of features unrelated to thermal stability in mesophilic enzymes, which subsequently compromised the model's ability to predict T_m. In the subsequent research, we will conduct testing on the Tm50 dataset. Additionally, when training DeepTM, we will utilize the OGT model mentioned in the previous section to predict the OGT values of each target protein.

We have chosen temperature intervals of [50, 60 ℃], [60, 80 ℃], and [80, 100 ℃] to represent the expected accuracy levels of DeepTM as good, fair, and poor, respectively. In these three temperature intervals, the root mean square errors on the test set are 4.14 ℃, 8.07 ℃, and 9.50 ℃, respectively. Based on the root mean square error results, it aligns with the T_m distribution in the Tm50 dataset. The [50, 60 ℃] interval has the highest number of sequences, totaling 4785, accounting for 61% of the total. The [60, 80 ℃] interval has a relatively balanced distribution of sequences, with a total of 2005. The [80, 100 ℃] interval has a relatively lower number of sequences, with only 1023, making up 13% of the total. To further enhance the model's predictive performance in the 80–100 ℃ temperature range, additional protein data with T_m values in this range may be needed.

In order to investigate the impact of OGT values of the host organisms on the protein's T_m values, we divided the test set into two subsets based on OGT: one subset exclusively comprising protein sequences with OGT ≥ 50 ℃, and the other subset exclusively comprising protein sequences with OGT < 50 ℃ (Fig. 4e). We recalculated three evaluation



(caption on next page)

Fig. 4. Prediction results of T_m and impact of OGT features on the performance of DeepTM. RMSE: Root mean square error. R^2: Coefficient of determination. P: Pearson correlation coefficient. n: Number of data points. T_m: Melting temperature. T_m^{Exp}: Experimental value of T_m. (a-b) The model trained on the training set of the entire T_m dataset. The abscissa represents the experimental values of T_m, while the ordinate represents the predicted values of T_m. The points on the diagonal represent cases where the experimental values of T_m and the corresponding predicted values are identical. (a) Comparison between the experimental values and predicted values of T_m in the training set, (b) Comparison between the experimental values and predicted values of T_m in the test set. (c-d) The model trained on the training set of the Tm50 dataset ($T_m \geq 50$ ℃). (c) Comparison between the experimental values and predicted values of T_m in the training set of the Tm50 dataset ($T_m \geq 50$ ℃), (d) Comparison between the experimental values and predicted values of T_m in the test set of the Tm50 dataset ($T_m \geq 50$ ℃). (e) Distribution of protein sequences in the test set ($T_m \geq 50$ ℃). The abscissa represents three different datasets, namely the test set ($T_m \geq 50$ ℃), test subset (OGT ≥ 50 ℃), and test subset (OGT < 50 ℃). The ordinate represents the number of sequences in each dataset. Each dataset is divided into two parts: data with experimental T_m values ≥ 70 ℃ and data with experimental T_m values < 70 ℃. The numbers inside the bars represent the number of sequences in each category. (f) Analysis of the model's performance on the test set ($T_m \geq 50$ ℃). The left ordinate represents the RMSE, while the right ordinate represents either R^2 or P. RMSE, R^2, and P represent the relationship between the experimental values and predicted values of T_m.

metrics, namely R^2, P, and RMSE, for these two subsets of the test set. Subsequently, we compared these metrics with the evaluation metrics of the model on the entire test set (Fig. 4f). In the entire test set, as well as the test subset (OGT ≥ 50 ℃) and test subset (OGT < 50 ℃), both the coefficient of determination (R^2) and Pearson correlation coefficient (P) exhibit a decreasing trend. The Pearson correlation coefficient (P) indicates that the predicted T_m values exhibit extremely strong correlation, strong correlation, and moderate correlation with the experimental T_m values in these three datasets, respectively. Although the number of proteins in the test subset (OGT ≥ 50 ℃) accounts for only 29.68% of the entire test set, the number of proteins in the test subset (OGT < 50 ℃) represents 70.32% of the entire test set. However, the former calculates higher values for both R^2 and P, indicating that the model performs better in the test subset (OGT ≥ 50 ℃) compared to the latter. However, The fact that the proteins in the test subset (OGT ≥ 50 ℃) all have T_m values ≥ 50 ℃ suggests that proteins may exhibit optimal activity when approaching the host's optimal growth temperature, which is consistent with the findings in the references [57,58]. In terms of root mean square error (RMSE), the test subset (OGT ≥ 50 ℃) exhibits the highest calculated RMSE of 7.74 ℃, while the test subset (OGT < 50 ℃) demonstrates the lowest calculated RMSE of 5.46 ℃. The latter shows a relative reduction of 29.46% in RMSE compared to the former. This could be attributed to the fact that within the test subset (OGT ≥ 50 ℃), there are 336 proteins with experimental T_m values ≥ 70 ℃, accounting for 85.28% of the proteins with T_m values ≥ 70 ℃ in the entire test set. According to Eq. (9), the error is amplified when calculating RMSE, resulting in a higher RMSE for the test subset (OGT ≥ 50 ℃) compared to the entire test set. However, the overall prediction error remains within 10 ℃.

Furthermore, we compare the computational results of our model on the Tm50 dataset with those of other approaches [15–17,19,21,22,59] (see Supplementary Table 1). The outcomes demonstrate the outstanding performance of our model across three evaluation metrics, namely R^2, P, and RMSE. Method 7 [22] in Supplementary Table 1 represents one of the recent research achievements. Our results, compared to their results based on dataset 1 [22], exhibit improvements of 4.11%, 1.16%, and 22.88% in the evaluation metrics R^2, P, and RMSE, respectively. Furthermore, when compared to their results based on dataset 2 [22], our model demonstrates enhancements of 31.03%, 14.47%, and 2.06% in R^2, P, and RMSE, respectively. Prior to this, the majority of researchers employed traditional machine learning methods and statistical approaches to investigate the T_m values of proteins. We are the first to apply the protein graph structure pattern representation to the prediction of protein T_m values. Moreover, we utilized an exceptionally large T_m dataset, which greatly contributed to the improvement in the model's performance.

3.4. Deciphering the factors in the sequence that contribute to thermal stability

To identify factors in protein sequences that contribute to protein thermal stability, we tested different combinations of features to train the model. We combined seven features, namely PSSM, HMM, SPD33, seven physicochemical properties (PP7), amino acid frequency, dipeptide frequency, and OGT, and analyzed the impact of each feature on the model's performance. Specifically, we trained the model using 15 different combinations of features (Supplementary Table 7) and compared their performance. Firstly, we utilized the model trained with a combination of all features as the baseline model to evaluate the performance of models trained with other feature combinations. Subsequently, we validated seven feature combinations by removing one feature at a time from the baseline feature combination. This was done to assess the impact of each feature on the model's performance. Due to the strong correlation between protein OGT and T_m, we have removed the OGT feature from each of the seven feature combinations, resulting in the formation of six new feature combinations. Lastly, building upon the baseline feature combination, we have computed the OGT feature twice to enhance its weight, and evaluated the impact of the OGT feature on the model performance. We utilized the Tm50 dataset comprising 7790 protein sequences, with a training-to-test ratio of 8:2. The performance of the models trained with different feature combinations was evaluated through 5-fold cross-validation. Figs. 5a and 5b assess the predictive capabilities of the model using the coefficient of determination (R^2) and the Pearson correlation coefficient (P) calculated by the model on the test set, respectively. The results indicate that the best model performance is achieved when using PSSM, HMM, SPD33, PP7, amino acid frequency, dipeptide frequency, OGT value, and protein contact map as features. We observed that incorporating the OGT value as a feature during model training generally enhanced performance. Therefore, we hypothesize that increasing the weight of the OGT feature might lead to even better performance. Contrary to expectations, on the basis of the baseline feature combination, adding the OGT feature once again and retraining the model did not improve the model's performance. Furthermore, among individual features, dipeptide frequency had the greatest impact on the model, followed by HMM features, and OGT value ranked third (Fig. 5c and Fig. 5d).

We opted to conduct a separate analysis of the dipeptide frequency, which had the most significant impact on the model's performance among the feature combinations, as well as another feature closely related to it, namely amino acid frequency. In order to further explore their relationship with protein thermal stability and their impact on the model's performance. In the test set, we utilized the best model for predictions and divided the test set into two groups based on the difference between predicted and experimental T_m values: a protein dataset where the difference between predicted and experimental T_m values is ≤ 10 ℃ and a protein dataset where the difference is > 10 ℃. Among them, the protein dataset where the difference between predicted and experimental T_m values is ≤ 10 ℃ accounts for 90.52% of the entire test set. Subsequently, we computed the frequency of 20 amino acids and 400 dipeptides for both protein datasets and the entire test set separately. We then calculated the average values for each of the three datasets, as illustrated in Fig. 6. In the dataset where the difference between predicted T_m values and experimental T_m values is ≤ 10 ℃, the distribution of amino acid frequencies is roughly similar to the overall amino acid frequency distribution of the test set. The differences in amino acid frequencies are less than 0.002. The results reveal that the

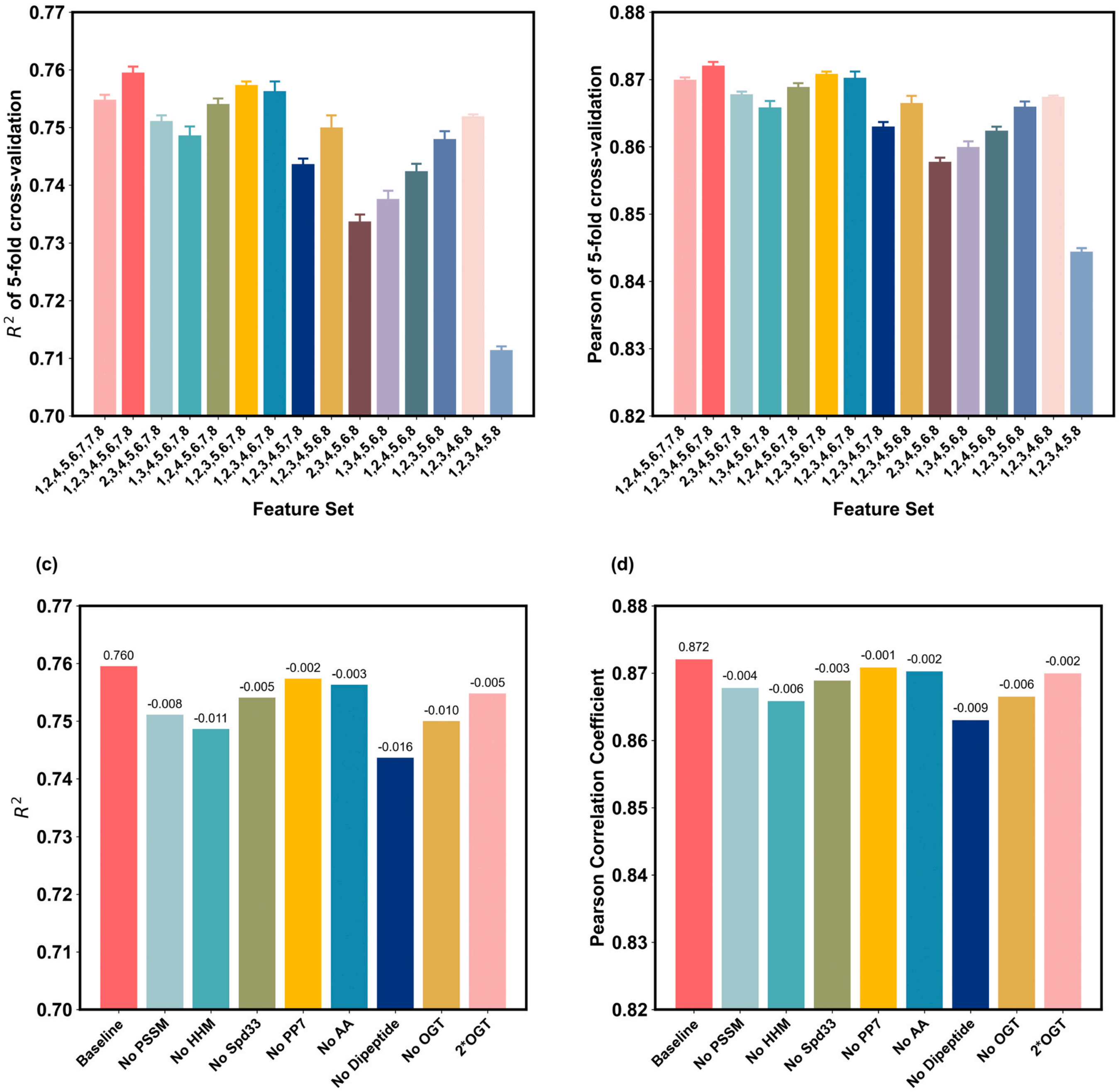
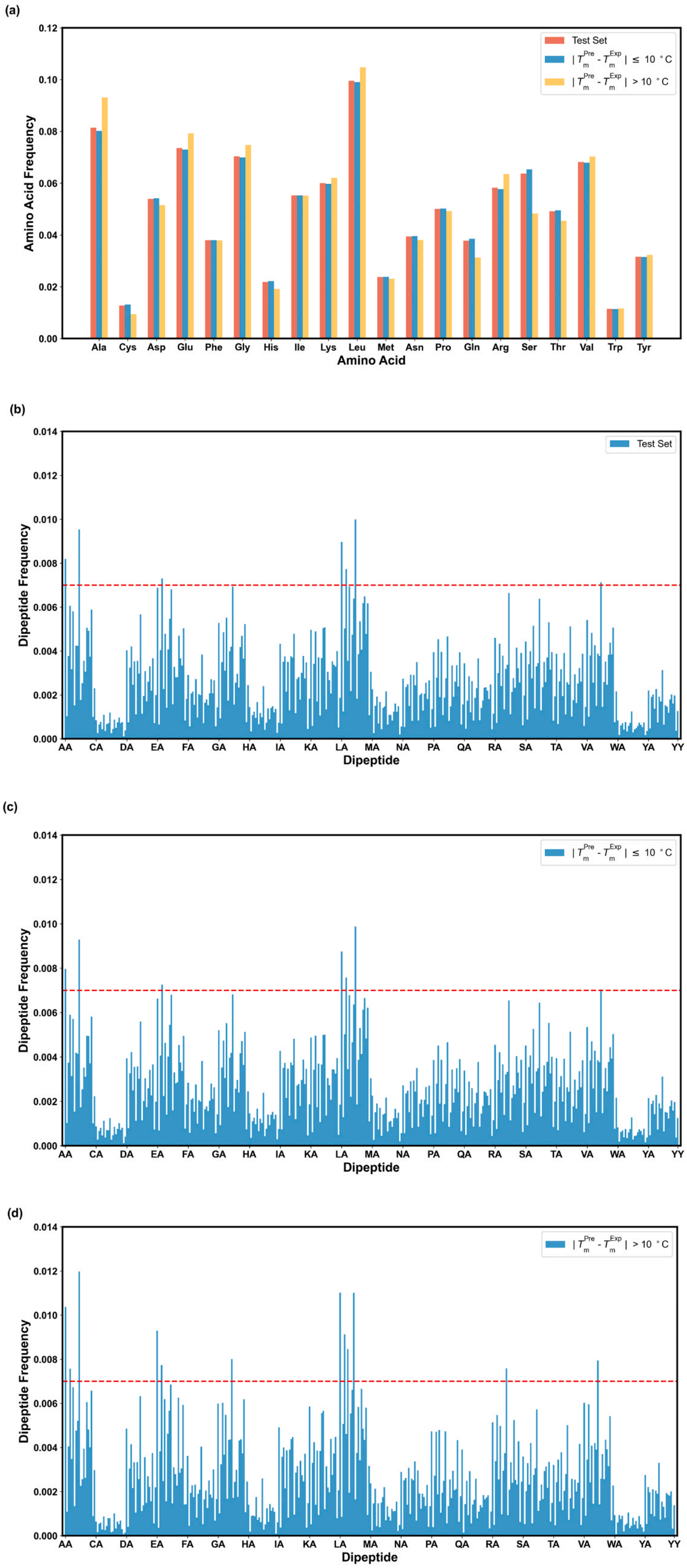


Fig. 5. Performance comparison of DeepTM models trained on different feature combinations using the training set ($T_m \geq 50$ ℃) and evaluated on the test set ($T_m \geq 50$ ℃). (a) R^2 scores of models trained on 15 different feature combinations evaluated on the test set ($T_m \geq 50$ ℃). The error bars represent the standard deviation of the R^2 scores obtained from 5-fold cross-validation. The abscissa represents different feature combinations, where the numbers 1–8 correspond to specific features: 1: PSSM, 2: HHM, 3: spd33, 4: PP7, 5: AA, 6: Dipeptide, 7: OGT, 8: Spotcon. The ordinate represents the R^2 scores obtained by the models. (b) The Pearson correlation coefficients (P) of models trained on 15 different feature combinations evaluated on the test set ($T_m \geq 50$ ℃). The error bars represent the standard deviation of the P scores obtained from 5-fold cross-validation. The ordinate represents the P scores obtained by the models. (c-d) Impact of individual features on the model ($T_m \geq 50$ ℃). (c) Comparison of R^2 scores on the test set ($T_m \geq 50$ ℃) for models trained with the addition or removal of a single feature, using the model trained with the second feature group in the feature combination as the baseline. The abscissa represents models trained with different feature combinations, arranged from left to right as follows: the baseline model, models trained with individual features removed from the feature combination, and the "2 *OGT" model, which indicates a model was trained by adding an additional OGT feature to the feature combination used for training the baseline model. The ordinate represents the R^2 score obtained by the model. The data labels above the bars represent the following: "Baseline" indicates the absolute R^2 score, while the others indicate the relative changes compared to the Baseline. Negative values indicate a decrease in the R^2 score, while positive values indicate an improvement in the R^2 score. (d) Comparison of P scores on the test set ($T_m \geq 50$ ℃) for models trained with the addition or removal of a single feature, using the model trained with the second feature group in the feature combination as the baseline. The ordinate represents the P score obtained by the model. The data labels above the bars represent the following: "Baseline" indicates the absolute P score, while the others indicate the relative changes compared to the Baseline. Negative values indicate a decrease in the P score, while positive values indicate an improvement in the P score.



(caption on next page)

Fig. 6. Impact of amino acid frequency and dipeptide frequency on protein thermal stability. (a) The occurrence frequency of amino acids in the test set ($T_m \geq$ 50 ℃) for all proteins. The test set is divided into two groups: the first group includes protein samples with a difference between predicted and experimental T_m values ≤ 10 ℃ (blue), and the second group includes samples with a difference > 10 ℃ (yellow). A comparison is made between the occurrence frequencies of amino acids in these two groups and the entire test set (orange). The abscissa represents the 20 types of natural amino acids, while the ordinate represents the frequency of occurrence of each amino acid. T_m: Melting temperature. (b) The frequency of occurrence of dipeptides in the test set ($T_m \geq$ 50 ℃) comprising all proteins. The abscissa represents 400 dipeptides arranged in ascending order based on their alphabetical abbreviation. Every 20 dipeptides are displayed with intervals on the abscissa. The ordinate represents the frequency of occurrence of dipeptides. The red dashed line represents a dipeptide frequency of 0.007. Dipeptides with frequencies greater than 0.007 include LL, AL, LA, AA, LE, EE, and VL. (c) The dipeptide frequency of proteins in the test set ($T_m \geq$ 50 ℃) with a difference between predicted and experimental T_m values ≤ 10 ℃. The dipeptides with frequencies greater than 0.007 include LL, AL, LA, AA, LE, EE, and VL. (d) The dipeptide frequency of proteins in the test set ($T_m \geq$ 50 ℃) with a difference between predicted and experimental T_m values > 10 ℃. The dipeptides with frequencies greater than 0.007 include LL, AL, LA, AA, LE, EE, VL, EA, LG, GL, RL, and AE.

frequencies of Leu, Ala, Glu, and Gly are greater than 0.07, whereas the frequencies of Trp, Cys, His, and Met are less than 0.03. In the dataset where the difference between predicted and experimental T_m values is > 10 ℃, the frequencies of amino acids such as Ala, Glu, Arg, Leu, and Gly were higher compared to the frequencies in the other two datasets. The amino acid frequency differences were greater than 0.003 and were arranged in descending order based on these differences. On the other hand, amino acids such as Ser, Gln, Thr, and Cys exhibited lower frequencies compared to the other two datasets. The differences in amino acid frequencies were greater than 0.003. The common characteristic of these amino acids is that they are all uncharged polar amino acids. This finding is consistent with the discoveries in the literature [57], indicating that amino acids such as Ala, Arg, Cys, Gly, Gln, and other uncharged polar amino acids significantly contribute to the thermal stability of proteins.

To analyze the correlation between the prediction accuracy of the model and the experimental T_m values of proteins in the dataset, we divided the test set into two groups based on the experimental T_m values: the hyperthermophilic group (T_m values ≥ 70 ℃) and the thermophilic group (50 ℃ ≤ T_m values < 70 ℃). We then calculated the amino acid frequencies for each group. Subsequently, we categorized the 20 amino acids based on the two different classification criteria for the two groups. The first group consists of two categories: a dataset where the difference between predicted and experimental T_m values is ≤ 10 ℃, and a dataset where the difference is > 10 ℃. The other group comprises two categories: the hyperthermophilic group and the thermophilic group. Each group includes these 20 natural amino acids, and based on the frequency of amino acids within the two categories of each group, the amino acids are classified into the category with the higher frequency (see Supplementary Table 9). We will sort the amino acids from left to right, in descending order, based on the difference in amino acid frequency between the two categories within each group. It is worth noting that the first three amino acid types in the dataset where the difference between predicted T_m values and experimental T_m values is ≤ 10 ℃ are exactly the same as the top three amino acid types in the thermophilic group. Additionally, the fourth amino acid type ranks fifth in the thermophilic group. There are 7 amino acid types common between the hyperthermophilic group and the dataset where the difference between predicted T_m values and experimental T_m values is > 10 ℃, accounting for 70% of the total amino acid types in the hyperthermophilic group. The distribution of amino acid frequencies corroborates our experimental findings, indicating that our model exhibits high accuracy in predicting the T_m values of thermophilic proteins. The RMSE is calculated to be 4.95 ℃. However, in predicting the T_m values of hyperthermophilic proteins, further improvement in the model's accuracy is needed, as the RMSE is measured to be 8.97 ℃.

To analyze the differences in dipeptide frequency distribution among the dataset where the difference between predicted and experimental T_m values is > 10 ℃, the dataset where the difference is ≤ 10 ℃, and the entire test set, we set a threshold of 0.007. Subsequently, we filtered out dipeptides with frequencies greater than this threshold in each of the three datasets (Fig. 6b-d). The results indicate that within the entire test set and the dataset where the difference between the predicted and experimental T_m is ≤ 10 ℃, the same dipeptide types exceed the threshold frequency. These include: Leu-Leu, Ala-Leu, Leu-Ala, Ala-Ala, Leu-Glu, Glu-Glu, and Val-Leu, arranged in descending order of dipeptide frequency. However, in the dataset where the difference between the predicted and experimental T_m is > 10 ℃, in addition to the aforementioned seven dipeptide types, it also includes: Glu-Ala, Leu-Gly, Gly-Leu, Arg-Leu, and Ala-Glu, likewise arranged in descending order of dipeptide frequency. These results indicate that our model performs well in learning dipeptide features, despite the potential negative impact of the presence of dipeptides such as Glu-Ala, Leu-Gly, Gly-Leu, Arg-Leu, and Ala-Glu on the model's predictive outcomes.

In our approach, we solely utilize protein sequences as input information without directly incorporating protein structure for predicting its thermal stability. However, we employ protein contact maps and descriptors of protein secondary structure to characterize the secondary structure features of proteins. Additionally, we utilize HMM, PSSM, and BLOSUM62 matrices to capture the evolutionary features of proteins. The interplay among these features is integrated into our neural network model.

3.5. *Prediction of T_m values for thermophilic plastic-degrading enzymes*

We collected a dataset comprising 22 thermally stable plastic-degrading enzymes [39] as a blind test dataset of protein sequences that have no overlap with the training and testing sets of the model. The dataset includes highly active and thermally stable PET plastic-degrading enzymes, namely LCC [60] and PETase [61], which have been discovered through current scientific research. These enzymes are used to validate the utility of the model. In the dataset of PET plastic-degrading enzymes, our model achieved a RMSE of 8.25 ℃ (Fig. 7a). Specifically, for protein YNPsite05_CeleraDRAFT_401410 [62], the experimental T_m value is 75.13 ℃, and the predicted T_m value is 73.29 ℃, resulting in a prediction error of 1.84 ℃. For PETase, the experimental T_m value is 56.8 ℃, and the predicted T_m value is 55.49 ℃, resulting in a remarkably low prediction error of only 1.31 ℃. Despite the difference between the OGT for these two proteins and their experimental T_m values, which are 17.53 ℃ and 29.5 ℃ respectively, our model still exhibits relatively low errors. This indicates that our model possesses good predictive capability even for proteins whose T_m values are not close to their OGT.

Our model performs excellently on the dataset of plastic-degrading enzymes, allowing us to predict the T_m values for enzymes selected from thermophilic microorganisms, such as AAZ54920.1, ADM47605.1, CAH17554.1, with a difference of around 2 ℃ between the predicted and experimental values. This outcome confirms our previous findings that our model exhibits high accuracy and low error in predicting the T_m values of thermophilic enzymes. Consequently, our model can be employed for the screening of novel thermophilic enzymes, demonstrating feasibility in practical applications. Further sequence feature analysis revealed (Fig. 7b, compared to Fig. 6) that the frequency distribution of amino acid types Arg, Ser, Thr, and Cys in PETase is similar to the frequency distribution in the dataset where the difference between predicted and experimental T_m values is ≤ 10 ℃. The frequency distribution of amino acid types Arg, Ser, Thr, as well as dipeptides Glu-Ala, Leu-Gly, Gly-Leu, Arg-Leu, and Ala-Glu in protein WP_117215036.1

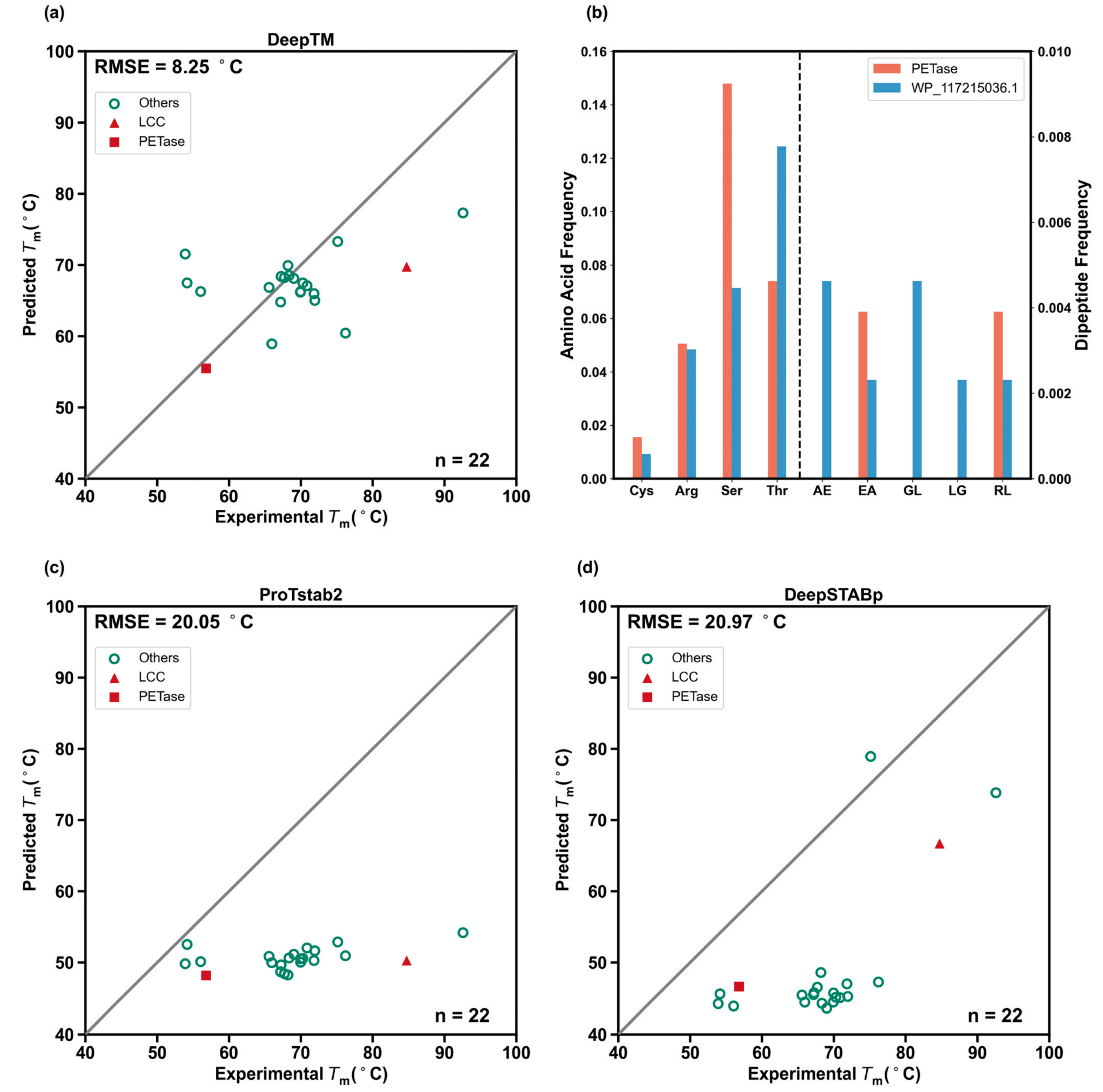


Fig. 7. Predicted T_m value of PET plastic-degrading enzymes. (a) Prediction of T_m values in the dataset of PET plastic-degrading enzymes using a model trained on a training set ($T_m \geq 50$ ℃), and a comparison between the experimental T_m values and the predicted T_m values. The abscissa represents the experimental T_m values, while the ordinate represents the predicted T_m values. Points along the diagonal line indicate instances where the experimental T_m values are equal to the predicted T_m values. RMSE: Root mean square error. n: Number of data points. Red triangle: LCC. Red rectangle: PETase. (b) The representative amino acid and dipeptide frequencies of protein PETase and protein WP_117215036.1 (NCBI). The abscissa represents amino acids or dipeptides, distributed on both sides of the dashed line. The left ordinate represents amino acid frequencies, while the right ordinate represents dipeptide frequencies. (c) The prediction results of the ProTstab2 model on the dataset of PET plastic-degrading enzymes, and the comparison between the experimental T_m values and the predicted T_m values. (d) The prediction results of the DeepSTABp model on the dataset of PET plastic-degrading enzymes, and the comparison between the experimental T_m values and the predicted T_m values.

also matches the frequency distribution of this dataset, consistent with the conclusions in the Predicted results for T_m section of this paper.

We conducted a comprehensive analysis of the PET plastic-degrading enzyme dataset, including amino acid frequencies (Supplementary Table 10), dipeptide frequencies (Supplementary Table 11), sequence similarity (Supplementary Table 12), and structural similarity (Supplementary Table 13). In the PET plastic-degrading enzyme dataset (Supplementary Table 2, Fig. 7a), approximately 73% of the sequences have experimental T_m values concentrated within the range of 65–75 ℃, with a high correlation between predicted and experimental values, maintaining an error margin within 7 ℃. The experimental T_m values of the remaining sequences are distributed in the ranges of 75–100 ℃ and 50–65 ℃. However, the correlation between their predicted and experimental T_m values is not significant, with an error range between

10 and 18 ℃, except for PETase. We compared proteins with T_m values between 75-100 ℃ and 50–65 ℃ to those within the range of 65–75 ℃. Across these three temperature ranges, amino acids exhibiting significant differences include Ala, Cys, and Gly (Supplementary Table 10). Furthermore, proteins in the 75–100 ℃ and 50–65 ℃ ranges show differences in one or more dipeptide frequencies compared to those in the 65–75 ℃ range (Supplementary Table 11). These disparities in amino acid and dipeptide frequencies may be contributing factors to the lower predictive correlation of T_m values for proteins within the 75–100 ℃ and 50–65 ℃ ranges. Within the 65–75 ℃ clustering, 53% of sequence similarity scores [63] exceed 90 (Supplementary Table 12). However, outside of these clusters, sequences WP_083947829.1 and WP_068752972.1, as well as 53% of sequences within the aforementioned clusters, exhibit similarity scores ranging from 70 to 90, yet their performance on DeepTM is suboptimal. We utilized TM-Align [64] to conduct structural similarity calculations for proteins with existing crystal structures in the PDB database (Supplementary Table 13). Specifically, this involved the proteins PETase, 5LUK_A, and LCC, with experimental T_m values distributed within the ranges of 50–65 ℃, 65–75 ℃, and 75–100 ℃, respectively. PETase and 5LUK_A exhibited relatively small prediction errors, at 1.31 ℃ and 0.51 ℃, respectively, while LCC showed a higher prediction error of 14.95 ℃. Nonetheless, their TM-Scores all exceeded 0.9, indicating a high level of structural similarity. Interestingly, the sequence similarity among these three proteins was relatively low. Hence, we deduce that the degree of sequence similarity does not directly impact the accuracy of T_m predictions, and likewise, the level of structural similarity cannot guarantee the accuracy of T_m predictions.

We directly compared the T_m prediction performance of DeepTM with the latest T_m prediction models, ProTstab2 [23] and DeepSTABp [24], on the dataset of PET plastic-degrading enzymes to assess the practicality of the models (Fig. 7a, Fig. 7c, Fig. 7d, and Supplementary Table 2). Since the dataset of plastic-degrading enzymes only consists of 22 protein sequences, we will solely utilize the RMSE as the evaluation criterion in this case. The results demonstrate that DeepTM exhibits a commendable generalization capability when applied to previously unseen datasets. Fig. 7a depicts the distribution of all data points clustered closely to the diagonal line, with a RMSE of 8.5 ℃. ProTstab2 achieved an RMSE of 20.05 ℃ on the dataset of plastic-degrading enzymes, while DeepSTABp obtained an RMSE of 20.97 ℃. In contrast, our model, DeepTM, demonstrated a reduction in RMSE for predicting T_m relative to ProTstab2 and DeepSTABp by 58.85% and 60.66%, respectively. Specifically, the predicted errors for protein PETase in DeepTM, DeepSTABp, and ProTstab2 are 1.31 ℃, 10.18 ℃, and 8.62 ℃, respectively. DeepTM demonstrates superior predictive performance in this case.

However, for proteins with T_m values exceeding 80 ℃, such as LCC and GxsBSedJan11_10009658, the performance of our model in terms of prediction is moderate. Specifically, the prediction error for LCC is 14.95 ℃, and the prediction error for GxsBSedJan11_10009658 is 15.26 ℃. This could be attributed to the fact that proteins with T_m values exceeding 80 ℃ accounted for only 13.04% of our dataset, which may have resulted in the model not fully capturing the relationship between the features related to the thermal stability of these proteins and their T_m values.

3.6. Prediction of T_m values for thermally stable proteins

We selected a set of proteins with a broader class as the second external validation dataset [40,41,65] (Supplementary Table 3), aiming to further verify the generalization performance of DeepTM. On this dataset, we used both DeepTM and ProTstab2 to predict the proteins' T_m values (Supplementary Table 3, Fig. 8, Supplementary Figure 6). The results demonstrated that DeepTM continues to perform accurately on the thermally stable protein dataset, showcasing strong generalization capabilities. DeepTM achieved an RMSE of 7.66 ℃ on this dataset, with the prediction error for 20 sequences being within 7 ℃, accounting for 69% of the total. This includes 12 membrane proteins, 5 antibodies, 1 transcription factor, and 2 esterase sequences. In contrast, ProTstab2 achieved an RMSE of 15.87 ℃, with only seven sequences having a

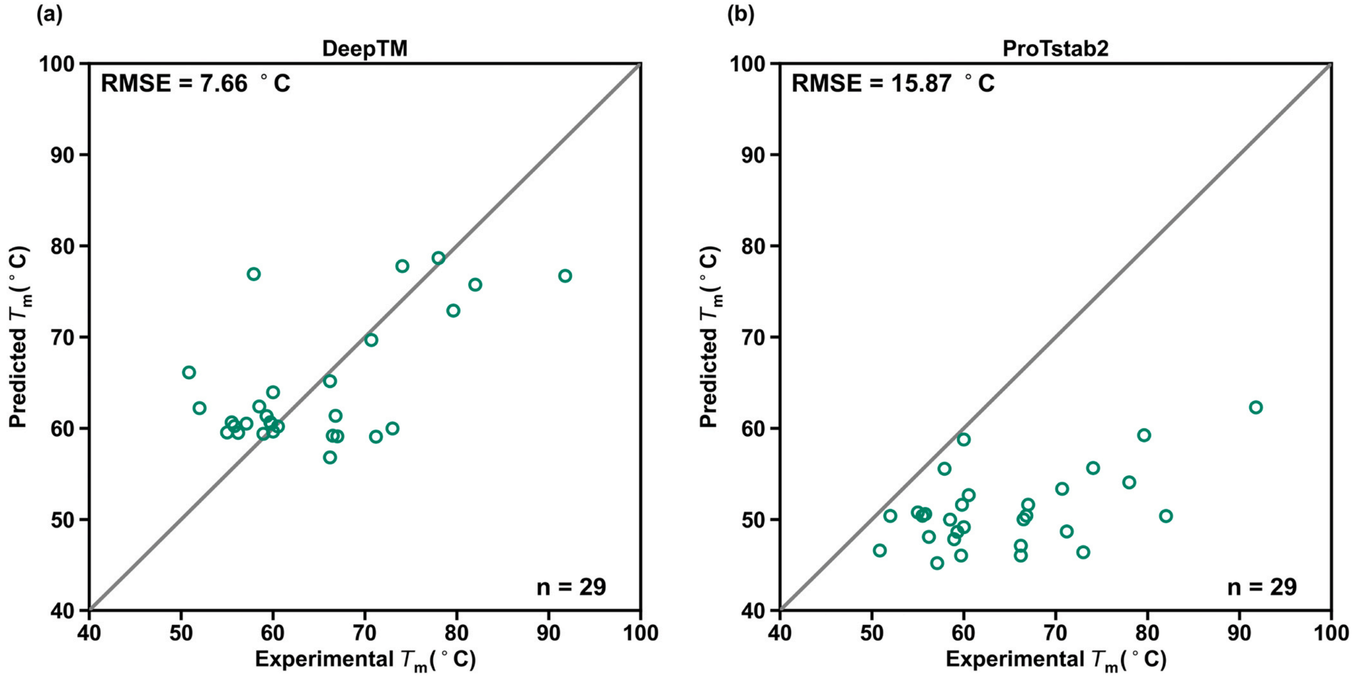


Fig. 8. Predicted T_m value of thermally stable proteins. (a) Prediction of T_m values in the dataset of thermally stable proteins using a model trained on a training set ($T_m \geq 50$ ℃), and a comparison between the experimental T_m values and the predicted T_m values. The abscissa represents the experimental T_m values, while the ordinate represents the predicted T_m values. Points along the diagonal line indicate instances where the experimental T_m values are equal to the predicted T_m values. RMSE: Root mean square error. n: Number of data points. (b) The prediction results of the ProTstab2 model on the dataset of thermally stable proteins, and the comparison between the experimental T_m values and the predicted T_m values.

prediction error within 7 ℃, representing 24% of the total. Seven proteins had prediction errors ranging from 20 to 32 ℃, also accounting for 24% of the total. For transcription factors and esterase sequences, their prediction errors exceeded 15 ℃. Relatively, DeepTM model reduced the root mean square error for T_m prediction in the thermally stable protein dataset by 51.73% compared to ProTstab2. Additionally, the residuals between the predicted T_m values of DeepTM and the experimental T_m values (Supplementary Figure 6) are distributed in the upper and lower range of 0 ℃, while the residuals of ProTstab2 are all above 0 ℃.

4. Conclusion

Here, we have developed an end-to-end deep learning model called DeepTM, based on graph convolutional neural networks and self-attention networks. The objective of DeepTM is to predict T_m values of proteins based only on sequences. We trained the model using a dataset consisting of 7790 sequences of thermophilic proteins. On the test set of 1550 samples, the model demonstrates a high coefficient of determination ($R^2 = 0.76$), a strong Pearson correlation coefficient ($P = 0.87$), and a low root mean square error (RMSE = 6.17 ℃). These findings robustly validate the precise depiction of thermal stability attributes of thermophilic proteins by our model. DeepTM, based on graph convolutional neural networks, incorporates protein contact maps as one of its features, enabling the model to predict protein melting temperatures with enhanced precision. The improved accuracy can be attributed to the ability of protein contact maps to directly represent the 2D structural characteristics of proteins and capture the underlying relationships between protein residue-residue pairs. In addition, we applied the concept of Stacking [66–68] from machine learning: using the OGT model as an excellent base model and DeepTM as the meta-model. Following this, we incorporated the output of the base model as one of the inputs for the meta-model training to enhance the predictive performance of the model. This approach not only avoids potential issues related to circular inference arising from using experimental OGT data but also effectively leverages features associated with OGT. This may have contributed to the improvement in model performance. Furthermore, we have also discovered that dipeptide frequencies, OGT values of the host organisms, and evolutionary information of proteins significantly influence the model's prediction of the melting temperature of thermophilic proteins. While we did not directly use experimentally obtained OGT data in the training process of DeepTM, relying solely on our OGT model to predict protein's OGT values, it is undeniable that incorporating predicted OGT features into the model significantly improved its performance (Fig. 5). However, assigning higher weights to these predicted OGT features did not further enhance the predictive capabilities of the model (Fig. 5). However, DeepTM does not account for differences in experimental conditions. In future research, we may explore incorporating the protonation states of titratable amino acids under different pH conditions based on their isoelectric points as an additional feature in the neural network.

We evaluated the performance of DeepTM, ProTstab2, and DeepSTABp in practical applications by validating them on blind test datasets containing 22 PET plastic-degrading enzymes and 29 thermally stable proteins with a broader classification. By comparing the root mean square error (RMSE) of these three algorithms on the PET plastic-degrading enzyme dataset, we observed that DeepTM (RMSE = 8.25 ℃) significantly outperformed ProTstab2 (RMSE = 20.05 ℃) and DeepSTABp (RMSE = 20.97 ℃). On the thermally stable protein dataset, we found that DeepTM (RMSE = 7.66 ℃) demonstrated a 51.73% reduction in RMSE compared to ProTstab2 (RMSE = 15.87 ℃). This indicates that DeepTM exhibits superior generalization performance on the blind test datasets. Furthermore, compared to DeepSTABp, DeepTM only requires the protein sequence as input for predicting protein melting temperature, eliminating the need for additional experimental conditions such as TPP and information about the protein's host optimal growth temperature. DeepTM has achieved a fully end-to-end T_m prediction process, making the prediction of T_m more convenient. As an underlying framework, DeepTM can be easily extended to other tasks involving the design of thermophilic proteins that require the utilization of protein sequences.

Author statement

We the undersigned declare that this manuscript entitled "DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences" is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process. He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

CRediT authorship contribution statement

Mengyu Li: Investigation, Methodology, Data curation, Software, Visualization, Writing – original draft. **Hongzhao Wang**: Data curation, Investigation, Methodology, Software, Visualization. **Zhenwu Yang**: Investigation, Validation, Visualization. **Longgui Zhang**: Conceptualization, Resources, Supervision. **Yushan Zhu**: Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the financial support from the National Key Research and Development of China (grant number: 2021YFC2103600), the National Natural Science Foundation of China (grant number: 21878170), and the Fundamental Research Funds for the Central Universities (buctrc202136 and JD2103).

Appendix A. Supplementary data

Supplementary data to this article can be found in the Supporting Information file.

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.11.006.

References

[1] Atalah J, Cáceres-Moreno P, Espina G, Blamey JM. Thermophiles and the applications of their enzymes as new biocatalysts. Bioresour Technol 2019;280: 478–88. https://doi.org/10.1016/j.biortech.2019.02.008.

[2] Nezhad NG, Rahman RNZRA, Normi YM, Oslan SN, Shariff FM, et al. Thermostability engineering of industrial enzymes through structure modification. Appl Microbiol Biotechnol 2022;106(13):4845–66. https://doi.org/10.1007/s00253-022-12067-x.

[3] Sharma S, Vaid S, Bhat B, Singh S, Bajaj BK. Chapter 17 - Thermostable enzymes for industrial biotechnology. In: Singh RS, Singhania RR, Pandey A, Larroche C, editors. Advances in Enzyme Technology. Elsevier; 2019. p. 469–95.

[4] Brown LR. Commercial challenges of protein drug delivery. Expert Opin. Drug Deliv. 2005;2(1):29–42. https://doi.org/10.1517/17425247.2.1.29.

[5] Wang G, Cao RY, Chen R, Mo L, Han JF, et al. Rational design of thermostable vaccines by engineered peptide-induced virus self-biomineralization under physiological conditions. Proc Natl Acad Sci 2013;110(19):7619–24. https://doi.org/10.1073/pnas.1300233110.

[6] Tiller KE, Tessier PM. Advances in antibody design. Annu Rev Biomed Eng 2015;17 (1):191–216. https://doi.org/10.1146/annurev-bioeng-071114-040733.

[7] Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci 2006;103(15):5869–74. https://doi.org/10.1073/pnas.0510098103.

[8] Kan SBJ, Lewis RD, Chen K, Arnold FH. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. Science 2016;354(6315): 1048–51. https://doi.org/10.1126/science.aah6219.

[9] Rigoldi F, Donini S, Redaelli A, Parisini E, Gautieri A. Review: Engineering of thermostable enzymes for industrial applications. APL Bioeng 2018;2(1). https://doi.org/10.1063/1.4997367.

[10] Finch AJ. Thermophilic proteins as versatile scaffolds for protein engineering. Microorganisms 2018;6(4):97. https://doi.org/10.3390/microorganisms6040097.

[11] Camps M, Herman A, Loh E, Loeb LA. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol 2007;42(5):313–26. https://doi.org/10.1080/10409230701597642.

[12] Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res 2006;34(1):D204–6. https://doi.org/10.1093/nar/gkj103.

[13] Consortium TU. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2022;51(D1):D523–31. https://doi.org/10.1093/nar/gkac1052.

[14] Zhou MYC, Li W. Comparation of three measuring methods for thermodynamic stability of protein. Anal Test Technol Instrum 2021;27(4):252–9. https://doi.org/10.16495/j.1006-3757.2021.04.004.

[15] M. Gorania, H. Seker, P.I. Haris. Predicting a protein's melting temperature from its amino acid sequence, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 1820–1823.

[16] Zhang G, Liu G, Fang B. A study on the recognition of thermophilic and mesophilic proteins based on support vector machine. Comput Appl Chem 2006;23(8):707–10. https://doi.org/10.16866/j.com.app.chem2006.08.005.

[17] Pucci F, Dhanani M, Dehouck Y, Rooman M. Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. PLoS One 2014;9(3):e91659. https://doi.org/10.1371/journal.pone.0091659.

[18] Dehouck Y, Folch B, Rooman M. Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. Protein Eng Des Sel 2008;21(4): 275–8. https://doi.org/10.1093/protein/gzn001.

[19] Ku T, Lu P, Chan C, Wang T, Lai S, et al. Predicting melting temperature directly from protein sequences. Comput Biol Chem 2009;33(6):445–50. https://doi.org/10.1016/j.compbiolchem.2009.10.002.

[20] Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 1999;82 (1):51–67. https://doi.org/10.1016/S0301-4622(99)00103-9.

[21] Pucci F, Kwasigroch JM, Rooman M. SCooP: an accurate and fast predictor of protein stability curves as a function of temperature. Bioinformatics 2017;33(21): 3415–22. https://doi.org/10.1093/bioinformatics/btx417.

[22] Li G, Buric F, Zrimec J, Viknander S, Nielsen J, et al. Learning deep representations of enzyme thermal adaptation. Protein Sci 2022;31(12):e4480. https://doi.org/10.1002/pro.4480.

[23] Yang Y, Zhao J, Zeng L, Vihinen M. ProTstab2 for prediction of protein thermal stabilities. Int J Mol Sci 2022;23(18):10798. https://doi.org/10.3390/ijms231810798.

[24] Jung F, Frey K, Zimmer D, Mühlhaus T. DeepSTABp: a deep learning approach for the prediction of thermal protein stability. Int J Mol Sci 2023;24(8):7444. https://doi.org/10.3390/ijms24087444.

[25] Pucci F, Rooman M. Towards an accurate prediction of the thermal stability of homologous proteins. J Biomol Struct Dyn 2016;34(5):1132–42. https://doi.org/10.1080/07391102.2015.1073631.

[26] Lihan M, Lupyan D, Oehme D. Target-template relationships in protein structure prediction and their effect on the accuracy of thermostability calculations. Protein Sci 2023;32(2):e4557. https://doi.org/10.1002/pro.4557.

[27] Ngo K, Bruno da Silva F, Leite VBP, Contessoto VG, Onuchic JN. Improving the thermostability of xylanase a from bacillus subtilis by combining bioinformatics and electrostatic interactions optimization. J Phys Chem B 2021;125(17):4359–67. https://doi.org/10.1021/acs.jpcb.1c01253.

[28] Liu C, Zhao J, Liu J, Guo X, Rao D, et al. Simultaneously improving the activity and thermostability of a new proline 4-hydroxylase by loop grafting and site-directed mutagenesis. Appl Microbiol Biotechnol 2019;103(1):265–77. https://doi.org/10.1007/s00253-018-9410-x.

[29] Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 2018;34(23): 4039–45. https://doi.org/10.1093/bioinformatics/bty481.

[30] Chen J, Zheng S, Zhao H, Yang Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. J Chemin- 2021;13(1):7. https://doi.org/10.1186/s13321-021-00488-1.

[31] Desai DK, Nandi S, Srivastava PK, Lynn AM. ModEnzA: Accurate identification of metabolic enzymes using function specific profile HMMs with optimised discrimination threshold and modified emission probabilities. Adv Bioinformatics 2011;2011:12. https://doi.org/10.1155/2011/743782.

[32] Chang J, Zhang C, Cheng H, Tan Y-W. Rational Design of Adenylate Kinase Thermostability through Coevolution and Sequence Divergence Analysis. Int J Mol Sci 2021;22(5):2768. https://doi.org/10.3390/ijms22052768.

[33] Kipf T.N., Welling M. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations (ICLR), 2017.

[34] Lu H, Diaz DJ, Czarnecki NJ, Zhu C, Kim W, et al. Machine learning-aided engineering of hydrolases for PET depolymerization. Nature 2022;604(7907): 662–7. https://doi.org/10.1038/s41586-022-04599-z.

[35] Li G, Rabe KS, Nielsen J, Engqvist MKM. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. ACS Synth Biol 2019;8(6):1411–20. https://doi.org/10.1021/acssynbio.9b00099.

[36] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17(3):282–3. https://doi.org/10.1093/bioinformatics/17.3.282.

[37] Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21): 3128–30. https://doi.org/10.1093/bioinformatics/btu500.

[38] Jarzab A, Kurzawa N, Hopf T, Moerch M, Zecha J, et al. Meltome atlas—thermal proteome stability across the tree of life. Nat Methods 2020;17(5):495–503. https://doi.org/10.1038/s41592-020-0801-4.

[39] Erickson E, Gado JE, Avilán L, Bratti F, Brizendine RK, et al. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. Nat Commun 2022;13(1):7850. https://doi.org/10.1038/s41467-022-35237-x.

[40] Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res 2021;49(D1):D420–4. https://doi.org/10.1093/nar/gkaa1035.

[41] Kulandaisamy A, Sakthivel R, Gromiha MM. MPTherm: database for membrane protein thermodynamics for understanding folding and stability. Brief Bioinforma 2021;22(2):2119–25. https://doi.org/10.1093/bib/bbaa064.

[42] Mount DW. Using BLOSUM in sequence alignments. Cold Spring Harb Protoc 2008; 3(6). https://doi.org/10.1101/pdb.top39.

[43] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 1992;89(22):10915–9. https://doi.org/10.1073/pnas.89.22.10915.

[44] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389.

[45] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–9. https://doi.org/10.1093/nar/gkaa1100.

[46] Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J Mol Biol 1994;235(5): 1501–31. https://doi.org/10.1006/jmbi.1994.1104.

[47] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinforma 2019;20(1):473. https://doi.org/10.1186/s12859-019-3019-7.

[48] Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170–6. https://doi.org/10.1093/nar/gkw1081.

[49] Hasan AKMM, Ahmed AY, Mahbub S, Rahman MS, Bayzid MS. SAINT-Angle: self-attention augmented inception-inside-inception network and transfer learning improve protein backbone torsion angle prediction. Bioinforma Adv 2023;3(1): vbad042. https://doi.org/10.1093/bioadv/vbad042.

[50] Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 2017;33(18):2842–9. https://doi.org/10.1093/bioinformatics/btx218.

[51] Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Mol Model Annu 2001;7(9):360–9. https://doi.org/10.1007/s008940100038.

[52] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, et al. PyTorch: an imperative style, high-performance deep learning library. NeurIPS 2019:32.

[53] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

[54] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv Prepr. 2014. https://doi.org/10.48550/arXiv.1412.6980.

[55] Xavier G, Antoine B, Yoshua B. Deep sparse rectifier neural networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics 2011.

[56] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. NeurIPS 2017:30.

[57] Vieille C, Zeikus Gregory J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65(1): 1–43. https://doi.org/10.1128/mmbr.65.1.1-43.2001.

[58] Engqvist MKM. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiol 2018;18(1):177. https://doi.org/10.1186/s12866-018-1320-7.

[59] Miotto M, Armaos A, Di Rienzo L, Ruocco G, Milanetti E, et al. Thermometer: a webserver to predict protein thermal stability. Bioinformatics 2022;38(7):2060–1. https://doi.org/10.1093/bioinformatics/btab868.

[60] Tournier V, Topham CM, Gilles A, David B, Folgoas C, et al. An engineered PET depolymerase to break down and recycle plastic bottles. Nature 2020;580(7802): 216–9. https://doi.org/10.1038/s41586-020-2149-4.

[61] Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate. Science 2016;351(6278): 1196–9. https://doi.org/10.1126/science.aad6359.

[62] Yang Y, Malten M, Grote A, Jahn D, Deckwer W-D. Codon optimized Thermobifida fusca hydrolase secreted by Bacillus megaterium. Biotechnol Bioeng 2007;96(4): 780–94. https://doi.org/10.1002/bit.21167.

[63] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23(21):2947–8. https://doi.org/10.1093/bioinformatics/btm404.

[64] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9. https://doi.org/10.1093/nar/gki524.

[65] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res 2023;51(D1):D488–508. https://doi.org/10.1093/nar/gkac1077.

[66] Wolpert DH. Stacked generalization. Neural Netw 1992;5(2):241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.

[67] Breiman L. Stacked regressions. Mach Learn 1996;24(1):49–64. https://doi.org/10.1007/BF00117832.

[68] van der Laan M.J., Polley E.C., Hubbard A.E. Super Learner. 2007;6(1). doi: 10.2202/1544–6115.1309.