# Environment and taxonomy shape the genomic signature of prokaryotic extremophiles

**Pablo Millán Arias[1,*,†], Joseph Butler[2,†], Gurjit S. Randhawa[3], Maximillian P. M. Soltysiak[2], Kathleen A. Hill[2], and Lila Kari[1]**

[1]School of Computer Science, University of Waterloo, Waterloo, ON, Canada
[2]Department of Biology, University of Western Ontario, London, ON, Canada
[3]School of Mathematical and Computational Sciences, University of Prince Edward Island, Charlottetown, PE, Canada
[*]Corresponding author, pmillana@uwaterloo.ca
[†]These first authors contributed computer science, respectively biology, aspects of this work

## ABSTRACT

This study presents a machine learning-based analysis supporting the hypothesis that microbial adaptations to extreme temperatures and pH conditions can result in a pervasive environmental component within their genomic signatures. To this end, an alignment-free method was used in conjunction with both supervised and unsupervised machine learning, to analyze genomic signatures extracted from a curated dataset of $\sim$700 extremophilic (temperature, pH) bacteria and archaea genomes. The dataset analyzed in this paper is, to the best of our knowledge, the largest and most comprehensive to date for the study of the genomic signatures of extremophilic organisms. The supervised learning identified specific sets of $k$-mers that are relevant for the identification of the environmental components in the genomic signatures of extremophiles. For $k = 3$, our findings are consistent with amino acid compositional biases and codon usage patterns in coding regions that were previously attributed to extreme environment adaptations. The unsupervised learning of unlabelled sequences, by multiple algorithms, identified some hyperthermophilic organisms with large genomic signature similarities in spite of belonging to different domains This computational finding was confirmed by supervised learning experiments under challenging training scenarios, and corroborated by the presence of shared phenotypic traits and characteristics of the isolating environments.

## Introduction

Life exists in the most unusual and extreme environments on our planet. Biodiversity exists in environments such as volcanoes, deep-sea trenches, and polar regions, that are characterized by extreme physical conditions (temperature, radiation, pressure, salinity, pH, chemical, etc.), and pose unique challenges to life[1]. Organisms that are able to survive, and sometimes thrive, in extreme conditions are known as *extremophiles*, characterized by their peculiar phenotypic adaptations and yet underexplored in their genome composition. The study of the genome organization and diversity of extremophiles is of particular interest, as it can provide insights into the mechanisms of adaptation and the evolution of biodiversity in extreme environmental conditions[2–4]. In particular, the study of microbial extremophiles has yielded important research reagents and gained popularity more recently due to their potential applications in biorefineries[5], as sources of industrially-relevant biocatalysts[6], and due to explorations of microbial dormancy and survivability in outer space[7,8]. This study uses several machine learning techniques and an alignment-free methodology to uncover evidence suggesting that microbial adaptations to extreme temperatures and pH conditions imprint a discernible environmental component in their genomic signatures.

As it has been previously observed, extremophiles have developed a wide range of structural, biochemical and metabolic strategies that support cell viability in high-stress environments, and there is evidence that the adaptive mechanisms arising with convergent evolution of extremophilic taxa can be observed at proteomic, transcriptomic and genomic levels[9]. At the proteomic level, diverse organisms living in certain extreme environments have a strong amino acid compositional bias, attributed in part to convergent proteomic adaptations[3,10,11]. At the genomic level, codon usage patterns in the genomes of different extremophilic organisms are linked to the physiochemical characteristics imposed by selective pressures experienced in their respective environments[2,12,13]. As an another example, high Guanine+Cytosine (GC) content is correlated with DNA stability in thermophiles of diverse taxa[14] and is the major factor influencing tRNA stability in hyperthermophiles[15], while the fraction of Adenine+Guanine (AG) content in coding DNA is correlated with the optimal growth temperature (OGT) in thermophiles[16–18]. These observations support the hypothesis of a correlation between adaptation to extreme environments and specific genome composition features.

Traditionally, genomic composition has been studied under the umbrella term "genomic signature"[19], used for a variety of quantitative measures, pervasive along a genome, that can be used to discriminate between genomes of different species[20]. In the last two decades, numerous studies have confirmed the effectiveness of alignment-free methods that use genomic signatures for the purpose of genome analysis[21], comparison[22], and sensitive taxonomic classification[23–25] even without supervision[26]. These findings confirmed the existence of a strong evolutionary signal that is present in (genome-wide, pervasive) genomic signatures, and offered a different perspective complementing alignment-based taxonomic comparisons and distinctions. In particular, $k$-mer (subwords of length $k$) frequencies have been widely used to classify organisms at different taxonomic levels, from Kingdom to species subtypes[23–25,27–30].

These findings suggested that differences and similarities in genomic signatures can be attributed to phylogeny. However, the possibility exists that there are other contributors to the differentiating power of genomic signatures[31]. Of special interest are the genomes of extremophile microbes of diverse taxa, which present a unique case marked by convergent adaptations to extreme physical environments. As shared phenotypic adaptations can also impact genome sequence composition (i.e., codon usage, nucleotide biases), genomic signature analyses may identify instances of convergent evolution. Prior work on thermophiles has also suggested some genomic composition biases, such as dinucleotide[19] or tetranucleotide[32] frequences, that are pervasive across the genome. In addition, other research suggests that amino acid compositional bias and codon usage patterns in coding regions may contribute to the existence of a detectable pervasive genomic signature, strong enough to differentiate between taxonomically related organisms that live at opposite environmental extremes[10,18].

This paper uncovers comprehensive quantitative evidence suggesting that adaptation to extreme temperatures or pH introduces a detectable environmental component in the genomic signature of microbial extremophiles. This hypothesis was tested using both supervised and unsupervised machine learning algorithms, on a prokaryote dataset consisting of 500,000 bp DNA fragments, each randomly selected from a genome of a bacterial or an archaeal organism adapted to extreme temperature conditions, or extreme pH conditions. Supervised machine learning has proved effective in using genomic signatures for taxonomic classification, on data that was problematic for traditional alignment-based algorithms due to its sparseness, complexity, and high dimensionality[24,25,28]. Thus, the first approach was to use supervised machine learning methods to learn the taxonomic and environmental components of genomic signatures. To this end, several supervised learning algorithms were trained on $k$-mer frequency genomic signatures labelled with either taxonomic labels or with environment category labels. Each classifier was then used to obtain a taxonomic classification (if it was trained using taxonomic labels), or an environment category classification (if it was trained using environment category labels). The classification accuracies

obtained were high for the taxonomic classification, and moderately high for the environment category classification, suggesting the presence of an environmental component in the genomic signature in addition to its taxonomic component. For further insight, interpretability tools of supervised learning were used to determine the features of the genomic signatures (specific *k*-mers) that were most relevant to the environment category classification, and our findings were compared with existing literature on codon usage and amino acid compositional bias in extremophiles.

The presence of an environmental component in the genomic signature was independently confirmed by unsupervised clustering analysis of data, whereby the first step was to assess several unsupervised learning algorithms for their ability to learn the taxonomic structure of unlabelled data. The most performant clustering algorithms were then used to identify several candidate organisms with similar genomic signatures in spite of large taxonomic differences. Of these, additional stringent tests based on supervised learning classifications, in challenging scenarios, identified the hyperthemophile bacterium *Thermocrinis ruber* and the hyperthermophile archaea *Pyrococcus furiosus*, *Thermococcus litoralis*, and *Pyrococcus chitonophagus*, whose genomic signatures were grouped together under all classification and clustering scenarios, by all machine learning algorithms used.

The dataset analyzed in this paper is, to the best of our knowledge, the largest and most comprehensive to date for the study of the genomic signatures of extremophiles. The main contributions of this paper are:

- A comprehensive machine learning analysis of a curated dataset, complete with literature references and annotations, of $\sim$700 high-quality microbial extremophile genomes (temperature and pH).
- A supervised-machine learning-based identification of subsets of 3-mers that contribute to the enviromental (temperature, and pH) component of the genomic signature of microbial extremophiles, together with an assessment of the relative importance of their contribution.
- An unsupervised clustering-based analysis of the genomic signatures extracted from the microbial genomes in the datasets, that supports the existence of an environmental component in the genomic signature of extremophiles.

The combination of supervised and unsupervised learning, also corroborated by the presence of shared biological characteristics, suggests the existence of an *environmental component* in the *k*-mer frequency genomic signature of organisms living in conditions of extreme temperatures or of extreme pH.

## Materials and methods

### Datasets

The data in this study were collected through a systematic literature search focused on identifying extremophilic microbes adapted to environments of extreme temperature and pH. The search was conducted on the PubMed Database (accessed September 2022) and Google Scholar (accessed September 2022) for primary research articles and reviews, and identified 768 microbial species or strains for which extremophilic characteristics were recorded. Subsequently, these species/strains were identified in the Genome Taxonomy Database (GTDB; release R207 April 8, 2022, Accessed February 2023), the gold-standard database for taxonomy[33], and only GTDB species representative genomes with reported completeness of over 95%, and contamination of under 5% were selected. Species/strains were mapped to their identified extremophilic characteristic(s), along with genome assembly numbers provided by GTDB for each given organism. The extremophilic characteristic(s) was validated for each organism by searching PubMed with the given strain/species name, and identifying a primary article/review and/or

reliable BacDive database (accessed February 2023) entry to confirm the accuracy of the characteristic(s). Entries lacking consistent observations related to the growth characteristics of the respective microbe were removed from the dataset.

For this study, we used the following definitions, based on the Optimal Growth Temperature (OGT), respectively Optimal Growth pH (OGpHP): Psychrophile (OGT of < 20ºC)[31], mesophile (OGT of 20-45ºC)[31], thermophile (OGT of 45-80ºC)[31], and hyperthermophile (OGT of > 80ºC)[31], acidophile (OGpH < pH 5)[31] and alkaliphile (OGpH > pH 9)[31]. The dataset was then curated for 154 descriptors, so as to be in accordance with the temperature and pH intervals used in the above definitions. Fourteen entries could not be validated and were discarded from the dataset.

This selection process resulted in 693 annotated high-quality extremophile microbial genome assemblies. These high-quality assemblies were then used to form two datasets according to two extremophilic characteristic(s), as follows. The first dataset, called the Temperature Dataset, is composed of 148 psychrophilic genomes (8 archaeal, 140 bacterial), 190 mesophilic genomes (84 archaeal, 106 bacterial), 183 thermophilic genomes (67 archaeal, 116 bacteria), and 77 hyperthermophilic genomes (70 archaeal, 7 bacterial) for a total of 598 organism genomes (229 archaeal, 369 bacterial) (Table 1). The second dataset, called the pH Dataset, is composed of 100 acidophilic genomes (39 archaeal, 61 bacterial) and 86 alkaliphilic genomes (30 archaeal, 56 bacterial), for a total of 186 organisms (69 archaeal, 117 bacterial) (Table 2). Note that 91 organisms were identified to belong to both the temperature and the pH datasets. The entire dataset is described in Supplementary Table S1, with assembly metadata provided in Supplementary Table S2. The proportions of both the pH and Temperature dataset described in terms of genus, organized by domain and by extremophilic category, are described in greater detail in Supplementary Data S1. As well, phylogenetic trees organized by domain (bacteria, archaea) and environmental category (temperature, pH) are accessible in Supplementary Data S2.

The selection of a genomic fragment $s$ to represent the genome of an organism is a process that has to consider several factors, including fragment length, taxonomic level, and computational complexity of the algorithms used. For methods that rely on $k$-mer frequency for sequence classification, some studies[34,35] suggest the relation $k = log_4(|s|)$, where $|s|$ is the minimum length of sequence $s$ that is necessary, in theory, to obtain statistical significance. However, in practice, longer sequences are needed. For example, another study[26] used sequence length of 500,000 bp in conjunction with $k = 6$ to cluster bacterial sequences at the family level, even though in theory a length of 4,096 bp would have sufficed for this value of $k$. In this study, each genome (assembly) was represented by a single 500,000 bp long DNA fragment, and values $k \in \{3,4,5,6\}$ were used.

More precisely, a representative DNA fragment was constructed from each DNA genome/assembly as follows. First, the contigs of the assembly were sorted by length, from the longest to the shortest. Then, if the longest contig was longer than 500,000 bp, then a random 500,000 bp fragment was selected as the representative DNA sequence for that genome. Otherwise, the sorted contigs were concatenated one by one, until the desired length of 500,000 bp was reached, and this became the DNA representative for that genome. To avoid spurious $k$-mers that arise from the concatenation of contigs, the $N$ character was added as a separator between contigs or contig fragments, but no $k$-mers that contain $N$ were considered when calculating $k$-mer counts. Also note that the inserted letters $N$ were not counted towards the length of the DNA sequence representing each genome/assembly. To eliminate the variable of the strand orientation of the uploaded DNA sequences, the final $k$-mer frequency vector of a sequence was computed as the sum between the vector of its $k$-mer counts and the corresponding vector of $k$-mer counts of its reverse complement[36].

**Table 1.** Composition of the Temperature Dataset: 598 DNA fragments from microbial genomes/species (369 DNA fragments from bacterial genomes, and 229 DNA fragments from archaeal genomes).

| Domain | Temperature Category | # Phyla | # Classes | # Orders | # Families | # Genera | # Species |
|---|---|---|---|---|---|---|---|
| Archaea | Psychrophiles | 2 | 4 | 4 | 5 | 7 | 8 |
| | Mesophiles | 4 | 6 | 7 | 20 | 45 | 84 |
| | Thermophiles | 6 | 11 | 14 | 21 | 41 | 67 |
| | Hyperthermophiles | 5 | 6 | 8 | 15 | 31 | 70 |
| Bacteria | Psychrophiles | 4 | 4 | 6 | 13 | 19 | 140 |
| | Mesophiles | 3 | 3 | 6 | 10 | 14 | 106 |
| | Thermophiles | 15 | 19 | 24 | 27 | 47 | 116 |
| | Hyperthermophiles | 5 | 5 | 5 | 5 | 5 | 7 |

**Table 2.** Composition of the pH Dataset: 186 DNA fragments from microbial genomes/species (117 DNA fragments from bacterial genomes, and 69 DNA fragments from archaeal genomes).

| Domain | pH Category | # Phyla | # Classes | # Orders | # Families | # Genera | # Species |
|---|---|---|---|---|---|---|---|
| Archaea | Acidophiles | 4 | 5 | 7 | 11 | 24 | 39 |
| | Alkaliphiles | 2 | 5 | 5 | 9 | 18 | 30 |
| Bacteria | Acidophiles | 10 | 12 | 13 | 13 | 32 | 61 |
| | Alkaliphiles | 12 | 14 | 25 | 30 | 36 | 56 |

## Sequence Classification Using Supervised Machine Learning

To test the hypothesis of the existence of a genome-wide pervasive environmental component in the genomic signature of extremophiles, the two previously described datasets (Temperature, and pH) were used to determine the average accuracy of the classification. For each dataset, computational experiments were performed using six different classifiers, and different values of $k$ (the length of the $k$-mers analyzed), as detailed below. In addition, for each computational experiment, three different scenarios for labelling the training dataset were analyzed, as follows:

(1) All DNA sequences used in training were labelled taxonomically, by their domain (bacteria or archaea),

(2) All DNA sequences used in training were labelled by their environment category (psychrophile, mesophiles, acidophile, etc.),

(3) All DNA sequences used in training were labelled according to labels sampled from a discrete uniform distribution. The discrete uniform distribution was $Unif(0,3)$ in the case of the Temperature Dataset (four possible labels), and respectively $Unif(0,1)$ in the case of the pH Dataset (two possible labels). This third scenario was introduced as a control, and it was expected to result in predictions of the correct label with a probability equal to the sampling probability for each dataset.

The six different classifiers used for these classification tasks were selected as being representative algorithms of four main categories in the classification of DNA sequences. Support Vector Machines (SVM) were selected as a representative of *Kernel Methods*, with a radial basis function kernel[37]. Random Forest was selected as a representative of *Tree-Based Methods*, with the Gini index as the classification criteria[38]. The third algorithm was an *Artificial Neural Network (ANN)*, with a simple and versatile architecture consisting of two fully connected layers, Linear (512 neurons) and Linear (64 neurons),

each one followed by a Rectified Linear Unit (ReLU) and a Dropout layer with a dropout rate of 0.5. Lastly, a *Digital Signal Processing* framework[25] was considered, whereby pairwise distances between numerical representations of DNA sequences are computed and then used in conjunction with either *Linear Discriminant* (MLDSP-1) or with *Quadratic SVM* (MLDSP-2), or with *Subspace Discriminant* (MLDSP-3) machine learning algorithms. All classifiers were tested with $k$-mer frequencies with values $k \in \{3,4,5,6\}$.

Two different types of computational experiments were performed for each of the two datasets (temperature and pH), supervised machine learning classifier (six classifiers), value of $k$ (3 through 6), and training data labelling (taxonomy, environment, random).

In the first type of tests, called *restriction-free*, the predictive power of the algorithms was simply tested using stratified 10-fold cross-validation, without any modification of the 10-fold cross-validation procedure, as follows. The dataset was split into 10 distinct subsets, called *folds*, and a model was trained using 9 of the folds as training data; the resulting model was validated on the remaining part of the data (i.e., it was used as a test set to compute a performance measure such as accuracy). The performance measure reported by 10-fold cross-validation was then the average of the classification accuracy for each of the 10 possible test sets.

The second type of tests, called *restricted* (or *non-overlapping genera*), was designed to address the possibility that a correct classification based on the environmental label may be influenced by a hidden taxonomic component. In other words, this was to determine whether a DNA sequence was classified correctly as a hyperthermophile due on its environmental characteristics, and not due to its similarity to DNA sequences of the same genus that belong to the hyperthermophile group. To achieve this, we adopted a modified 10-fold cross-validation approach, whereby we ensured that each genus appeared exactly once in each of the 10 folds, while preserving in every fold the percentage of samples from each training category. In this modified approach, if a DNA sequence is in the test set, then no other sequence of the same genus was included in the training set, thus disentangling the environmental component from the taxonomic component of the genomic signature (at the genus level).

As an independent method for assessing the environmental components of the genomic signature, we employed interpretability tools for machine learning methods. Global interpretability tools were preferred as they are useful in understanding the general mechanisms in the data through a global importance measure. Given the high correlation between the $k$-mers, the mean decrease in impurity (MDI) for Random Forest was selected as a $k$-mer global importance measure, and then used to learn the actual $k$-mers that were relevant to the environment-based classification. (This measure was preferred over the widely adopted global-agnostic Permutation Feature Importance method as this is not suitable for handling colinear features.) The methodology used to determine the relevant $k$-mers is as follows. First, a one-vs-all classifier was trained for each environment present in the dataset, using stratified 10-fold cross-validation. Second, the MDI algorithm was used to compute the global importance of each $k$-mer in each fold, and the average taken over all folds was used to create a ranked list of $k$-mers, in decreasing order of their contribution. Finally, for each environmental condition, the “most relevant subset of $k$-mers" was computed, by determining the subset of the ranked $k$-mer list that was sufficient to classify the dataset with the same classification accuracy as when *all* $k$-mers were used in that classification.

### Unsupervised Learning for Sequence Clustering

In unsupervised learning, no labels are provided to the DNA sequences in the dataset, and various algorithms are used to cluster similar sequences and explore the structure of the space of the genomic signatures of the dataset.

Two groups of tests with unsupervised learning algorithms were performed in this study: parametric

clustering algorithms (that take the number of expected clusters as an input parameter), and non-parametric clustering algorithms (that determine automatically the number of clusters). In the first group, three parametric clustering algorithms were used: $k$-means, Gaussian Mixture Model, $k$-medoids, and DeLUCS[26]. The computation of the cluster-label assignments for each sequence in the Temperature and the pH Datasets was performed with various values of the parameter `n_clusters` (the expected number of clusters) in each algorithm, `n_clusters` $\in \{2,4,8\}$ for the Temperature Dataset, and respectively `n_clusters` $\in \{2,4\}$ for the pH Dataset, based on the number of potential true clusters in each dataset.

For each dataset, the strength of each of the two components of the signature (taxonomic, environmental) was assessed by comparing the clustering accuracies of two scenarios, the first where the clustering was assessed against the true taxonomic groups, and the second when the clustering was assessed against the true environmental groups. In each case, the performance was evaluated using the unsupervised clustering accuracy metric, defined as:

$$ACC = \frac{\sum_{i=1}^{n} \mathbf{1}[l_i = f(c_i)]}{n}, \tag{1}$$

where $n$ is the total number of sequences and, for each DNA sequence $x_i$, $1 \leq i \leq n$, we have that $l_i$ is its true taxonomic label, $c_i$ is the cluster label found by the algorithm, $f(c_i)$ is the taxonomic label assigned to $c_i$ by the optimal mapping $f$, and $\mathbf{1}[k = i] \in \{0,1\}$ is an indicator function evaluating to 1 iff $k = i$.

In the second type of test, we assessed whether the clusters of each dataset at the lowest possible taxonomic level (genus) can be discovered by non-parametric clustering algorithms. For this purpose, we used two non-parametric clustering algorithms, HDBSCAN[39] and iterative medoids, combined with three different dimensionality reduction techniques: Variational autoencoders (VAE), Deep Contrastive Learning (CL) and Uniform map approximation (UMAP). We also used *i*DeLUCS, which is semi-parametric, in the sense that its parameter `n_clusters` (herein $= 300$) represents the upper limit of the number of clusters found by the algorithm. These seven clustering algorithms were used to recover the lowest taxonomic groups. The following metrics were defined to assess the quality of the found clusters: the *completeness* of each cluster (defined as the number of occurrences of the most common genus present in the cluster divided by the total number of sequences of that genus in the dataset), and the *contamination* of each cluster (defined as the number of sequences that belong to the most common genus in the cluster, divided by the cluster size). The overall quality of each clustering algorithm was then calculated as the total number of clusters that are at least 0.5 complete, and at most 0.5 contaminated.

## Results

### Supervised machine learning analysis of the Temperature Dataset and the pH Dataset

#### *Supervised classification by taxonomy, environment category, and random label assignment*

Several supervised machine learning computational tests were performed to classify the Temperature Dataset and pH Dataset, respectively, using *(1)* taxonomy labels (domain), *(2)* environment category labels, and *(3)* randomly assigned environment category labels (four for the Temperature Dataset, respectively two for the pH Dataset). In the restriction-free case, classifications using taxonomy labels for training resulted in classification accuracies of over 97.49% for the temperature dataset and over 94.18% for the pH dataset at $k = 6$ across all six classification models. In the restriction-free case, classifications using environment category labels for training (hyperthermophile, thermophile, mesophile, and psychrophile for the Temperature Dataset, and acidophile and alkaliphile for the pH Dataset) resulted in classification accuracies of over 77.59% for the temperature dataset and over 84.95% for the pH dataset at $k = 6$ across all six classification models. In the restricted case (whereby, for any test sequence, the sequences of

the same genus were excluded from the training set) classifications using taxonomy labels for training resulted in accuracies scores similar to the restriction-free case (over 95.30% for the temperature dataset and over 91.90% for the pH dataset at $k = 6$ across all six classification models). Using environment category labels for training in the restricted case resulted in medium-range classification accuracies (over 61.90% for the temperature dataset and over 79.24% for the pH dataset at $k = 6$ across all six classification models). Classifications using randomly assigned labels for training (assigned randomly, in proportion to the dataset composition) resulted in the expected low accuracies (restriction-free case: over 22.91% for the temperature dataset and over 45.70% for the pH dataset at $k = 6$ across all six classification models; restricted case: over 24.10% for the temperature dataset and over 47.30% for the pH dataset at $k = 6$ across all six classification models).

**Table 3.** Classification accuracies of six supervised learning classifiers trained on the Temperature Dataset and pH Dataset, in the *restriction-free* scenario, for three different label assignments (taxonomy, environment category, and random label assignment), and four different values of $k \in \{3,4,5,6\}$. The classification accuracy in each cell is calculated through stratified 10-fold cross-validation.

| Dataset | k-value | Class Labelling Type | Classification Model Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | RBF SVM | Random Forest | ANN | MLDSP-1 | MLDSP-2 | MLDSP-3 |
| Temperature | k=3 | Taxonomy | 98.82 | 97.99 | 97.32 | 92.64 | 96.82 | 92.64 |
| | | Environment | 82.11 | 81.59 | 75.41 | 71.91 | 74.58 | 71.24 |
| | | Random | 23.58 | 25.08 | 27.76 | 25.59 | 26.09 | 24.58 |
| | k=4 | Taxonomy | 99.50 | 98.33 | 98.66 | 98.16 | 97.16 | 98.16 |
| | | Environment | 83.29 | 84.11 | 82.28 | 78.43 | 75.08 | 80.43 |
| | | Random | 25.06 | 23.74 | 27.59 | 25.42 | 26.92 | 23.58 |
| | k=5 | Taxonomy | 99.50 | 98.16 | 99.33 | 97.32 | 97.32 | 98.16 |
| | | Environment | 83.27 | 84.76 | 83.29 | 69.23 | 77.26 | 81.77 |
| | | Random | 24.08 | 20.23 | 23.07 | 26.09 | 25.42 | 24.25 |
| | k=6 | Taxonomy | 99.50 | 98.50 | 99.33 | 99.16 | 97.49 | 98.83 |
| | | Environment | 83.46 | 83.94 | 84.12 | 79.60 | 77.59 | 82.44 |
| | | Random | 27.24 | 22.91 | 26.58 | 28.09 | 25.59 | 24.25 |
| pH | k=3 | Taxonomy | 97.34 | 94.09 | 96.78 | 94.62 | 96.24 | 94.62 |
| | | Environment | 90.94 | 90.94 | 90.38 | 81.18 | 83.87 | 80.11 |
| | | Random | 44.15 | 52.72 | 55.91 | 45.16 | 46.77 | 44.62 |
| | k=4 | Taxonomy | 97.87 | 96.29 | 96.81 | 93.01 | 95.16 | 97.85 |
| | | Environment | 90.44 | 88.80 | 91.58 | 84.95 | 86.02 | 89.78 |
| | | Random | 49.42 | 47.84 | 49.01 | 44.62 | 44.62 | 47.85 |
| | k=5 | Taxonomy | 98.42 | 96.81 | 95.79 | 95.70 | 96.24 | 98.92 |
| | | Environment | 91.55 | 88.30 | 87.81 | 88.17 | 86.02 | 90.32 |
| | | Random | 55.35 | 53.77 | 52.13 | 48.39 | 46.24 | 46.24 |
| | k=6 | Taxonomy | 98.42 | 94.71 | 94.18 | 98.92 | 96.77 | 98.39 |
| | | Environment | 91.99 | 88.30 | 86.70 | 92.47 | 84.95 | 92.47 |
| | | Random | 47.81 | 49.06 | 50.06 | 50.00 | 45.70 | 46.77 |

More specifically, six supervised machine learning algorithms were used to classify the two datasets, based on the taxonomy as well as the environment category labels (temperature, respectively pH), for several $k$-mer lengths ($k \in \{3,4,5,6\}$). The classification tests were performed under two scenarios, *(a) restriction-free*, using stratified 10-fold cross-validation, and *(b) restricted*, that is, using stratified 10-fold cross-validation with non-overlapping genera.

Under the restriction-free scenario (see Table 3) to address the potential bias in the datasets because of the different class sizes, stratified 10-fold cross-validation was used, which maintains the original

proportion of different classes while creating different folds. This prevents the models from being exposed more to the over-represented classes and hence minimizes the bias during model training.

For the Temperature Dataset, the classification accuracy scores obtained (across all six algorithms and all four $k$-values) were in the range of 92.64% to 99.50% when using the taxonomy labels, and in the range of 69.23% to 84.12% when using the temperature category labels. Using the highest value for $k$, namely $k = 6$, improved the minimum accuracy scores to 97.49% when taxonomy labels were used, and to 77.59% when the temperature category labels were used. In the control test, the same sequences were classified, but this time they were randomly assigned one of the four existing temperature category labels. In this scenario, accuracies fell in the expected range of 20.23% to 28.09%, proportional to the probability that a sequence has one of the four temperature category labels.

For the pH Dataset, the classification accuracy scores obtained (across all six algorithms and all four $k$-values) were in the range of 93.01% to 98.92% when using the taxonomy labels, and the range of 80.11% to 92.47% when using the pH category labels. Using the highest value for $k$, namely $k = 6$, improved the minimum accuracy scores to 94.18% when the taxonomy labels were used, and to 84.95% when the pH category labels were used. In the control test, the same sequences were classified, but this time each sequence was randomly assigned one of the two existing pH category labels. In this scenario, the accuracies fell in the expected range of 44.15% to 55.91%, proportional to the probability that a sequence has one of the two pH category labels.

A restricted scenario was also analyzed (see Table 4), to minimize the potential taxonomic bias that may have been introduced in the environment-based classification, due to the presence of sequences belonging to the related species in both the training and the testing folds. In this restricted scenario, non-overlapping genera stratified 10-fold cross-validation was used, as follows. The partitioning of the dataset into the 10-folds was performed such that all the sequences belonging to a particular genus were assigned to the same fold to ensure that models were never exposed to the testing sequence genus during training. It is important to note that these restrictions may force the models to underperform, because they may have never been trained on features similar to that of the testing sequences.

For the Temperature Dataset, classification accuracy scores obtained in this restricted scenario (across all six algorithms and all four $k$-values) were in the range of 87.00% to 99.20% when using the taxonomy labels, and the range of 44.50% to 74.17% when using the temperature category labels. Also in this scenario, using the highest value of $k$, namely $k = 6$, improved the minimum accuracy to 95.30% when using taxonomy labels, and to 61.90% when using temperature category labels. In the control test, the same sequences were randomly assigned one of the four temperature category labels, the accuracies fell in the range of 21.40% to 27.90%.

For the pH Dataset, classification accuracy scores obtained in this restricted scenario (across all six algorithms and all four $k$-values) were in the range of 88.20% to 98.42% when using the taxonomy labels, and in the range of 73.10% to 88.83% when using the pH category labels. Similar to the restriction-free scenario, using the highest value of $k = 6$ improved the minimum accuracy scores to 91.90% when using the taxonomy labels, and to 79.24% when using the pH category labels. In the control test, the same sequences were classified, but each sequence was randomly assigned one of the two pH category labels. In this case, the accuracies fell in the expected range, from 41.74% to 55.96%.

Note that, for both the Temperature Dataset and the pH Dataset, the classification accuracies in the restricted scenario were slightly lower that in the restriction-free scenario, across the board (for both the taxonomic classification, and environment category classification). This could be partly attributed to the decrease in the amount of training data in the restricted scenario. However, as seen in Table 4, the environment category classification accuracies in the restricted scenario were still significantly higher than those for the random label assignment scenario.

**Table 4.** Classification accuracies of six supervised learning classifiers trained on the Temperature Dataset and pH Dataset, in the *restricted* scenario, for three different label assignments (taxonomy, environment category, and random label assignment), and four different values of $k \in \{3,4,5,6\}$. The classification accuracy in each cell is calculated through stratified 10-fold cross-validation with *non-overlapping genera*.

| Dataset | k-value | Class Labelling Type | Classification Model Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | RBF SVM | Random Forest | ANN | MLDSP-1 | MLDSP-2 | MLDSP-3 |
| Temperature | k=3 | Taxonomy | 98.82 | 95.13 | 97.14 | 87.00 | 94.60 | 87.00 |
| | | Environment | 65.57 | 63.25 | 58.10 | 44.80 | 53.30 | 44.50 |
| | | Random | 24.93 | 21.40 | 26.12 | 26.60 | 27.10 | 26.60 |
| | k=4 | Taxonomy | 99.16 | 96.13 | 97.81 | 95.00 | 94.50 | 97.20 |
| | | Environment | 70.55 | 63.75 | 63.29 | 54.00 | 56.70 | 59.90 |
| | | Random | 25.94 | 26.60 | 27.22 | 26.40 | 26.90 | 25.40 |
| | k=5 | Taxonomy | 99.16 | 96.13 | 98.82 | 92.50 | 94.50 | 97.20 |
| | | Environment | 72.21 | 64.13 | 66.89 | 50.00 | 62.70 | 65.40 |
| | | Random | 26.74 | 23.23 | 22.49 | 24.20 | 26.80 | 26.40 |
| | k=6 | Taxonomy | 99.16 | 96.47 | 97.81 | 99.20 | 95.30 | 98.00 |
| | | Environment | 74.17 | 65.48 | 67.88 | 61.90 | 64.70 | 67.90 |
| | | Random | 24.20 | 26.74 | 24.59 | 24.90 | 24.10 | 27.90 |
| pH | k=3 | Taxonomy | 97.37 | 93.54 | 96.78 | 88.70 | 92.50 | 88.20 |
| | | Environment | 79.24 | 86.73 | 84.04 | 73.70 | 76.30 | 74.20 |
| | | Random | 54.91 | 48.57 | 55.96 | 43.50 | 54.30 | 44.10 |
| | k=4 | Taxonomy | 97.37 | 96.20 | 96.78 | 88.20 | 92.50 | 94.10 |
| | | Environment | 81.43 | 83.51 | 85.61 | 73.10 | 79.60 | 80.60 |
| | | Random | 46.83 | 41.74 | 47.31 | 52.70 | 48.40 | 49.50 |
| | k=5 | Taxonomy | 97.89 | 97.28 | 96.23 | 94.60 | 92.50 | 96.80 |
| | | Environment | 80.91 | 88.83 | 83.01 | 77.40 | 79.60 | 83.90 |
| | | Random | 46.73 | 54.83 | 52.44 | 46.80 | 50.00 | 50.00 |
| | k=6 | Taxonomy | 98.42 | 96.23 | 95.73 | 97.30 | 91.90 | 96.80 |
| | | Environment | 83.54 | 86.70 | 79.24 | 81.70 | 80.10 | 86.60 |
| | | Random | 53.15 | 48.69 | 55.62 | 47.30 | 50.50 | 52.70 |

### *Sets of $k$-mers relevant to environment category classifications computed by interpretablity tool of supervised learning algorithm*

Of the six supervised classifiers used in the previous section, in this section we use the Mean Decrease in Impurity (MDI) algorithm for the Random Forest classifier to compute a global measure of feature relevance. This serves as an interpretability tool that provides insight into the relative contribution of each feature ($k$-mer) to the successful classification.

To this end, we first conducted 10-fold cross-validation on a One-vs-All Random Forest classifier, which achieves a specific accuracy for each environment category. In the four computational experiments associated with the Temperature Dataset, the psychrophile category was correctly separated from the other sequences in the Temperature Dataset with 86.31% accuracy, the mesophile category with 71.14% accuracy, the thermophile category with 75.22% accuracy, and the hyperthermophile category with 89.62% accuracy. Similarly, in the two computational experiments associated with the pH Dataset, the alkaliphile category was separated from the other sequences in the pH Dataset with 86.45% accuracy, and the acidophile category with 83.76% accuracy.

We then used the trained models obtained in these computational experiments in conjunction with Random Forest's interpretability tool, the MDI algorithm, to compute a *global importance* measure

for each $k$-mer ($k \in \{3,4,5,6\}$) to determine their relative contribution to the One-vs-All environment-based classification. This global importance can then be visualized using the frequency chaos game representation ($fCGR_k$) to identify potential patterns, as seen Figure 1. A visual inspection of Figure 1 suggests that different 6-mers are relevant in distinguishing each environment category of DNA sequences.

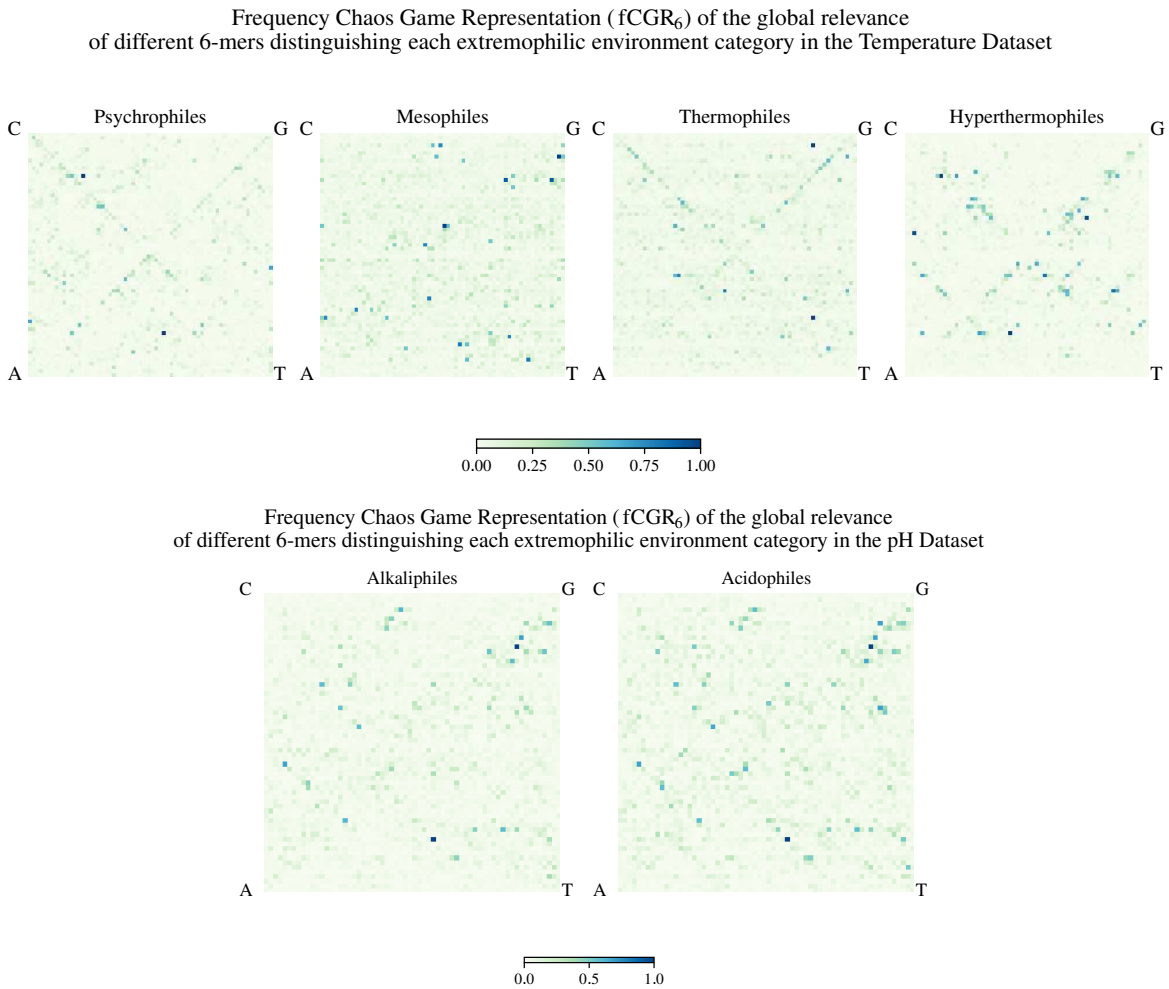


**Figure 1.** Frequency Chaos Game Representation ($fCGR_k$) of the global importance of various 6-mers in the classification of DNA sequences of each environment category from the rest of the dataset. The top panel shows the $fCGR_k$ for the Temperature Dataset, and the bottom panel shows the $fCGR_k$ for the pH Dataset. The $fCGR$s were generated using $k = 6$. The colour and intensity of each pixel represent the relative relevance (importance) of its corresponding 6-mer (dark blue pixels represent the most relevant 6-mers, etc., as described in the colour bar legend).

As the next step, we concretely identify specific $k$-mers that are relevant for each of the One-vs-All Random Forest environment category classifications. Since the majority of biological studies of genome composition of microbial extremophiles focus on codon usage patterns or amino acid compositional biases, we selected the parameter $k = 3$. Here, the 3-mers do not correspond directly to codons, as these 3-mers are not taken only from coding sequences, or only in translation frame. However, a portion of the 3-mers obtained will be codons. Consequently, a correspondence between the relevant $k$-mers found by this method and codon usage patterns in extremophiles found by biological studies would provide further confirmation of their relevance.

For each environment category, this method identified the most relevant subset of 3-mers, that is, the subset of 3-mers that was sufficient to obtain the same classification accuracy as in the setup where all possible 3-mers were used. For example, the frequencies of four distinct 3-mers (AGG, AGA, CTC, and CTA) were sufficient to obtain the same classification accuracy for hyperthermophile category vs. all other temperature categories, as when all possible 3-mers were used for this purpose. Tables 5 and 6 contain a complete list of relevant 3-mers and comparisons with the relevant literature. Although there is not a one-to-one correspondence between these 3-mers and the codons and amino acids reported in the literature, there are some evident similarities, such as the relevance of AGG and AGA for hyperthermophiles and the relevance of CAA for psychrophiles (Supplementary Tables S3 and S4).

To explore whether "relevance" indicates overrepresentation of 3-mers, underrepresentation of 3-mers, or neither, we computed histograms displaying the deviation from the mean, for each environment category in each dataset, see Figure 2 and Figure 3. As seen in Figure 2, GGA which was determined by our method as being relevant in distinguishing both psychrophiles and thermophiles from the rest, is underrepresented in psychrophiles, but overrepresented in thermophiles. This finding is consistent with previously reported codon usages[18]. However, in general, the fact that a subset of 3-mers was determined to be significant for accurate supervised learning classification does not necessarily indicate over- or under-representation of any of its 3-mers, individually.

### Unsupervised clustering of the Temperature and pH Datasets (unlabelled DNA sequences)

The supervised learning computational experiments tested the existence of an environmental signature in a restriction-free scenario, and a restricted scenario where sequences from the same genus as the test sequence were absent from the training set. The latter experiment was performed to decrease the potential influence of the taxonomic component on the results of environment category classification.

It should be noted that the datasets considered in this study are not comprehensive, since the discovery and sequencing of genomes of extremophilic organisms is an ongoing difficult process given the challenging environments in which they are found, which are difficult to reproduce in order to culture and further characterize microbial extremophiles[40]. In particular, the datasets' sparsity and sampling bias do not allow computational experiments in restricted scenarios such as the one above, at taxonomic levels higher than the genus level. More concretely, such restrictions would eliminate many of the labelled sequences from the training set, rendering it insufficient for supervised learning purposes.

To address this challenge, in this section we explore the genomic signatures of the Temperature Dataset and pH Dataset through an unsupervised clustering approach. In unsupervised clustering, no taxonomic or environment category labels for DNA sequences are used during the entire process of learning, and ground-truth labels are used exclusively for evaluation of the quality of clustering (if applicable).

In a first set of tests, we applied *parametric* unsupervised algorithms for the task of clustering both datasets according to different values of the parameter `n_cluster`, which stands for the number of clusters. When compared to the highest taxonomic level (domain), the ACC measure for the clustering assignments computed by each algorithm suggest that for the Temperature Dataset, all of them can partly cluster sequences according to their real taxonomic labels at `n_clusters` = 2, with *i*DeLUCs (68%) outperforming the others by a small margin (see Table 7 for accuracies). For the pH Dataset, all algorithms are completely unsuccessful at separating by domain (see Table 8 for accuracies). For values of the parameter `n_clusters` greater than 2, the accuracy increases for both datasets, but the increase is more significant for the pH Dataset where the ACC increments by$\sim$ 30% which suggests that there is a good separation due to the environmental characteristic within each domain in the pH Dataset. When compared to the environmental labelling, the value of the unsupervised clustering accuracy is lower than the one of the taxonomy counterpart which confirms the strength of the taxonomic component of the signature over

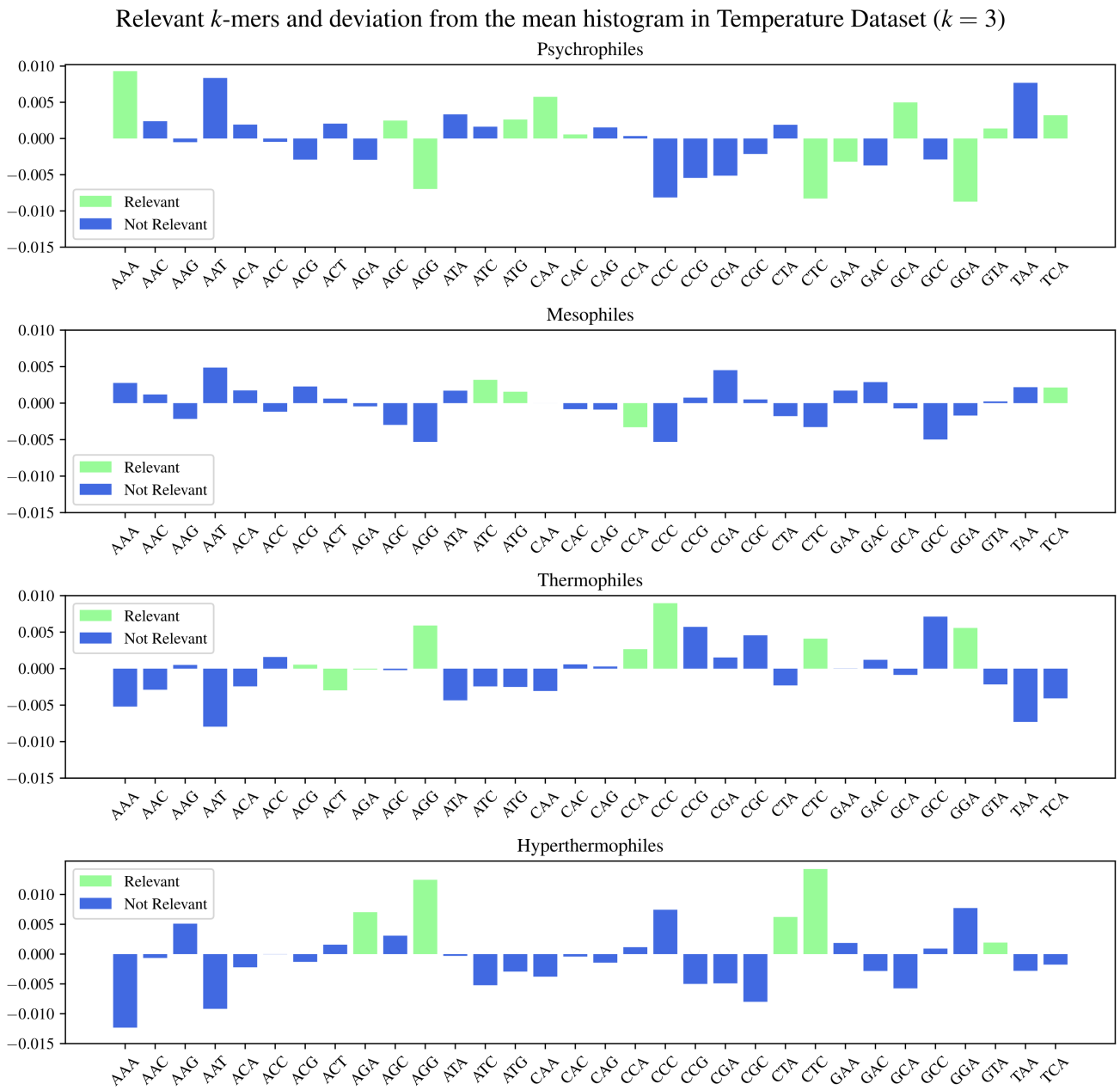


**Figure 2.** Histograms of 3-mers relevant for the separation of DNA sequences of each environment category from the rest of the sequences in the Temperature dataset. The height of each bar represents the difference with respect to the mean value for each 3-mer, and the color represents the relevance (green indicates 3-mers relevant for the classification, and blue indicates 3-mers that are not relevant for the classification).

the environmental component.

In a second set of tests, the six different *non-parametric* algorithms (the number of clusters is not given as a parameter) and the semi-parametric algorithm *i*DeLUCS were employed to cluster both datasets, and the obtained clusters were compared with the genera in the datasets. Figure 4 compares the clusters obtained by each algorithm with the genera, and VAE + Iterative Medoids is the method that performs

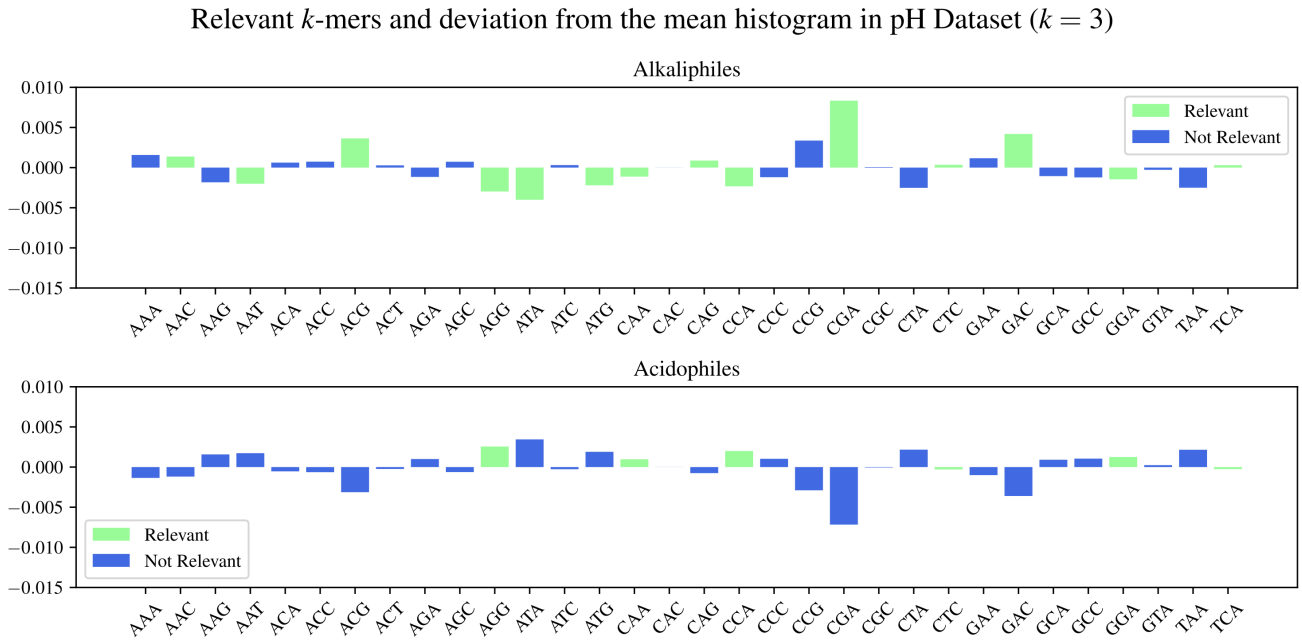


**Figure 3.** Histograms of 3-mers relevant for the separation of DNA sequences of each environment category from the rest of the sequences in the pH Dataset. The height of each bar represents the difference with respect to the mean value for each 3-mer, and the color represents the relevance (green indicates 3-mers relevant for the classification, and blue indicates 3-mers that are not relevant for the classification).

best. Indeed, VAE + Iterative Medoids recovers 61 out of 93 true genera with more than two sequences in the Temperature Dataset, and 18 out of 37 true genera with more than two sequences in the pH Dataset.

Based on this analysis, the five algorithms that were able to recover at least 20% of the total number of genera were VAE+HDBSCAN, CL+HDBSCAN, VAE+IM, UMAP+HDBSCAN, *i*DeLUCS for Temperature, respectively VAE+HDBSCAN, CL+HDBSCAN, VAE+IM, CL+IM, *i*DeLUCS for pH. These algorithms were thus selected as source information for subsequent analysis, since they performed best when compared to true genera groupings.

## Microbial extremophiles from different domains, with similar genomic signatures

Following the selection of the five top performing unsupervised clustering algorithms in the previous section, the clusters discovered by these algorithms were used in conjunction with a majority voting scheme, to determine concrete "candidates," that is, concrete examples of taxonomically different organisms that were clustered together presumably due to the environmental component of their genomic signatures. This computational process identified a list of pairs of hyperthermophilic, alkaliphilic, and acidophilic *candidate sequences*, each belonging to a different taxonomic domain but nevertheless grouped together by the majority of the aforementioned unsupervised clustering algorithms.

Of these candidates, we then proceeded to select sequences for which the unexpected results of the clustering could be independently confirmed by *(i)* supervised machine learning for the prediction of environment category, by *(ii)* supervised machine learning for the prediction of the taxonomic labels, and by *(iii)* biological factors. In these experiments, both thermophiles and hyperthermophiles were treated as part of a single environment category called "high-temperature," so as to enhance the rigor of the confirmation procedure, given the lack of definitive knowledge regarding the precise threshold that separates one environmental category from the other.

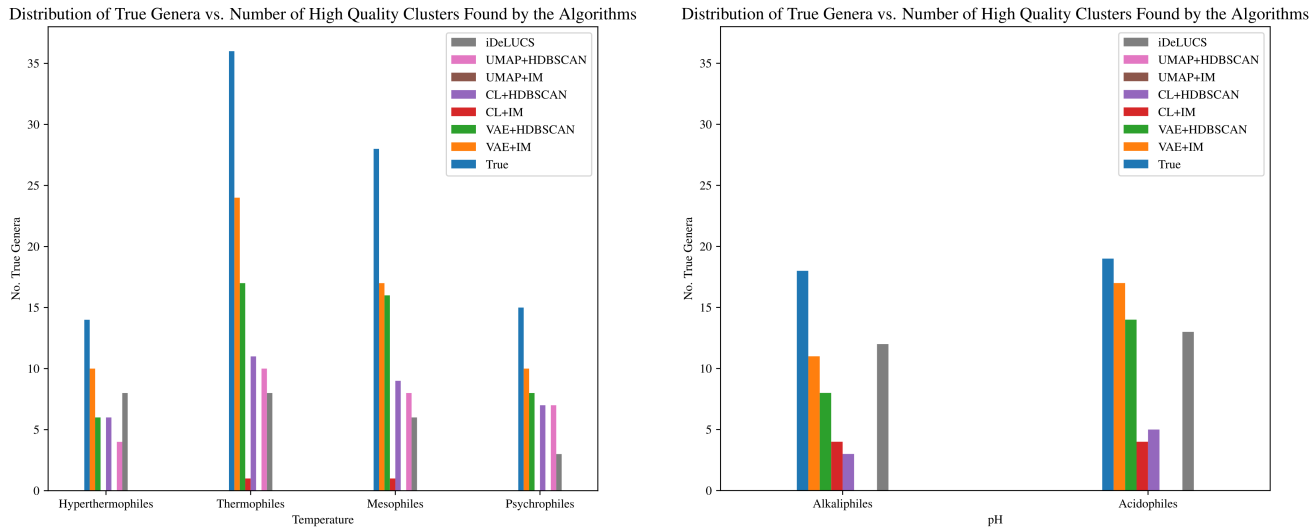


**Figure 4.** Recovery of genera by clustering algorithms for the Temperature Dataset (left) and the pH Dataset (right), where the number of genera with over 50% completeness and under 50% contamination are shown. Only genera with more than one representative are included. The results demonstrate the effectiveness of deep learning-based clustering methodologies in recovering the true genera, as well as the importance of selecting appropriate algorithms for specific datasets

The experimental design aimed to devise challenging scenarios that would clearly demonstrate the presence of the environmental component in the genomic signature of each candidate. To this end, for each candidate test sequence, a challenge training set was created by selecting all DNA sequences of organisms from the opposite domain (i.e., archaea or bacteria), as well as sequences within the same domain but under a different environment category. The classifiers were then trained to perform different tasks.

In experiments *(i)*, a classifier was trained to predict the environment category of a candidate test sequence. For instance, if the test sequence was of a hyperthermophilic bacteria, the training set comprised all archaeal sequences together with the mesophilic and psychrophilic bacterial sequences. The objective was to determine if the hyperthermophilic bacterial test sequence would be assigned the correct label "high-temperature," despite the absence of high-temperature bacterial sequences in the training set. If this were the case, it would indicate that the correct temperature label assignment was due to the similarity of this bacterial sequence to other high-temperature archaeal sequences in the dataset, further suggesting that the environmental component overrides the taxonomical component in the genomic signature of the candidate sequence.

In experiments *(ii)*, a classifier was trained to predict the domain of each candidate test sequence. For example, if the candidate test sequence was of hyperthermophilic archaeal sequence, the training set comprised all bacteria sequences together with the mesophilic and psychrophilic archaeal sequences. The objective was to determine if the hyperthermophilic archaeal sequence would be assigned the incorrect label "Bacteria." If this were indeed the case, it would indicate that the assignment of this archaeal sequence to domain Bacteria was likely based on its similarity to the sequences of the high-temperature bacterial sequences, further suggesting that the environmental component overrides the taxonomic component of the candidate sequence.

All candidate sequences generated by the unsupervised clustering experiment underwent both computational experiments *(i)* and *(ii)*. Of these, the following four sequences were assigned by the majority

of the classifiers (SVM, Random Forest, ANN, MLDSP) to the correct environment category in experiment *(i)*, and to the incorrect domain in experiment *(ii)*: the bacterial sequence *Thermocrinis ruber* – Accession ID: GCA_000512735.1, and the three archaeal sequences, *Pyrococcus furiosus DSM 3638 (formerly Pyrococcus sp000211475 )* – Accession ID: GCA_000007305.1, *Thermococcus litoralis DSM 5473 (formerly Thermococcus litoralis NS-C)* – Accession ID: GCA_000246985.3, and *Pyrococcus chitonophagus (formerly known as Thermococcus chitonophagus)* – Accession ID: GCA_002214605.1. Note that the current release of Genome Taxonomy Database (GTDB release R214 April 28, 2023) defines *Thermococcus litoralis* as a strain type of species *Thermococcus alcaliphilus*. In this study, we refer to it as "*Thermococcus litoralis*," given its classification in the database version used for creation of the dataset.

Indeed, in experiments *(i)*, all environment-trained classifiers correctly predicted these four microbial sequences as belonging to the high-temperature environment category, in spite of the fact that all genomic sequences used to train the classifier to predict temperature conditions were from a different domain. Moreover, in experiments *(ii)* all taxonomy-trained classifiers erroneously predicted the genomic sequences of these microbial extremophiles as belonging to a different domain, likely due to their extremophilic characteristic.

For biological corroboration (iii), a literature search was undertaken in an attempt to correlate the candidate species in the context of phenotypic traits and the characteristics of the isolating environments. It was determined that few phenotypic traits were congruent between candidates, including gram negative cell walls, OGpH falling within the neutrophilic range (pH 5.0 to 9.0) for each candidate, presence of intergenic sequences, and emissions of light hydrocarbons from the nearby environment[41–48]. However, several more phenotypic traits display dissimilarities between organisms, as described in Supplementary Table S5.

Thus, the particular environments each of the species was initially isolated from were analyzed in greater detail, and it was found that two Joint Genome Institute's Genomes OnLine Database-derived (JGI-GOLD) ecosystem classifiers describe the isolating environment of all 4 species, as follows: ID 4027 for *P. furiosus* and *P. chitonophagus*, and ID 3991 for *T. litoralis* and *T. ruber*[42,49,50]. The descriptors for these classifiers are "aquatic marine hydrothermal vent" and "aquatic thermal hot springs" respectively[49]. Although these environments are classified differently by JGI-GOLD, as ID 3991 and ID 4027 respectively, the descriptors accurately describe these environments due to the presence of hydrothermal systems[46,51]. Note that *T. litoralis* (ID 3991) has been recently isolated from the Guaymas Basin, albeit from a geographic site of the Guaymas basin that was different from the isolation site of *P. chitonophagus*[52] (Supplementary Table S6).

For additional insight, for each dataset we inspected the pairwise distance matrix of all DNA fragments, generated by ML-DSP[25] used with $k = 6$. The pairwise distance matrix of the Temperature Dataset revealed that the DNA fragment with the shortest distance from that of *Thermocrinis ruber* (bacterium) belonged to *Thermococcus_A litoralis* (archaea) with a distance value of 0.0327 (the distance ranges between 0 and 1, with 0 the minimum distance, between identical sequences, and 1 the maximum distance).

## Discussion

This paper utilizes six supervised machine learning algorithms to output a highly accurate classification of extremophile prokaryotic genome sequences based on taxonomy, and a moderately high accurate classification of the same sequences with respect to their home environment, through identifying a pervasive environmentally-related genomic signature. These results support the hypothesis of the existence of a unique genome-wide signature in microbial organisms, specific to their adapted extreme environmental conditions such as extreme temperature or pH.

The results of supervised learning confirm that a distinct genomic signature, in the form of a distinct $k$-mer frequency profile, is associated with each environment category. In other words, if the bacteria and archaea sequences in the training set are labelled by environment category, then the algorithms will assign a new sequence to its correct environment category regardless of its domain being bacteria or archaea. Higher classification accuracies obtained using the taxonomic and environment category-labelled datasets were both above those attained for the random-labelled dataset. These findings are consistent with taxonomic and environment category classifications not being attributable to chance, but instead consistent with and in support of the hypothesis of the presence of both a taxonomic and an environmental component in the genomic signatures.

The high classification accuracies obtained in the restriction-free supervised taxonomic classifications suggest that the taxonomic component of the genomic signature is the strongest. The relatively lower classification accuracies obtained in the restricted (non-overlapping genera) scenario of environment category classifications were still higher than those in the control “random label" scenario. This suggests a weaker, but still detectable, environmental (temperature, respectively pH) component in the genomic signature of microbial extremophiles in this dataset.

Note that, while the subsets of 3-mers relevant for the environment category classification that were identified by the MDI algorithm provide insights into the relations between genomic signatures and extreme environmental conditions, caution should be taken when interpreting the results. This is because the experiment prioritized classification accuracy, and the identified subsets of relevant 3-mers may partially reflect correlation between taxonomy and environment. In other words, especially due to the bias and sparsity of both datasets, it is likely that some taxonomic information may also have influenced the process of computational discovery of these subsets of relevant 3-mers. This being said, the overlap between aforementioned subsets of relevant 3-mers and codon usage patterns and amino acid compositional biases found to be associated with extreme environments in the biological literature, still suggests a detectable environmental component to genomic signatures in temperature and pH microbial extremophiles. Future work should also explore the possibility of multiple environmental components influencing the in the genomic signatures of polyextremophiles.

The use of unsupervised learning algorithms for the identification of signatures holds significant value as these algorithms effectively discover clusters of genomic fragments possessing comparable genomic signatures, without influence arising from human annotations. Since the precise definition of the term "genomic signature" entails differentiation from genetically distant organisms, a high-performing clustering algorithm should primarily yield clusters corresponding to the true genera within the dataset. That being said, ascertaining causality for fragments assigned to erroneous clusters proves challenging, given the potential for similar genomic signatures to coincide with taxonomic information at a lower level, as well as the inherent systematic errors in each algorithm. For that reason, in this study the identification of pairs exhibiting an environmental component in their genomic signature, based on the clustering assignments, relies predominantly on the consensus of the high-performing clustering algorithms. Furthermore, only pairs of fragments originating from organisms in different domains were retained. Additional confirmation steps by supervised learning in challenging scenarios were applied to the remaining pairs, and four hyperthermopilic exemplars successfully passed all these stringent tests. It is thus possible that other candidates from the list identified by unsupervised clustering could be viable, such as pairs for which only some of the supervised tests yielded successful results e.g., the acidophilic organisms *Thermoanaerobacterium thermosaccharolyticum* (bacterium) and *Caldisphaera lagunensis* (archaea) in the pH Dataset. Further analysis is needed to confirm such additional pairs, by, e.g., an analysis that utilizes, as a genome representative, multiple randomly selected DNA fragments combined into a single genomic signature.

The dataset in this paper is, to the best of our knowledge, the largest and most comprehensive to date for the study of the genomic signatures of extremophilic organisms. Larger and more balanced extremophilic datasets, combined with extensive literature and database searches, could yield more nuanced bioinformatic analyses in future studies.

## Conclusion

This paper utilizes deep learning algorithms to output both a highly accurate classification of extremophile prokaryotic genome sequences based on taxonomy, and a moderately high accurate classification of the same sequences with respect to their home environment, through identifying a pervasive environmentally-related genomic signature. These results are confirmed by unsupervised clustering analysis of the data, and support the hypothesis of the existence of a unique genome-wide signature in microbial organisms, specific to their adaptations to extreme environmental conditions such as extreme pH or temperature.

## References


1. Rothschild, L. J. & Mancinelli, R. L. Life in extreme environments. *Nature* **409**, 1092–1101, DOI: 10.1038/35059215 (2001).
2. Wang, Q., Cen, Z. & Zhao, J. The survival mechanisms of thermophiles at high temperatures: An angle of omics. *Physiology* **30**, 97–106, DOI: 10.1152/physiol.00066.2013 (2015).
3. Saunders, N. F. W. *et al.* Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.* **13**, 1580—1588, DOI: 10.1101/gr.1180903 (2003).
4. Raymond-Bouchard, I. *et al.* Conserved genomic and amino acid traits of cold adaptation in subzero-growing Arctic permafrost bacteria. *FEMS Microbiol. Ecol.* **94**, DOI: 10.1093/femsec/fiy023 (2018).
5. Turner, P., Mamo, G. & Karlsson, E. N. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb. Cell Factories* **6**, 9, DOI: 10.1186/1475-2859-6-9 (2007).
6. Zeldes, B. M. *et al.* Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. *Front. Microbiol.* **6**, DOI: 10.3389/fmicb.2015.01209 (2015).
7. Horne, W. H. *et al.* Effects of Desiccation and Freezing on Microbial Ionizing Radiation Survivability: Considerations for Mars Sample Return. *Astrobiology* **22**, 1337–1350, DOI: 10.1089/ast.2022.0065 (2022).
8. Ott, E. *et al.* Molecular repertoire of *Deinococcus radiodurans* after 1 year of exposure outside the International Space Station within the Tanpopo mission. *Microbiome* **8**, 150, DOI: 10.1186/s40168-020-00927-5 (2020).
9. Xu, S., Wang, J., Guo, Z., He, Z. & Shi, S. Genomic Convergence in the Adaptation to Extreme Environments. *Plant Commun.* **1**, 100117, DOI: 10.1016/j.xplc.2020.100117 (2020).
10. Friedman, R., Drake, J. W. & Hughes, A. L. Genome-Wide Patterns of Nucleotide Substitution Reveal Stringent Functional Constraints on the Protein Sequences of Thermophiles. *Genetics* **167**, 1507–1512, DOI: 10.1534/genetics.104.026344 (2004).
11. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLOS Comput. Biol.* **3**, 1–11, DOI: 10.1371/journal.pcbi.0030005 (2007).

**12.** Singer, G. A. & Hickey, D. A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39–47, DOI: 10.1016/S0378-1119(03)00660-7 (2003).

**13.** Lobry, J. & Necşulea, A. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* **385**, 128–136, DOI: https://doi.org/10.1016/j.gene.2006.05.033 (2006). Evolutionary Genomics.

**14.** Bao, Q. *et al.* A complete sequence of the *T. tengcongensis* genome. *Genome Res* **12**, 689–700 (2002).

**15.** Saunders, N. F. W. *et al.* Mechanisms of thermal adaptation revealed from the genomes of the Antarctic archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res* **13**, 1580–1588 (2003).

**16.** Barik, S. Evolution of protein structure and stability in global warming. *Int. J. Mol. Sci.* **21**, DOI: 10.3390/ijms21249662 (2020).

**17.** Karlin, S., Mrázek, J. & Campbell, A. M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**, 3899–3913, DOI: 10.1128/jb.179.12.3899-3913.1997 (1997).

**18.** Khan, M. F. & Patra, S. Deciphering the rationale behind specific codon usage pattern in extremophiles. *Sci. Reports* **8**, 15548, DOI: 10.1038/s41598-018-33476-x (2018).

**19.** Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283 – 290, DOI: 10.1016/S0168-9525(00)89076-9 (1995).

**20.** Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391–1399 (1999).

**21.** Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **18**, 186, DOI: 10.1186/s13059-017-1319-7 (2017).

**22.** Karamichalis, R., Kari, L., Konstantinidis, S., Kopecki, S. & Solis-Reyes, S. Additive methods for genomic signatures. *BMC Bioinforma.* **17**, 313, DOI: 10.1186/s12859-016-1157-8 (2016).

**23.** Solis-Reyes, S., Avino, M., Poon, A. & Kari, L. An open-source k-mer based machine learning tool for fast and accurate sub-typing of HIV-1 genomes. *PLoS One* **13**, e0206409 (2018).

**24.** Liang, Q., Bible, P. W., Liu, Y., Zou, B. & Wei, L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics Bioinforma.* **2**, DOI: 10.1093/nargab/lqaa009 (2020).

**25.** Randhawa, G. S., Hill, K. A. & Kari, L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genomics* **20**, DOI: 10.1186/s12864-019-5571-y (2019).

**26.** Millán Arias, P., Alipour, F., Hill, K. A. & Kari, L. DeLUCS: Deep learning for unsupervised clustering of DNA sequences. *PLOS ONE* **17**, e0261531, DOI: 10.1371/journal.pone.0261531 (2022).

**27.** Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560, DOI: 10.1038/s41587-020-00777-4 (2021).

**28.** Fiannaca, A. *et al.* Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinforma.* **19**, 198, DOI: 10.1186/s12859-018-2182-6 (2018).

**29.** Zhang, P., Jiang, Z., Wang, Y. & Li, Y. CLMB: Deep Contrastive Learning for Robust Metagenomic Binning. In *Research in Computational Molecular Biology*, 326–348 (Springer International Publishing, Cham, 2022).

**30.** Randhawa, G. S. *et al.* Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLOS ONE* **15**, e0232391 (2020).

**31.** Merino, N. *et al.* Living at the extremes: Extremophiles and the limits of life in a planetary context. *Front. Microbiol.* **10**, DOI: 10.3389/fmicb.2019.00780 (2019).

**32.** Dexter Dyer, B., Kahn, J., M., LeBlanc & D., M. Classification and regression tree (cart) analyses of genomic signatures reveal sets of tetramers that discriminate temperature optima of archaea and bacteria. *Archaea* **2**, 159–167, DOI: 10.1155/2008/829730 (2008).

**33.** Tsai, M.-H., Liu, Y.-Y., Soo, V.-W. & Chen, C.-C. A New Genome-to-Genome Comparison Approach for Large-Scale Revisiting of Current Microbial Taxonomy. *Microorganisms* **7**, 161, DOI: 10.3390/microorganisms7060161 (2019).

**34.** de la Fuente, R., Díaz-Villanueva, W., Arnau, V. & Moya, A. Genomic signature in evolutionary biology: A review. *Biology* **12**, DOI: 10.3390/biology12020322 (2023).

**35.** Bonnici, V. & Manca, V. Informational laws of genome structures. *Sci. Reports* **6**, 28840, DOI: 10.1038/srep28840 (2016).

**36.** Kislyuk, A., Bhatnagar, S., Dushoff, J. & Weitz, J. S. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinforma.* **10**, 316, DOI: 10.1186/1471-2105-10-316 (2009).

**37.** Vapnik, V. *The Nature of Statistical Learning Theory* (Springer: New York, 2000).

**38.** Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 278–282 vol.1, DOI: 10.1109/ICDAR.1995.598994 (1995).

**39.** McInnes, L., Healy, J. & Astels, S. HDBSCAN: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205, DOI: 10.21105/joss.00205 (2017).

**40.** Zhu, D. *et al.* Recent Development of Extremophilic Bacteria and Their Application in Biorefinery. *Front. Bioeng. Biotechnol.* **8**, 483, DOI: 10.3389/fbioe.2020.00483 (2020).

**41.** Fiala, G. & Stetter, K. O. *Pyrococcus furiosus* sp. nov., represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100 °C. *Arch. Microbiol.* **145**, 56–61, DOI: 10.1007/BF00413027 (1986).

**42.** Huber, R. *et al.* *Thermococcus chitonophagus* sp. nov., a novel, chitin-degrading, hyperthermophilic archaeum from a deep-sea hydrothermal vent environment. *Arch. Microbiol.* **164**, 255–264, DOI: 10.1007/BF02529959 (1995).

**43.** Neuner, A., Jannasch, H. W., Belkin, S. & Stetter, K. O. *Thermococcus litoralis* sp. nov.: A new species of extremely thermophilic marine archaebacteria. *Arch. Microbiol.* **153**, 205–207, DOI: 10.1007/BF00247822 (1990).

**44.** Huber, R. *et al.* *Thermocrinis ruber* gen. nov., sp. nov., a Pink-Filament-Forming Hyperthermophilic Bacterium Isolated from Yellowstone National Park. *Appl. Environ. Microbiol.* **64**, 3576–3583, DOI: 10.1128/AEM.64.10.3576-3583.1998 (1998).

**45.** Capaccioni, B., Tassi, F. & Vaselli, O. Organic and inorganic geochemistry of low temperature gas discharges at the Baia di Levante beach, Vulcano Island, Italy. *J. Volcanol. Geotherm. Res.* **108**, 173–185, DOI: 10.1016/S0377-0273(00)00284-5 (2001).

**46.** Orlando, V., Franco, T., Dario, T., Robert, P. J. & Antonio, C. Submarine and Inland Gas Discharges from the Campi Flegrei (Southern Italy) and the Pozzuoli Bay: Geochemical Clues for a Common Hydrothermal-Magmatic Source. *Procedia Earth Planet. Sci.* **4**, 57–73, DOI: 10.1016/j.proeps.2011.11.007 (2011).

**47.** Bazylinski, D. A., Farrington, J. W. & Jannasch, H. W. Hydrocarbons in surface sediments from a Guaymas Basin hydrothermal vent site. *Org. Geochem.* **12**, 547–558, DOI: 10.1016/0146-6380(88)90146-5 (1988).

**48.** Clifton, C., Walters, C. & Simoneit, B. Hydrothermal petroleums from Yellowstone National Park, Wyoming, U.S.A. *Appl. Geochem.* **5**, 169–191, DOI: 10.1016/0883-2927(90)90047-9 (1990).

**49.** Mukherjee, S. *et al.* Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.* **51**, D957–D963, DOI: 10.1093/nar/gkac974 (2023).

**50.** Belkin, S., Wirsen, C. O. & Jannasch, H. W. A new sulfur-reducing, extremely thermophilic eubacterium from a submarine thermal vent. *Appl. Environ. Microbiol.* **51**, 1180–1185, DOI: 10.1128/aem.51.6.1180-1185.1986 (1986).

**51.** Sherman, L. *et al.* Mercury isotopic composition of hydrothermal systems in the Yellowstone Plateau volcanic field and Guaymas Basin sea-floor rift. *Earth Planet. Sci. Lett.* **279**, 86–96, DOI: 10.1016/j.epsl.2008.12.032 (2009).

**52.** Zhou, Z. *et al.* Genome- and Community-Level Interaction Insights into Carbon Utilization and Element Cycling Functions of *Hydrothermarchaeota* in Hydrothermal Sediment. *mSystems* **5**, e00795–19, DOI: 10.1128/mSystems.00795-19 (2020).

## Acknowledgements

The authors thank Flora Li and Mohammed Essa for their assistance in the process of data collection. This work has been supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grants [RGPIN-2023-03663 to L.K, R3511A12 to K.A.H, and RGPIN-2022-03547 to G.R], and Compute Canada Research Platforms & Portals Grant [616 to K.A.H].

## Author contributions statement

L.K, K.A.H., M.S. designed the computational experiments. J.B. collected and curated the data, P.M.A. and G.R. conducted the computational experiments. J.B, K.A.H, M.S. provided biological interpretation of the findings. P.M.A., L.K., K.A.H., J.B., and G.R. contributed to the writing of the manuscript, and all authors analysed the results and reviewed the manuscript.

## Additional information

**Supplementary Information** accompanying this paper will be made available online, upon publication.

**Competing Interests:** The authors declare no competing interests.

**Table 5.** Relevant 3-mers associated with temperature-adapted extremophilic genomic signatures, that are supported by existing literature. The encoded amino acids are described by the IUPAC abbreviation, followed by the codons responsible for the respective amino acid. The associated 3-mer's matching the described codons are described for each class in their respective column, followed the reference for previously published observations supporting that association. Italicized 3-mers are those that match previous observations for the frequency of amino acids/codons in each class, while non-italicized 3-mers refer to those that do not match previous observations for the frequency of amino acids/codons in each class. Bolded 3-mers are observed with increased frequency in the genomic signature of each class, while non-bolded 3-mers are observed with decreased frequency. Referenced articles are described in Supplementary Table S3. Codons are listed as equivalent DNA, which are encoded into mRNA.

| Amino Acid | Codons | Psychrophiles | Mesophiles | Thermophiles | Hyperthermophiles |
|---|---|---|---|---|---|
| Ala | GCT<br>GCC<br>GCA<br>GCG | ***GCA [1]*** | | | |
| Arg | CGT<br>CGC<br>CGA<br>CGG<br>AGA<br>AGG | *AGG [2]* | | ***AGG*** [10]<br>AGA [10] | ***AGG [14]***<br>***AGA [14]*** |
| Gln | CAA<br>CAG | ***CAA [4]*** | | | |
| Glu | GAA<br>GAG | *GAA [5]* | | | |
| Gly | GGT<br>GGC<br>GGA<br>GGG | GGA [1,3,6] | | ***GGA [12]*** | |
| His | CAT<br>CAC | **CAC [2]** | | | |
| Ile | ATT<br>ATC<br>ATA | | **ATC [7]** | | |
| Leu | TTA<br>TTG<br>CTT<br>CTC<br>CTA<br>CTG | *CTC [4]* | | ***CTC [10]*** | ***CTC [3]***<br>***CTA [3]*** |
| Lys | AAA<br>AAG | **AAA [3]** | | | |
| Met | ATA<br>ATG | ***ATG [3]*** | ***ATG [8]*** | | |
| Pro | CCT<br>CCC<br>CCA<br>CCG | | CCA [9] | ***CCA [11]***<br>***CCC [11]*** | |
| Ser | TCT<br>TCC<br>TCA<br>TCG<br>AGT<br>AGC | ***AGC [6]***<br>***TCA [6]*** | **TCA [7]** | | |
| Thr | ACT<br>ACC<br>ACA<br>ACG | | | *ACT* [4, 13]<br>**ACG** [4, 13] | |
| Val | GTT<br>GTC<br>GTA<br>GTG | ***GTA [1]*** | | | ***GTA [14]*** |

**Table 6.** Relevant 3-mers associated with the genomic signatures of pH-adapted extremophilic genomic signatures, that are supported by previous observations. The encoded amino acids are described by the IUPAC abbreviation, followed by the codons responsible for the respective amino acid. The associated 3-mers matching the described codons are described for each class in their respective column, followed the reference for previously published observations supporting that association. Italicized 3-mers are those that match previous observations for the frequency of amino acids/codons in each class, while non-italicized 3-mers are those that do not match previous observations for the frequency of amino acids/codons in each class. Bolded 3-mers are observed with increased frequency in the genomic signature of each class, while non-bolded 3-mers are observed with decreased frequency. Referenced articles are described in Supplementary Table S4. Codons are described as equivalent DNA, which are encoded into mRNA.

| Amino Acid | Codons | Alkaliphiles | Acidophiles |
|---|---|---|---|
| Arg | CGT<br>CGC<br>CGA<br>CGG<br>AGA<br>AGG | AGG [2]<br>***CGA [1]*** | **AGG [1]** |
| Asn | AAT<br>AAC | ***AAC [1,3]***<br>AAT [1] | |
| Asp | GAT<br>GAC | ***GAC [1]*** | |
| Gln | CAA<br>CAG | ***CAG [1]*** | |
| Glu | GAA<br>GAG | *CAA [1]* | ***CAA [1]*** |
| Gly | GGT<br>GGC<br>GGA<br>GGG | *GGA [1]* | ***GGA [1]*** |
| Ile | ATT<br>ATC<br>ATA | *ATA [1]* | |
| Leu | TTA<br>TTG<br>CTT<br>CTC<br>CTA<br>CTG | ***CTC [1]*** | *CTC [1]* |
| Met | ATA<br>ATG | *ATG [1]* | |
| Pro | CCT<br>CCC<br>CCA<br>CCG | CCA [1] | ***CCA [3]*** |
| Ser | TCT<br>TCC<br>TCA<br>TCG<br>AGT<br>AGC | **TCA [1]** | TCA [1] |
| Thr | ACT<br>ACC<br>ACA<br>ACG | **ACG [1]** | |

**Table 7.** Comparison of different parametric clustering algorithms using unsupervised clustering accuracy computed with taxonomic and environmental labellings on the Temperature Dataset

| Temperature Dataset | | | | | |
|---|---|---|---|---|---|
| **n_clusters** | **Labelling** | **k-means** | **k-medoids** | **GMM** | ***i*DeLUCS** |
| 2 | Tax | 63% | 63% | 63% | 68% |
| | Env | 36% | 36% | 36% | 38% |
| 4 | Tax | 63% | 77% | 68% | 75% |
| | Env | 34% | 40% | 38% | 48% |
| 8 | Tax | 87% | 82% | 77% | 81% |
| | Env | 50% | 49% | 53% | 56% |

**Table 8.** Comparison of different parametric clustering algorithms using unsupervised clustering accuracy computed with taxonomic and environmental labellings on the pH Dataset

| pH Dataset | | | | | |
|---|---|---|---|---|---|
| **n_clusters** | **Labelling** | **k-means** | **k-medoids** | **GMM** | ***i*DeLUCS** |
| 2 | Tax | 52.0% | 52.0% | 51.0% | 56.0% |
| | Env | 50.0% | 50.0% | 51.0% | 50.0% |
| 4 | Tax | 78.0% | 80.0% | 76.0% | 87.0% |
| | Env | 63.0% | 74.0% | 67.0% | 75.0% |