JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

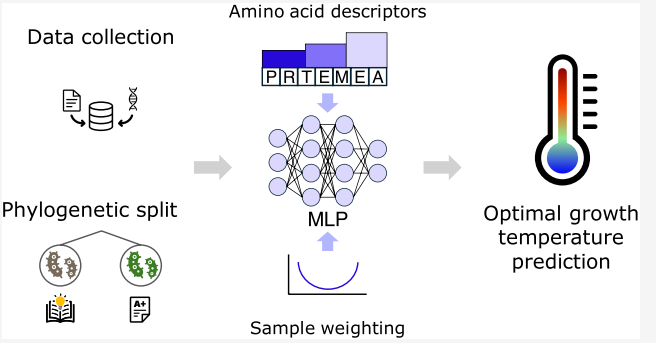
OGTFinder: A Curated Growth Temperature Data Set and Its Application To Predict Optimal Growth Temperatures of Bacteria and Archaea

Sophie Colette, Jaldert François, Bart De Moor, and Vera van Noort*

Cite This: J. Chem. Inf. Model. 2026, 66, 4903−4912 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The optimal growth temperature (OGT) of organisms is valuable in bioprospecting enzymes that work under extreme conditions. Existing OGT prediction models achieve high accuracy but mainly capture trends of overrepresented groups in the training set including organisms that thrive at moderate temperatures and those from well-described taxa. In this study, we incorporated weighted scoring and phylogenetic splits to improve the generalizability of the prediction models. We first built a new growth temperature data set comprising more than 15,000 species distributed over all three domains of life, with special attention to include OGT and extreme temperature data. We then trained machine learning models on the prokaryotic OGT data using proteome-averaged amino acid descriptors. The best-performing model was the multilayer perceptron (MLP) with a test RMSE of 5.49 °C and an R^2 of 0.84. The most important proteome features were related to backbone flexibility and charged residues, as well as surface accessibility. The MLP model is integrated in the command line tool OGTFinder and available under MIT license at: https://github.com/SC-Git1/OGTFinder.



INTRODUCTION

Bioprospecting identifies novel molecules, enzymes, and microbes relevant to biological and biochemical research and industrial applications. Of particular interest for the bioprospecting of microbes and enzymes are industrial process parameters, such as temperature. Temperature affects the internal energy of the cell and therefore directly impacts every cellular component.[1,2] For example, temperature has a large influence on protein stability, with higher temperatures leading to denaturation, protein aggregation, and eventual loss of function, while low temperatures lead to reduced enzyme kinetics and eventual inactivation.[3,4] Each organism has an optimal growth temperature (OGT), at which it achieves its maximal growth rate. According to their OGT, organisms are broadly classified into psychrophiles (OGT <15 °C), mesophiles (15 °C < OGT <45 °C), thermophiles (45 °C < OGT <80 °C), and hyperthermophiles (OGT >80 °C), although these temperature boundaries are not fixed.[5] OGT correlates well with enzyme optimal catalytic temperatures (r = 0.76),[6] and current prediction models of these temperatures rely heavily on OGT as the input parameter.[7−9] Thus, the OGT indirectly informs species selection for bioprospecting enzymes. Beyond its practical applications, the OGT annotation is valuable for investigating general trends in thermal adaptation or specific trends through comparative analysis.

Experimentally determining OGT is limited to culturable organisms, which are biased toward certain phyla and only account for a small fraction (roughly 3%) of all bacteria and archaea.[10] In addition, it is a laborious and time-consuming process, requiring incubation and growth analysis at various temperatures, which is impractical if you want to screen a large number of organisms to select organisms of interest.[11,12] Initial approximation of growth temperature from environmental data of metagenomics can be inaccurate due to local temperature gradients,[13] or influx of micro-organisms.[14] Furthermore, organisms can tolerate temperatures far from their OGT, such as facultative thermophiles whose OGT is larger than 45 °C, but can also grow at moderate temperatures.[15] Similarly, psychrotrophs are cold-tolerant but have an OGT above 15 °C.[16] OGT predictions can thus (1) enhance the culturability of organisms and (2) avoid the high cost and other challenges associated with the experimental determination of optimal growth conditions.

OGT has been found to correlate with properties of multiple omics layers, some of which are interdependent. For example,

Received: December 8, 2025
Revised: March 30, 2026
Accepted: March 31, 2026
Published: April 13, 2026



Downloaded via KU LEUVEN on May 26, 2026 at 10:33:40 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications
© 2026 The Authors. Published by American Chemical Society

increased genomic purine loading (A + G) at higher OGTs mainly results from shifts in amino acid frequencies.[17] Genomic features significantly correlating with OGT include ApG dinucleotide frequency,[17] genome size,[18] and intergenic region lengths.[19] Genome-wide mononucleotide frequencies correlate weakly with OGT,[11,17] and the GC content only shows a significant association after phylogenetic correction.[20,21] In contrast, a strong correlation exist for the GC content of rRNA and tRNA stems.[20] DNA-level and RNA-level relationships have been leveraged for OGT prediction such as in the CnnPOGTP model, which uses genomic k-mer distributions ($k = 2-8$),[22] and in tRNA-thermometer, which predicts OGT from tRNA sequences.[23] However, most OGT prediction models are partially or completely trained on protein-level features, including amino acid composition, sequence, and functional information such as Pfam domains.[5,7,11,24−27] Thermophilic adaptation has been shown to prefer both charged and bulky hydrophobic amino acids,[11,17,28,29] while depleting glutamine (Q), asparagine (N), and aspartate (D) frequencies, as these amino acids are unstable at higher temperatures.[28,30] In addition, thermophiles are enriched in protein domains related to, among others, CRISPR-CAS, spermidine and spermine biosynthesis, and SAM decarboxylase.[25,26] Finally, thermophilicity class information can also be inferred from KEGG metabolic pathways.[27]

Due to its importance, predicting OGT has become a popular research topic with a continuous release of dedicated databases, such as TEMPURA[31] and ThermoBase.[32] Despite this progress, current OGT prediction models face several limitations. One key issue involves the phylogenetic signal in OGT data.[21] Random train-test splits may cause models to overfit on taxonomic patterns, resulting in overoptimistic test performances. This limits the model's ability to generalize to distantly related organisms and gain biological insights.[23] Moreover, mesophiles are often overrepresented in training data, reducing model performance at the extremes, especially at low temperatures.[7,11] While some studies have attempted to address this bias through undersampling of mesophiles,[22,23] others did not.[7,11] Third, the lack of extensive model explanation for most OGT prediction models has limited the identification of the most predictive features and their coupling back to underlying biological mechanisms. Finally, models depending on rRNA/tRNA sequences, genome size, or protein function annotation have a limited applicability when genome sequence or annotations are incomplete.

Here, we present OGTFinder, an OGT prediction model based on physicochemical amino acid properties from the predicted proteomes. Our approach includes phylogenetically informed data splits, hyperparameter tuning with a weighted RMSE for better predictions in extremophiles, and the SHAP framework for model interpretation.[33] OGTFinder is available as a command line tool at https://github.com/SC-Git1/OGTFinder.

MATERIALS AND METHODS

Construction of the Growth Temperature Data Set

Temperature Data Retrieval. Records of growth temperatures, along with their species name and, when available, the source-specific unique identifier, were retrieved from nine different resources. OGT data were downloaded from three databases (ThermoBase,[32] TEMPURA,[31] aciDB[34]) and the Supporting Information of Lyubetsky et al.[19] In addition, cultivation temperatures, i.e., any temperature suitable for growth, were obtained from three culture collection centers (CCUG,[1] CECT,[2] NIES[3]) and MediaDB[35] through Python web scraping. Finally, temperature records were extracted from the bacterial metadatabase BacDive using the API.[36] The data extraction was performed from February to April, 2023.

Taxonomy Mapping. Extracted records were mapped to a unique taxonomic identifier, the NCBI TaxId, based on organism name(s) with the help of the NCBI Datasets Python API and the Taxonomy browser. If no matching TaxId could be found, manual annotation was attempted by looking up the cross-referencing culture collection numbers or through other metadata. The preferred TaxId rank was "strain". Records that could be identified only down to the genus level were removed. Any out-of-use TaxIds were replaced. At this stage, records without a source-defined unique identifier were given one constructed from the source name and a sequential integer corresponding to their position in the final file (e.g., MediaDB0).

Temperature Data Postprocessing. In the case of BacDive, records were of type minimal, maximal, optimum, or growth. Minimal and maximal temperatures were removed as these correspond to conditions of no growth. For the remaining databases, all records were uniformly annotated as either "optimum" when the source explicitly reports OGT,[19,31,32,34] or else "growth" for cultivation temperature. Two records of *Levilactobacillus acidifarinae* and *Levilactobacillus zymae*, both with an OGT of 7.5 °C in BacDive, were excluded because the annotations could not be verified against the original publication.

Data Set Composition. The data set consists of 34,744 growth temperature records, of which 12,864 are optimal growth temperatures and 21,880 are cultivation temperatures. Each record consists of a source identifier with the associated NCBI TaxId, organism name, taxonomic rank, taxonomic lineage (superkingdom, phylum, class, order, family, genus, species), temperature, and temperature type ("growth" or "optimum"). The data set contains mainly bacteria, and almost no eukaryotic OGT data (Table 1). For machine learning, only a subset of prokaryotic OGTs was used. Other records, i.e., eukaryotic and cultivation temperature records, were retained to (1) assess the difference in data quality between cultivation temperature and OGT and (2) provide a resource of growth conditions for more microorganisms, including extremophiles, for future studies.

Table 1. Distribution of TaxIds and Species with Growth Temperature Annotation by Type and Superkingdom

	cultivation		optimal growth	
#	TaxIds	species	TaxIds	species
archaea	290	269	756	624
bacteria	7019	6284	9798	9402
eukaryota	1758	1734	27	27

Machine Learning

Genome Extraction and Annotation. All TaxIds linked to prokaryotic OGTs were provided to the NCBI Datasets command line tool[37] to retrieve their reference genome or, if unavailable, all their associated genomes from NCBI with the "summary" command in March−May 2023. Assemblies present in both RefSeq and GenBank were deduplicated by retaining only the RefSeq version. The species level was used if no genomes were available at the lowest taxonomic level. High-quality genomes were selected based on (1) the average nucleotide identity (ANI) report and (2) the assembly level. The ANI report contains the result of an ANI-based analysis that evaluates whether the assigned taxonomy of an assembly is correct. All genome assemblies with a "taxonomy-check-status" of "inconclusive" or "failed" were removed, leaving only unassessed genomes and genomes with a correct taxonomic identity. Next, for each TaxId, only assemblies with the highest assembly level were kept following the order "Complete Genome > Chromosome > Scaffold > Contig". The 6592 remaining assemblies with available OGT annotation ranging 4−106 °C were downloaded using the NCBI Datasets command line tool and annotated with Prokka[38] (version 1.14.6) using the default installation.

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

The superkingdom and genus names were provided to Prokka as input parameters. NCBI ftp links for the genomes are provided in Data S1.

Feature Calculation. Existing OGT prediction models based on proteome information were trained on amino acid or dipeptide composition.[7,11,24] In this work, we used 58 descriptors that capture key amino acid characteristics derived from dimensionality reduction on physicochemical or structural properties. The hypothesis is that the additional information in the descriptors compared to merely composition can further improve the model and facilitate the biological interpretation. The selected amino acid descriptors belong to the descriptor sets z-Scales,[39] T-scale,[40] ST-scale,[41] MS-WHIM,[42] VHSE,[43] Kidera,[44] Cruciani,[45] Blosum,[46] and Fasgai.[47] The descriptors were extracted with the R package "peptides"[48] (version 2.4.4) and averaged over the entire proteome excluding the top 1% longest sequences as they can exhibit unrealistically large lengths. In addition, superkingdom information was coded into a dummy variable. To remove data imbalance originating from the variable number of records per TaxId, the data were summarized to one observation per TaxId, represented by the mean proteome descriptors and the median OGT, after first averaging any reported temperature ranges. The final machine learning data set consists of 6401 observations.

Data Splitting and Model Evaluation. We performed a phylogenetically informed split by randomly assigning genera to either the training set (89%; $n = 5716$) or the test set (11%; $n = 685$) using scikit-learn's (version 1.5.1)[49] train_test_split with test_size = 0.102. Hyperparameter tuning and performance evaluation were carried out using 5-fold group stratified cross-validation (GroupKFold) on the training set. Genus was used as the grouping variable to ensure that all samples from the same genus were assigned to the same fold, thereby preventing phylogenetic data leakage. The test set was reserved for an external model evaluation. To prevent information leakage, feature normalization was performed within each fold with the function StandardScaler from scikit-learn.[49] Model performance was compared using the overall RMSE as well as the RMSE computed on the top and bottom 5%, 10%, and 20% most extreme observations within each fold. For evaluation on the held-out test set, a final model was retrained on the entire training set by using the best hyperparameters. If applicable, scalers were fit on the entire training set and then applied to the held-out test set. The hyperparameter spaces and selected parameters after tuning are listed in Data S2.

Classical Machine Learning Algorithms. We trained 7 nonlinear models (kernel ridge, Gaussian process, Support Vector Regression (SVR), Multilayer Perceptron (MLP), Random Forest, Adaptive Boosting (AdaBoost), and k-nearest neighbors (KNN)) and 1 linear model (ridge) as baseline using the package scikit-learn.[49] Hyperparameter tuning was performed in Optuna[50] (version 3.6.1) using 5-fold cross-validation with a TPE sampler over 100 trials with a weighted root-mean-square (RMSE) loss function to address the temperature imbalance in the data set. The weights were calculated on the entire training set as follows: for observation i with OGT value x

- Calculate a weight proportional to the inverse of the probability density function:

$$\text{weight}(x) = \frac{1}{\text{KDE}_{\text{Gaussian}}(x) + \delta} \quad (1)$$

with $\delta = 0.01$, and $\text{KDE}_{\text{Gaussian}}(x)$ estimated on the training data with the function `scipy.stats.gaussian_kde`.

- Min-max normalize.

Deep Learning

Deep learning models were trained on the 58 amino acid descriptors without superkingdom. In addition to feature standardization, target values (OGT) were independently standardized using scikit-learn's `StandardScaler`. Nested cross-validation was implemented to ensure fair hyperparameter tuning and performance evaluation. Additionally, the same reweighting was applied for deep learning models as for the scikit-learn implementation for underrepresented temperature ranges. Hyperparameters were optimized using Optuna with a TPE sampler and median pruner (10 startup trials and 20 warmup steps). Each model was optimized with 100 trials, minimizing the mean validation error across the 5 folds. Learning rate (log-uniform over $[10^{-5}, 10^{-2}]$), dropout rate (uniform over $[0.05, 0.5]$), and weight decay (log-uniform over $[10^{-6}, 10^{-2}]$) were optimized for all architectures. Furthermore, architecture specific parameters such as layer count, hidden dimensionality, and gating mechanism type were tuned independently per model. Models were trained with a batch size of 32 and a maximum of 500 epochs with early stopping (30 epochs). The learning rate was halved after 15 epochs without improvements, with a minimum of 10^6.

A total of 19 deep learning architectures were implemented and benchmarked, including the architectures: Attention MLP, Baseline MLP, DANets, FT-Transformer, GANDALF, Gated MLP, Hopular, LassoNet, NODE, RealMLP, ResNet, MLP, ResNet, SAINT, SNN, TabM, TabNet, TabR, VIME, and Wide & Deep. These architectures were adapted from tabular deep learning literature.[51−58] For a detailed overview, see Table S1.

Ensemble Methods

In addition to individual deep learning architectures, we implemented 7 stacking ensemble models combining several scikit-learn regressors trained on the 58 amino acids descriptors and superkingdom (Table S1). All base learners were scaled with a `StandardScaler`, optimized with Optuna and trained from scratch. Additionally, sample reweighting was applied to account for under-represented temperature ranges. The stacking architecture comprised two levels: a set of base learners generating out-of-fold (OOF) predictions and an elastic net meta learner trained on those OOF predictions to produce the final output. We evaluated 3 stacking configurations, each combining 3−7 base learners with complementary inductive biases, drawn from kernel methods (SVR, kernel ridge, Gaussian process), tree-based ensembles (Categorical Boosting (CatBoost), Histogram-Based Gradient Boosting, Random Forest, Gradient Boosting, AdaBoost), gradient boosting libraries (eXtreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM)), linear models (ridge, Bayesian ridge, stochastic gradient descent), neural networks (MLP), and instance-based methods (KNN) from packages scikit-learn (version 1.7.2),[49] xgboost (version 3.1.1),[59] catboost (version 1.2.8),[60] lightgbm (version 4.6.0),[61] keras (version 3.13.2),[62] and tensorflow (version 2.20.0).[63] Hyperparameters were optimized using Optuna with a TPE sampler, with 288 trials per base learner and 300 trials for the meta-learner, minimizing weighted RMSE across the five cross-validation folds. After optimization, each base learner was retrained on the full training set using optimal parameters, followed by meta learner fitting on the aggregated OOF predictions.

Model Interpretation

The descriptors most important to the model were identified using the SHAP (SHapley Additive exPlanations) framework. SHAP values, expressed in target units (°C), are estimates of a feature's marginal contribution to deviations of the predictions from the expected value. SHAP values were calculated with the Python package shap[33] (version 0.46.0) using the default KernelExplainer and the background set to 20 K-means samples.

While SHAP values identify influential features, biological insight relies on the interpretability of those features. Some amino acid descriptors applied in this work lack a biological explanation. Therefore, we investigated the correlations of the descriptors with 586 physico- and biochemical amino acid properties from the AAindex,[64] and two additional sources.[65,66] The indices and their hierarchical categorization were retrieved from the Aanalysis package.[67]

Statistical Significance Testing

To examine whether performances between models were statistically significant, we employed a paired bootstrap resampling test following the methodology of Keohn et al.[68,69] We evaluated the top 10 models based on cross-validation performance, a linear (ridge) model as baseline, and the Tome and Sauer models. The top 10 models included four deep learning architectures (DANets, Attention MLP, Wide&-Deep, Baseline MLP), three scikit-learn models (MLP, SVR, Gaussian process), and three stacking ensemble models (kernel_trio, classic_trio, broad_classic). These models were evaluated on the held-out

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

test set of $N = 685$ samples. For each of the 78 unique model pairs, we evaluated overall RMSE difference and RMSE difference on the top 10th percentile of OGT values.

For each pairwise comparison, $B = 100{,}000$ bootstrap replicates were drawn by sampling N observations with replacement from the test set. The same observations were used for both models in each replicate to preserve the paired structure. RMSE and top 10 RMSE were computed for both models on each bootstrap sample, and the number of wins for each model were recorded and stored. Tied scores for a given bootstrap sample were counted as 0.5 for both models. A two-sided p-value was computed with $p = 2 \times \min(\hat{p}, 1 - \hat{p})$, with $\hat{p}$ the proportion of bootstrap replicates in which model A achieved a better (lower) score than model B. To control the false discovery rate across the 78 comparisons within for each metric, p-values were adjusted using the Benjamini–Hochberg (BH) method. A significance threshold of $\alpha = 0.05$ was applied to the BH adjusted p-values.[70]

Correlation Analysis

The Pearson correlation coefficient was used, unless stated otherwise. Correlations were classified into strong ($|r| \leq 0.7$), moderate ($0.7 < |r| \leq 0.5$), and weak ($0.5 < |r| \leq 0.3$). Correlations for observations with OGT ≥37 °C correspond to the top 20% of OGT values.

Principal Component Analysis

PCA was performed on the normalized values for the 58 amino acid descriptors. Additional explanatory features (genomic GC content, amino acid frequencies, amino acid usage bias (AAUB)) were extracted with custom Python scripts. The genomic GC content was calculated as the proportion of canonical DNA bases guanine (G) and cytosine (C) in the genome sequence after removal of ambiguous or unspecified bases. Amino acid frequencies were calculated as the individual relative abundances of the 20 canonical amino acids in the predicted proteome. The AAUB was then defined as the standard deviation across these 20 amino acid frequencies.[71]

Stratified Subsampling

To mitigate potential bias from data imbalance across taxa, PCA and correlation analysis findings were validated on taxonomically reduced subsets. Within each taxon of a given taxonomic rank (genus, family, order, class, and phylum), observations were normalized, after which Euclidean distances among the observations were computed. The medoid, i.e., the observation with the minimum total distance to all other members, was selected as the representative for that taxon. In the case of ties, the first instance in the data set was chosen.

Phylogenetic Signal

The phylogenetic signal in the amino acid descriptors and OGT per superkingdom was measured with Pagel's lambda[72] on a time-scaled phylogenetic species tree retrieved from TimeTree.[73] The tree contained 298 of the 399 archaea and 3830 of the 5889 bacteria in our training set. Species with only strain-level information were represented by averaged values. Any zero-length branches were trimmed. The final newick tree was imported with the treeio package,[74] and Pagel's lambda was calculated with the function `phylosig` of the R package phytools (version 2.4–4).[75]

Benchmark

Most previously published OGT prediction frameworks do not include taxonomy-aware data splitting, resulting in models prone to overly optimiztic estimations of performance. Constructing an independent test set for benchmarking is challenging due to differences in training data sets and the need to include species that are distantly related to all training data to ensure a fair comparison. Therefore, we retrained existing models using their original features and machine learning algorithms on our training set and evaluated them on our test set in terms of R^2, RMSE, and weighted RMSE using custom scripts. We included the two published OGT frameworks based on proteome information.[7,11] First, Sauer and Wang (2019)[11] trained several multiple linear regression models on genomic and genome-derived (ORF, protein, tRNA, rRNA) features. To ensure compatibility with incomplete genomes, which lack genome size information and sometimes RNA, only models excluding these features were considered. Second, Tome uses dipeptide frequencies and the SVR algorithm to predict OGT.[7]

RESULTS

Cultivation versus Optimal Growth Temperature

First, we examined how the distributions of the reported cultivation temperatures and the OGT relate. Kernel density estimations of the median cultivation temperature per species showed distinct peaks at common culturing temperatures (CCTs) of Eukaryota and Bacteria (Figure S1). This bias was also present in the kernel density estimation of bacterial OGT, albeit to a lesser extent. Next, to assess the extent to which cultivation temperatures deviated from optimal growth temperatures, we performed an analysis similar to Wang et al.[6] The Pearson correlation between the median cultivation temperature and the median OGT for the 2598 species with available data was 0.90 (Figure 1). 14.2% of the species had a deviation bigger than 5 °C, which is about the lower bound of RMSE for current OGT prediction models (Figure S2). In 2.6% of the cases ($n = 67$), the deviation was bigger than 10 °C.

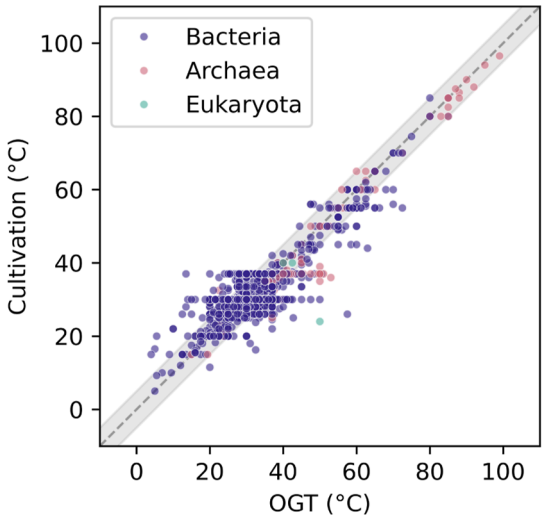


Figure 1. Absolute differences between optimal growth and cultivation median temperatures for 2598 species. Notably, at CCTs 30 and 37 °C, the range of deviations for cultivation temperatures versus OGT exceeds 5 °C (highlighted in gray), indicating significantly higher data quality for OGT measurements.

Nucleotide Bias Dominates Variation in Amino Acid Descriptors, Followed by Environmental Factors

To identify the directions of highest variance and investigate the clustering patterns in the machine learning data set, we performed PCA. The first principal component (PC1) explained 65.68% of the total variance in the features and showed a near-perfect negative correlation ($r = -0.97$) with the GC content (Figure 2A). Although no genome features were included, the GC content was strongly reflected in the frequencies of amino acids with GC-rich codons: proline ($r = 0.95$), glycine ($r = 0.95$), alanine ($r = 0.96$), and arginine ($r = 0.96$), as well as AT-rich amino acids lysine ($r = -0.96$), isoleucine ($r = -0.95$), asparagine ($r = -0.95$), tyrosine (-0.90), and phenylalanine ($r = -0.88$). Methionine ($r = -0.47$) was the exception among AT-rich amino acids. These trends remained present in subsets of

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903–4912

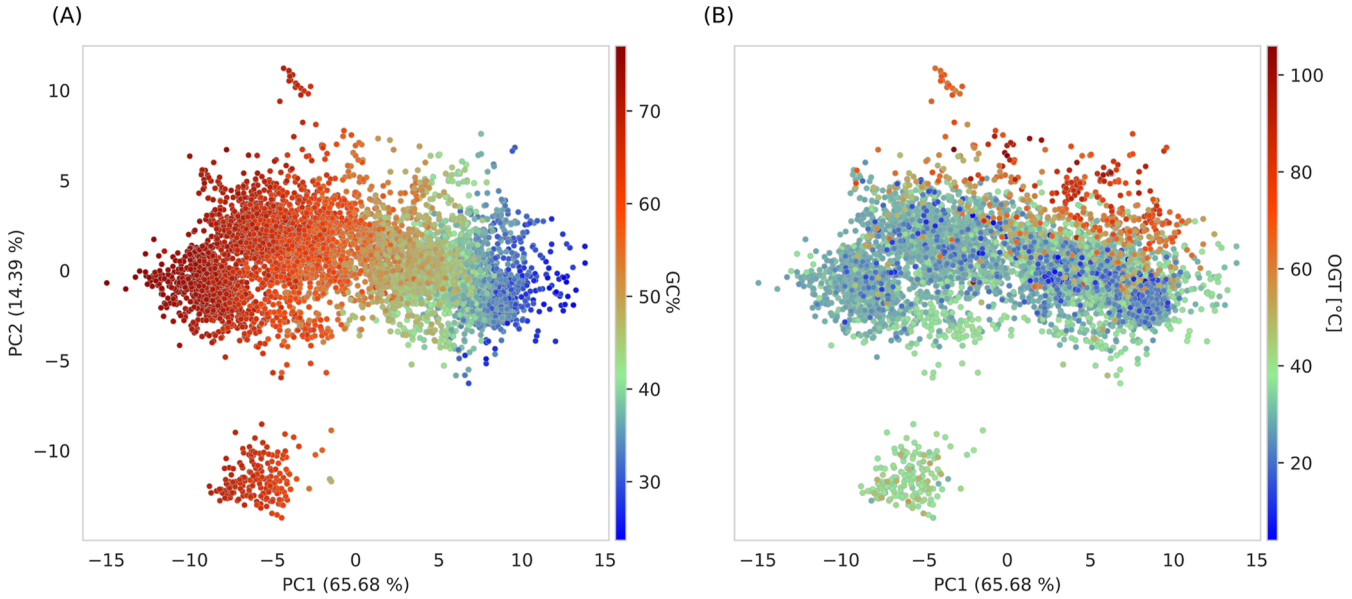


Figure 2. First two principal components (PCs) for the 6401 TaxIds in the machine learning data set, colored by (A) the average GC percentage, and (B) the median OGT. PC1 follows the direction of decreasing GC percentage. High PC2 values indicate thermophilicity. The outlying observations with low PC2 values belong to the archaeal class Halobacteria.

representatives at taxonomic ranks ranging from genus up to phylum (Figure S3 and Table S2).

The second principal component (PC2), explaining 14.39% of the variance, separated organisms living under extreme conditions, namely, the archaeal class Halobacteria at low values and, to a lesser extent, the thermophiles at high values (Figure 2B). PC2 correlated strongly negative with aspartate (r = −0.80), moderately negative with threonine (r = −0.59), and moderately positive with the fraction of leucine (r = 0.66). The correlations weakened on the class or phylum level (Table S3) but the same three amino acids continued to show the strongest correlations with PC2.

Amino Acid Usage Bias Captures Both GC Content and Adaptation to the Environment

In our data set, the AAUB, quantified by the standard deviation of the 20 common amino acid frequencies,[71] was strongly correlated with GC content in GC-rich genomes (r = 0.94) and weakly correlated in AT-rich genomes (r = −0.49). The strength increased for AT-rich genomes after subsampling at higher taxonomic ranks (Table S4). The relationship between AAUB and OGT depended heavily on the GC content (Figures S4 and S2). GC-rich genomes showed no or weak Spearman correlations, whereas for AT-rich genomes, those correlations were moderate to strong, although for OGT <37 °C, this only became apparent at the class and phylum levels (Table S4).

Model Performance Evaluation

From cross-validation, the best models were the classical models SVR and MLP (Figure S5), the deep learning models DANets, Attention MLP and Wide&Deep (Figure S6), and the stacking ensemble model "classic_trio" with base learners as the top three performing classical algorithms (SVR, MLP, and Gaussian process) (Figure S7). The stacking and deep learning models had lower RMSEs at the upper and lower quantiles compared with the classical models, with a higher overall RMSE (Figure 3). In all cases, the RMSE increased near the extremes, especially at low temperatures.

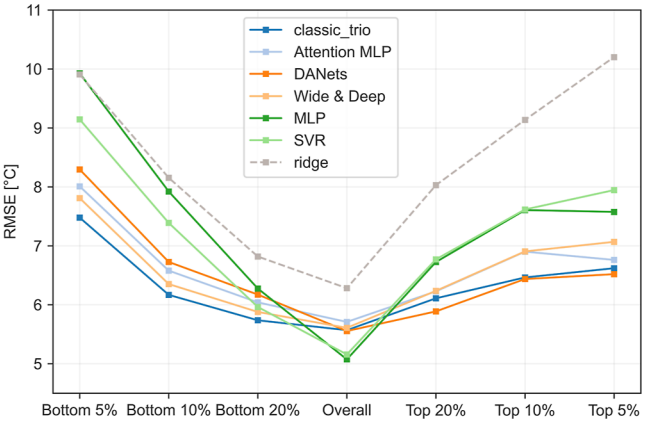


Figure 3. Cross-validation RMSEs for the top performing classical machine learning (SVR, MLP), deep learning (DANets, Attention MLP and Wide&Deep), and ensemble (classic_trio: SVR, MLP, Gaussian process) models compared to a ridge linear model baseline. The percentages indicate the upper and lower 5%, 10%, and 20% quantiles. The best nonlinear models systematically outperform the linear baseline, especially at higher temperatures. A trade-off appears for the deep learning and ensemble methods where improvements at the extremes are accompanied by a declining overall RMSE.

Model Interpretation

Our OGT prediction models were trained on superkingdom (bacteria or archaea) and 58 proteome-averaged amino acid descriptors summarizing key physicochemical and/or structural amino acid properties within their respective descriptor sets.[39−47] To identify the most influential features, SHAP analysis was performed on the MLP, SVR, and Gaussian process models as well as on the stacking ensemble model with these algorithms as its base models. According to the SHAP analysis, these models shared the amino acid descriptors B6, MSWHIM3, and F1 among their most important features (Figures S8−S11). B6 is a descriptor associated with backbone dynamics, MSWHIM3 separates positively charged amino acids, and F1 captures hydrophobicity and buriability, in contrast to accessible

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

surface area (Supporting Information, Data S3, Table S5). Both B6 and F1 had an inverse relation with OGT. The superkingdom, included as a dummy variable in the training set, had only a small contribution in comparison.

Benchmark: Comparison to Existing Frameworks

To assess the performance of our models against the two existing prediction frameworks that also use proteome-based features,[7,11] we retrained their models on our training set and evaluated them on our external test set. From our models, we selected four deep learning architectures (DANets, Attention MLP, Wide&Deep, and Baseline MLP), three classic models (MLP, SVR, Gaussian process), three stacking ensemble models (kernel_trio, classic_trio, broad_classic), and the ridge model as baseline.

Paired bootstrap significance testing with $B = 10^5$ and Benjamini−Hochberg correction on the held-out test set showed clear performance differences between models (Data S4). For overall RMSE, Tome achieved the best score (5.34 °C), followed by MLP (5.49 °C). MLP was significantly better than SVR (5.84 °C, $p_{BH} = 2.23 \times 10^{-4}$), Wide&Deep (6.27 °C, $p_{BH} < 0.001$), DANets (6.39 °C, $p_{BH} < 0.001$), and Sauer (6.87 °C, $p_{BH} < 0.001$), but it was not significantly different from Tome ($p_{BH} = 0.390$). Hence, both Tome and MLP showed the comparable best overall RMSE performance. Furthermore, the best ensemble models showed worse overall performance compared with Tome and MLP.

However, when evaluating predictions on the top 10% most extreme temperatures, model rankings shifted: Wide&Deep (7.62 °C), DANets (8.59 °C), and MLP (8.69 °C) were the three best models, and none of their pairwise differences were statistically significant (Wide&Deep vs DANets: $p_{BH} = 0.0503$; Wide&Deep vs MLP: $p_{BH} = 0.2569$; DANets vs MLP: $p_{BH} = 0.8074$). Relative to Tome (9.15 °C), Wide&Deep was significantly better ($p_{BH} = 0.0233$), while DANets and MLP were not significantly different from Tome ($p_{BH} = 0.360$ and 0.522, respectively). Sauer (12.83 °C) was significantly worse than Wide&Deep, DANets, and MLP on this extreme subset (all $p_{BH} < 0.001$). Additionally, the top ensemble models performed worse for predictions at extreme compared to DANets, Wide&Deep, MLP and Tome. These results suggest MLP and TOME are strongest overall, while Wide&Deep, DANets, and MLP are statistically comparable on the hardest high-temperature cases and score best in identifying heat-adapted organisms. Based on this result, we selected the MLP model to be implemented in the OGTFinder package. Per temperature bin performances of our MLP model compared to Tome are provided in (Table 2). A comprehensive overview of the test set predictions of our MLP model and the retrained models (Tome and Sauer) is provided in Data S5.

Table 2. Test RMSE [°C] of the OGTFinder Framework (MLP) Compared to Tome on the Bottom and Top OGT Values

	bottom 5%	bottom 10%	bottom 20%	top 20%	top 10%	top 5%
	<24.0 °C	<25.0 °C	<28.0 °C	>37.5 °C	>60.0 °C	>70.0 °C
Tome	9.55	7.23	5.13	8.10	9.15	7.22
OGTFinder	10.85	8.21	5.69	7.64	8.69	6.19

DISCUSSION

Cultivation Temperatures as Proxy for Optimal Growth Temperatures

In the past, OGT prediction models were trained on cultivation temperatures,[7,11] before switching to OGT.[22,24] To evaluate whether cultivation temperatures, i.e., any temperature at which growth has been reported, can serve as reliable estimations of the OGT, we constructed a new growth temperature data set containing the OGTs and cultivation temperatures for 10,581 and 9067 unique NCBI TaxIds, respectively. The data set spans all three domains of life, although prokaryotes are heavily overrepresented. Some temperature data redundancy is expected because of, for example, the requirement for novel species to be deposited in multiple culture collection centers (Rule 30 (3b))[76] and shared references among OGT databases.

Comparison between the two types of temperature annotation showed that they are strongly correlated and have similar biases toward CCTs. While OGT has its formal definition, in practice, OGT is derived from a subset of tested conditions with varying granularity, leading to differences in OGT data quality.[24] The bias toward CCTs is thus likely, at least in part, due to the selection of evaluation temperatures during the experimental setup. Likewise, cultivation temperatures are not arbitrarily chosen but set to conditions supporting reliable and fast growth, which may explain why many cultivation temperatures are close to those of the OGTs. However, because of the sometimes large deviations of cultivation temperatures from the values of the OGT, we opted to include only the OGT data in our analyses.

The Relation between Amino Acid Usage Bias and OGT Changes with the GC Content

Variation in the amino acid descriptors was dominated by the GC content, and secondarily by environmental conditions, consistent with earlier findings of Hickey and Singer,[77] who reported nucleotide bias as the main driver of amino acid composition, followed by temperature. In addition, we observed a distinct clustering of archaeal class Halobacteria along PC2, likely due to salinity adaptations, as other archaea clustered together with bacteria (Figures S12 and S13). Indeed, PC2 correlated most strongly with the aspartate fraction, which combines two trends in extremophilic adaptation. Thermophiles avoid aspartate due to its chemical instability at high temperatures.[28] In contrast, halophiles prefer negatively charged amino acids because of their stabilizing interactions with the environment and as a negative design mechanism promoting repulsion in misfolded proteins.[29] The AAUB, capturing the degree of preference toward certain amino acids, combines signals of GC content and OGT, with moderate to strong correlations to OGT only in the case of AT-rich genomes.

Main Trends in Amino Acid Adaptations

The most influential descriptors (B6, MSWHIM3, and F1) highlight three well-established thermoadaptive strategies. First, low B6 values, corresponding to decreased backbone dynamics, were associated with a high OGT, indicating that thermophiles favor more rigid protein structures. As kinetic energy increases with higher temperature, more stability is needed to retain the protein structure, which is essential for its function.

Second, MSWHIM3 separates positively charged residues from other, with the lowest values for the negatively charged amino acid aspartate (D) and the polar amino acids serine (S) and threonine (T) (Table S5). Higher fractions of charged

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

residues are typically associated with thermophilic adaptation. The two main reasons are (1) increasing protein thermostability by stabilizing the core through interaction with the solvent and salt bridges at the protein surface[29,78] and (2) disfavoring misfolded states through negative design.[29] At the same time, thermophiles reduce their proportion of the thermolabile aspartate.[79]

Third, F1 captures an amino acid's hydrophobicity and buriability in contrast to the accessible surface area (Data S3). Its negative correlation with OGT suggests that thermophiles favor hydrophilic, solvent-accessible amino acids. An increased accessible surface area promotes solvent interaction, which improves thermostability by preventing solvent molecules from entering the protein.[78] Positive F1 values are associated with sulfur-containing amino acids methionine (M) and cysteine (C), and aromatic amino acids tryptophan (W) and phenylalanine (F). Concerning the sulfur-containing amino acids, no significant correlations were previously found with OGT.[11,28] However, in our data set, the fraction "CM" has consistently weak to moderate negative correlations with OGT for range ≥37 °C in AT-rich genomes, but not in GC-rich genomes, which may have previously obscured the relation (Table S6). Finally, aromatic amino acids have been associated with thermoadaptation as the $\pi-\pi$ interactions contribute to thermostability. However, the overall correlation in our data set between OGT and the fraction $Y + W + F$ is low (Table S6). Furthermore, PCA coloring shows a strong phylogenetic signal (Figures S13 and S14).

Limitations

Findings Reflect Proteome-Wide Adaptations. Since amino acid descriptors were calculated from proteome-averaged amino acid frequencies, their correlations with OGT reflect the thermoadaptation strategy of globular proteins which constitute the majority of proteins in prokaryotes.[17] Minor protein classes such as membrane proteins containing beta-barrel fold or proteins applying complementary strategies such as disulfide bridges have been reported to have distinct amino acid composition signatures of thermoadaptation.[17]

Generalizability to Other Taxa. In our machine learning data set, both the OGT and amino acid descriptors contain significant evolutionary signal under the Brownian motion model of evolution (Table S7). In the case of OGT, Pagel's lambda was 0.96 ($p < 0.001$) for Archaea and 0.92 ($p < 0.001$) for Bacteria. Therefore, data leakage is likely under a random training-test split. In reality, the best model is a trade-off between model generalizability and performance.[23] Here, we kept observations at the same genus level together.

Performance on Psychrophiles. All of our models capture thermophilic trends better than psychrophilic ones, and this trend can also be seen for the Tome model. The reason is 2-fold. First, the data set consists of only 0.9% ($n = 58$) psychrophiles compared to 88,9% ($n = 5689$) mesophiles and 10.2% ($n = 656$) (hyper)thermophiles. Given the few psychrophilic observations, the model does not focus enough on accurately predicting these during training. To partially overcome this, we opted for weighted RMSE. Second, psychrophiles have weaker signals in their amino acid composition, making it more challenging for the model to predict them accurately.[79]

OUTLOOK

Recently, embeddings from protein language models (pLMs) have proven useful for predicting various protein properties.[80] However, two key challenges remain. First, pLMs offer limited biological insight.[81] Second, pLMs remain computationally expensive, especially in the absence of server-grade GPUs. Our analysis shows that calculations on large CPUs take hours and require substantial RAM (Data S6). Thus, calculating embeddings for the complete proteome of each species of interest is beyond the capacity of a typical microbiology laboratory. Amino acid composition-based features provide a practical alternative that is more than 1000 times faster, enabling proteome-wide predictions and maximizing biological interpretability.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c03033.

NCBI download links for the genomes in the machine learning data set; initial hyperparameter spaces and selected hyperparameters; amino acid descriptor interpretations; paired bootstrap results of the benchmark; detailed benchmark results with NCBI taxid, name, lineage, OGT measured, OGT predicted errors, and OGT predictions for the benchmark set; and CPU timings for proteome-level ESM2 model embeddings versus amino acid descriptors (XLSX)

Kernel density estimation of temperature data per type; deviations between optimal growth and cultivation temperature; PCA on the machine learning data set and its taxonomically reduced subsets (genus, family, order, class, phylum); AAUB distribution in the PCA plot; cross-validation performances of the classical, deep learning, and ensemble models; top 20 features from SHAP analysis of 4 models (MLP, SVR, Gaussian process, and the stacking ensemble thereof); taxonomic distribution in the PCA plot; aromatic amino acid fraction (Y + W + F) distribution in the PCA plot; overview of ensemble methods and deep learning architectures; GC-content correlations with PC1 and AT/GC-rich amino acid fractions; PC2 correlations with OGT and amino acid fractions (Asp, Leu, Thr); AAUB correlations with the GC content and OGT; per amino acid contributions to the key descriptors; aromatic and sulfur-containing amino acid fraction correlations with OGT and GC contents; and phylogenetic signals in OGT and machine learning features (PDF)

AUTHOR INFORMATION

Corresponding Author

Vera van Noort − *Department of Microbial and Molecular Systems, Computational Systems Biology, KU Leuven, Leuven 3001, Belgium; Institute of Biology Leiden (IBL), Leiden University, Leiden 2333 BE, The Netherlands;* orcid.org/0000-0002-8436-6602; Email: vera.vannoort@kuleuven.be

Authors

Sophie Colette − *Department of Microbial and Molecular Systems, Computational Systems Biology, KU Leuven, Leuven 3001, Belgium; Department of Biosystems, Laboratory of Gene Technology, KU Leuven, Leuven 3001, Belgium;* orcid.org/0009-0004-2193-3309

Jaldert François − *Department of Microbial and Molecular Systems, Computational Systems Biology, KU Leuven, Leuven*

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

3001, Belgium; Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Leuven 3001, Belgium

Bart De Moor − *Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Leuven 3001, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c03033

Notes

The growth temperature data set is hosted on Zenodo (10.5281/zenodo.15282729). OGTFinder is available on GitHub and archived on Zenodo (10.5281/zenodo.15282983). The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by KU Leuven project C1 "ACES" [C16/20/001] and VLAIO project "ENZYMARES" [HBC.2021.0076]. J.F. holds a fellowship strategic basic research (grant number: 1SE4625N) of the Research Foundation—Flanders (FWO, Belgium). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government.

■ ADDITIONAL NOTES

[1]https://www.ccug.se/.
[2]https://www.uv.es/uvweb/spanish-type-culture-collection/en/spanish-type-culture-collection-1285872233521.html.
[3]https://mcc.nies.go.jp/.

■ REFERENCES

(1) Ramón, A.; Esteves, A.; Villadóniga, C.; Chalar, C.; Castro-Sowinski, S. A general overview of the multifactorial adaptation to cold: biochemical mechanisms and strategies. *Braz. J. Microbiol.* **2023**, *54*, 2259−2287.

(2) Wang, Q.; Cen, Z.; Zhao, J. others The survival mechanisms of thermophiles at high temperatures: an angle of omics. *Physiology* **2015**, *30*, 97−106.

(3) Plaza Del Pino, I. M.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M. Lower kinetic limit to protein thermal stability: A proposal regarding protein stability in vivo and its relation with misfolding diseases. *Proteins Struct. Funct. Genet.* **2000**, *40*, 58−70.

(4) D'Amico, S.; Claverie, P.; Collins, T.; Georlette, D.; Gratia, E.; Hoyoux, A.; Meuwis, M.-A.; Feller, G.; Gerday, C. Molecular basis of cold adaptation. *Phil. Trans. Roy. Soc. Lond. B Biol. Sci.* **2002**, *357*, 917−925.

(5) Jensen, D. B.; Vesth, T. C.; Hallin, P. F.; Pedersen, A. G.; Ussery, D. W. Bayesian prediction of bacterial growth temperature range based on genome sequences. *BMC Genomics* **2012**, *13*, S3.

(6) Wang, X.; Zong, Y.; Zhou, X.; Xu, L.; He, W.; Quan, S. Artificial Intelligence-Powered Construction of a Microbial Optimal Growth Temperature Database and Its Impact on Enzyme Optimal Temperature Prediction. *J. Phys. Chem. B* **2024**, *128*, 2281−2292.

(7) Li, G.; Rabe, K. S.; Nielsen, J.; Engqvist, M. K. M. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth. Biol.* **2019**, *8*, 1411−1420.

(8) Gado, J. E.; Beckham, G. T.; Payne, C. M. Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4098−4107.

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

(10) Lewis, W. H.; Tahon, G.; Geesink, P.; Sousa, D. Z.; Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **2021**, *19*, 225−240.

(11) Sauer, D. B.; Wang, D.-N. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics* **2019**, *35*, 3224−3231.

(12) Yang, F.; Wang, H.; Chen, L.-Q.; Zhou, N.; Lu, J.-J.; Pu, X.-X.; Wan, B.; Wang, L.; Liu, S.-J. Clostridium lapidicellarium sp. nov. and Clostridium moutaii sp. nov., two species isolated from fermentation cellar-producing sauce-flavour Chinese baijiu. *Int. J. Syst. Evol. Microbiol.* **2024**, *74*..

(13) Khachane, A. N.; Timmens, K. N.; Martins dos Santos, V. A. P. Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. *Nucleic Acids Res.* **2005**, *33*, 4016−4022.

(14) Kurokawa, M.; Higashi, K.; Yoshida, K.; Sato, T.; Maruyama, S.; Mori, H.; Kurokawa, K. Metagenomic Thermometer. *DNA Res.* **2023**, *30*, dsad024.

(15) Bergey, D. H. Thermophilic bacteria. *J. Bacteriol.* **1919**, *4*, 301−306.

(16) Morita, R. Y. Psychrophilic bacteria. *Bacteriol. Rev.* **1975**, *39*, 144−167.

(17) Zeldovich, K. B.; Berezovsky, I. N.; Shakhnovich, E. I. Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLoS Comput. Biol.* **2007**, *3*, No. e5.

(18) Sabath, N.; Ferrada, E.; Barve, A.; Wagner, A. others Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **2013**, *5*, 966−977.

(19) Lyubetsky, V. A.; Zverkov, O. A.; Rubanov, L. I.; Seliverstov, A. V. Optimal Growth Temperature and Intergenic Distances in Bacteria, Archaea, and Plastids of Rhodophytic Branch. *BioMed Res. Int.* **2020**, *2020*, 3465380.

(20) Galtier, N.; Lobry, J. R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **1997**, *44*, 632−636.

(21) Hu, E.-Z.; Lan, X.-R.; Liu, Z.-L.; Gao, J.; Niu, D.-K. A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics* **2022**, *23*, 110.

(22) Wang, S.; Li, G.; Liao, Z.; Cao, Y.; Yun, Y.; Su, Z.; Tian, X.; Gui, Z.; Ma, T. CnnPOGTP: a novel CNN-based predictor for identifying the optimal growth temperatures of prokaryotes using only genomic *k*-mers distribution. *Bioinformatics* **2022**, *38*, 3106−3108.

(23) Cimen, E.; Jensen, S. E.; Buckler, E. S. Building a tRNA thermometer to estimate microbial adaptation to temperature. *Nucleic Acids Res.* **2020**, *48*, 12004−12015.

(24) Barnum, T. P.; Crits-Christoph, A.; Molla, M.; Carini, P.; Lee, H. H.; Ostrov, N. Predicting microbial growth conditions from amino acid composition. *biorxiv* **2024**.

(25) Koblitz, J.; Reimer, L. C.; Pukall, R.; Overmann, J. Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. *Commun. Biol.* **2025**, *8*, 897.

(26) Liu, H.; Zhu, G.; Chen, L.; Ye, H.; Zhang, Y.; Han, G. Machine learning prediction of bacterial optimal growth temperature from protein domain signatures reveals thermoadaptation mechanisms. *BMC Genomics* **2025**, *26*, 957.

(27) Weber Zendrera, A.; Sokolovska, N.; Soula, H. A. Functional prediction of environmental variables using metabolic networks. *Sci. Rep.* **2021**, *11*, 12192.

(28) Villain, E.; Fort, P.; Kajava, A. V. Aspartate-phobia of thermophiles as a reaction to deleterious chemical transformations. *BioEssays* **2022**, *44*, 2100213.

(29) Amangeldina, A.; Tan, Z. W.; Berezovsky, I. N. Living in trinity of extremes: Genomic and proteomic signatures of halophilic, thermophilic, and pH adaptation. *Curr. Res. Struct. Biol.* **2024**, *7*, 100129.

(30) Klipcan, L.; Safro, I.; Temkin, B.; Safro, M. Optimal growth temperature of prokaryotes correlates with class II amino acid composition. *FEBS Lett.* **2006**, *580*, 1672−1676.

(31) Sato, Y.; Okano, K.; Kimura, H.; Honda, K. TEMPURA: Database of Growth TEMPeratures of Usual and RAre Prokaryotes. *Microb. Environ.* **2020**, *35*, n/a, ,

(32) DiGiacomo, J.; McKay, C.; Davila, A. ThermoBase: A database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms. *PLoS One* **2022**, *17*, No. e0268253.
(33) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
(34) Neira, G.; Cortez, D.; Jil, J.; Holmes, D. S. AciDB 1.0: a database of acidophilic organisms, their genomic information and associated metadata. *Bioinformatics* **2020**, *36*, 4970−4971.
(35) Richards, M. A.; Cassen, V.; Heavner, B. D.; Ajami, N. E.; Herrmann, A.; Simeonidis, E.; Price, N. D. MediaDB: A Database of Microbial Growth Conditions in Defined Media. *PLoS One* **2014**, *9*, No. e103548.
(36) Reimer, L. C.; Sardà Carbasse, J.; Koblitz, J.; Ebeling, C.; Podstawka, A.; Overmann, J. Bac *Dive* in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.* **2022**, *50*, D741−D746.
(37) O'Leary, N. A.; Cox, E.; Holmes, J. B.; Anderson, W. R.; Falk, R.; Hem, V.; Tsuchiya, M. T. N.; Schuler, G. D.; Zhang, X.; Torcivia, J.; et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **2024**, *11*, 732.
(38) Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068−2069.
(39) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.
(40) Tian, F.; Zhou, P.; Li, Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.* **2007**, *830*, 106−115.
(41) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids* **2010**, *38*, 805−816.
(42) Zaliani, A.; Gancia, E. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525−533.
(43) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A new set of amino acid descriptors and its application in peptide QSARs. *Pept. Sci.* **2005**, *80*, 775−786.
(44) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.* **1985**, *4*, 23−55.
(45) Cruciani, G.; Baroni, M.; Carosati, E.; Clementi, M.; Valigi, R.; Clementi, S. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemom.* **2004**, *18*, 146−155.
(46) Georgiev, A. G. Interpretable Numerical Descriptors of Amino Acid Space. *J. Comput. Biol.* **2009**, *16*, 703−723.
(47) Liang, G.; Li, Z. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci.* **2007**, *26*, 754−763.
(48) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R Journal* **2015**, *7*, 4.
(49) Pedregosa, F. others Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(50) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. O.. A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Anchorage AK: USA, 2019, pp 2623−2631.
(51) Holzmüller, D., Grinsztajn, L., Steinwart, I. Code and Data for: Better by default: Strong pre-tuned MLPs and boosted trees on tabular data [NeurIPS. arXiv v2] 2024,
(52) Joseph, M.; Raj, H. GANDALF: Gated Adaptive Network for Deep Automated Learning of Features. *arXiv* **2024**, arXiv:2207.08548.
(53) Schäfl, B.; Gruber, L.; Bitto-Nemling, A.; Hochreiter, S. Hopular: Modern Hopfield Networks for Tabular Data. *arXiv* **2022**, arXiv:2206.00664.
(54) Chen, J.; Liao, K.; Wan, Y.; Chen, D. Z.; Wu, J. DANets: Deep Abstract Networks for Tabular Data Classification and Regression. *arXiv* **2022**, arXiv:2112.02962.
(55) Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting Deep Learning Models for Tabular Data. *arXiv* **2023**, arXiv:2106.11959.
(56) Somepalli, G.; Goldblum, M.; Schwarzschild, A.; Bruss, C. B.; Goldstein, T. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *arXiv* **2021**, arXiv:2106.01342.
(57) Arik, S. O.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *arXiv* **2020**, arXiv:1908.07442.
(58) Cheng, H.-T.; et al. Wide & Deep Learning for Recommender Systems. *arXiv* **2016**, arXiv:1606.07792.
(59) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: San Francisco California USA, 2016, pp 785−794.
(60) Dorogush, A. V.; Ershov, V.; Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
(61) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146−3154.
(62) Chollet, F.; others Keras. https://keras.io, 2015.
(63) Abadi, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2015**, arXiv:1603.04467.
(64) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **2007**, *36*, D202−D205.
(65) Koehler, J.; Woetzel, N.; Staritzbichler, R.; Sanders, C. R.; Meiler, J. A unified hydrophobicity scale for multispan membrane proteins. *Proteins* **2009**, *76*, 13−29.
(66) Lins, L.; Thomas, A.; Brasseur, R. Analysis of accessible surface of residues in proteins. *Protein Sci.* **2003**, *12*, 1406−1417.
(67) Breimann, S.; Kamp, F.; Steiner, H.; Frishman, D. AAontology: An Ontology of Amino Acid Scales for Interpretable Machine Learning. *J. Mol. Biol.* **2024**, *436*, 168717.
(68) Koehn, P. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural language Processing*, 2004; pp 388−395.
(69) Berg-Kirkpatrick, T.; Burkett, D.; Klein, D. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural language Processing and Computational Natural language Learning*, 2012; pp 995−1005.
(70) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B Stat. Methodol.* **1995**, *57*, 289−300.
(71) Bohlin, J.; Brynildsrud, O.; Vesth, T.; Skjerve, E.; Ussery, D. W. Amino Acid Usage Is Asymmetrically Biased in AT- and GC-Rich Microbial Genomes. *PLoS One* **2013**, *8*, No. e69878.
(72) Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **1999**, *401*, 877−884.
(73) Kumar, S.; Suleski, M.; Craig, J. M.; Kasprowicz, A. E.; Sanderford, M.; Li, M.; Stecher, G.; Hedges, S. B. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **2022**, *39*, msac174.
(74) Wang, L.-G.; Lam, T. T.-Y.; Xu, S.; Dai, Z.; Zhou, L.; Feng, T.; Guo, P.; Dunn, C. W.; Jones, B. R.; Bradley, T.; Zhu, H.; Guan, Y.; Jiang, Y.; Yu, G. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **2020**, *37*, 599−603.
(75) Revell, L. J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* **2024**, *12*, No. e16505.
(76) Oren, A.; Arahal, D. R.; Göker, M.; Moore, E. R. B.; Rossello-Mora, R.; Sutcliffe, I. C. International Code of Nomenclature of Prokaryotes. Prokaryotic Code (2022 Revision). *Int. J. Syst. Evol. Microbiol.* **2023**, *73*..
(77) Hickey, D. A.; Singer, G. A. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* **2004**, *5*, 117.

https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912

(78) Hait, S.; Mallik, S.; Basu, S.; Kundu, S. Finding the generalized molecular principles of protein thermal stability. *Proteins: Struct., Funct., Bioinf.* **2020**, *88*, 788−808.
(79) Yang, L.-L.; Tang, S.-K.; Huang, Y.; Zhi, X.-Y. Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. *Genome Biol. Evol.* **2015**, *7*, 3426−3433.
(80) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **2024**, *15*, 7407.
(81) Gujral, O.; Bafna, M.; Alm, E.; Berger, B. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proc. Natl. Acad. Sci. U. S. A.* **2025**, *122*, No. e2506316122.



https://doi.org/10.1021/acs.jcim.5c03033
J. Chem. Inf. Model. 2026, 66, 4903−4912