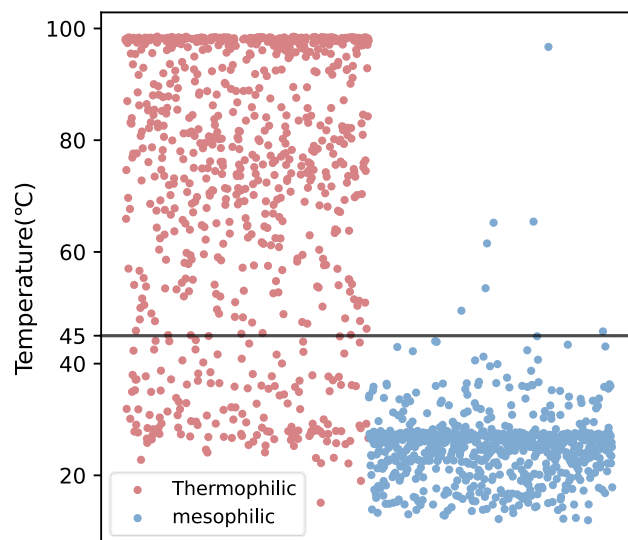
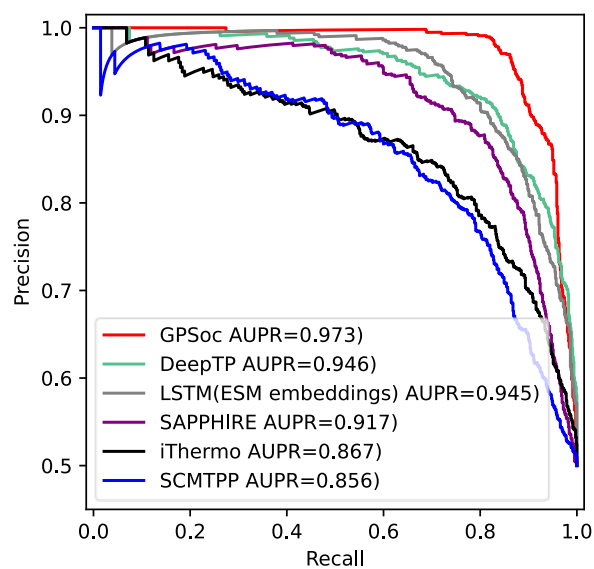


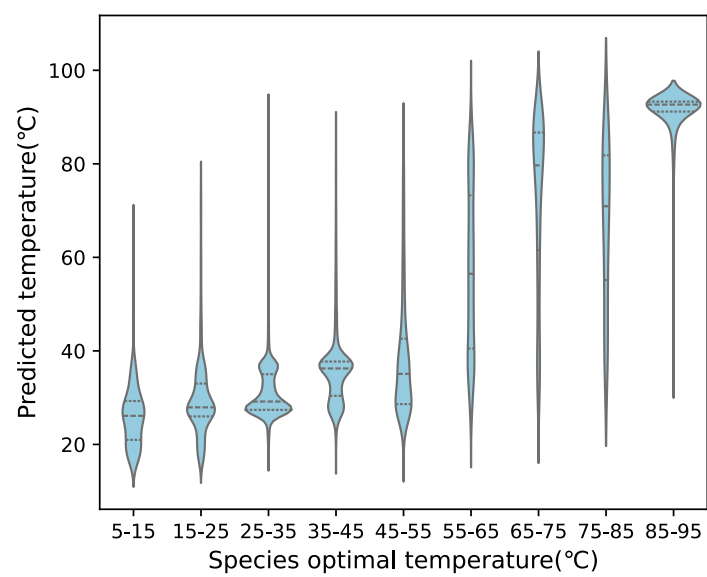
Supplementary



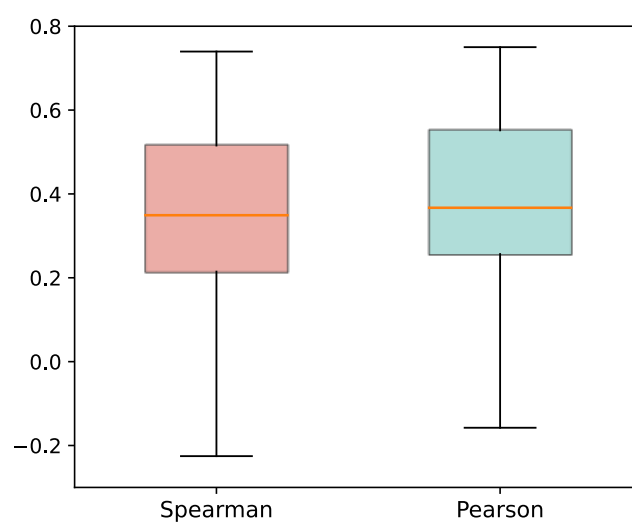
Supplementary Figure 1. The prediction of GeoPoc on the independent test set.



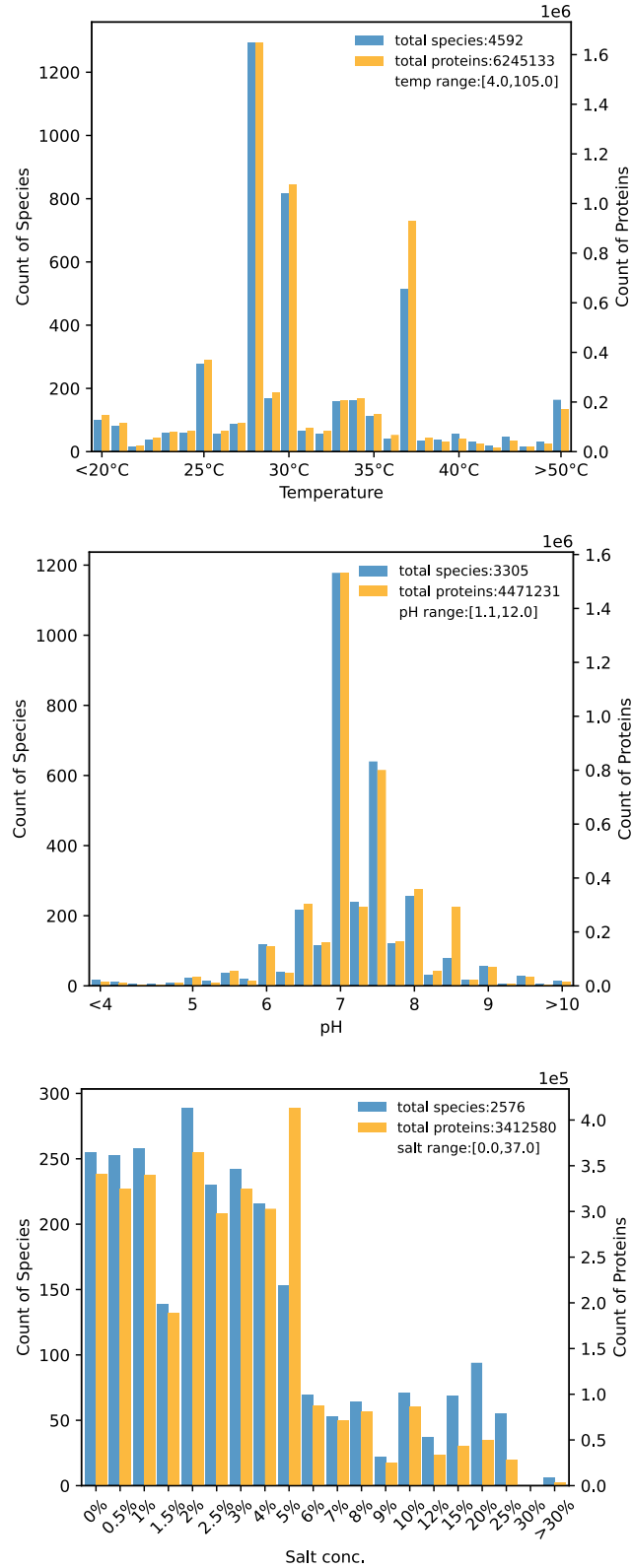
Supplementary Figure 2. Precision-Recall curves of GPSoc and comparison methods in the independent test.



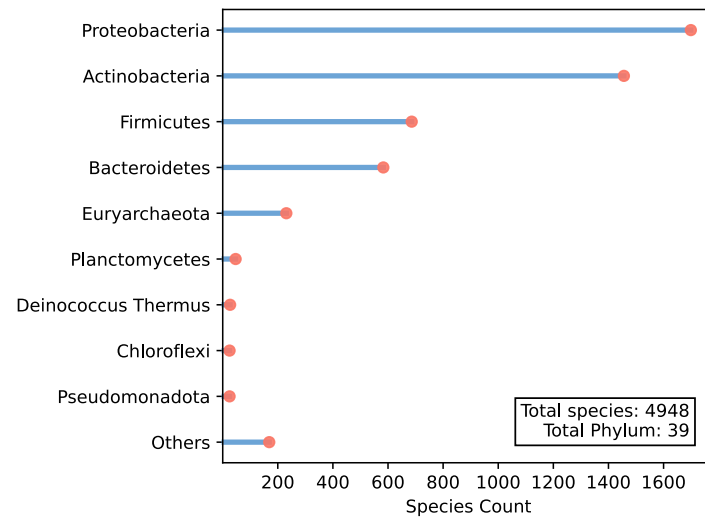
Supplementary Figure 3. Species-level validation. The distribution of predicted protein temperatures for each temperature interval across species.



Supplementary Figure 4. The distribution of Pearson and Spearman correlation coefficient between the conservation and importance of the 52 protein sequence regions from two protein family



Supplementary Figure 5. The species and proteins distribution of temperature, pH, and salt concentration in the data set without removing homologous protein.



Supplementary Figure 6. The distribution of species in different Phylum.

Supplementary Table 1: Compare different dataset sizes, tasks, and protein identities. Note: TPPs denotes thermophilic proteins, MPPs denotes mesophilic proteins, and HPPs denotes hyperthermophilic proteins.

Datasets	Number of Data	Task	Protein identity
Zhang[1]	3521 TPPs, 4895 MPPs	classification	-
Wu[2]	352 HPPs, 580 TPPs, 878 MPPs	classification	-
Gromiha[3]	1609 TPPs, 3075 MPPs	classification	<40%
Lin and Chen[4]	915 TPPs, 793 MPPs	classification	<40%
Charoenkwan[5]	1823 TPPs, 3124 MPPs	classification	<40%
Zhao[6]	9521 TPPs, 9521 MPPs	classification	<40%
GeoPoc	102,615 proteins	regression	<25%

Supplementary Table 2. Performance comparison of GPSoc and ablation methods on the test set.

	Models	PCC
5-fold CV	LSTM(one-hot)	0.40±0.03
	LSTM(ESM embeddings)	0.66±0.04
	GPSoc(w/o geometric features)	0.71±0.01
	GPSoc(w/o attention pooling)	0.74±0.01
	GPSoc	0.75±0.01
Test	LSTM(one-hot)	0.42
	LSTM(ESM embeddings)	0.68
	GPSoc(w/o geometric features)	0.73
	GPSoc(w/o attention pooling)	0.76
	GPSoc	0.77

Supplementary Table 3. Performance comparison of different methods on the independent balanced test set in temperature prediction.

Models	MCC	Accuracy
TMPpred	0.418	0.708
SCMTPP	0.545	0.761
iThermo	0.583	0.791
SAPPHIRE	0.657	0.821
DeepTP	0.746	0.873
LSTM(ESM embeddings)	0.730	0.863
GPSoc	0.823	0.905

Supplementary Table 4. Performance comparison of GPSoc with the ablation method of GPSoc on 5-fold CV and test in pH and salt concentration prediction, respectively.

Datasets	Models	AUPR	f1	MCC
pH 5-fold CV	LSTM(ESM embedding)	0.33±0.03	0.31±0.01	0.01±0.01
	GPSoc(w/o geometric features)	0.53±0.01	0.46±0.02	0.25±0.02
	GPSoc(w/o attention pooling)	0.54±0.01	0.52±0.01	0.27±0.01
	GPSoc	0.55±0.01	0.54±0.01	0.30±0.01
pH test	LSTM(ESM embedding)	0.33	0.31	0.01
	GPSoc(w/o geometric features)	0.52	0.45	0.23
	GPSoc	0.54	0.52	0.28
	GPSoc	0.56	0.55	0.32
Salt 5-fold CV	LSTM(ESM embedding)	0.41±0.03	0.20±0.02	0.01±0.01
	GPSoc(w/o geometric features)	0.55±0.01	0.45±0.02	0.36±0.01
	GPSoc(w/o attention pooling)	0.56±0.01	0.50±0.02	0.36±0.02
	GPSoc	0.57±0.01	0.51±0.01	0.37±0.01
Salt test	LSTM(ESM embedding)	0.42	0.19	0.01
	GPSoc(w/o geometric features)	0.56	0.46	0.37
	GPSoc(w/o attention pooling)	0.57	0.49	0.37
	GPSoc	0.59	0.51	0.40

Supplementary Table 5. Each amino acid's average importance and frequency thermophilic proteins.

amino acid	importance	frequency
L	0.574	0.1028431102737261
A	0.561	0.08568712534085626
V	0.577	0.07472491525536779
G	0.536	0.07425395868864115
E	0.594	0.07126060051022227
I	0.589	0.06983323091574663
R	0.578	0.05964502518776638
K	0.593	0.05841137418108222
S	0.567	0.056127060833726525
D	0.575	0.04956788865009174
T	0.571	0.048946423180575016
P	0.559	0.04591449528332939
F	0.581	0.041452297812748944
N	0.574	0.03696893049649668
Y	0.584	0.034487418586718506
Q	0.573	0.030265339365676023
M	0.583	0.02188469046060654
H	0.569	0.01711219525209661
W	0.566	0.01281761655957528
C	0.565	0.007796303164949928

Reference

- [1] Zhang G, Fang B. LogitBoost classifier for discriminating thermophilic and mesophilic proteins[J]. Journal of biotechnology, 2007, 127(3): 417-424.
- [2] Wu L C, Lee J X, Huang H D, et al. An expert system to predict protein thermostability using decision tree[J]. Expert Systems with Applications, 2009, 36(5): 9007-9014.
- [3] Gromiha M M, Suresh M X. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms[J]. Proteins: Structure, Function, and Bioinformatics, 2008, 70(4): 1274-1279.
- [4] Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique[J]. Journal of microbiological methods, 2011, 84(1): 67-70.
- [5] Charoenkwan P, Chotpatiwetchkul W, Lee V S, et al. A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides[J]. Scientific Reports, 2021, 11(1): 23782.
- [6] Zhao J, Yan W, Yang Y. DeepTP: a deep learning model for thermophilic protein prediction[J]. International Journal of Molecular Sciences, 2023, 24(3): 2217.