Bioinformatics Advances, 2024, **00**, vbae103
https://doi.org/10.1093/bioadv/vbae103
Advance Access Publication Date: 13 July 2024
Original Article

OXFORD

Sequence analysis

TemBERTure: advancing protein thermostability prediction with deep learning and attention mechanisms

Chiara Rodella [1,2,‡], **Symela Lazaridi** [1,2,‡], **Thomas Lemmin** [1,*]

[1]Institute of Biochemistry and Molecular Medicine (IBMM), University of Bern, Bern CH-3012, Switzerland
[2]Graduate School for Cellular and Biomedical Sciences (GCB), University of Bern, Bern CH-3012, Switzerland

*Corresponding author. Institute of Biochemistry and Molecular Medicine (IBMM), University of Bern, Bühlstrasse 28, Bern, CH-3012, Switzerland. E-mail: thomas.lemmin@unibe.ch

‡These authors contributed equally.

Associate Editor: Michael Gromiha

Abstract

Motivation: Understanding protein thermostability is essential for numerous biotechnological applications, but traditional experimental methods are time-consuming, expensive, and error-prone. Recently, deep learning (DL) techniques from natural language processing (NLP) was extended to the field of biology, since the primary sequence of proteins can be viewed as a string of amino acids that follow a physicochemical grammar.

Results: In this study, we developed TemBERTure, a DL framework that predicts thermostability class and melting temperature from protein sequences. Our findings emphasize the importance of data diversity for training robust models, especially by including sequences from a wider range of organisms. Additionally, we suggest using attention scores from Deep Learning models to gain deeper insights into protein thermostability. Analyzing these scores in conjunction with the 3D protein structure can enhance understanding of the complex interactions among amino acid properties, their positioning, and the surrounding microenvironment. By addressing the limitations of current prediction methods and introducing new exploration avenues, this research paves the way for more accurate and informative protein thermostability predictions, ultimately accelerating advancements in protein engineering.

Availability and implementation: TemBERTure model and the data are available at: https://github.com/ibmm-unibe-ch/TemBERTure.

1 Introduction

Biocatalysts have become integral to numerous industrial processes, driving e.g. advances in pharmaceutical, food, and biofuel productions (Himmel *et al.* 2007, Kuddus 2018, Singh *et al.* 2016). In these applications, protein thermostability plays a crucial role (Adams and Kelly 1995, Bommarius *et al.* 2006). Proteins that can endure high temperatures are essential for accelerating and enhancing chemical reactions, leading to reduced production costs (Singh *et al.* 2016). However, exposure to elevated temperatures can cause denaturation and loss of biological activity (Matsuura *et al.* 2015), underscoring the importance of improving our understanding of protein thermostability.

Despite notable progress in experimental techniques for measuring protein thermostability, the process remains time-consuming and challenging to scale up, resulting in limited data on protein thermostability (Stourac *et al.* 2021). Currently, ProThermDB is the largest dataset of experimental thermodynamic data for protein stability (Nikam *et al.* 2021), encompassing a comprehensive collection of 32 000 proteins, of which 38% are wild-type sequences and 51% single point mutations. Recently, novel experimental techniques have emerged that allow for the determination of the thermal stability of proteins across the entire genome of a cell. These techniques involve the integration of mass spectrometry with limited proteolysis (Leuenberger *et al.* 2017), or liquid chromatography (Jarzab *et al.* 2020). In addition to experimental techniques, the growth temperature of organisms is commonly employed as a proxy for protein thermostability (Ahmed *et al.* 2022a, Chakravarty and Varadarajan 2002, Lin and Chen 2011, Modarres *et al.* 2018, Vieille and Zeikus 2001).

Protein thermostability is a complex interplay between a protein's intrinsic properties, encoded in its amino acid sequence and structure, and extrinsic factors such as pH, solvent conditions, and the presence of stabilizing agents. Extrinsic factors can be important *in vivo*. However, large datasets of *in vivo* thermostability data along with the metadata are still currently lacking, thus making *in vivo* thermostability modeling difficult. On the other hand, understanding and predicting the inherent stability encoded in a protein's sequence and structure is crucial, in particular, for biotechnological applications. By optimizing intrinsic thermostability, proteins become less reliant on specific external conditions, increasing their versatility and applicability in diverse biotechnological settings.

Statistical comparisons of thermophilic and non-thermophilic protein sequences have identified key features associated with thermostability, including higher proportions of hydrophobic and charged residues and specific dipeptide motifs of thermophilic proteins (Fukuchi and Nishikawa 2001, Vieille and

Received: April 30, 2024; **Revised:** June 14, 2024; **Editorial Decision:** July 8, 2024; **Accepted:** July 12, 2024

Zeikus 2001, Ding et al. 2004, Liang et al. 2005, Zhou et al. 2008). A higher occurrence of hydrogen bonds, salt bridges, disulfide bonds, and hydrophobic interactions is also observed in thermophilic proteins (Haney et al. 1997, Sadeghi et al. 2006, Bleicher et al. 2011, Bashirova et al. 2019).

Extensive research has led to the development of several machine learning models aimed at predicting protein thermostability, treating it as a classification task (Gromiha and Xavier Suresh 2007, Zhang and Fang 2006, 2007, Wu et al. 2009, Lin and Chen 2011, Nakariyakul et al. 2012, Tang et al. 2017, Charoenkwan et al. 2021, 2022). Early models such as Thermopred employed a support vector machines (SVM) classifier trained on a dataset of 793 non-thermophilic and 915 thermophilic protein sequences (Lin and Chen 2011), which became a foundation for training subsequent models (Nakariyakul et al. 2012, Tang et al. 2017). An expanded version of this dataset, consisting of 1368 thermophilic and 1443 non-thermophilic proteins, was utilized for training the iThermo model, a multi-layer perceptron (MLP) (Ahmed et al. 2022a) and the Sapphire framework, a staking-based ensemble model (Charoenkwan et al. 2022). Other models have approached the problem as a regression task to directly predict the melting temperature (Yang et al. 2019, 2022).

Transformer-based models such as bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018) have improved natural language processing (NLP). By considering proteins as a string of amino acids, NLP can be applied to biology and more specifically to protein modeling and classification. ProtTrans (Elnaggar et al. 2022), a family of models including protBERT, leverages transformers to extract protein characteristics from sequence data. BertThermo (Pei et al. 2023) uses the protBERT embeddings with classical machine learning models for thermophilicity classification, whereas DeepSTABp incorporates ProtTrans-XL embeddings and growth temperature to predict protein melting temperature (Jung et al. 2023). Similarly, TemStaPro (Pudžiuvelytė et al. 2024) is an ensemble of models incorporating ProtT5-XL (Elnaggar et al. 2022) embeddings to feed-forward densely connected neural network models, and ProLaTherm (Haselbeck et al. 2023) integrates the encoder part of a T5-3B (Raffel et al. 2020) model with ProtT5-XL (Elnaggar et al. 2022) as the feature extractor.

To overcome the shortcomings of present model approaches, we developed TemBERTure, a deep learning package for protein thermostability prediction. It consists of three components: (i) TemBERTure$_{DB}$, a large-curated database of thermophilic and non-thermophilic sequences, (ii) TemBERTure$_{CLS}$, a classifier, and (iii) TemBERTure$_{Tm}$, a regression model, which predicts, respectively, the thermal class (non-thermophilic or thermophilic) and melting temperature of a protein, based on its primary sequence. Both models are built upon the existing protBERT-BFD language model (Elnaggar et al. 2022) and fine-tuned through an adapter-based approach (Houlsby et al. 2019, Poth et al. 2023). Our findings demonstrate the remarkable capability of deep learning to differentiate protein classes based on their sequences. However, they also highlight its limitations due to the current lack of available data. Despite these limitations, the insights gained from the attention scores within these models offer promising clues to unraveling the underlying mechanisms of protein thermostability and thus can suggest new research directions in biotechnology and protein engineering.

2 Methods

This section is composed of four main parts. Part 1 outlines the workflow for establishing comprehensive curated databases of thermophilic and non-thermophilic protein sequences sourced from various experiments and data collection, with TemBERTure$_{DB}$ as the primary training resource and two additional databases used for bias and generalization assessment. Parts 2 and 3 describe the architecture and training of TemBERTure$_{CLS}$ and TemBERTure$_{Tm}$, respectively. And finally, Part 4 provides the technical details used for the analyses.

2.1 Database creation

2.1.1 TemBERTure$_{DB}$

TemBERTure$_{DB}$ leveraged data from the Meltome Atlas experiment (Jarzab et al. 2020). We obtained pre-processed protein sequences from the ProtStab2 dataset (Yang et al. 2022). These sequences were supplemented by retrieving all sequences from UniProtKB (The UniProt Consortium 2023) corresponding to the same 13 organisms as in the Meltome Altas. To address the class imbalance between thermophilic and non-thermophilic sequences, we enriched the thermophilic dataset by sourcing additional data from the BacDive database (Reimer et al. 2022). Here, we classified sequences based on the growth temperature of their respective organisms: thermophilic (>60°C) and non-thermophilic (<30°C). Protein sequences were retrieved for each organism from the NCBI database (Sayers et al. 2022). Ambiguous and short (<30 amino acids) sequences were excluded. MMseqs (Hauser et al. 2016) was then employed to cluster the sequences within each dataset, using a threshold of 50% for thermophilic and 80% for non-thermophilic. To further address the class imbalance, we augmented the non-thermophilic dataset with challenging examples. These examples were retrieved from non-thermophilic organisms (BacDive) and exhibited high sequence similarity (80% < identity < 95%) to the thermophilic sequences (Fig. 1A). The final TemBERTure$_{DB}$ was stored as an SQL database facilitating efficient data retrieval for downstream analyses (Supplementary Table S1).

2.1.2 BacDIVE

Within the BacDive database, organisms were classified based on growth temperature: thermophilic (>60°C) and non-thermophilic (<30°C). Protein sequences were then retrieved for each organism from the NCBI database, and ambiguous or short sequences (<30 amino acids) were excluded. Given the substantial disparity between the number of non-thermophilic and thermophilic sequences, we used MMseqs in cascading mode to cluster the non-thermophilic sequences. We then undersampled the centroids (representatives of each cluster) to align with the number of thermophilic centroids identified using MMseqs with a 50% identity threshold (Supplementary Table S2).

2.1.3 Meltome

We leveraged data curated within TemBERTure$_{DB}$ and excluded the non-thermophilic counterparts of the high-similarity sequence pairs retrieved from the BacDive database (Supplementary Table S3).

2.1.4 Splitting

For model training, we partitioned the datasets into an 80:10:10 ratio for the training, validation, and test sets, respectively. To mitigate any potential information leakage

between sets, all sequences were clustered with MMseqs at a 50% identity threshold. Centroids and their corresponding clusters were then assigned to the same split.

For the regression task, we exclusively used the initial Meltome dataset. Melting temperatures were categorized into temperature bins of 10°C, and 10 data points from each temperature bin were randomly selected for both the test and validation sets. To address the imbalance in the distribution of melting temperatures within the training set, we implemented a combination of undersampling and oversampling techniques. Temperature bins with an abundance of data points (40–55°C) were undersampled, whereas bins with a scarcity of data points (20–40 and 60–90°C) were oversampled. This approach ensured a balanced number of data points across all temperature bins.

2.2 TemBERTure$_{CLS}$

TemBERTure$_{CLS}$ (Fig. 1B) is a sequence-based classifier that takes the amino acid sequence as input and outputs the corresponding thermal class of the protein along with its associated score. It was built on top of the pre-trained protBERT-BFD model (Elnaggar *et al.* 2022), a BERT model composed of 30 layers, 16 heads, and 1024 hidden layers and trained on over 2 billion protein sequences from the BFD100 (Steinegger and Söding 2018, Steinegger *et al.* 2019) dataset. In order to reduce the number of trainable parameters and enhance the efficiency of the training process, we opted for an adapter-based fine-tuning technique (Houlsby *et al.* 2019, Poth *et al.* 2023), where light weight bottleneck layers are inserted between each transformer layer.

TemBERTure$_{CLS}$ was thus implemented as a BertAdapterModel with Pfeiffer adapters (Pfeiffer *et al.* 2021) configuration using the PyTorch framework *via* adapters (Houlsby *et al.* 2019, Poth *et al.* 2023) library. It was initiated with the proBERT-BFD (Elnaggar *et al.* 2022) weights through the HuggingFace API (Wolf *et al.* 2019) and the Pfeiffer adapter architecture layers were added after the feed-forward block of each transformer layer (Vaswani *et al.* 2017, Wolf *et al.* 2020). In this way, we reduced the number of trainable parameters from 420 million to 5 million.

2.2.1 Training

Protein sequences were tokenized at the amino acid level utilizing the protBERT-BFD (Elnaggar *et al.* 2022) tokenizer, with all sequences truncated to a maximum length of 512. For each dataset, a separate hyperparameter search was carried out to optimize the training and architecture of the model (Supplementary Table S4). This hyperparameter search was performed using W&B Sweeps (Biewald 2020) grid hyperparameter search. The adapter training was carried out for a maximum of 20 epochs for each dataset with a batch size of 16, using AdamW optimizer (Loshchilov and Hutter 2017) with default Hugging Face (Wolf *et al.* 2019) configuration. The model that achieved the lowest validation loss was then saved for evaluation. To ensure model robustness, the final configuration of each model was trained three times under identical conditions, varying only the random seed. This approach allowed us to assess the model's independence from specific random seeds and to confirm its reliability across different runs. All models were trained on a single NVIDIA A100 80G GPU.

2.3 TemBERTure$_{Tm}$

TemBERTure$_{Tm}$ is a sequence-based regression model designed to predict the protein melting temperature (Tm) directly from its amino acid sequence. This model has the same

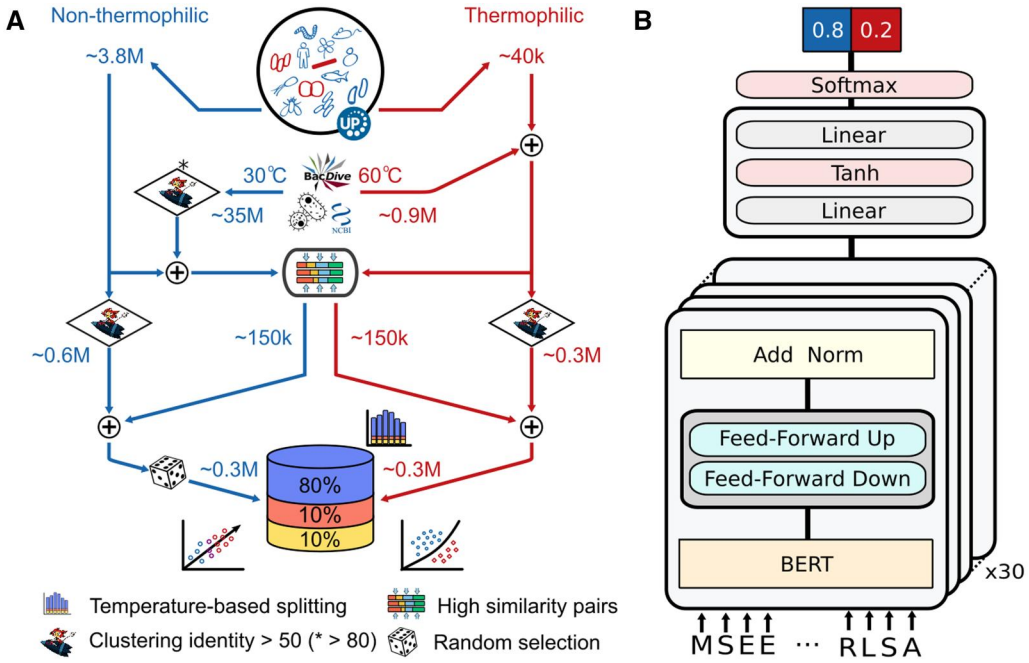

Figure 1. TemBERTure database creation and model architecture. (A) TemBERTure$_{DB}$ creation pipeline: protein sequences from organisms within the Meltome Atlas were retrieved from the UniProt database and categorized based on their thermophilicity (right: thermophilic, left: non-thermophilic). Additional sequences were then collected from BacDive (Reimer *et al.* 2022) and NCBI (Sayers *et al.* 2022) databases at various temperature thresholds to augment the dataset. The final database comprises approximately 0.3 million each for thermophilic and non-thermophilic proteins, further divided into training, testing, and validation sets that are representative of the temperature distribution. (B) TemBERTure$_{CLS}$ model architecture was based on the proBERT-BFD framework, with lightweight bottleneck adapter layers inserted between each transformer layer (shown in gray). The model takes a protein sequence as input and outputs a score indicating the classification score of the sequence being thermophilic or non-thermophilic.

underlying architecture configuration and tokenization as TemBERTure$_{CLS}$, with a regression head. Leveraging the pre-trained protBERT-BFD model, we adopted again an adapter-based fine-tuning technique to reduce trainable parameters.

2.3.1 Training

The model was trained on a curated dataset created specifically for predicting protein melting temperatures, based on TemBERTure$_{DB}$. All sequences are truncated to a maximum length of 512. The training was carried out for a maximum of 200 epochs for each run with a batch size of 16 and using AdamW optimizer (Loshchilov and Hutter 2017) with default Hugging Face (Wolf *et al.* 2019) values. We conducted, with W&B Sweeps (Biewald 2020), an extensive search to identify the optimal configuration of the regression head (Supplementary Table S5). We then explored various weight initialization approaches for the model. In addition to random initialization, we investigated transfer learning from TemBERTure$_{CLS}$ at different training stages. This involved introducing classifier weights at 25%, 50%, 75%, and 100% of the first epoch, along with weights from the fully trained classifier. To assess model stability and consistency across random initializations, all models were trained three times with different random seeds. For each configuration, the model achieving the lowest validation loss was saved for further evaluation. All training runs utilized a single NVIDIA A100 80G GPU.

2.4 Analyses

2.4.1 Ensemble evaluation for melting temperature prediction

To improve prediction accuracy, we evaluated different ensembles of models on the validation set. We built these ensembles by selecting subsets of the initial 18 models. These 18 models encompassed all distinct initialization methods (random and transfer learning with TemBERTure$_{CLS}$ weights) and their replicates. We investigated three ensemble approaches: greedy algorithm, weighted ensemble, and a method leveraging TemBERTure$_{CLS}$. Additionally, we experimented with various averaging techniques (standard deviation and interquartile range, IQR) to combine predictions and identify the optimal value for each data point. Overall, these ensemble strategies aimed to harness the strengths of multiple models and achieve a configuration effective across a broad temperature range. Detailed descriptions are provided in the Extended Methods in the Supplementary Information.

2.4.2 High attention score

The IQR method was used to identify amino acids within a protein sequence with a high attention score (HAS). We calculated a threshold by adding 1.5 times the IQR to the third quartile ($Q3$) of the attention scores. Attention scores exceeding this threshold are flagged as outliers, indicating a noticeably HAS and potentially significant influence on the model's decisions.

3 Results

3.1 TemBERTure$_{DB}$

To train our deep learning models for predicting protein thermostability, we curated TemBERTure$_{DB}$, a comprehensive dataset built upon the Meltome Atlas (Jarzab *et al.* 2020) that includes data for over 48 000 proteins across 13 different species (Fig. 1A). We further enriched it with all protein sequences from UniProtKB for each organism (The UniProt Consortium 2023). This initially resulted in a highly imbalanced dataset with only 44 000 sequences from thermophilic organisms (growth temperature above 60°C) compared to 4.3 million sequences from non-thermophilic organisms (growth temperatures: 16–36°C). To address this imbalance, we incorporated thermophilic proteomes from BacDive, adding 0.9 million sequences (Reimer *et al.* 2022). However, the thermophilic dataset remained biased toward bacterial and archaeal sequences. Therefore, we included similar bacterial sequences (<30°C growth temperature) with high identity (>80%) to thermophiles. This added valuable non-thermophilic examples outside the target class, for a more challenging training set (Supplementary Table S1).

To ensure that both classes contained diverse protein families and folds, we clustered each class separately using MMseqs (Hauser *et al.* 2016), resulting in a balanced dataset of 300 000 sequences per class. We partitioned it into training, validation, and test sets at an 80:10:10 ratio, ensuring that sequences with high similarity remained within the same split, to avoid information leakage. To enhance model learning and generalization, pairs of highly similar sequences from different classes were exclusively reserved for training, effectively bridging the gap between thermophilic and non-thermophilic sequences (Supplementary Table S6).

3.2 TemBERTure$_{CLS}$

TemBERTure$_{DB}$ served as the training dataset for TemBERTure$_{CLS}$, a sequence-based classifier designed to predict the thermal class of a protein solely from its amino acid sequence (Fig. 1B). TemBERTure$_{CLS}$ is a binary classifier, where the thermophilic class is defined as a sequence coming from organisms with a growth temperature above 60°C. TemBERTure$_{CLS}$ leveraged protBERT-BFD, a pre-trained protein language model (Elnaggar *et al.* 2022), and utilized adapter layers (Houlsby *et al.* 2019, Poth *et al.* 2023) for efficient task-specific learning. This approach offers faster (up to 50%) and more robust training (avoiding catastrophic forgetting) than full fine-tuning, thus enabling rapid model experimentation and optimization without sacrificing performance.

TemBERTure$_{CLS}$ achieved an overall accuracy of 0.89, an $F1$-score of 0.9, and a Matthews correlation coefficient (MCC) of 0.78, with balanced predictive performance across both classes (0.88 and 0.90 $F1$-score for non-thermophilic and thermophilic sequence, respectively). Low standard deviation across multiple trained models confirms robust training. We, therefore, chose to retain the initially trained model as the final TemBERTure$_{CLS}$ model. When comparing the performance of TemBERTure$_{CLS}$ to state-of-the-art models, we observed that many of the latter tend to overpredict the non-thermophilic class (Fig. 2). Despite achieving a competitive average precision of 0.79 for thermophilic sequences, their recall fell below 0.7, resulting in numerous misclassifications of non-thermophilic proteins. This highlights the limitations in the generalizability of current methods (Supplementary Table S7).

To assess the generalization of TemBERTure$_{CLS}$, we tested it on the widely used iThermo dataset (Ahmed *et al.* 2022a) and the TemStaPro test set (Pudžiuvelytė *et al.* 2024). After removing similar sequences (over 50% identity), the final test sets contained 65 and 1495 thermophilic sequences and 505 and 10 849 non-thermophilic sequences for iThermo and

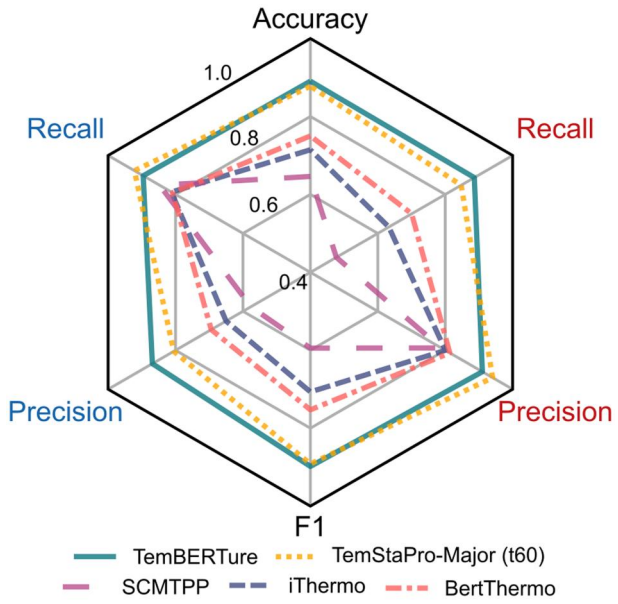

Figure 2. Comparison of TemBERTure$_{CLS}$ with state-of-the-art models on the TemBERTure$_{DB}$ test set. Recall and precision are shown separately for thermophilic (right) and non-thermophilic (left) thermal categories.

TemStaPro, respectively. This substantial reduction in dataset size resulted in highly imbalanced non-overlapping sets. Consequently, we evaluated TemBERTure$_{CLS}$ performance using the macro-averaged *F*1-score, recall, and precision (Supplementary Table S8). TemBERTure$_{CLS}$ maintained high accuracy, achieving 86% on iThermo and 83% on TemStaPro. To explore TemBERTure$_{CLS}$ ability to perform on sequences from novel organisms, we created a new test set with sequences from organisms in the BacDive database (Reimer *et al.* 2022). Although non-thermophilic sequence precision remained high (0.81), precision for thermophilic sequences dropped (0.74), suggesting limitations in generalizing to completely new organisms.

To further investigate this observation, we trained separate models, with the same architecture as TemBERTure$_{CLS}$, with two distinct datasets: one derived from BacDive (Reimer *et al.* 2022), focusing solely on bacterial and archaeal organisms, and another one from the Meltome Atlas (Jarzab *et al.* 2020), augmented solely with thermophilic sequences (Supplementary Tables S2 and S3). Each model performed well on the dataset derived from the same source as its training data. However, performances dropped significantly when tested on the other datasets (Fig. 3A). These variations were less pronounced for the thermophilic class, most likely because all datasets used BacDive for selecting thermophilic organisms. In contrast, the non-thermophilic class exhibited greater performance variations. The BacDive-trained model's performance dropped significantly, when tested on the TemBERTure$_{DB}$ or Meltome$_{DB}$ data (almost random classifications), whereas TemBERTure$_{CLS}$ and the Meltome-trained model maintained comparable performance across all datasets, indicating the necessity of using diverse training datasets to improve generalizability. To assess potential data leakage between training and test sets, we clustered TemBERTure$_{DB}$ test sequences based on their maximum identity to training set sequences (Fig. 3B). TemBERTure$_{CLS}$ performance remained consistent across all identity ranges for the non-thermophilic class. A decrease in performance was observed specifically within the thermophilic class for sequences with

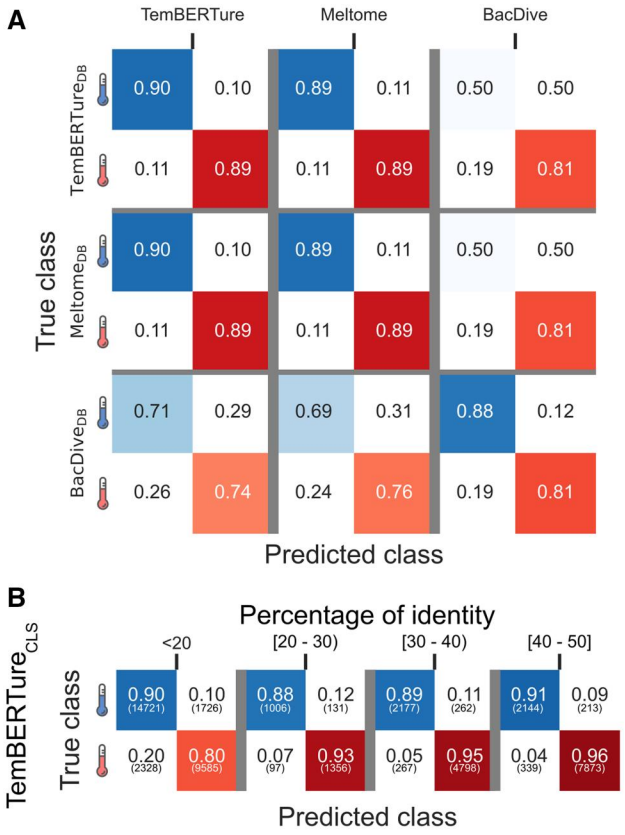

Figure 3. TemBERTure database creation and model architecture. (A) Confusion matrix comparing the performance of the TemBERTure$_{CLS}$ model with models trained on data derived from only BacDive and Meltome. The evaluation is performed on three separate test sets: TemBERTure$_{DB}$, BacDive$_{DB}$, and Meltome$_{DB}$ test sets. Each cell in the matrix represents the proportion of predictions made by a specific model on a specific test set. (B) TemBERTure$_{CLS}$ performance on TemBERTure$_{DB}$ test sequences clustered by maximum identity to training data. Shades of blue (top row) indicates correct predictions for the non-thermophilic category, while shades of red (bottom row) represent the performance for thermophilic sequences. Off-diagonal entries indicate instances of misclassification.

less than 20% identity. However, the performance remained comparable to the one previously observed when using a test set from a different source than the training data. This could be attributed to either overfitting to specific training data patterns or the inherent difficulty of classifying these sequences (e.g. orphan proteins).

3.3 TemBERTure$_{Tm}$

Building on these promising TemBERTure$_{CLS}$ results, we developed TemBERTure$_{Tm}$, to predict protein melting temperature (Tm) from its primary sequence. Extracting the readily available protein melting temperature data from the Meltome Atlas, we again leveraged protBERT-BFD and adapter layers for training TemBERTure$_{Tm}$. Even though the model achieved a seemingly high Pearson correlation of 0.78, a more detailed analysis revealed a clear limitation (Fig. 4A). The predicted temperatures displayed a surprising bimodal distribution, concentrated around non-thermophilic (below 60°C) and thermophilic (above 80°C) ranges. This suggests a bias toward classifying temperatures into these broad categories rather than accurately predicting the melting points. This bias agrees with the weak correlation within each class (0.41 for non-thermophilic, −0.33 for thermophilic) and high

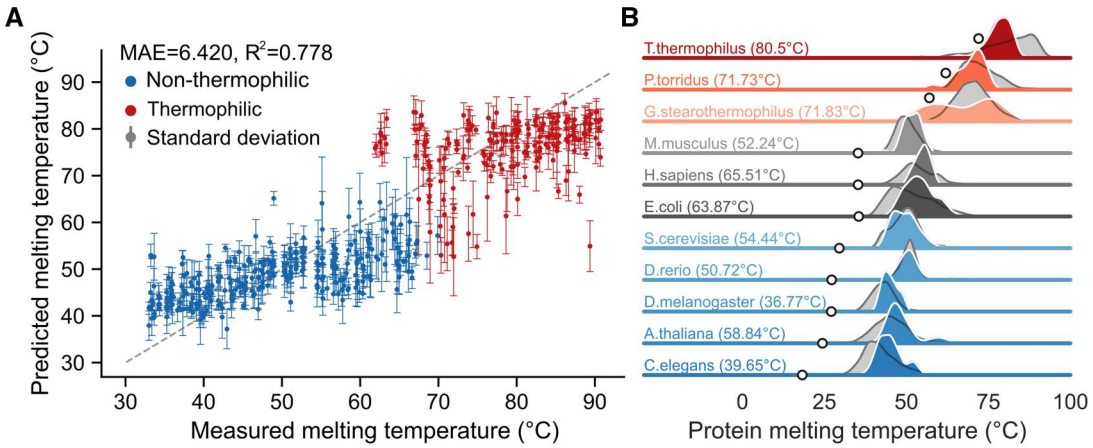

Figure 4. Predicted melting temperatures. (A) Scatter plot comparing the measured melting temperatures to predicted melting temperature. Each point is colored base on the thermal category (blue: non-thermophilic and red: thermophilic). The dashed gray line represents a perfect prediction. Standard deviations are calculated from the predictions of three replicates. (B) Distributions of melting temperature for various organisms, represented by a colored gradient ranging from red (high growth temperature) to blue (low growth temperature). The measured melting temperature distributions are shown in gray, while the predicted distributions using TemBERTure$_{Tm}$ are shown in color. Gray circles mark the growth temperatures of each organism and the temperatures noted in parentheses indicating the average melting temperatures of the organism's proteome.

accuracy (82%) of TemBERTure$_{Tm}$ as a classifier using a 70°C threshold. Moreover, TemBERTure$_{Tm}$ displayed significant variability among replicates trained with different random seeds, suggesting instability and limitations within the training process.

Given the limited size (around 30 000 sequences) of the Meltome Atlas dataset, we explored transfer learning. We hypothesized that pre-trained adapter weights from TemBERTure$_{CLS}$, which captured thermal class features, could improve TemBERTure$_{Tm}$ regression performance. Our approach involved replacing the random initialization of the adapter layers with weights from various stages of the classification training process. Since TemBERTure$_{Tm}$ prediction followed a bimodal distribution, we chose different training stages for the adapter weights, aiming to balance leveraging learned thermal features and enabling the regression to move beyond this bias. However, this approach did not yield any significant improvements in performance.

In order to improve the performance, we explored diverse ensembling strategies (see Extended Methods in Supplementary Information). First, we established an upper bound on achievable performance using an oracle approach. From all TemBERTure$_{Tm}$ variations, the oracle selected the prediction from all TemBERTure$_{Tm}$ variations that was closest to the experimentally measured melting temperature. This yielded a best-case scenario with an MAE of 2.64°C and an R^2 of 0.94 on the test set, highlighting the potential of the underlying approach. However, the ensemble techniques only led to a marginal change in performance (Supplementary Table S9). A more promising approach involved leveraging thermal class information. We first predicted a protein's class (non-thermophilic or thermophilic) using TemBERTure$_{CLS}$ to predict the thermal class (non-thermophilic or thermophilic) of the protein sequence. Then, we selected a subset of best performing TemBERTure$_{Tm}$ models for each class. This resulted in a combination of five models for non-thermophilic predictions (all transfer learning) and two models for thermophilic predictions (Supplementary Table S9), i.e. one with random weights and one with partial first-epoch weights. This highlights the importance of incorporating class information, achieving a decrease in MAE (6.31°C) and an increase in R^2 (0.78) on the test set compared to other ensembling techniques.

Despite limitations in predicting individual melting points, TemBERTure$_{Tm}$ showed promise in capturing broader thermal properties. We used the model to predict the melting temperatures of unmeasured proteins from organisms within the Meltome Atlas. Interestingly, the predicted distribution mirrored the known distribution of measured melting temperatures across diverse organisms (Fig. 4B). This suggests that, although TemBERTure$_{Tm}$ has some difficulties in predicting individual values, it still might capture underlying patterns related to protein thermostability across species.

3.4 Interpretability

To explore the intricate relationships between amino acid properties and thermostability, we conducted an analysis of the attention mechanisms in the TemBERTure$_{CLS}$ model. Attention mechanisms offer an interpretable scoring function, highlighting segments of the input sequence that are most important for a particular prediction by assigning them higher scores. In the context of TemBERTure$_{CLS}$, this would allow for a comprehensive identification of crucial amino acids and regions within a sequence that may influence the thermostability prediction. We defined HAS regions as exceeding the IQR of attention values across the entire sequence. All analyses were performed using the first replica of TemBERTure$_{CLS}$.

3.4.1 Effect of fine-tuning

To investigate the impact of fine-tuning on the model's attention patterns, we compared the frequencies of HAS amino acids between the pre-trained protBERT-BFD model and TemBERTure$_{CLS}$. We hypothesized that changes in HAS frequencies might correlate with features linked to thermostability. Although the overall attention scores remained remarkably similar between the two models, we observed a shift in the frequency of HAS for specific amino acids (Fig. 5A). For thermophilic proteins, leucine, arginine, and alanine appeared more frequently as HAS, whereas the frequency only increased for leucine in non-thermophilic sequences (Supplementary Fig. S1).

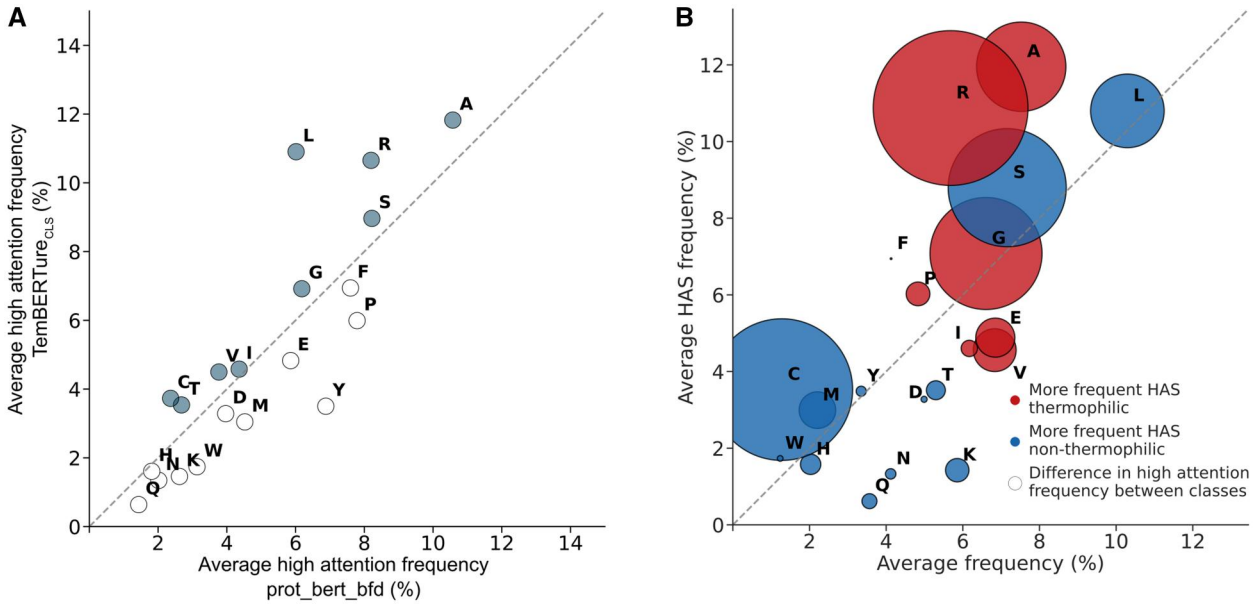

Figure 5. Frequency of high attention score (HAS) by amino acid. (A) Scatter plot comparing the frequency of HAS amino acids of the pre-trained proBERT-BFD model to TemBERTure$_{CLS}$. Each point represents an amino acid and is colored in gray if the frequency of HAS increased in TemBERTure$_{CLS}$. (B) Bubble plot comparing the frequency of each amino acid in the test set to its HAS frequency. Red bubble indicates that the frequency of HAS is higher for thermophilic and blue bubbles for non-thermophilic. Each bubble is scaled to the difference in frequency between both classes.

3.4.2 Amino acids enrichment

We conducted a more in-depth analysis by comparing the enrichment levels of each amino acid within the protein sequences with their natural occurrence frequencies. We calculated the background frequency of each amino acid in the TemBERTure$_{DB}$ test set and compared it to the frequency at which they appeared as HAS (Fig. 5B and Supplementary Fig. S2). This analysis revealed distinct patterns between thermophilic and non-thermophilic proteins. For example, we observed an increase in HAS frequency for several hydrophobic residues, such as alanine, phenylalanine, and leucine, which potentially reflect their role in stabilizing the protein core through tight packing. Interestingly, cysteine, which is known for forming stabilizing disulfide bridges and coordinating metals (Pace and Weerapana 2014), received higher attention in non-thermophiles. Glutamine and asparagine, susceptible to deamidation at high temperatures (Ahern and Klibanov 1985, Tomazic and Klibanov 1988, Rahimzadeh *et al.* 2012), showed decreased HAS, in agreement with their expected scarcity in these organisms. TemBERTure$_{CLS}$ also showed a clear preference for different charged amino acids, with an increase in HAS for arginine and a decrease in HAS for lysine. However, it is crucial to underscore the potential complexity in interpreting HAS scores. An increase in HASs might suggest functional importance; however, their interpretation requires caution due to dependence on the local amino acid environment. Conversely, decreased HAS for specific amino acids might not indicate a negative impact, but rather reflect the model's focus on their specific critical interactions within the protein structure.

3.4.3 Structural analysis

In order to gain some structural insights from the attention scores, we analysed 17 pairs of homologous thermophilic and non-thermophilic proteins correctly classified by TemBERTure$_{CLS}$. These pairs shared moderate sequence similarity (identity score: 0.28–0.54). Although the overall attention patterns between homologous proteins showed some correlation, the HAS amino acids exhibited more variability. Between homologous proteins, the model assigned a similar number of HAS to both conserved and non-conserved amino acids (Fig. 6A and Supplementary Fig. S3). Interestingly, the specific amino acids receiving HAS often differed between homologs, even in conserved regions. This is further supported by the presence of many HAS within insertion regions, highlighting the model's ability to focus on regions beyond the conserved core for thermostability prediction.

To understand how TemBERTure$_{CLS}$ leverages structural information beyond sequence similarity, we mapped the attention scores directly onto protein structures (Fig. 6B and C, and Supplementary Fig. S4). Higher attention scores localized similarly across homologs, regardless of sequence entropy (Fig. 6D and Supplementary Fig. S5). Notably, higher attention scores often resided in helical regions and in the protein core, potentially revealing the prioritization of structurally important elements for predicting thermostability.

4 Discussion

Protein thermostability is crucial for various applications in biotechnology and biology. Traditional experimental methods for assessing it are laborious, expensive, and prone to errors. Here, we developed a new set of tools which allowed us to explore the potential of deep learning models to predict protein thermostability. Our study highlights the crucial role that data diversity plays in training robust models. We observed significant performance improvement with datasets encompassing a wider range of sequences from various organisms. Conversely, insufficient diversity, as seen in the BacDive derived dataset, led to models that struggled with challenging test sets. This emphasizes the need for a holistic approach to data curation, in order to ensure balanced representation of diverse species in the training data.

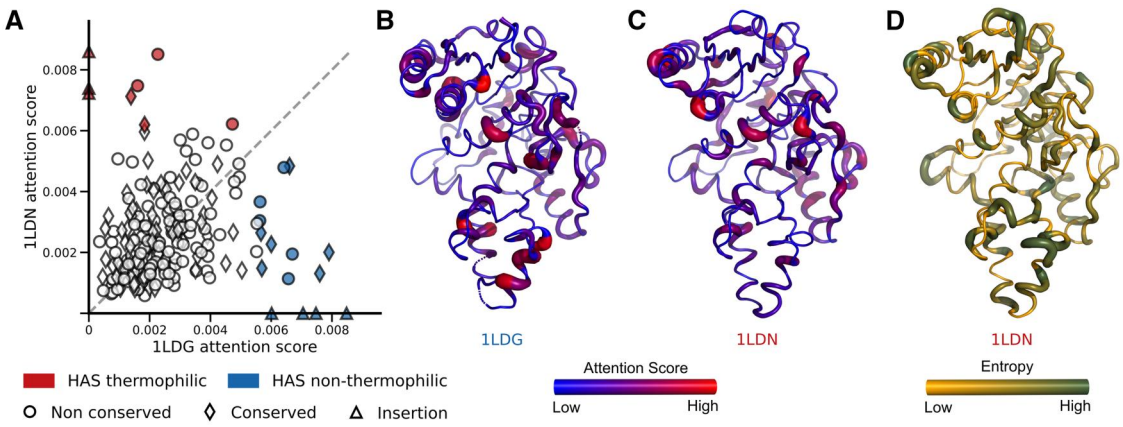

Figure 6. Representative structural analysis of attention score. (A) Scatter plot comparing the attention scores assigned by the TemBERTure$_{CLS}$ model to individual amino acids in two homologous protein structures (PDB ID: 1LDN [thermophilic] and 1LDG [non-thermophilic]) with 46% sequence identity. Each marker represents an amino acid, categorized by its conservation level: circles for non-conserved, diamonds for conserved, and triangles for insertions. HAS amino acids in the thermophilic structure are highlighted in red, while those in the non-thermophilic counterpart are highlighted in blue. (B and C) Cartoon representation of both protein structures. The width and color indicate the attention score values, with regions with higher attention scores appearing thicker and redder. (D) Cartoon representation of 1LDN colored based on the entropy at each amino acid position. Higher entropy (green, thicker regions) indicates greater sequence variability.

Although the Meltome Atlas presents an impressive number of melting temperatures, it suffers from certain biases, in particular, the data primarily represents non-thermophilic organisms with a temperature gap between 60 and 70°C. And yet, TemBERTure$_{Tm}$'s predictions, while not accurate for absolute melting temperatures, still captured the overall distribution of melting temperatures observed across different species in the dataset. This suggests the model might have prioritized recognizing the species origin of the sequence rather than intrinsic thermostability features. This agrees with previous findings showing that sequence embeddings from language models can already capture these broad differences between thermophilic and non-thermophilic organisms (Pudžiuvelytė *et al.* 2024). Additionally, the presence of thermostable proteins within non-thermophilic proteomes further underscores the limitations of using growth temperature alone as a thermostability proxy.

Various statistical approaches have attempted to identify important changes in amino acid composition linked to thermostability (Schäfer *et al.* 1996, Kumar *et al.* 2000, Vieille and Zeikus 2001, Sadeghi *et al.* 2006, Folch *et al.* 2008, Folch *et al.* 2010, Ahmed *et al.* 2022b). However, such analyses heavily depend on dataset curation, leading to contradictory results. Furthermore, while certain biophysical properties of residues may elucidate their prevalence in thermostable proteins, thermophilicity is a multifaceted attribute influenced by the positioning and microenvironment of amino acids within the protein. This study presents the concept of leveraging attention scores to gain more nuanced insights into protein thermostability. Even though we observed some global trends consistent with previous analyses (e.g. enrichment of specific amino acids), TemBERTure$_{CLS}$ also highlighted the value of analyzing these interactions within the context of the 3D protein structure. However, our findings suggest that the present attention scores still need to be refined, since they capture both thermostability-related features and organism-specific characteristics. Further research is needed to refine them for a more precise understanding of protein thermostability.

In conclusion, this study shed light on the limitations of current approaches for predicting protein thermostability and introduced new avenues for exploration. It highlighted the importance of using diverse training data, thus extending the analysis beyond single-species, and exploiting important features of the models, such as attention scores. Although our study demonstrates the importance of careful data splitting strategies, the precise impact of different sequence identity thresholds warrants further investigation. These findings can be expected to lay the groundwork for future research to develop even more robust and informative methods for predicting protein thermostability.

Author contributions

Chiara Rodella study design, model design and training, article drafting, critical revision and final approval of the manuscript. Symela Lazaridi study design, data acquisition, data analysis, article drafting, critical revision and final approval of the manuscript. Thomas Lemmin conception of the idea, study design, article drafting, critical revision, and final approval of the manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the Swiss National Science Foundation (SNSF: PCEFP3_194606).

References

Adams MWW, Kelly RM. Enzymes from microorganisms in extreme environments. *Chem Eng News Archive* 1995;**73**:32–42. https://doi.org/10.1021/cen-v073n051.p032

Ahern TJ, Klibanov AM. The mechanism of irreversible enzyme inactivation at 100°C. *Science* 1985;**228**:1280–4. https://doi.org/10.1126/science.4001942

Ahmed Z, Zulfiqar H, Khan AA *et al.* iThermo: a sequence-based model for identifying thermophilic proteins using a multi-feature fusion strategy. *Front Microbiol* 2022a;**13**:790063. https://doi.org/10.3389/fmicb.2022.790063

Ahmed Z, Zulfiqar H, Tang L *et al.* A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *Int J Mol Sci* 2022b;**23**:10116. https://doi.org/10.3390/ijms231710116

Bashirova A, Pramanik S, Volkov P *et al.* Disulfide bond engineering of an endoglucanase from penicillium verruculosum to improve its thermostability. *Int J Mol Sci* 2019;**20**:1602. https://doi.org/10.3390/ijms20071602

Biewald L. Experiment tracking with weights and biases. https://www.wandb.com/. 2020.

Bleicher L, Prates ET, Gomes TCF *et al.* Molecular basis of the thermostability and thermophilicity of laminarinases: X-ray structure of the hyperthermostable laminarinase from *Rhodothermus marinus* and molecular dynamics simulations. *J Phys Chem B* 2011;**115**:7940–9. https://doi.org/10.1021/jp200330z

Bommarius AS, Broering JM, Chaparro-Riggers JF *et al.* High-throughput screening for enhanced protein stability. *Curr Opin Biotechnol* 2006;**17**:606–10. https://doi.org/10.1016/j.copbio.2006.10.001

Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry* 2002;**41**:8152–61. https://doi.org/10.1021/bi025523t

Charoenkwan P, Chotpatiwetchkul W, Lee VS *et al.* A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Sci Rep* 2021;**11**:23782. https://doi.org/10.1038/s41598-021-03293-w

Charoenkwan P, Schaduangrat N, Moni MA *et al.* SAPPHIRE: a stacking-based ensemble learning framework for accurate prediction of thermophilic proteins. *Comput Biol Med* 2022;**146**:105704. https://doi.org/10.1016/j.compbiomed.2022.105704

Devlin J, Chang M-W, Lee K *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–86, Minneapolis, Minnesota. Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/N19-1423.

Ding Y, Cai Y, Zhang G *et al.* The influence of dipeptide composition on protein thermostability. *FEBS Lett* 2004;**569**:284–8. https://doi.org/10.1016/j.febslet.2004.06.009

Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381

Folch B, Dehouck Y, Rooman M. Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophys J* 2010;**98**:667–77. https://doi.org/10.1016/j.bpj.2009.10.050

Folch B, Rooman M, Dehouck Y. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. *J Chem Inf Model* 2008;**48**:119–27. https://doi.org/10.1021/ci700237g

Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J Mol Biol* 2001;**309**:835–43. https://doi.org/10.1006/jmbi.2001.4718

Gromiha MM, Xavier Suresh M. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins: Struct, Funct, Bioinform* 2007;**70**:1274–9. https://doi.org/10.1002/prot.21616

Haney P, Konisky J, Koretke KK *et al.* Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus*. *Proteins* 1997;**28**:117–30. https://doi.org/10.1002/(SICI)1097-0134(199705)28:1<117::AID-PROT12>3.0.CO;2-M

Haselbeck F, John M, Zhang Y *et al.* Superior protein thermophilicity prediction with protein language model embeddings. *NAR Genom Bioinform* 2023;**5**:lqad087. https://doi.org/10.1093/nargab/lqad087

Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* 2016;**32**:1323–30. https://doi.org/10.1093/bioinformatics/btw006

Himmel ME, Ding S-Y, Johnson DK *et al.* Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 2007;**315**:804–7. https://doi.org/10.1126/science.1137016

Houlsby N, Giurgiu A, Jastrzebski S *et al.* Parameter-efficient transfer learning for NLP. In: *International Conference on Machine Learning*, pp. 2790–99. PMLR, 2019.

Jarzab A, Kurzawa N, Hopf T *et al.* Meltome atlas—thermal proteome stability across the tree of life. *Nat Methods* 2020;**17**:495–503. https://doi.org/10.1038/s41592-020-0801-4

Jung F, Frey K, Zimmer D *et al.* DeepSTABp: a deep learning approach for the prediction of thermal protein stability. *Int J Mol Sci* 2023;**24**:7444. https://doi.org/10.3390/ijms24087444

Kuddus, M. (ed.), *Enzymes in Food Technology: Improvements and Innovations*. Singapore: Springer Singapore, 2018. https://doi.org/10.1007/978-981-13-1933-4

Kumar S, Tsai C-J, Nussinov R. Factors enhancing protein thermostability. *Protein Eng* 2000;**13**:179–91. https://doi.org/10.1093/protein/13.3.179

Leuenberger P, Ganscha S, Kahraman A *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 2017;**355**:eaai7825. https://doi.org/10.1126/science.aai7825

Liang H-K, Huang C-M, Ko M-T *et al.* Amino acid coupling patterns in thermophilic proteins. *Proteins: Struct, Funct, Bioinform* 2005;**59**:58–63. https://doi.org/10.1002/prot.20386

Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. *J Microbiol Methods* 2011;**84**:67–70. https://doi.org/10.1016/j.mimet.2010.10.013

Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *International Conference on Learning Representations*, 2017.

Matsuura Y, Takehira M, Joti Y *et al.* Thermodynamics of protein denaturation at temperatures over 100°C: cutA1 mutant proteins substituted with hydrophobic and charged residues. *Sci Rep* 2015;**5**:15545. https://doi.org/10.1038/srep15545

Modarres HP, Mofrad MR, Sanati-Nezhad A *et al.* ProtDataTherm: a database for thermostability analysis and engineering of proteins. *PLoS One* 2018;**13**:e0191222. https://doi.org/10.1371/journal.pone.0191222

Nakariyakul S, Liu Z-P, Chen L. Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. *Amino Acids* 2012;**42**:1947–53. https://doi.org/10.1007/s00726-011-0923-1

Nikam R, Kulandaisamy A, Harini K *et al.* ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res* 2021;**49**:D420–4. https://doi.org/10.1093/nar/gkaa1035

Pace NJ, Weerapana E. Zinc-binding cysteines: diverse functions and structural motifs. *Biomolecules* 2014;**4**:419–34. https://doi.org/10.3390/biom4020419

Pei H, Li J, Ma S *et al.* Identification of thermophilic proteins based on sequence-based bidirectional representations from transformer-embedding features. *Appl Sci* 2023;**13**:2858. https://doi.org/10.3390/app13052858

Pfeiffer J, Kamath A, Rücklé A *et al.* AdapterFusion: non-destructive task composition for transfer learning. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online. Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.eacl-main.39.

Poth C, Sterz H, Paul I *et al.* Adapters: a unified library for parameter-efficient and modular transfer learning. In: *Conference on Empirical Methods in Natural Language Processing*, 2023.

Pudžiuvelytė I, Olechnovič K, Godliauskaite E *et al.* TemStaPro: protein thermostability prediction using sequence representations from

protein language models. *Bioinformatics* 2024;**40**:btae157. https://doi.org/10.1093/bioinformatics/btae157
Raffel C, Shazeer N, Roberts A *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;**21**:140:5485–551.
Rahimzadeh M, Khajeh K, Mirshahi M *et al.* Probing the role of asparagine mutation in thermostability of bacillus KR-8104 α-Amylase. *Int J Biol Macromol* 2012;**50**:1175–82. https://doi.org/10.1016/j.ijbiomac.2011.11.014
Reimer LC, Sardà Carbasse J, Koblitz J *et al.* BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res* 2022;**50**:D741–6. https://doi.org/10.1093/nar/gkab961
Sadeghi M, Naderi-Manesh H, Zarrabi M *et al.* Effective factors in thermostability of thermophilic proteins. *Biophys Chem* 2006;**119**:256–70. https://doi.org/10.1016/j.bpc.2005.09.018
Sayers EW, Bolton EE, Brister JR *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022;**50**:D20–6. https://doi.org/10.1093/nar/gkab1112
Schäfer T, Bönisch H, Kardinahl S *et al.* Three extremely thermostable proteins from sulfolobus and a reappraisal off He 'traffic rules'. *Biol Chem Hoppe-Seyler* 1996;**377**:505–12. https://doi.org/10.1515/bchm3.1996.377.7-8.505
Singh R, Kumar M, Mittal A *et al.* Microbial enzymes: industrial progress in 21st century. *3 Biotech* 2016;**6**:174. https://doi.org/10.1007/s13205-016-0485-8
Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* 2019;**16**:603–6. https://doi.org/10.1038/s41592-019-0437-4
Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;**9**:2542–8. https://doi.org/10.1038/s41467-018-04964-5
Stourac Jan J, Dubrava M, Musil J *et al.* FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res* 2021;**49**:D319–24. https://doi.org/10.1093/nar/gkaa981
Tang H, Cao R-Z, Wang W *et al.* A two-step discriminated method to identify thermophilic proteins. *Int J Biomath* 2017;**10**:1750050. https://doi.org/10.1142/S1793524517500504
The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052
Tomazic SJ, Klibanov AM. Why is one *Bacillus* alpha-amylase more resistant against irreversible thermoinactivation than another? *J Biol Chem* 1988;**263**:3092–6. https://doi.org/10.1016/S0021-9258(18)69039-8
Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems 6000–6010*. Red Hook, NY, USA: Curran Associates Inc., 2017.
Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 2001;**65**:1–43. https://doi.org/10.1128/MMBR.65.1.1-43.2001
Wolf T, Debut L, Sanh V *et al.* HuggingFace's transformers: state-of-the-art natural language processing. arXiv, 2019:1910.03771, preprint: not peer reviewed.
Wolf T, Debut L, Sanh V *et al.* Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6.
Wu L-C, Lee J-X, Huang H-D *et al.* An expert system to predict protein thermostability using decision tree. *Expert Syst Appl* 2009;**36**:9007–14. https://doi.org/10.1016/j.eswa.2008.12.020
Yang Y, Ding X, Zhu G *et al.* ProTstab—predictor for cellular protein stability. *BMC Genomics* 2019;**20**:804. https://doi.org/10.1186/s12864-019-6138-7
Yang Y, Zhao J, Zeng L *et al.* ProTstab2 for prediction of protein thermal stabilities. *Int J Mol Sci* 2022;**23**:10798. https://doi.org/10.3390/ijms231810798
Zhang G, Fang B. Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. *Protein Pept Lett* 2006;**13**:965–70. https://doi.org/10.2174/092986606778777560
Zhang G, Fang B. LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J Biotechnol* 2007;**127**:417–24. https://doi.org/10.1016/j.jbiotec.2006.07.020
Zhou X-X, Wang Y-B, Pan Y-J *et al.* Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids* 2008;**34**:25–33. https://doi.org/10.1007/s00726-007-0589-x

Bioinformatics Advances, 2024, 00, 1–10
https://doi.org/10.1093/bioadv/vbae103
Original Article