Bioinformatics, 2022, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Machine learning for optimal growth temperature prediction of prokaryotes using amino acid descriptors

Sophie Colette[1,2], Jaldert François[1,3], Bart De Moor[3] and Vera van Noort[1,4,*]

[1]KU Leuven, Department of Microbial and Molecular Systems, Computational Systems Biology, Leuven 3001,Belgium, [2]KU Leuven, Department of Biosystems, Laboratory of Gene Technology, Leuven 3001,Belgium, [3]KU Leuven, Department of Electrical Engineering, ESAT-STADIUS, Leuven 3001, Belgium and [4]Leiden University, Institute of Biology Leiden (IBL), Leiden 2333 BE, The Netherlands
[*]Corresponding author. vera.vannoort@kuleuven.be
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

The optimal growth temperature of organisms is valuable in bioprospecting enzymes that work under extremophilic temperature conditions. Existing prediction models achieve high accuracy, but they mainly capture the trends of the mesophiles (OGT = 15-45°C) and taxonomies that are abundantly present in the training set. In this study, we investigated the use of a weighted root mean square deviation (RMSE) and phylogenetic splits to improve the generalizability of the prediction models trained on amino acid descriptors. To do this, we first built a new OGT database of more than 10,000 species distributed over 51 phyla of Bacteria, Archaea and Eukaryota with special attention to include extreme temperature data. Then, we trained machine learning models on 6,401 observations with available genomes. The best performance was from the multi-layer perceptron with an RMSE of 5.05°C and an R^2 of 0.81. The most important model descriptors were related to charged residues, as well as bulky, hydrophobic residues.

Key words: machine learning, interpretable AI, Optimal growth temperature, temperature adaptation, prokaryotes

Introduction

Bioprospecting identifies novel molecules, enzymes, and microbes relevant to biological and biochemical research and industrial applications. Of particular interest are process parameters such as temperature for bioprospecting of enzymes and microbes. Each organism is adapted to a specific temperature range that influences its growth rate (Corkrey et al., 2016). At the optimal growth temperature (OGT), an organism achieves its maximal growth rate. Organisms are broadly classified based on their OGT 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 (Jensen et al., 2012).

Temperature affects the internal energy of the cell, and changes in temperature directly impact every cellular component (Ramón et al., 2023; Wang and others, 2015). For instance, 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 (Plaza Del Pino et al., 2000; D'Amico et al., 2002).

OGT data is valuable to understand general trends in thermal adaptation, or specific ones through comparative analysis. Correlations have been found between OGT and properties of the genome (Kawashima et al., 2000; Hu et al., 2022; Zeldovich et al., 2007; Sabath et al., 2013), rRNA and tRNA (Galtier and Lobry, 1997; Hurst and Merchant, 2001; Khachane and others, 2005), codon usage (Lobry and Chessel, 2003; Sauer and Wang, 2019), proteome (Zeldovich et al., 2007; Kurokawa et al., 2023; Sauer and Wang, 2019; Tekaia and Yeramian, 2006; Villain et al., 2022), and metabolic network (Weber Zendrera et al., 2019; Farrell et al., 2025). Furthermore, OGT correlates with enzyme optimal catalytic temperatures (r = 0.76; Wang et al. 2024), and current models for predicting enzyme optimal catalytic temperature rely heavily on OGT (Li et al., 2019; Gado et al., 2020; Li et al., 2022). Thus, OGT can indirectly estimate optimal enzyme catalytic temperatures and inform species selection for bioprospecting enzymes.

Experimentally determining OGT is a laborious and time-consuming process, as it requires incubation and growth analysis at various temperatures (Sauer and Wang, 2019; Yang et al., 2024). This method is limited to culturable organisms, which are biased towards certain phyla and only account for a small fraction (roughly 3%) of all prokaryotes (Lewis et al., 2021). Culturability depends on the range of conditions tested, and a lack of knowledge about an organism's environmental requirements, including temperature preference, can impede its culturability (Lewis et al., 2021). In addition, environmental data from metagenomics can be inaccurate due to local temperature gradients (Khachane and others, 2005), or influx of micro-organisms (Kurokawa et al., 2023). Finally, organisms can tolerate temperatures far from their OGT, such

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

as facultative thermophiles whose OGT is larger than 45°C, but can also grow at moderate temperatures (Bergey, 1919). Similarly, psychrotrophs are cold-tolerant but have an OGT above 15 °C (Morita, 1975). OGT predictions can thus (1) enhance the culturability of organisms and (2) help identify their optimal growth conditions.

Given the challenges of experimental determination and the importance of OGT in thermal adaptation research, enzyme prospecting, and increasing culturability, predicting OGT has become a popular research topic. Several OGT prediction models have been developed based on genomic (Wang et al., 2022), tRNA (Cimen et al., 2020), proteomic features (Li et al., 2019; Kurokawa et al., 2023; Barnum et al., 2024), or a combination thereof (Jensen et al., 2012; Sauer and Wang, 2019). Furthermore, the continuous release of dedicated databases, such as TEMPURA (Sato et al., 2020) and ThermoBase (DiGiacomo et al., 2022), further underscores the ongoing interest in understanding optimal growth conditions, especially OGT.

Despite the progress, current OGT prediction models face several limitations. One key issue involves the phylogenetic signal in OGT data Hu et al. (2022). Random train-test splits may cause models to overfit on taxonomic patterns, resulting in overly optimistic test performances, and limiting their ability to generalize to distantly related organisms and gain biological insights Cimen et al. (2020). Moreover, mesophiles are often overrepresented in the training data, reducing model performance at the extremes, especially at low temperatures Sauer and Wang (2019); Li et al. (2019). While some studies have attempted to address this bias through undersampling of mesophiles Wang et al. (2022); Cimen et al. (2020), others did not Li et al. (2019); Sauer and Wang (2019). Finally, these models have often provided limited insights into the underlying biological mechanisms.

Here, we developed an OGT prediction model based on physicochemical amino acid properties Osorio et al. (2015) from predicted proteome information. Our approach includes phylogenetically informed data splits, weighted RMSE for better predictions in extremophiles, and the SHAP framework for model interpretation Lundberg and Lee (2017).

Materials and Methods

Optimal growth temperature database construction

Data retrieval

Optimal growth temperature (OGT) data was downloaded from four databases (ThermoBase (DiGiacomo et al., 2022), TEMPURA (Sato et al., 2020), aciDB (Neira et al., 2020)), MediaDB (Richards et al., 2014)) and the supplementary material of Lyubetsky et al. (2020). In addition, growth temperature data was obtained from three culture collection centers (CCUG [1], CECT [2], NIES [3]) through Python web scraping from February to April 2023. Finally, temperature records were extracted from the bacterial metadatabase BacDive using the API in [DATE] (Reimer et al., 2022).

[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/

Data preprocessing

All records were mapped to a unique taxonomic identifier, the NCBI TaxId, based on organism name(s) and 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 only be identified down to the genus level were removed. Any out-of-use TaxIds were replaced.

Temperature data cleaning

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 "optimum" or, if no indication of optimality was present, "growth" for cultivation growth temperature. Temperature ranges were averaged. 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 paper.

Genome extraction and annotation

All prokaryotic TaxIds were mapped to their reference genome or, if unavailable, all their genomes in NCBI using the NCBI datasets command line tool. 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 for prokaryotes 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, if there were still multiple genome assemblies for a given TaxId, only assemblies with the highest assembly level were kept following the order 'Complete Genome > Chromosome > Scaffold > Contig'.

All 10,652 assemblies left after filtering were downloaded using the NCBI Datasets command line tool. Genomes were annotated with Prokka (version 1.14.6, Seemann 2014) using the default installation. The taxonomic domain and genus names were provided.

Machine learning

Features

Previous OGT prediction models used amino acid (AA) composition to describe the proteome (Sauer and Wang 2019; Li et al. 2019; Barnum et al. 2024). Here, we extracted 58 amino acid descriptors from the descriptor sets z-Scales (Sandberg et al., 1998), T-scale (Tian et al., 2007), ST-scale (Yang et al., 2010), MS-WHIM (Zaliani and Gancia, 1999), VHSE (Mei et al., 2005) Kidera (Kidera et al., 1985), Cruciani (Cruciani et al., 2004), Blosum (Georgiev, 2009) and Fasgai (Liang and Li, 2007) with the R package "peptides" (version 2.4.4, Osorio et al. 2015). We summarized them on the proteome level by averaging them across all proteins. The descriptors capture key amino acid characteristics derived from dimensionality reduction on physicochemical and/or structural properties. The hypothesis is that the additional information in the descriptors, compared to the composition, further improves the model. In addition, taxonomic domain information was coded into a dummy variable.

Dataset composition

While in the past, OGT prediction models were trained on cultivation temperatures (Li et al., 2019; Sauer and Wang, 2019), most recent models have been trained on OGT data (Wang et al., 2022; Barnum et al., 2024). Here, we only used OGT data for modelling. Still, the cultivation temperature records have two purposes: (1) to compare with OGT and (2) to provide temperature growth conditions for more microorganisms, including extremophiles, for further studies.

Our machine learning dataset contains one observation per TaxId with the proteome descriptors averaged over its genomes, the taxonomic domain, and the median OGT. We performed a phylogenetically informed train-test split (90.0%-10.0%), keeping observations of the same family together (Barnum et al., 2024). Hyperparameter tuning and performance evaluation were carried out using 5-fold cross-validation on the training set, while the test set was reserved for the final model comparison.

Machine learning algorithms

We trained 9 linear models (lasso, ridge, Bayesian ridge, elastic net, LARS, LARS-lasso, Huber, PLS, and stochastic gradient descent), 5 non-linear models (kernel ridge, Gaussian process, SVR, multi-layer perceptron and k-nearest neighbor), and the ensemble model histogram gradient boosting using the packages scikit-learn (version 1.5.1, Pedregosa et al. 2011). Hyperparameter tuning was performed in Optuna (version 3.6.1, Akiba et al. 2019) using 5-fold cross-validation over 200 trials with a weighted root-mean-square (RMSE) loss function to address the temperature imbalance in the dataset. The hyperparameter search spaces are given in supplementary material [todo]. The weights were calculated in two steps. For observation i with OGT value x:

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

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

   With $\delta = 0.01$, and $\text{KDE}_{\text{Gaussian}}(x)$ estimated on the training data with *scipy.stats.gaussian_kde*.
2. Min-max normalize

Model interpretation

The most influential descriptors to the model were identified using the SHAP (SHapley Additive exPlanations) framework. SHAP values, expressed in target units (°C), are estimates of each feature's marginal contribution to deviations of the predictions from the expected value. SHAP values were calculated with the Python package shap (version 0.46.0, Lundberg and Lee 2017) using the default KernelExplainer and the background set to 10 K-means samples.

Amino acid descriptor interpretations

Some amino acid descriptors applied in this work lack a clear biological explanation. Therefore, we investigated the correlation of key descriptors with 586 physico- and biochemical amino acid properties from the AAindex (Kawashima et al., 2007), and two additional sources (Koehler et al., 2009; Lins et al., 2003). The indices and their hierarchical categorization were retrieved from the AAanalysis package (Breimann et al., 2024).

Table 1. Distribution of species and genera by taxonomic domain.

	Cultivation		Optimal growth	
#	TaxIds	Species	TaxIds	Species
Archaea	290	269	756	624
Bacteria	7,026	6,287	9,801	9,405
Eukaryota	1,758	1,734	27	27

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 ($0.5 < |r| \leq 0.3$). To address the temperature imbalance, the dataset was sorted by OGT and divided into 5 equal-sized bins (n = 1,280), with correlations calculated within each temperature bin. The fifth bin, comprising the top 20% of OGT values, contains temperatures of 37°C or higher.

Phylogenetic signal

The taxonomic domain-specific phylogenetic signal in the amino acid descriptors and OGT was measured with Pagel's lambda (Pagel, 1999) on a time-scaled phylogenetic tree retrieved from TimeTree (Kumar et al., 2022). The tree contained 298 of the 399 archaea and 3,941 of the 5,889 bacteria in our training set. Any zero-length branch lengths were trimmed. the final newick tree was imported with the the treeio package (Wang et al., 2020), and Pagel's lambda was calculated with the function phylosig of the R package phytools (version 2.4-4, Revell 2024).

Results and Discussion

Data

Over 34,000 records were collected and assigned to a taxid, of which 21,881 were cultivation temperatures and 12,866 optimal growth temperatures (Table 1). The dataset contains mainly Bacteria, and almost no eukaryotic OGT data. Some records are expected to be duplicates because of, for example, the requirement for novel species to be deposited in multiple culture collection centers (Rule 30 (3b), Oren et al. 2023) and shared references among optimal growth temperature databases. The OGTs range from 2°C to 114°C. The GC-content is rather balanced with 57% of entries having a GC-content of 50% or higher.

Cultivation versus optimal growth temperature

OGT data is biased towards common culturing temperatures

In the past, OGT prediction models were trained on cultivation temperature data (Li et al., 2019; Sauer and Wang, 2019) before switching to OGT (Wang et al., 2022; Barnum et al., 2024). While cultivation temperature refers to any temperature at which an organism can grow, OGT is the condition that leads to the highest growth rate. However, in practice, the OGT is derived from a subset of tested conditions, with subsets varying in granularity, leading to differences in OGT data quality (Barnum et al., 2024). In our dataset, based on the kernel density estimations of the median temperature per species TaxId, the cultivation temperatures of Eukaryota and

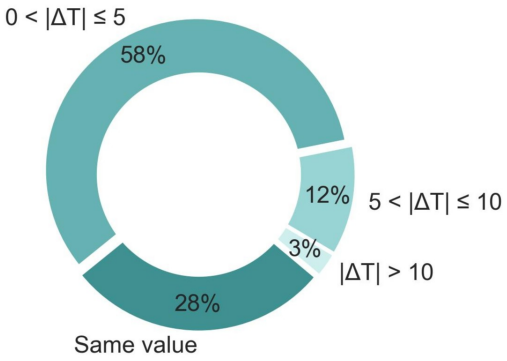


Fig. 1. Distribution of the absolute deviations between median cultivation temperature and median OGT for 2,598 species.

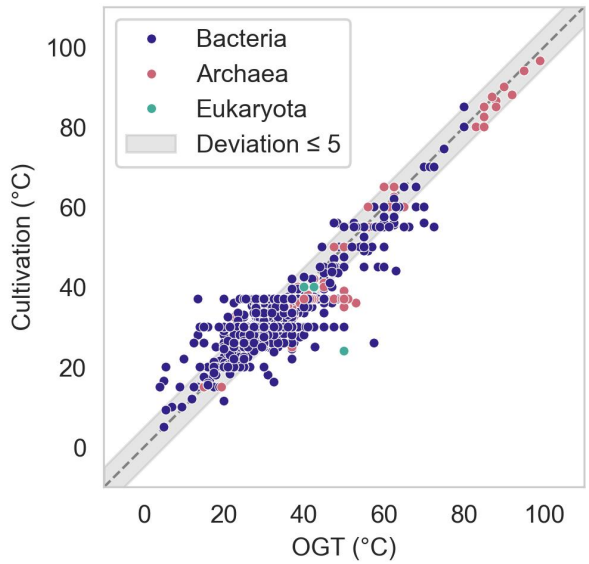


Fig. 2. Absolute differences between optimal growth and cultivation temperatures for 2,598 species. Notably, at common culturing temperatures 30°C and 37°C, the range of deviations for cultivation temperatures versus OGT exceed 5°C, indicating significantly higher data quality for OGT measurements.

Bacteria show distinct peaks at common culturing temperatures (Supplementary Figure S1). The reduced, although remaining bias toward CCTs in bacterial OGT data is likely due to the selection of evaluation temperatures during the experimental setup.

We also compared the median cultivation temperature with the median OGT for the 2,598 species for which both were available. The overall Pearson correlation was 0.90. 14.6% of the species had a deviation larger than 5°C, which is the lower bound of RMSE for current OGT prediction models. In 2.6% of the cases (n = 67), the deviation was larger than 10°C (Figure 1, 2).

Because of the sometimes large deviations of cultivation temperature from OGT, only the latter was considered for model training. In addition, given the negligible amount of eukaryotes with OGT data, we, from this point onwards, focus on the prokaryotic subset.

PCA on prokaryotic OGT

Nucleotide bias follows the direction of maximum variance

The first principal component (PC1) from principal component analysis (PCA) explains 65.68% of the total variance in the features. PC1 shows a near-perfect inverse correlation (r = -0.97) with GC-content (Figure 3A).

The GC-content is 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), glutamine (r = -0.95), tyrosine (-0.90) and phenylalanine (r = -0.88). Methionine (r = -0.47) is the exception among AT-rich amino acids.

Differences in environmental conditions define PC2

The second principal component, explaining 14.39% of the variance, separates organisms living in extreme conditions, namely, the archaeal class *Halobacteria* at low values and to a lesser extent, the thermophiles at high values (3B). The distinct grouping of *Halobacteria* is not due to a taxonomic domain split, as other archaea cluster together with bacteria (Supplementary Figure S2)

PC2 correlates strongly negative with the fraction of aspartate (r = -0.80) and moderately positive with the fraction of leucine (r = 0.66). The aspartate fraction combines two trends in extremophilic adaptation. Thermophiles avoid aspartate due to its chemical instability at high temperatures (Villain et al., 2022). In contrast, halophiles prefer negatively charged amino acids as a negative design mechanism promoting repulsion in misfolded proteins (Amangeldina et al., 2024).

Amino acid usage bias captures both GC-content and adaptation to environment

The main driver of amino acid composition is nucleotide bias, followed by temperature (Hickey and Singer, 2004). In our dataset, the amino acid usage bias, measured by the standard deviation of the 20 common amino acid frequencies, is strongly correlated with GC-content in GC-rich genomes (GC% >= 50%: r = 0.94; GC% < 50%: r = -0.49). This trend remains across temperature bins. AAUB correlates moderately with OGT only for AT-rich genomes and in the top 20% of the dataset (r = 0.57) (Supplementary Figure S3).

Phylogenetic signals in the training data

Both OGT and the descriptors contain significant evolutionary signal under the Brownian motion model of evolution (Supplementary Table S1). For Archaea, Pagel's lambda was 0.96 ($p < 0.001$), and for Bacteria 0.92 ($p < 0.001$). Therefore, under a random training-test split, data leakage occurs and an alternative phylogeny-based split should be performed. In reality, the best model is a trade-off between generalization and performance (Cimen et al., 2020). Here, we kept observations of the same genus level together.

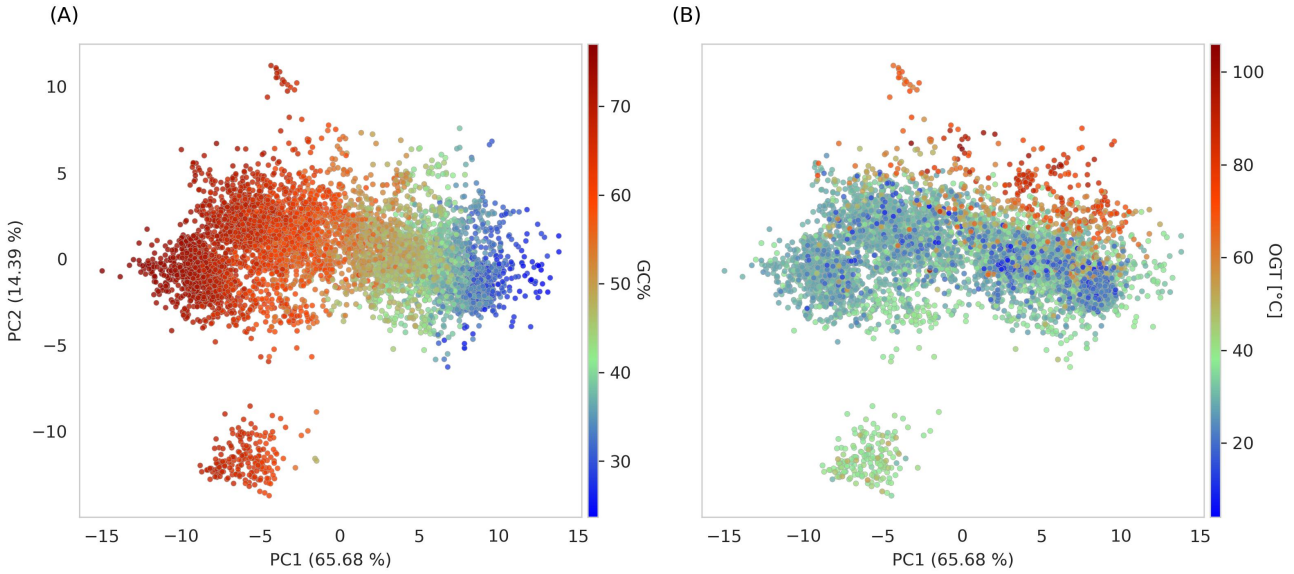


Fig. 3. First two principal components (PCs) for the 6,401 taxids in the OGT dataset, colored by (A) the average GC percentage, and (B) the median optimal growth temperature. PC1 follows the direction of decreasing GC percentage. High PC2 values indicate thermophilicity. The outlying observations with low PC2 values belong to the class Halobacteria.

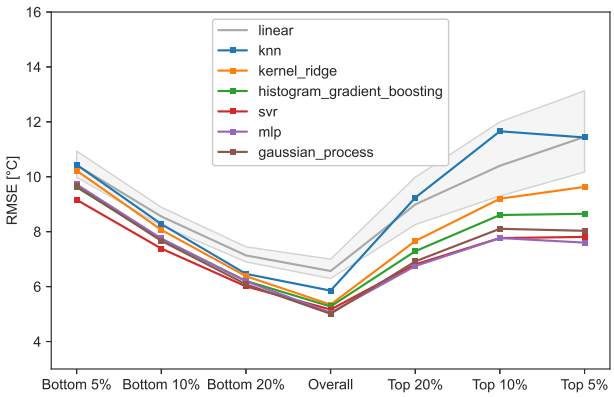


Fig. 4. Cross-validation RMSE for the individual non-linear models compared to a linear model background (CI = 95%). The percentages indicate the upper and lower 5%, 10%, and 20% quantiles. Improvement over the linear models is stronger at higher temperatures

Machine learning results

Non-linear models show the most improvement in (hyper)thermophiles

Non-linear models have consistently better cross-validation RMSEs compared to the linear models (RMSE = 6.57°C) overall and specifically in the thermophilic region 4. The best model based on overall metrics is the Gaussian process model with RSME = 5.01 and R^2 = 0.815, followed by the multi-layer perception (mlp) model (RMSE = 5.05°C and R^2 = 0.812).

Poor fit for psychrophiles

All models capture thermophilic trends better than the psychrophilic ones. The reason is two-fold. First, the dataset consists for 88,9% (n = 5,689) of mesophiles, followed by 10.2% (n = 656) (hyper)thermophiles and only 0.9% (n = 58) psychrophiles. Given the few psychrophilic observations, we opted for weighted RMSE instead of under- or oversampling of the dataset. Secondly, psychrophiles have weaker signals in their amino acid composition (Yang et al., 2015).

Model interpretation

According to the SHAP analysis on the mlp, svr and kernel ridge model, the most important descriptors are consistently K6, followed by T4. Both K6 and T4 have an inverse relation with OGT. The taxonomic domain, included as a dummy variable in the training set, has only a small contribution in comparison.

K6

The descriptor K6 is a mixture of amino acid properties (Kidera et al., 1985), and shows strongest negative correlations with AA indices related to typical amino acid compositions of nuclear, membrane or anchor proteins. K6 thus indicate a general deviating composition. Lowest K6 values are observed for the positively charged lysine and hydrophobic leucine, which also correlated moderately with PC2 capturing environmental factors. An increase in charged residues including lysine, has previously been described as a trend in thermoadaptation (Villain et al., 2022).

In contrast, the fractions of sulfur-containing amino acids cysteine and methionine contribute most negatively to K6. Previously, no significant correlations were found between these amino acids (C, M) and OGT (Villain et al., 2022; Sauer and Wang, 2019). However, in our dataset, the fraction "CM" is negatively correlated to OGT in the top 20% for AT-rich genomes (r = -0.45), but not in GC-rich genomes (r = -0.09), which may have previously obscured the relation.

T4

Finally, T4 is negatively correlated to side chain volume, with lowest values for W,Y,F, and H. More bulky, hydrophobic residues have previously been shown to be associated with

adaption to high temperatures as the bulkiness leads to denser packing of the protein core (Smole et al., 2011).

Comparison to other OGT prediction models

Comparing OGT prediction models directly requires an independent test set, which is challenging due to different training datasets and the need for distantly related species because of the strong phylogenetic signal. Furthermore, comparison of RMSE and R^2 is unreliable as many models did not include a phylogenetically informed split, resulting in overoptimistic estimates. Still, our mlp model has an RMSE of 5.05°C, which is similar to that of the best existing prokaryotic OGT prediction model (RMSE = 5.17, Sauer and Wang 2019), while including a phylogenetic split on the genus level.

Conclusion

In this work we constructed a new (optimal) growth dataset containing OGT values for 10,584 unique taxids from all three domains of life, although eukaryotes are heavily underrepresented. For the 6,401 taxids with available genomes we built OGT prediction models. The best model was the multi-layer perception with an RMSE of 5.05°C and an R^2 equal to 0.81. This performance is similar to that of the state-of-the-art prokaryotic OGT prediction model, even though we included a phylogenetic split on the genus level to reduce data leakage between training and test set and thus. The most important predictors were related to charged residues (K6), and bulky hydrophobic residues (T4). Moreover, fractions of the sulfur-containing amino acids cyteine and methionine were significantly decreased in (hyper)thermophiles compared to mesophiles, but only significantly in AT-rich genomes.

Competing interests

No competing interest is declared.

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 PhD 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.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Publisher: arXiv Version Number: 1.

Amangeldina, A., Tan, Z. W., and Berezovsky, I. N. (2024). Living in trinity of extremes: Genomic and proteomic signatures of halophilic, thermophilic, and pH adaptation. *Current Research in Structural Biology*, 7:100129.

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

Bergey, D. H. (1919). Thermophilic bacteria. *Journal of Bacteriology*, 4(4):301–306. Publisher: American Society for Microbiology.

Breimann, S., Kamp, F., Steiner, H., and Frishman, D. (2024). AAontology: An Ontology of Amino Acid Scales for Interpretable Machine Learning. *Journal of Molecular Biology*, 436(19):168717.

Cimen, E., Jensen, S. E., and Buckler, E. S. (2020). Building a tRNA thermometer to estimate microbial adaptation to temperature. *Nucleic Acids Research*, 48(21):12004–12015.

Corkrey, R., McMeekin, T. A., Bowman, J. P., Ratkowsky, D. A., Olley, J., and Ross, T. (2016). The Biokinetic Spectrum for Temperature. *PLOS ONE*, 11(4):e0153343.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E. B., Zaharova, B., Lavinha, J., Vona, G., Aman, R., Calì, F., Akar, N., Richards, M., Torroni, A., Novelletto, A., and Scozzari, R. (2004). Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. *The American Journal of Human Genetics*, 74(5):1014–1022.

D'Amico, S., Claverie, P., Collins, T., Georlette, D., Gratia, E., Hoyoux, A., Meuwis, M.-A., Feller, G., and Gerday, C. (2002). Molecular basis of cold adaptation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1423):917–925.

DiGiacomo, J., McKay, C., and Davila, A. (2022). ThermoBase: A database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms. *PLOS ONE*, 17(5):e0268253.

Farrell, A. A., Nesbo, C. L., and Zhaxybayeva, O. (2025). Bacterial Growth Temperature as a Horizontally Acquired Polygenic Trait. *GENOME BIOLOGY AND EVOLUTION*, 17(1):evae277. Num Pages: 20 Place: Oxford Publisher: Oxford Univ Press Web of Science ID: WOS:001393354700001.

Gado, J. E., Beckham, G. T., and Payne, C. M. (2020). Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. *Journal of Chemical Information and Modeling*, 60(8):4098–4107.

Galtier, N. and Lobry, J. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, 44(6):632–636.

Georgiev, A. G. (2009). Interpretable Numerical Descriptors of Amino Acid Space. *Journal of Computational Biology*, 16(5):703–723.

Hickey, D. A. and Singer, G. A. (2004). [No title found]. *Genome Biology*, 5(10):117.

Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J., and Niu, D.-K. (2022). A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics*, 23(1):110.

Hurst, L. D. and Merchant, A. R. (2001). High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society B: Biological Sciences*, 268(1466):493–497.

Jensen, D. B., Vesth, T. C., Hallin, P. F., Pedersen, A. G., and Ussery, D. W. (2012). Bayesian prediction of bacterial growth temperature range based on genome sequences. *BMC Genomics*, 13(Suppl 7):S3.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(Database):D202–D205.

Kawashima, T., Amano, N., Koike, H., Makino, S.-i., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000). Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proceedings of the National Academy of Sciences*, 97(26):14257–14262.

Khachane, A. N. and others (2005). Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. *Nucleic Acids Research*, 33(13):4016–4022.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55.

Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C. R., and Meiler, J. (2009). A unified hydrophobicity scale for multispan membrane proteins. *Proteins*, 76(1):13–29.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., and Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8):msac174.

Kurokawa, M., Higashi, K., Yoshida, K., Sato, T., Maruyama, S., Mori, H., and Kurokawa, K. (2023). Metagenomic Thermometer. *DNA Research*, 30(6):dsad024.

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., and Ettema, T. J. G. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, 19(4):225–240. Publisher: Nature Publishing Group.

Li, G., Buric, F., Zrimec, J., Viknander, S., Nielsen, J., Zelezniak, A., and Engqvist, M. K. M. (2022). Learning deep representations of enzyme thermal adaptation. *Protein Science*, 31(12):e4480.

Li, G., Rabe, K. S., Nielsen, J., and Engqvist, M. K. M. (2019). Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synthetic Biology*, 8(6):1411–1420.

Liang, G. and Li, Z. (2007). 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 & Combinatorial Science*, 26(6):754–763.

Lins, L., Thomas, A., and Brasseur, R. (2003). Analysis of accessible surface of residues in proteins. *Protein Science : A Publication of the Protein Society*, 12(7):1406–1417.

Lobry, J. R. and Chessel, D. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *Journal of Applied Genetics*, 44(2):235–261.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lyubetsky, V. A., Zverkov, O. A., Rubanov, L. I., and Seliverstov, A. V. (2020). Optimal Growth Temperature and Intergenic Distances in Bacteria, Archaea, and Plastids of Rhodophytic Branch. *acidb*, 2020:1–10.

Mei, H., Liao, Z. H., Zhou, Y., and Li, S. Z. (2005). A new set of amino acid descriptors and its application in peptide QSARs. *Peptide Science*, 80(6):775–786.

Morita, R. Y. (1975). Psychrophilic bacteria. *Bacteriological Reviews*, 39(2):144–167.

Neira, G., Cortez, D., Jil, J., and Holmes, D. S. (2020). AciDB 1.0: a database of acidophilic organisms, their genomic information and associated metadata. *Bioinformatics*, 36(19):4970–4971.

Oren, A., Arahal, D. R., Göker, M., Moore, E. R. B., Rossello-Mora, R., and Sutcliffe, I. C. (2023). International Code of Nomenclature of Prokaryotes. Prokaryotic Code (2022 Revision). *International Journal of Systematic and Evolutionary Microbiology*, 73(5a).

Osorio, D., Rondon-Villarreal, P., and Torres, R. (2015). Peptides: A Package for Data Mining of Antimicrobial Peptides. *R JOURNAL*, 7(1):4–14. Num Pages: 11 Place: Wien Publisher: R Foundation Statistical Computing Web of Science ID: WOS:000357431900002.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12(null):2825–2830.

Plaza Del Pino, I. M., Ibarra-Molero, B., and Sanchez-Ruiz, J. M. (2000). Lower kinetic limit to protein thermal stability: A proposal regarding protein stability in vivo and its relation with misfolding diseases. *Proteins: Structure, Function, and Genetics*, 40(1):58–70.

Ramón, A., Esteves, A., Villadóniga, C., Chalar, C., and Castro-Sowinski, S. (2023). A general overview of the multifactorial adaptation to cold: biochemical mechanisms and strategies. *Brazilian Journal of Microbiology*, 54(3):2259–2287.

Reimer, L. C., Sardà Carbasse, J., Koblitz, J., Ebeling, C., Podstawka, A., and Overmann, J. (2022). Bac *Dive* in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Research*, 50(D1):D741–D746.

Revell, L. J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, 12:e16505.

Richards, M. A., Cassen, V., Heavner, B. D., Ajami, N. E., Herrmann, A., Simeonidis, E., and Price, N. D. (2014). MediaDB: A Database of Microbial Growth Conditions in Defined Media. *PLOS ONE*, 9(8):e103548. Publisher: Public Library of Science.

Sabath, N., Ferrada, E., Barve, A., and Wagner, A. (2013). Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biology and Evolution*, 5(5):966–977.

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998). New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *Journal of Medicinal Chemistry*, 41(14):2481–2491.

Sato, Y., Okano, K., Kimura, H., and Honda, K. (2020). TEMPURA: Database of Growth TEMPeratures of Usual and RAre Prokaryotes. *Microbes and Environments*, 35(3):n/a.

Sauer, D. B. and Wang, D.-N. (2019). Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics*, 35(18):3224–3231.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Smole, Z., Nikolic, N., Supek, F., Šmuc, T., Sbalzarini, I. F., and Krisko, A. (2011). Proteome sequence features carry signatures of the environmental niche of prokaryotes. *BMC Evolutionary Biology*, 11(1):26.

Tekaia, F. and Yeramian, E. (2006). Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics*, 7(1):307.

Tian, F., Zhou, P., and Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, 830(1-3):106–115.

Villain, E., Fort, P., and Kajava, A. V. (2022). Aspartate-phobia of thermophiles as a reaction to deleterious chemical transformations. *BioEssays*, 44(1):2100213.

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., and Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, 37(2):599–603.

Wang, Q. and others (2015). The survival mechanisms of thermophiles at high temperatures: an angle of omics. *Physiology*, 30(2):97–106.

Wang, S., Li, G., Liao, Z., Cao, Y., Yun, Y., Su, Z., Tian, X., Gui, Z., and Ma, T. (2022). CnnPOGTP: a novel CNN-based predictor for identifying the optimal growth temperatures of prokaryotes using only genomic k -mers distribution. *Bioinformatics*, 38(11):3106–3108.

Wang, X., Zong, Y., Zhou, X., Xu, L., He, W., and Quan, S. (2024). Artificial Intelligence-Powered Construction of a Microbial Optimal Growth Temperature Database and Its Impact on Enzyme Optimal Temperature Prediction. *The Journal of Physical Chemistry B*, 128(10):2281–2292. Publisher: American Chemical Society.

Weber Zendrera, A., Sokolovska, N., and Soula, H. A. (2019). Robust structure measures of metabolic networks

that predict prokaryotic optimal growth temperature. *BMC Bioinformatics*, 20(1):499.

Yang, F., Wang, H., Chen, L.-Q., Zhou, N., Lu, J.-J., Pu, X.-X., Wan, B., Wang, L., and Liu, S.-J. (2024). Clostridium lapidicellarium sp. nov. and Clostridium moutaii sp. nov., two species isolated from fermentation cellar-producing sauce-flavour Chinese baijiu. *International Journal of Systematic and Evolutionary Microbiology*, 74(11).

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., and Li, Z. (2010). ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids*, 38(3):805–816.

Yang, L.-L., Tang, S.-K., Huang, Y., and Zhi, X.-Y. (2015). Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. *Genome Biology and Evolution*, 7(12):3426–3433.

Zaliani, A. and Gancia, E. (1999). MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *Journal of Chemical Information and Computer Sciences*, 39(3):525–533.

Zeldovich, K. B., Berezovsky, I. N., and Shakhnovich, E. I. (2007). Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLoS Computational Biology*, 3(1):e5.