Learning temperature-aware representations from millions of annotated protein sequences

Mingchen Li[1,2,3,4]† **Liang Zhang**[4]† **Zilan Wang**[5] **Bozitao Zhong**[6]
Pan Tan[2,4] **Jiabei Cheng**[2,4] **Bingxin Zhou**[4] **Liang Hong**[2,3,4] **Huiqun Yu**[1]*
[1] East China University of Science and Technology, Shanghai, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] Shanghai-Chongqing Institute of Artificial Intelligence, Chongqing, China
[4] School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai China
[5] Nanyang Technological University, Singapore
[6] The Chinese University of Hong Kong, Hong Kong, China
Emails: limc.19980301@gmail.com, yhq@ecust.edu.cn

Abstract

Temperature plays a dominant environmental role in determining the efficiency of protein function. Accurately predicting the thermal stability of proteins is crucial for fundamental biology, drug discovery, and protein engineering. Here, we introduce **ThermoFormer**, a transformer-based protein language model that learns both temperature-aware representation and sequence patterns. Specifically, we first build a large-scale dataset comprising more than 96 million protein sequences annotated with their optimal growth temperature (OGT). ThermoFormer is pre-trained with a supervised OGT prediction task and an unsupervised masked language modeling (MLM) task on the dataset. We evaluated the performance of ThermoFormer on the pre-training and the performance of transferring ThermoFormer to other temperature prediction datasets, including two melting temperature (TM) datasets and an optimal catalytic temperature (OCT) dataset. The results show that ThermoFormer is able to achieve state-of-the-art performance in both OGT, TM, and OCT prediction tasks, outperforming previous unsupervised pre-trained models. In addition, we have also shown that ThermoFormer enables zero-shot temperature prediction, i.e., even without further fine-tuning, ThermoFormer can still achieve comparable performance. We believe that ThermoFormer can serve as a foundational model for encoding protein sequences with temperature-aware representations, providing better transfer ability for temperature-related downstream tasks. The datasets, model weights, and source codes are available at https://github.com/ginnm/ThermoFormer.

1 Introduction

Temperature is a fundamental environmental factor that affects protein function [1, 2]. Accurately predicting temperature from protein sequences is essential. There are three main types of temperature related to protein functionality: optimal growth temperature (OGT) [3], melting temperature (TM) [4], and optimal catalytic temperature (OCT) [5]. Their detailed definitions are in Table 1. Compared to OGT data, TM and OCT data are more difficult to obtain [6]. Experiments to determine the TM or OCT of a protein are relatively complex, and currently, only tens of thousands of data points have been accumulated [7, 8, 9]. However, obtaining the OGT of proteins is relatively simple, as measuring

*Corrsespongding Author; †Equal contribution.

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

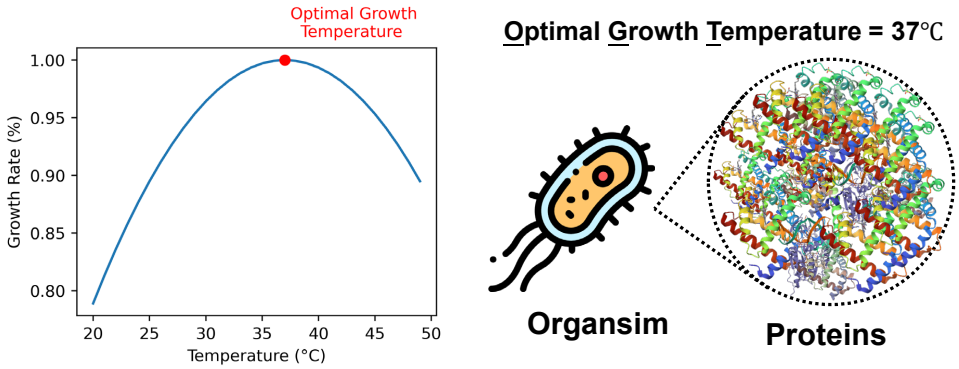

Figure 1: Optimal growth temperature refers to the temperature at which an organism exhibits its highest growth rate. The OGT of proteins is that of the host organism.

Name	Definition	Abbreviation
Optimal Growth Temperature	The temperature at which a protein's host organism achieves the highest growth rate.	OGT
Melting Temperature	The temperature at which a protein unfolds and loses its functionality.	TM
Optimal Catalytic Temperature	The temperature at which a protein enzyme exhibits its highest catalytic activity.	OCT

Table 1: Different types of protein temperature involved in this work.

the OGT of a microorganism provides the OGT of all its proteins (See in Figure 1). Moreover, it has been observed that the OGT of a protein is positively correlated with its TM and OCT [10, 11]. This is because proteins from an organism are expected to be functional at its optimal growth temperature (OGT). Thus, an intuitive idea is to pre-train a protein representation model on OGT data and then transfer the model to the prediction of TM and OCT.

To this end, we first collect a large-scale protein dataset containing more than 96 million protein sequences. All these proteins are annotated with OGT labels and have a unique ID in the UniProtKB database. Then, we propose ThermoFormer, a Transformer-based model including 690 million parameters, and pre-train it on the OGT-labelled dataset. The pre-training process consists of two tasks: a supervised task involving learning to predict the OGT of a protein and an unsupervised task focusing on learning to understand protein sequences through masked language modeling (MLM). Through this hybrid approach of supervised and unsupervised pre-training, ThermoFormer learns contextual representations of amino acids and temperature-aware protein sequence representations. The later experiments show that the unsupervised task can improve the performance of the supervised task. Another interesting finding is that ThermoFormer exhibits zero-shot temperature prediction capabilities, meaning it can predict TM and OCT directly without further fine-tuning.

In summary, the main contributions of this work are:

1. We collect a large-scale protein dataset containing over 96 million protein sequences with annotated OGT labels, which is approximately 32 times larger than the previous largest protein dataset with OGT labels.
2. We present ThermoFormer, a Transformer-based model that is pre-trained on the large-scale dataset with both supervised OGT prediction and unsupervised MLM tasks, enabling it to learn temperature-aware representations of proteins.
3. We evaluate ThermoFormer on temperature-related downstream tasks, including two TM and an OCT datasets, demonstrating its state-of-the-art performance and zero-shot prediction capability.

We suggest that ThermoFormer can serve as a foundational model in the field of protein temperature prediction since it has learned temperature-sensitive representations.

2 Related Work

2.1 Protein Temperature Prediction

Protein temperature prediction is a classic problem in machine learning. Previous methods are based on statistical inference [12, 13], random forest [14], LightGBM [15], decision tree [16], etc., which learn artificial protein features related to melting or optimal catalytic temperatures. Many end-to-end deep learning models, such as CNN [17, 6] and RNN [18], have also been proposed to predict protein temperatures from one-hot encodings of protein sequences directly. Recently, with the success of pre-training on the field of natural language processing (NLP), pre-trained protein language models (PLMs) have emerged. They are often Transformer-based [19] and learn on millions of protein sequences through BERT-like masked language modeling (MLM) [20, 21, 22, 23], GPT-like causal language modeling (CLM) [24, 25, 26], or T5-like encoder-decoder model [27, 28, 29]. CLM is mainly used for protein generation, while MLM excels in protein representation and downstream task fine-tuning [30] and is more suitable for temperature prediction. And there are also pre-trained PLMs incorporating protein structure information [31, 32, 33, 34]. We can utilize a PLM to encode protein sequences or structures into a hidden vector and then utilize an additional regression model to learn the mapping function between the hidden vector and protein temperatures [35, 36, 37, 38]. Since these models are pre-trained on massive protein sequences or structures, they typically achieve higher accuracy.

2.2 Optimal Growth Temperature Prediction for Pre-training

The optimal growth temperature refers to the most favorable environmental temperature that supports the growth and reproduction of a specific organism. Leveraging optimal growth temperature prediction for pre-training is motivated by the following: Numerous protein sequences are annotated with optimal growth temperatures, while protein sequences labeled TM and OCT are scarce. Second, there is a positive correlation between OGT and other temperatures, such as TM and OCT [38, 39, 16]. This is because proteins can function with the highest efficiency and stability near their optimal growth temperature. Prior work, DeepET [6], a CNN-based model, has also shown that representations from learning from OGT prediction tasks can be effectively transferred to other temperature-related tasks. However, it only used 3 million protein sequences for pre-training, which is only 3% of ours.

3 Method

Figure 2 shows the overall workflow of this work. We collect a protein dataset annotated with optimal growth temperature (Figure 2A) and pre-train ThermoFormer on the dataset with a supervised OGT prediction task and unsupervised MLM task (Figure 2B). After pre-training, ThermoFormer can be utilized to learn the temperature prediction downstream task (Figure 2C).

3.1 Pre-training Dataset Collection

We first collected a dataset containing 21,498 micro-organisms and their corresponding optimal growth temperatures (OGT) from literature [40]. Then we utilized the taxon IDs of these organisms to search for the proteins contained within them in the UniProt [41] protein database and annotated these proteins with the OGT label of the organisms they belong to, resulting in 96.4 million annotated protein sequences from 14,612 organisms. The protein sequences containing non-standard amino acid residues and longer than 2,048 are removed to ensure training efficiency. We split the pre-training dataset into a validation set, a mix-species test set, and a cross-species test set. The validation set and the cross-test set each contain 100 types of micro-organisms that are different from those in the training set. The mix-species test set contains 500,000 sequences randomly split from the training set. The statistical information of our dataset and splits is shown in Table 2.

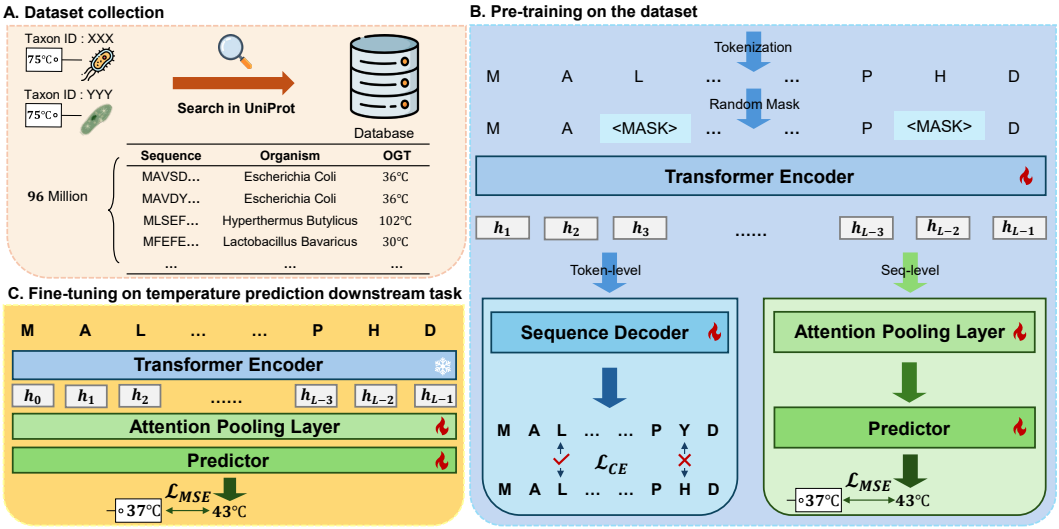

Figure 2: Overview of this work. (A) We first collect a large-scale dataset comprising 96 million protein sequences annotated with optimal growth temperature. (B) Then we propose ThermoFormer, which consists of a Transformer encoder and a sequence decoder for unsupervised MLM pre-training and a predictor for supervised OGT prediction. The OGT prediction task enables it to learn temperature-sensitive representation. (C) We can utilize ThermoFormer to perform fine-tuning for learning TM or OCT.

Splitting	# Organsims	# Sequences	OGT				Seqeunce Lengths			
			Min.(°C)	Max.(°C)	Avg.(°C)	Std.(°C)	Min.	Max.	Avg.	Std.
Training	14,412	95,038,959	3	103	31.25	6.14	32	2048	357	245
Validation	100	502,979	15	70	30.27	5.6	32	2048	354	235
Cross-Test	100	475,199	10	57	29.31	5.2	32	2048	360	257
Mix-Test	9,363	500,000	3	103	31.24	6.13	32	2048	257	245
Total	14,612	96,017,137	3	103	31.23	6.15	32	2048	357	245

Table 2: Statistics of the Pre-training Dataset.

3.2 Model Architecture and Pre-training

ThermoFormer is a pre-trained Transformer model. It contains four components: a transformer-based encoder for extracting residue-level representations, an attention-based pooling layer for aggregating the residue-level representation into sequence-level representation, a sequence decoder for MLM pre-training, and a predictor for OGT prediction. These components are detailed below:

Transformer-based encoder. The Transformer-based encoder encodes the protein sequences into a sequence of hidden states. Let $\boldsymbol{s} = (\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_L) \in \mathcal{R}^{L \times V}$ denote a protein sequence, where $\boldsymbol{r}_i \in \mathcal{R}^V$ is the one-hot encoding of the i_{th} residue, L is the length of the protein and V is the residue vocab size. The transformer encoder learns to map $\boldsymbol{s}$ into a hidden state:

$$(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_L) = F_\theta(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_L) \tag{1}$$

where F_θ is the Transformer encoder and θ denotes its parameters, $\boldsymbol{h}_i \in \mathcal{R}^d$ is the contextual embedding of $\boldsymbol{r}_i$ and d is the hidden dimension.

Sequence Decoder. The sequence decoder learns to recover the masked token from the hidden states. It contains two position-wise dense layers with GELU activation unit and a layer normalization layer [42]:

$$\forall i \in 1, ..., L, \boldsymbol{y}_i = \boldsymbol{W}_2^T F_{LN}(\sigma(\boldsymbol{W}_1^T \boldsymbol{h}_i)) + \boldsymbol{b} \tag{2}$$

where $\boldsymbol{W}_1 \in \mathcal{R}^{d\times d}$, $\boldsymbol{W}_2 \in \mathcal{R}^{d\times V}$ and $\boldsymbol{b} \in \mathcal{R}^V$ are learnable parameters, F_{LN} is the layer normalization function and σ is the GELU [43] activation function. $\boldsymbol{y}_i \in \mathcal{R}^V$ is the probability distribution of predicted i_{th} residue. And we utilize cross-entropy as the loss function:

$$L_{CE} = -\mathbb{E}\left[\log \boldsymbol{y}_i[\boldsymbol{y}_i^*]\right] \tag{3}$$

where $\boldsymbol{y}_i^*$ represents the true residue for the i-th token in the sequence, and $\boldsymbol{y}_i[\boldsymbol{y}_i^*]$ denotes the predicted probability for the correct residue.

Attention-based Pooling Layer. The attention-based pooling layer learns to aggregate the hidden states $(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_L)$ into a global hidden state for further adaption on sequence-level task. The weights of hidden states $(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_L)$ are computed by a projection-soft-max layer that produces a weighted vector $\boldsymbol{c}$:

$$\begin{aligned}(\hat{\boldsymbol{h}}_1, \hat{\boldsymbol{h}}_2, ..., \hat{\boldsymbol{h}}_N) &= F_{LN}(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N) \\ s_i &= \frac{e^{\boldsymbol{W}_a \hat{\boldsymbol{h}}_i + \boldsymbol{b}_a}}{\sum_{n=1}^{L} e^{\boldsymbol{W}_a \hat{\boldsymbol{h}}_n + \boldsymbol{b}_a}}, \forall i \in 1, ..., L \\ \boldsymbol{c} &= \sum_{n=1}^{L} s_i \cdot \boldsymbol{h}_i\end{aligned} \tag{4}$$

where s_i is the attention weight of the i_th residue and $\boldsymbol{W}_a$ and $\boldsymbol{b}_a$ are the learnable parameters of the attention pooling layer. Then, a multi-layer perceptron with two dense layers and GELU activation is employed to transform the weighted vector $\boldsymbol{c}$. The first dense layer maps $\boldsymbol{c}$ to the same dimension as the Feed-Forward Network (FFN) layer of the Transformer encoder, which in our implementation is four times the size of the hidden layer. The second dense layer maps the output of the first layer back to the original dimension. Between the first and second dense layers, there is a GELU activation function. Additionally, there is a residual connection between the output of the second dense layer and the output of the attention layer:

$$\boldsymbol{r} = \boldsymbol{c} + \boldsymbol{W}_4(\sigma(\boldsymbol{W}_3\boldsymbol{c} + \boldsymbol{b}_3)) + \boldsymbol{b}_4, \tag{5}$$

, where $\boldsymbol{W}_3, \boldsymbol{W}_4, \boldsymbol{b}_3$, and $\boldsymbol{b}_4$ are learnable parameters layers, σ is the GELU activation function. The output hidden state $\boldsymbol{r}$ is the representation of the whole sequence.

Predictor. The predictor learns to predict a temperature value $T \in \mathcal{R}$ from the sequence representation $\boldsymbol{r}$. It has two dense layers and a $Tanh$ activation function:

$$\hat{T} = \boldsymbol{c} + \boldsymbol{W}_6(\sigma_t(\boldsymbol{W}_5\boldsymbol{r} + \boldsymbol{b}_5)) + \boldsymbol{b}_6 \tag{6}$$

, where $\boldsymbol{W}_5, \boldsymbol{W}_6, \boldsymbol{b}_5$, and $\boldsymbol{b}_6$ are learnable parameters, σ_t is the Thanh activation function, and $\hat{T}$ is the predicted temperature. We utilize the mean square error (MSE) criterion as the loss function:

$$L_{MSE} = \mathbb{E}\left[(\hat{T} - T)^2\right] \tag{7}$$

where $\mathbb{E}[\cdot]$ denotes the expectation, $\hat{T}$ is the predicted temperature, and T is the ground truth temperature.

Joint Loss Function. The Pre-training loss function is the sum of L_{CE} and L_{MSE}. Since we have observed that L_{MSE} has a significantly different magnitude compared to L_{CE}, with values ranging from 0-1000 initially and stabilizing at 0-100 later. We multiplied L_{MSE} by 0.01 to maintain numerical stability. The final joint loss function is:

$$L = \beta L_{MSE} + L_{CE} \tag{8}$$

3.3 Supervised Fine-tuning Paradigm

As shown in Figure 2C, we can further fine-tune ThermoFormer on other temperature prediction tasks. For the temperature prediction task, we removed the sequence decoder during the fine-tuning stage, while the rest of the parts remained. The parameters were inherited from those of the pre-trained model. The Transformer encoder was kept frozen, and we only fine-tuned the parameters of the attention-based pooling layer and the predictor to reduce the training cost. The subsequent experiments demonstrate that this transfer learning approach effectively enhances the convergence speed and accuracy of the Transformer model during training.

4 Experiments

4.1 Model Pre-training

We pre-trained ThermoFormer on the OGT training dataset, using the validation set during the training process to monitor overfitting. After training, we selected the model that performed best on the validation set and tested it on two test sets. We utilized PyTorch and Hugging-Face Transformers API[2] to implement ThermoFormer. The transformer encoder comprises 33 layers and 20 attention heads, with 650 million parameters and an embedding size of 1280. The encoder is compatible with ESM-2 [23], so we load the ESM-2 checkpoint [3] as the initialization of our Transformer encoder, except we replace the naive attention layer with flash attention layer [44]. The learning rate was set to 0.0001. We train ThermoFormer on a DGX server equipped with eight NVIDIA A800 GPUs. The micro-batch size per GPU is 4096 tokens, and the gradient accumulation step is 32. The max training step is set to 250k, and a cosine schedule with 3000 linear warm-up steps was used.

4.2 Fine-tuning on Temperature-related Tasks

Dataset	Source	Training	Test	Total
TM-Cell	[8]	2255	251	2506
TM-Atlas	[7]	33719	3714	37433
OCT	[9]	1756	190	1902

Table 3: Statistics of the temperature-related down-stream datasets.

We evaluated ThermoFormer on three temperature-related datasets.

- **TM-Cell**. TM-Cell contains 2506 proteins with melting temperatures from three species: E. coli, S. cerevisiae, and T.thermophilus. The data is measured by Leuenberger et al. [8] and split by Li et al. [6] The dataset includes 2255 training samples and 251 test sequences.
- **OCT**. OCT includes the optimal catalytic temperatures of 1902 enzymes from the BRENDA database [9]. The dataset was randomly split into training (1712 enzymes) and test (190 enzymes) sets based on a 90–10 ratio.
- **TM-Atlas**. TM-Atlas consists of 48,000 proteins from 13 species. The data is measured by Mega et al. [7] and split by Li et al. [6] The splitting statistic is shown in Table 3, including 2255 training samples and 251 test sequences.

The statistics of this dataset are shown in Table 3. For the task of temperature prediction, we utilize root-mean-square-error (RMSE), Coefficient of Determination(R^2), Pearson correlation coefficient (ρ_p), and Spearman rank correlation coefficient (ρ_s) as evaluation metrics to measure the differences between predicted values and gold truth. We use 5-fold cross-validation for the assessment, where in each iteration, 20% of the training set was selected as the validation set, which was not involved in training and was only used to choose the best training epoch. The remaining 80% of the sequences were used for training. The model parameters from the checkpoint of the epoch that performed best on the validation set were then used to test on the test set. The performance metrics reported are the averages of the 5-fold cross-validation, and the error is represented by the variance of the performance.

For the MLM task, we also record the perplexity of the model on the validation and test sets, defined as the power of the natural logarithm of the cross-entropy loss. The lower the perplexity, the better the model's ability to reconstruct the sequence.

Split	Model	OGT Prediction for Proteins				MLM
		RMSE(°C)↓	ρ_p ↑	R^2 ↑	ρ_s ↑	Perplexity
Validation	ThermoFormer	**2.88**	**0.87**	**0.73**	**0.61**	4.95
	ThermoFormer (-MLM)	2.98	0.86	0.72	0.60	-
Cross-Test	ThermoFormer	**3.10**	**0.80**	**0.64**	**0.76**	5.23
	ThermoFormer (-MLM)	3.18	0.79	0.63	0.74	-
Mix-Test	ThermoFormer	**3.10**	**0.86**	**0.75**	**0.82**	4.73
	ThermoFormer (-MLM)	3.20	0.85	0.73	0.81	-

Table 4: Performance of ThermoFormer on the pre-training validation set and test set. ThermoFormer(-MLM) is solely trained on the OGT prediction task.

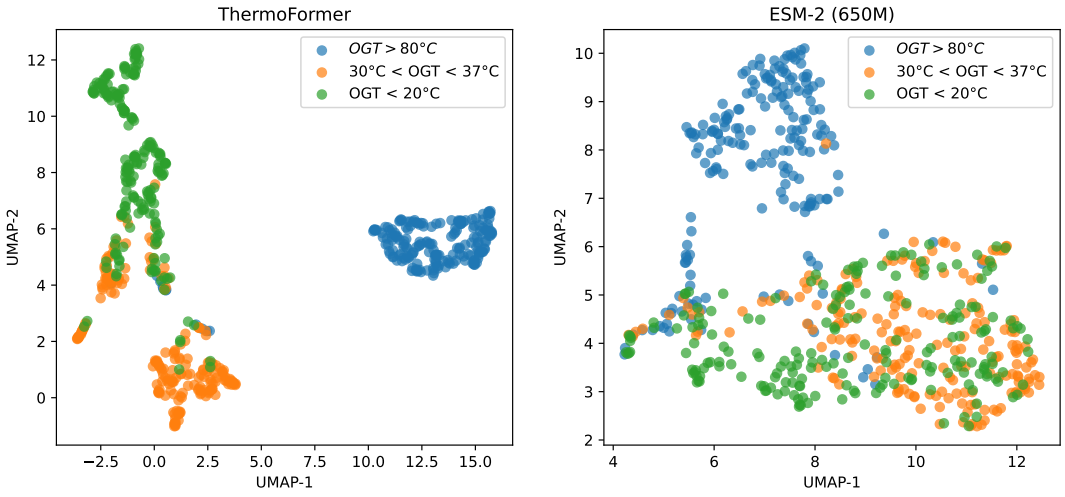

Figure 3: UMAP Projection of Protein Representations of ThermoFormer and ESM-2.

4.3 Results

4.3.1 Impact of OGT Pre-training on ThermoFormer Representations.

Table 4 presents the results of the performance of the pre-training where ThermoFormer is the base model whose pre-training process includes both OGT Prediction and MLM tasks, while ThermoFormer (-MLM) does not include the MLM task. The results show that ThermoFormer can predict the OGT of proteins accurately. For the validation set, the error (RMSE) is only 2.88°C, and the Pearson correlation between the predicted and actual OGT reaches 0.87. For the cross-species test set, the error (RMSE) is 3.10°C, and the Pearson correlation between the predicted and actual OGT reaches 0.80, indicating that ThermoFormer is capable of generalizing across different species. For the mix-species test set, the error (RMSE) is also 3.10°C, and the Pearson correlation between the predicted and actual OGT reaches 0.86, which is higher than it on the cross-species test set, demonstrating that ThermoFormer has better performance within the same species. Figure 3 also demonstrates the difference in learned representation between ThermoFormer and the unsupervised ESM-2. Compared to ESM-2, ThermoFormer effectively separates proteins across different temperature ranges.

Another notable point is that incorporating the MLM task can improve the accuracy of OGT prediction since ThermoFormer outperforms ThermoFormer (-MLM) across all metrics. This indicates that the

[2] `https://huggingface.co/docs/transformers/index`
[3] `https://huggingface.co/facebook/esm2_t33_650M_UR50D`

MLM task is beneficial to the OGT prediction task and enhances the generalization capability of the model.

4.3.2 ThermoFormer Fine-tuning Performance on Temperature-related Tasks

Dataset	Model	RMSE(°C)↓	ρ_p ↑	R^2 ↑	ρ_s ↑
TM-Cell	DeepET	11.13	0.79	0.35	0.72
	ProtT5	7.55 (±0.41)	0.86 (±0.003)	0.74 (±0.006)	0.73 (±0.010)
	Ankh	9.03 (±0.25)	0.79 (±0.005)	0.62 (±0.008)	0.72 (±0.007)
	ThermoFormer (-OGT)	7.51 (±0.48)	0.86 (±0.003)	0.74 (±0.026)	0.73 (±0.005)
	ThermoFormer	**7.04 (±0.15)**	**0.88 (±0.006)**	**0.77 (±0.011)**	**0.74 (±0.013)**
TM-Atlas	DeepET	6.30	0.76	0.58	0.55
	ProtT5	5.12 (±0.21)	0.81 (±0.017)	0.66 (±0.020)	0.62 (±0.007)
	Ankh	6.65 (±0.16)	0.67 (±0.012)	0.42 (±0.024)	0.44 (±0.011)
	ThermoFormer (-OGT)	5.59 (±0.16)	0.77 (±0.008)	0.59 (±0.016)	0.55 (±0.002)
	ThermoFormer	**4.80 (±0.17)**	**0.84 (±0.006)**	**0.70 (±0.012)**	**0.64 (±0.014)**
OCT	DeepET	12.21	0.76	0.57	0.62
	ProtT5	12.44 (±0.18)	0.76 (±0.006)	0.55 (±0.023)	0.72 (±0.014)
	Ankh	13.50 (±0.35)	0.69 (±0.002)	0.47 (±0.053)	0.63 (±0.002)
	ThermoFormer (-OGT)	11.89 (±0.17)	0.78 (±0.010)	0.59 (±0.010)	0.70 (±0.005)
	ThermoFormer	**11.23 (±0.22)**	**0.81 (±0.009)**	**0.63 (±0.014)**	**0.76 (±0.009)**

Table 5: Performance of ThermoFormer and baseline models on temperature-related fine-tuning tasks.

Table 5 shows the supervised fine-tuning results of ThermoFormer and other baseline models on temperature-related downstream tasks. ThermoFormer represents the complete model, while ThermoFormer(-OGT) is the model without the OGT prediction pre-training task, containing only the unsupervised MLM prediction task. The metric score in the table is the average from five-fold cross-validation, with the standard deviation in the bracket. It can be seen that ThermoFormer outperforms the ThermoFormer(-OGT) across all the datasets and metrics. This suggests that the representations learned by ThermoFormer are better suited for transfer to temperature-related downstream tasks. Therefore, we can conclude that supervised pre-training on large-scale OGT data enables the model to learn temperature-related representations, leading to improved learning capability and performance on temperature-related downstream tasks.

4.3.3 Comparison of ThermoFormer with Other Temperature Prediction Models.

We compare the performance of ThermoFormer to other models, DeepET [6], ProtT5 [28], and Ankh [29] on the three downstream temperature prediction tasks. DeepET is a CNN-based model trained on OGT prediction tasks and then transferred to TM and OCT prediction tasks. For DeepET[4], we used the TM prediction checkpoint provided by the authors for testing; therefore, there is no standard deviation. For the unsupervised models ProtT5[5] and Ankh[6], we used their encoder to encode the protein sequences into hidden states, followed by the same Attention Pooling module and Predictor module as ThermoFormer to ensure a fair comparison. The learning hyper-parameters are the same as those of ThermoFormer. The comparison results are shown in Table 5. The results show that ThermoFormer performs best across all three temperature prediction datasets and evaluation metrics.

[4] `https://doi.org/10.5281/zenodo.6351465`

[5] `https://huggingface.co/ElnaggarLab/ankh-base`

[6] `https://huggingface.co/Rostlab/prot_t5_xl_uniref50`

4.3.4 Zero-shot Temperature Prediction Performance of ThermoFormer

It has been observed that the OGT of protein exhibits a positive correlation with its thermal stability (TM and OCT). This correlation suggests that the OGT predicted by ThermoFormer may be directly regarded as TM or OCT, eliminating the need for separate fine-tuning of these temperature datasets. This approach can be referred to as "zero-shot temperature prediction," as it does not require further fine-tuning on specific TM or OCT datasets, leveraging the generalizability of the OGT prediction model. To validate this, we conduct experiments on the TM-Cell and OCT datasets to obtain the zero-shot temperature prediction performance of ThermoFormer. We also test the zero-shot temperature prediction performance of DeepET, as it can also predict the OGT of protein. The results are shown in 6. To validate this, we conducted experiments using the TM-Cell and OCT datasets to evaluate

Dataset	Model	RMSE(°C)↓	ρ_p ↑	ρ_s ↑
TM-Cell	ThermoFormer	**20.54**	**0.87**	**0.76**
	DeepET	23.81	0.75	0.69
OCT	ThermoFormer	**19.97**	**0.73**	**0.51**
	DeepET	21.26	0.66	0.40

Table 6: Zero-shot temperature prediction performance of ThermoFormer and DeepET.

the zero-shot temperature prediction performance of ThermoFormer. Additionally, we assessed the zero-shot temperature prediction capabilities of DeepET, as it can also predict the optimal growth temperature (OGT) of proteins. The results, presented in Table 6, demonstrate that ThermoFormer achieves acceptable accuracy in temperature prediction, even without fine-tuning. While DeepET also performs zero-shot temperature predictions, its accuracy is lower than that of ThermoFormer.

4.3.5 OGT Prediction For Organisms

Given that ThermoFormer is able to predict the OGT of proteins, it is natural to ask whether it can also be used to predict the OGT of entire organisms. The answer is affirmative. Specifically, for a given

Dataset	# Organisms	RMSE(°C)↓	ρ_p ↑	R^2 ↑	ρ_s ↑
Validation	100	2.51	0.96	0.92	0.87
Cross-Test	100	3.10	0.89	0.79	0.85
Mix-Test	9363	3.67	0.92	0.84	0.82

Table 7: Performance of ThermoFormer in predicting the optimal growth temperature for organisms.

organism, ThermoFormer predicts the OGT of all its proteins, and the average of these predictions is taken as the organism's OGT. We evaluated the performance of this approach on the pre-training datasets, with the results presented in Table 7. The results show that ThermoFormer demonstrates the capability to predict the OGT of organisms effectively.

5 Discussion and Conclusion

Temperatures are one of the key factors determining their function. The work aims to develop an end-to-end protein model specifically designed to capture the temperature-aware representations of proteins. The most commonly used temperatures for proteins are melting temperature and optimal catalytic temperature. However, due to the complexity of wet lab experiments, data for TM and OCT are scarce. In contrast, optimal growth temperature data is relatively more accessible, and previous studies have shown a positive correlation between OGT and both TM and OCT [10, 11]. Therefore, we propose to first pre-train the protein representation model on OGT data to learn temperature-related representations of proteins and later transfer these representations to specific prediction tasks for TM or OCT. To achieve this, we present a pre-training dataset containing over 96 million proteins labeled with OGT, 32 times larger than the largest OGT-annotated dataset one [6]. Next, we introduce

ThermoFormer, a Transformer-based model trained on this dataset using unsupervised masked language modeling and supervised OGT prediction tasks. The hybrid pre-training approach enables the model to learn protein temperature-aware representations. Compared to other protein language models, ThermoFormer significantly improves downstream temperature-related tasks, proving the effectiveness of supervised learning on OGT data. Additionally, ThermoFormer enables zero-shot temperature prediction capabilities, meaning the OGT predicted by it can be directly regarded as TM or OCT. Experimental results show acceptable accuracy in these predictions. Moreover, ThermoFormer can predict the optimal growth temperature of organisms. Precisely, by predicting the OGT of all proteins within an organism and then averaging them, we can estimate the optimal growth temperature of the organism. In conclusion, ThermoFormer is a language model capable of efficiently learning protein temperature-aware representations, and it can serve as a foundation model for efficiently transferring to various downstream tasks related to protein temperature prediction.

Acknowledgement

This work was supported by the grants from the National Science Foundation of China (Grant Number 12104295), the Computational Biology Key Program of Shanghai Science and Technology Commission (23JS1400600), Shanghai Jiao Tong University Scientific and Technological Innovation Funds (21X010200843), and Science and Technology Innovation Key R&D Program of Chongqing (CSTB2022TIAD-STX0017), the Student Innovation Center at Shanghai Jiao Tong University, and Shanghai Artificial Intelligence Laboratory.

References

[1] George N. Somero. Proteins and temperature. *Annual Review of Physiology*, 57(Volume 57,):43–68, 1995.

[2] Peter A Fields, Yunwei Dong, Xianliang Meng, and George N Somero. Adaptations of protein structure and function to temperature: there is more than one way to 'skin a cat'. *The Journal of Experimental Biology*, 218(12):1801–1811, 2015.

[3] David B Sauer and Da-Neng Wang. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics*, 35(18):3224–3231, 01 2019.

[4] Wayne J Becktel and John A Schellman. Protein stability curves. *Biopolymers: Original Research on Biomolecules*, 26(11):1859–1877, 1987.

[5] Roy M Daniel and Michael J Danson. Temperature and the catalytic activity of enzymes: A fresh understanding. *FEBS letters*, 587(17):2738–2743, 2013.

[6] Gang Li, Filip Buric, Jan Zrimec, Sandra Viknander, Jens Nielsen, Aleksej Zelezniak, and Martin KM Engqvist. Learning deep representations of enzyme thermal adaptation. *Protein Science*, 31(12):e4480, 2022.

[7] Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature methods*, 17(5):495–503, 2020.

[8] Pascal Leuenberger, Stefan Ganscha, Abdullah Kahraman, Valentina Cappelletti, Paul J Boersema, Christian von Mering, Manfred Claassen, and Paola Picotti. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science*, 355(6327):eaai7825, 2017.

[9] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl_1):D431–D433, 2004.

[10] Yves Dehouck, Benjamin Folch, and Marianne Rooman. Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein engineering, design & selection*, 21(4):275–278, 2008.

[11] Lucas Sawle and Kingshuk Ghosh. How do thermophilic proteins and proteomes withstand high temperature? *Biophysical journal*, 101(1):217–227, 2011.

[12] Tienhsiung Ku, Peiyu Lu, Chenhsiung Chan, Tsusheng Wang, Szuming Lai, Pingchiang Lyu, and Naiwan Hsiao. Predicting melting temperature directly from protein sequences. *Computational biology and chemistry*, 33(6):445–450, 2009.

[13] Fabrizio Pucci, Jean Marc Kwasigroch, and Marianne Rooman. Scoop: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, 33(21):3415–3422, 2017.

[14] Yang Yang, Xuesong Ding, Guanchen Zhu, Abhishek Niroula, Qiang Lv, and Mauno Vihinen. Protstab–predictor for cellular protein stability. *BMC genomics*, 20:1–9, 2019.

[15] Yang Yang, Jianjun Zhao, Lianjie Zeng, and Mauno Vihinen. Protstab2 for prediction of protein thermal stabilities. *International Journal of Molecular Sciences*, 23(18):10798, 2022.

[16] Japheth E Gado, Gregg T Beckham, and Christina M Payne. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *Journal of Chemical Information and Modeling*, 60(8):4098–4107, 2020.

[17] Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

[18] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[19] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[21] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[22] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[25] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[26] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[28] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[29] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

[30] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *bioRxiv*, pages 2024–06, 2024.

[31] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[32] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures. *arXiv preprint arXiv:2303.06275*, 2023.

[34] Yang Tan, Mingchen Li, Bingxin Zhou, Bozitao Zhong, Lirong Zheng, Pan Tan, Ziyi Zhou, Huiqun Yu, Guisheng Fan, and Liang Hong. Simple, efficient, and scalable structure-aware adapter boosts protein language models. *Journal of Chemical Information and Modeling*, 2024.

[35] Mengyu Li, Hongzhao Wang, Zhenwu Yang, Longgui Zhang, and Yushan Zhu. Deeptm: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. *Computational and Structural Biotechnology Journal*, 21:5544–5560, 2023.

[36] Florian Haselbeck, Maura John, Yuqi Zhang, Jonathan Pirnay, Juan Pablo Fuenzalida-Werner, Rubén D Costa, and Dominik G Grimm. Superior protein thermophilicity prediction with protein language model embeddings. *NAR Genomics and Bioinformatics*, 5(4):lqad087, 2023.

[37] Hongdi Pei, Jiayu Li, Shuhan Ma, Jici Jiang, Mingxin Li, Quan Zou, and Zhibin Lv. Identification of thermophilic proteins based on sequence-based bidirectional representations from transformer-embedding features. *Applied Sciences*, 13(5):2858, 2023.

[38] Felix Jung, Kevin Frey, David Zimmer, and Timo Mhlhaus. Deepstabp: a deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences*, 24(8):7444, 2023.

[39] Gang Li, Kersten S Rabe, Jens Nielsen, and Martin KM Engqvist. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS synthetic biology*, 8(6):1411–1420, 2019.

[40] Martin KM Engqvist. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC microbiology*, 18:1–14, 2018.

[41] UniProt Consortium. The universal protein resource (uniprot). *Nucleic acids research*, 36(1):D190–D195, 2007.

[42] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.

[43] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[44] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.