MiGenPro: A linked data workflow for phenotype-genotype prediction of microbial traits using machine learning.

Mike Loomans[1,*], Maria Suarez-Diez[1], Peter J. Schaap[1,2], Edoardo Saccenti[1], and Jasper J. Koehorst[1,2]

[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands
[2]UNLOCK, Wageningen University & Research and Delft University of Technology, The Netherlands

August 21, 2025

Abstract

Availability of microbial genomic data and development of machine learning methods create a unique opportunity to establish associations between genetic information and phenotypes. Here we present a computational workflow for Microbial Genome Prospecting (MiGenPro) that combines phenotype and genomic information. MiGenPro serves as a workflow for the generation of machine learning models that predict microbial traits from genome sequences. Microbial genomes have been consistently annotated and features were stored in a semantic framework. The data was used to train machine learning models and successfully predicted microbial traits such as motility, Gram stain, optimal temperature range, and sporulation capabilities. To ensure robustness, five-fold cross-validation was implemented and demonstrated consistent model performance across iterations and without overfitting. Effectiveness was further validated through comparison with existing models, showing comparable accuracy, with modest variations attributed to differences in datasets rather than methodology. Classification can be further explored using feature importance characterisation to identify biologically relevant genomic features. MiGenPro provides an interoperable workflow to build models and predict phenotypes from microbes based on their genome.

1 Introduction

Advances in experimental and computational techniques for DNA extraction, sequencing, assembly, and structural and functional annotation have made microbial genotypes increasingly available. As a result of the increasing throughput of these approaches, millions of genome sequences are stored in databases using highly structured formats that collect both the genetic sequences and the result of annotation tools that inform on genomic features. The potential to generate and store sequence data dwarfs the availability of phenotype information. The characterization of the microbial phenotypes linked to the generated sequences still lacks behind as dedicated experiments are often required to characterize each aspect such as optimal temperature range, substrate utilization potential or potential for sporulation among others.

Yet, exploitation of microbes hinges on the microbial phenotype. Consider the global biotechnology sector that is expanding in the coming years[1]. The expansion of this industry is coupled with developments in strain optimization and the discovery of strains with industrial potential[2, 3]. Selection of industrial strains often considers increased performance of the cell factory and relevant traits include alongside desired metabolic pathways, thermotolerance, osmotolerance, and general robustness[4]. For example, the production of polyhydroxyalkanoates in *Schlegelella thermodepolymerans* benefits from high temperatures as it decreases the risk of contamination[5]. Therefore, exploring these phenotypes can accelerate biotechnological progress. Similarly, bioremediation strategies could profit from accurate prediction of phenotypes to aid the development of microbial communities that promote soil health.

These applications demonstrate the potential of methods to accurately predict microbial phenotypes from genome sequences. The use of public genome repositories and annotation pipelines to predict prokaryotic phenotypes has been widely practiced[6–10]. However, these approaches are challenged by the lack of consistently annotated genomes in a format that supports automated querying and recovery of strain associated phenotypes.

Additionally, these methods also struggle with species-specific genetic elements as these limit the resolution of the prediction tools.

Recently the work of Koblitz *et al.* 2025 illustrates the potential of machine learning approaches when applied to the task of predicting phenotype information from genome data[9]. It also highlights the need of uniform data annotation and preserving the association between genotype and phenotype data in the training datasets. Here we illustrate how to combine semantic technologies for data storage and linking and machine learning models to predict phenotypes from genome data. Currently relevant and publicly available databases in the life sciences provide the means to retrieve their data using machine readable data formats (such as JSON) . Here, we provide a workflow to exploit this technology that can be adapted and extended to any relevant phenotype for which training data are available. The approach can also be used to test the performance of alternative prediction approaches. The Microbial Genome Prospecting (MiGenPro) phenotype prediction workflow encapsulates the steps required to go from a collection of genome sequences and phenotype information to machine learning models.

These steps include data retrieval using REST API, the use of the Genome biology ontology language (GBOL) together with SPARQL and RDF for data storage and standardized workflows for uniform genome data annotation[11, 12]. We have showcased the effectiveness of the MiGenPro workflow with the prediction of physiological characteristic using various machine learning models trained on data automatically retrieved from the BacDive repository[13]. The predictions are complimented with the addition of methods for feature selection that extract biological knowledge from the developed models.

2 Materials and methods

MiGenPro workflow for phenotype prediction.

Modules in the MiGenPro workflow contribute to two main tasks, data formatting and phenotype prediction as illustrated in Figure 1. The modular design ensures that any step can be run independently provided that the input data is available.

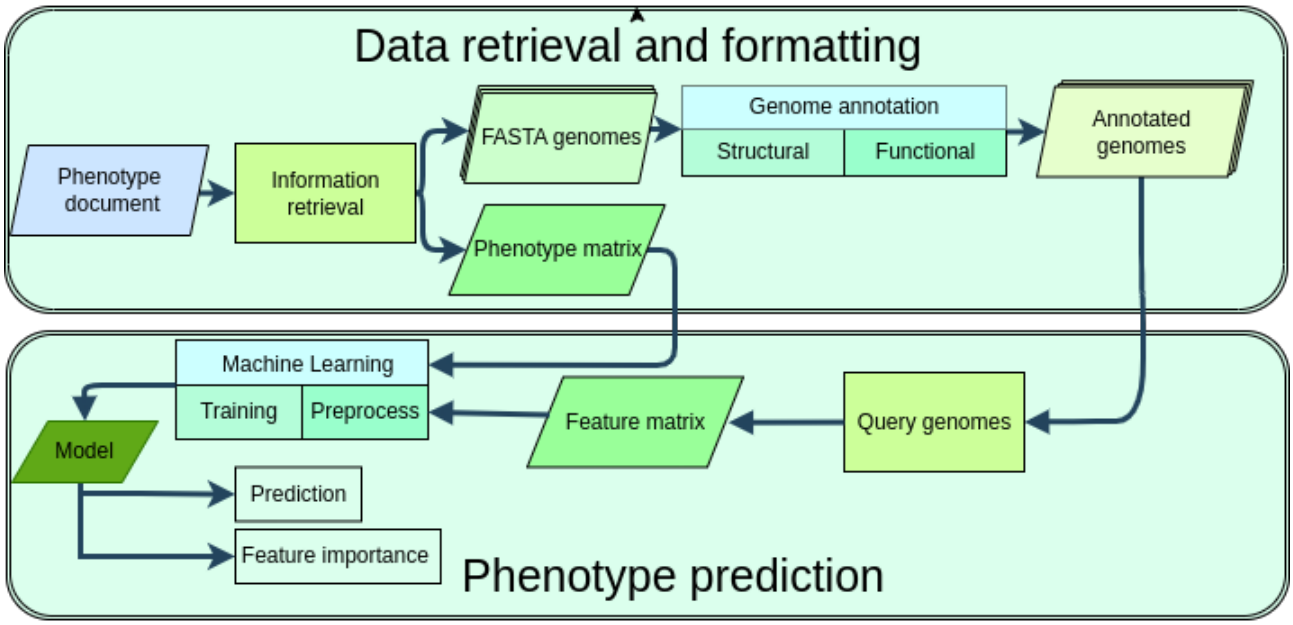


Figure 1: The workflow starts with by loading a phenotype graph, which is queried using SPARQL to generate a list of genome IDs and a matrix containing the phenotypes. These IDs are used to obtain corresponding FASTA files and the phenotype matrix links each genome ID to its respective phenotype. The FASTA genomes are annotated with SAPP (Semantic Annotation Platform with Provenance) that runs InterProScan and prodigal and outputs them in GBOL format. The annotated genomes are queried with SPARQL to extract genomic features. The phenotype and feature matrices are loaded into Python where a series of classes prepare the data and train the models. The resulting model can predict phenotypes, and biological relevance is determined by assessing feature importance.

2.1 Data formatting

Retrieval of phenotype graph Here we use the term phenotype graph to refer to a document describing the data available in the database(s) from which data are to be retrieved. This document should contain information on genome identifiers and respective phenotypes. Here we retrieved the phenotype graph containing Genome identifiers alongside selected phenotypes namely, Gram stain, motility, oxygen requirement, optimal growth temperature, and spore formation from the BacDive database. The BacDive database REST API was used to retrieve this information in JSON format. The JSON results were transformed into a linked data structure, JSON-LD, using SAPP (v2.0)[14].

Information retrieval: Query phenotypes and genomes

The linked data structure was converted into a HDT (Header Dictionary Triple) file using SAPP (v2.0) and subsequently queried using the SPARQL integration within SAPP using the queries in supplementary file 1[12]. Phenotype data were parsed into a columnar format using Python (v3.12.5) with the pandas library (v2.1.1), creating a matrix. This matrix contained for each phenotype a set of rows denoting the various genomes with NCBI genome identifiers and columns for the features. The matrix was filtered and only physiological characteristics for which 500 annotated genomes were available were considered. For species for which multiple genomes were available , only 10 genome sequences were considered (random selection) to avoid bias related to oversampling aided by the use of the ete4 (v.3.1.3)[15]. The genome data in FASTA format was retrieved using the NCBI Datasets conda package[16].

Genome annotation Genome data were then structurally and functionally annotated using a standardised workflow written in the Common Workflow Language[17, 18]. This workflow contains the following SAPP modules: Genome conversion to RDF, Prodigal (v2.6.3) for gene prediction and InterProScan (v5.53-87.0) for functional annotation [14, 19, 20]. The annotated results are represented in RDF following the GBOL (Genome Biology Ontology) structure using the HDT format. Each genome is encapsulated in a single HDT file.

2.2 Phenotype prediction

Query genomes and feature matrix construction Protein domain information was extracted from each genome separately using SPARQL queries (supplementary file 2). The resulting protein domains were merged into a single matrix containing all protein domain frequencies across the genomes considered for the selected phenotype.

For each phenotype, the Scikit-learn VarianceThreshold function was used to calculate the variance of protein domains. By setting a variance threshold of 1 redundant protein domains were filtered out, addressing the multi-dimensional nature of the dataset. These filtered protein domains were then used as features for machine learning models.

Machine learning data preprocessing The phenotype and feature matrices were used as starting point for training machine learning models. The dataset was split into a 70% train and 30% test with the scikit-learn train_test_split function[21]. The train_test_split function randomly splits the data into two subsets, ensuring that the distribution of the target variable is preserved in both the training and test sets. Training data was oversampled using SMOTEN (Synthetic Minority Over-sampling Technique for Nominal) oversampling from the imbalanced-learn library (v0.12.0) [22]. This technique was used for categorical oversampling with a maximum oversampling factor of two to balance difference in frequency amongst phenotypes.

2.3 Machine Learning training

Machine learning methods based on decision trees were selected due to their interpretability. The algorithms, random forest, decision trees, and gradient boosting were imported from the scikit-learn (v1.5.2) library in Python [21]. Parameter optimisation was performed per phenotype and classifier using successive halving grid search method with the grid space defined in supplementary file 3. A successive halving grid search iteratively eliminates poor parameter combinations through two successive resource allocation stages.

Decision tree classifiers used a maximum depth of 15 for spore formation and oxygen requirements, 10 for Gram stain prediction, optimal temperature range, and optimal growth pH, and 5 for motility prediction.

Random forest classifiers were trained using a maximum depth of 30 for all phenotypes except motility and spore formation which were trained with a maximum depth of 10 and 20 respectively[23]. 1000 trees were used to train random forest classifiers that predicted oxygen and Gram stains with the other models being trained

Table 1: Number of genomes used per phenotype for training and testing models after filtering for a minimum of 500 genomes per phenotype and application of the 10 genomes per species cut-off

Phenotype	**Count**
gram	10939
negative	8320
positive	2619
motility	4939
no	2621
yes	2318
oxygen	6258
aerobe	3347
anaerobe	1212
facultative anaerobe	1154
microaerophile	545
spore	4552
no	3314
yes	1238
temperature	11348
mesophilic	6209
psychrophilic	4565
thermophilic	574

with 500 trees. A minimum sample split of 5 was used for the prediction of spore formation, motility, and Gram stain prediction.

Gradient boosting classifiers utilised a maximum tree depth of 3 for all phenotypes, with minimum sample splits of 10 for oxygen requirements and spore formation, 5 for the prediction of Gram stains, and 2 for motility, optimal temperature range, and optimal pH. The number of estimators was 1000 for spore formation and 500 for motility, oxygen requirements, temperature, pH, and Gram stain prediction[24].

Model assessment and feature importance Models were evaluated on accuracy, Matthew coefficient, F1 score, and AUC on a held-out test set compromising 30% of data. Performance metrics were visualised using matplotlib (v3.8.2). The models feature importance was inferred using average Gini values from fivefold cross validation with the scikit learn library. The feature importance results from each model were then ordered by rank sum product. Where for each iteration the gini scores were ordered and then ranked, these ranks were subsequently summed; the importance of each feature over multiple iterations was then inferred using these products of ranked sums.

3 Results and discussion

The MiGenPro workflow offers a modular approach to building machine learning models that predict microbial phenotypes from genomic data. By integrating various linked data tools and techniques, the workflow streamlines the process from data retrieval and integration all the way to model training and evaluation. A key advantage of leveraging linked data is the ability to seamlessly query public SPARQL endpoints to access experimentally validated phenotypic metadata from resources like BacDive[13]. This interoperability significantly reduces manual preprocessing and fosters the reuse of high-quality, curated data across studies.

Performance of trained models Model performance for decision tree, random forest and gradient boosting classifiers are shown in Table 2. The comparison to different studies shows that models reach similar bottlenecks in terms of performance due to the datasets being used. It is to be noted that there are big differences in the training and testing datasets used for each study and the results were directly taken from the respective studies.

An overview of the model performance scores can be found in Table 3. Performance indicators for all methods and all considered physiological characteristics are high, indicating the suitability of the approach. The standard deviations observed from five random train-test splits remain low, indicating stable outcomes. Among the three model types tested, decision tree classifiers consistently underperform compared to the more sophisticated classifiers such as random forest and gradient boosting. This is most clearly seen in the prediction of oxygen requirements, where all methods achieve the worst performance.

Source	Accuracy
gram	
Felderbauer (2015)	0.99 ± 0.010
Koblitz (2025)	0.96 ± 0.000
Ligner (2010)	0.97 ± 0.000
This study decision tree (2025)	0.99 ± 0.000
This study gradient boosting (2025)	1.00 ± 0.000
This study random forest (2025)	1.00 ± 0.000
Wiemann (2016)	0.99 ± 0.010
motility	
Felderbauer (2015)	0.87 ± 0.060
Koblitz (2025)	0.98 ± 0.000
Ligner (2010)	0.93 ± 0.000
This study decision tree (2025)	0.90 ± 0.002
This study gradient boosting (2025)	0.95 ± 0.001
This study random forest (2025)	0.95 ± 0.000
Wiemann (2016)	0.95 ± 0.020
oxygen resistance	
Felderbauer (2015)	0.90 ± 0.050
Koblitz (2025)	0.94 ± 0.000
Ligner (2010)	0.97 ± 0.000
This study decision tree (2025)	0.92 ± 0.001
This study gradient boosting (2025)	0.97 ± 0.001
This study random forest (2025)	0.94 ± 0.000
Wiemann (2016)	0.97 ± 0.010
spore	
Felderbauer (2015)	0.93 ± 0.060
Koblitz (2025)	0.96 ± 0.000
Ligner (2010)	0.95 ± 0.000
This study decision tree (2025)	0.79 ± 0.005
This study gradient boosting (2025)	0.95 ± 0.001
This study random forest (2025)	0.94 ± 0.001
Wiemann (2016)	0.95 ± 0.050
temperature	
Felderbauer (2015)	-
Koblitz (2025)	0.99 ± 0.000
Ligner (2010)	-
This study decision tree (2025)	0.97 ± 0.000
This study gradient boosting (2025)	0.99 ± 0.000
This study random forest (2025)	0.99 ± 0.000
Wiemann (2016)	-

Table 2: In-study reported accuracy metrics for the Gram stain, oxygen requirement, motility and spore formation phenotypes. Accuracy measurements were taken from the original studies and were not assessed using the same dataset. Temperature prediction was absent for studies other than Koblitz *et al.* 2025.

Feature selection and biological interpretation Although the performance difference between random forests and gradient-boosting machines is dependent on the variables used[25]. Tree-based methods can be analysed using Gini importance to evaluate feature importance. The Gini index evaluates how much each feature contributes to the performance of the classifier and can be used to select key features associated to the classification.

As an example the 10 most relevant protein domains for predicting the potential to produce spores are provided in Table 4).

The rank sum product method aggregates feature importance across five iterations by assigning inverse ranks (with 5 = most important) to each feature and computing their cumulative sum. Lower total scores indicate higher overall importance, as features consistently ranked as critical across iterations accumulate smaller sums.

In the analysis of protein domains for spore formation prediction using a gradient boosting classifier, several key domains emerged. The biological function and annotation was retrieved from Uniprot using the Uniprot

Table 3: Classifier performance for each phenotype using, accuracy, are under the receiver operating characteristic (AUC), F1 score, and Matthews correlation coefficient (MCC) metrics to assess model performance. Mean and standard deviation values are derived from a fivefold cross validation of all models DT (Decision Tree), GB (Gradient Boosting), RF (Random Forest).

Classifier	**F1 score**	**Accuracy**	**Matthews**	**AUC**
Gram				
DT	0.99±0.00	0.99±0.00	0.98±0.00	0.98±0.00
GB	1.00±0.00	1.00±0.00	0.99±0.00	1.00±0.00
RF	1.00±0.00	1.00±0.00	0.99±0.00	1.00±0.00
Motility				
DT	0.90±0.00	0.90±0.00	0.80±0.00	0.91±0.00
GB	0.95±0.00	0.95±0.00	0.91±0.00	0.98±0.00
RF	0.95±0.00	0.95±0.00	0.91±0.00	0.99±0.00
Oxygen requirement				
DT	0.79±0.00	0.79±0.00	0.66±0.01	0.87±0.01
GB	0.95±0.00	0.95±0.00	0.92±0.00	0.98±0.00
RF	0.94±0.00	0.94±0.00	0.90±0.00	0.99±0.00
Spore formation				
DT	0.97±0.00	0.97±0.00	0.92±0.00	0.95±0.00
GB	0.99±0.00	0.99±0.00	0.98±0.00	1.00±0.00
RF	0.99±0.00	0.99±0.00	0.97±0.00	1.00±0.00
Optimal temperature range				
DT	0.92±0.00	0.92±0.00	0.86±0.00	0.90±0.00
GB	0.97±0.00	0.97±0.00	0.95±0.00	0.97±0.00
RF	0.94±0.00	0.94±0.00	0.89±0.00	0.99±0.00

domain	**rank sum**	**description**	**reference**
PF14098	8	Acid-soluble spore protein 1	Cabrera-Hernandez & Setlow 2000
PF03323	15	Associated with GerA spore germination protein	Paidhungat *et al.* 2000
PF10628	37	Required for outer spore coat protein E	Costa *et al.* 2007
PF00269	47	Involved in spore resistance to UV	Setlow 1992
PF03029	64	ATP-binding domain found in multiple domains	PF03029 Pfam entry[30]

Table 4: Gini feature importance for spore formation prediction with gradient boosting classifier visualised in the product rank sum over five iterations.

API[31].

PF14098 was identified as the most critical domain. This domain is hypothesized to encode a small, acid-soluble spore protein 1, which plays a role in spore resilience[26]. PF03323 emerged as the second most critical domain for spore formation prediction. This domain is associated with the GerA spore germination protein in *Bacillus subtilis*[27]. PF10628 is the third most important domain and is classified as a domain required for an outer spore coat protein E in *Bacillus subtilis*[28]. The analysis of feature importance of the models confirms its biological significance, as illustrated by this example.

4 Discussion and conclusion

MiGenPro enables data curation and machine learning approaches for microbial traits Previous work has illustrated the potential of machine learning models on high quality curated datasets to predict bacterial phenotypes[6–9]. The application of this workflow requires genomes combined with a database or matrix with phenotype of these genomes.

Here we provide a workflow to further extend these analysis to less high-quality datasets and any other case. Consistent annotation of genome data prior to comparative genomic studies has been flagged as essential to prevent bias and artifacts due to differences in the tools used for genome annotation [32]. MiGenPro enables automation of these tasks, thereby facilitating comparative genomics. Here we have illustrated the use of MiGenPro using data from BacDive, however it can be extended to any other data source. The starting point is a phenotype graph file that associates physiological measurements to standardized genome identifiers such as the

ones in NCBI. Once this starting point is available, the workflow provides the tools to retrieve and canonicalize the data and enable development of machine learning models. By standardising genomic data as linked data RDF triples, a high level of interoperability is obtained. This interoperability allows for the development of phenotype prediction learning models. The quality of genomes was not compared to model performance; however, it stands to reason that only high-quality genomes would yield accurate results, as machine learning methods could use the lack of a protein domain to infer a phenotype as well[33].

The effectiveness of this workflow has been demonstrated through successful prediction of microbial traits, motility, Gram stain, optimal temperature range, optimal growth pH, and spore formation. These predictions, derived from genomic features, illustrate the workflow's ability to handle complex biological data and provide insights.

Model performance is not artificially inflated The results from five-fold cross-validation show no significant deviations in feature importances across iterations, with importance remaining consistent throughout the process seen in Table 4. This suggests that oversampling, an artificial inflation of model performance by overfitting data, is not occurring. A similar trend can be seen in regards to the observed standard deviation in figure 3.

The consistent results across test sets further support the robustness of the model and the reliability of the predictions made using the current datasets. The biological functions associated with the highest-ranking features support are biologically relevant see 4. Therefore, the model demonstrates resilience against overfitting, ensuring that the predictions are generalisable to new, unseen data.

The Gini feature importance was used to assess the rank of these features, Gini is only applicable to tree-based methods. The workflow is modular and the user could use other methods for the evaluation of feature importance, such as Shapley additive explanation values[34].

Model performance matches previous phenotype prediction methods Cross-study comparison of machine learning approaches (Decision Trees, Random Forests, Gradient Boosting) in Figure 2 reveals marginal performance variations (Δ accuracy $< 5\%$), suggesting algorithm choice is less critical than data quality for phenotype prediction. From Figure 3, it can be inferred that the accuracy on the datasets used does not vary by more than approximately 5%.

The primary distinction between these studies lies in the datasets employed, thereby underscoring the need for a workflow capable of easily training models on new data. Comparing the methods themselves is challenging, as the accuracy metric is limited when evaluating model performance on imbalanced test datasets. The accuracy metrics reported were directly sourced from the respective studies. For a comprehensive comparison of models developed throughout last 15 years, the accuracy metric was selected as this was one of the few performance metrics available[6–9]. A curiosity here is the case of the motility dataset that classifies all types of motility, gliding, swimming, twitching under the same header as being motile[9, 13]. These subgroups of motility are not evenly distributed and have varying levels of complexity resulting in lower than expected performance metrics.

5 Data availability

Project code repository: https://gitlab.com/pig-paradigm/migenpro/
Operating system: Platform independent
License: MIT

6 Funding

M.L. and J.J.K acknowledge the Novo Nordisk Foundation for their financial contribution to the PIG-PARADIGM project, funded by (Grant no. NNFSA210073688).
P.J.S. and J.J.K acknowledge the Dutch Research Council (NWO) and Wageningen University & Research for their financial contribution to the Unlock initiative (NWO: 184.035.007).

7 Supplementary

Supplementary file 1

`https://gitlab.com/pig-paradigm/migenpro/-/tree/publication/migenpro/resources/sparql_phenotype`

Supplementary file 2

https://gitlab.com/pig-paradigm/migenpro/-/blob/47212f85218e50d09404e6f7cac0c5dae67e91d3/migenpro/resources/sparql_genome/DomainCopyNumber.sparql

Supplementary file 3

https://gitlab.com/pig-paradigm/migenpro/-/blob/47212f85218e50d09404e6f7cac0c5dae67e91d3/tests/resources/param_grid.json

References

1. Levy, V. *2024 Global Life Sciences Sector Outlook* tech. rep. (Deloitte, May 2024).
2. Chaudhary, R. *et al.* Microbial Cell Factories: Biodiversity, Pathway Construction, Robustness, and Industrial Applicability. en. *Microbiology Research* **15.** Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, 247–272. ISSN: 2036-7481. https://www.mdpi.com/2036-7481/15/1/18 (2024) (Mar. 2024).
3. Abbate, E. *et al.* Optimizing the strain engineering process for industrial-scale production of bio-based molecules. *Journal of Industrial Microbiology & Biotechnology* **50,** kuad025. ISSN: 1367-5435. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10548853/ (2024) (Sept. 2023).
4. Olsson, L., Rugbjerg, P., Torello Pianale, L. & Trivellin, C. Robustness: linking strain design to viable bioprocesses. *Trends in Biotechnology* **40,** 918–931. ISSN: 0167-7799. https://www.sciencedirect.com/science/article/pii/S016777992200004X (2024) (Aug. 2022).
5. Kourilova, X. *et al.* Production of polyhydroxyalkanoates (PHA) by a thermophilic strain of *Schlegelella thermodepolymerans* from xylose rich substrates. *Bioresource Technology* **315,** 123885. ISSN: 0960-8524. https://www.sciencedirect.com/science/article/pii/S0960852420311573 (2025) (Nov. 2020).
6. Lingner, T., Mühlhausen, S., Gabaldón, T., Notredame, C. & Meinicke, P. Predicting phenotypic traits of prokaryotes from protein domain frequencies. *BMC Bioinformatics* **11,** 481. ISSN: 1471-2105. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2955703/ (2024) (Sept. 2010).
7. Weimann, A. *et al.* From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. eng. *mSystems* **1,** e00101–16. ISSN: 2379-5077 (2016).
8. Feldbauer, R., Schulz, F., Horn, M. & Rattei, T. Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics* **16,** S1. ISSN: 1471-2105. https://doi.org/10.1186/1471-2105-16-S14-S1 (2023) (Oct. 2015).
9. Koblitz, J., Reimer, L. C., Pukall, R. & Overmann, J. Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. en. *Communications Biology* **8.** Publisher: Nature Publishing Group, 897. ISSN: 2399-3642. https://www.nature.com/articles/s42003-025-08313-3 (2025) (June 2025).
10. Sauer, D. B. & Wang, D.-N. *Using machine learning to predict quantitative phenotypes from protein and nucleic acid sequences* en. Pages: 677328 Section: New Results. Oct. 2022. https://www.biorxiv.org/content/10.1101/677328v3 (2024).
11. Dam, J. C. J. v., Koehorst, J. J., Vik, J. O., Schaap, P. J. & Suarez-Diez, M. *Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining* en. Pages: 184747 Section: New Results. Sept. 2017. https://www.biorxiv.org/content/10.1101/184747v1 (2024).
12. *SPARQL 1.1 Overview* https://www.w3.org/TR/sparql11-overview/ (2024).
13. Schober, I. *et al.* BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research* **53,** D748–D756. ISSN: 1362-4962. https://doi.org/10.1093/nar/gkae959 (2025) (Jan. 2025).
14. Koehorst, J. J. *et al.* SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. eng. *Bioinformatics (Oxford, England)* **34,** 1401–1403. ISSN: 1367-4811 (Apr. 2018).
15. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* **33,** 1635–1638. ISSN: 0737-4038. https://doi.org/10.1093/molbev/msw046 (2025) (June 2016).

16. O'Leary, N. A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. en. *Scientific Data* **11.** Publisher: Nature Publishing Group, 732. ISSN: 2052-4463. https://www.nature.com/articles/s41597-024-03571-y (2025) (July 2024).

17. Crusoe, M. R. *et al.* Methods included: standardizing computational reuse and portability with the Common Workflow Language. *Commun. ACM* **65,** 54–63. ISSN: 0001-0782. https://dl.acm.org/doi/10.1145/3486897 (2025) (May 2022).

18. *Microbial (meta-) genome annotation* https://workflowhub.eu/workflows/1170/ (2025).

19. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119. ISSN: 1471-2105. https://doi.org/10.1186/1471-2105-11-119 (2023) (Mar. 2010).

20. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. eng. *Nucleic acids research* **49,** D344–D354. ISSN: 1362-4962. https://europepmc.org/articles/PMC7778928 (2024) (Jan. 2021).

21. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. ISSN: 1533-7928. http://jmlr.org/papers/v12/pedregosa11a.html (2024) (2011).

22. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16,** 321–357. ISSN: 1076-9757 (June 2002).

23. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32. ISSN: 0885-6125. https://doi.org/10.1023/A:1010933404324 (2025) (Oct. 2001).

24. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29.** Publisher: Institute of Mathematical Statistics, 1189–1232. ISSN: 0090-5364, 2168-8966. https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full (2025) (Oct. 2001).

25. Nawar, S. & Mouazen, A. M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. en. *Sensors* **17.** Publisher: Multidisciplinary Digital Publishing Institute, 2428. ISSN: 1424-8220. https://www.mdpi.com/1424-8220/17/10/2428 (2025) (Oct. 2017).

26. Cabrera-Hernandez, A. & Setlow, P. Analysis of the regulation and function of five genes encoding small, acid-soluble spore proteins of *Bacillus subtilis*. *Gene* **248,** 169–181. ISSN: 0378-1119. https://www.sciencedirect.com/science/article/pii/S0378111900001256 (2025) (May 2000).

27. Paidhungat, M., Setlow, B., Driks, A. & Setlow, P. Characterization of spores of Bacillus subtilis which lack dipicolinic acid. eng. *Journal of Bacteriology* **182,** 5505–5512. ISSN: 0021-9193 (Oct. 2000).

28. Costa, T. *et al.* The timing of cotE expression affects Bacillus subtilis spore coat morphology but not lysozyme resistance. eng. *Journal of bacteriology* **189,** 2401–2410. ISSN: 1098-5530. https://europepmc.org/articles/PMC1899386 (2025) (Mar. 2007).

29. Setlow, P. I will survive: protecting and repairing spore DNA. *Journal of Bacteriology* **174.** Publisher: American Society for Microbiology, 2737–2741. https://journals.asm.org/doi/10.1128/jb.174.9.2737-2741.1992 (2025) (May 1992).

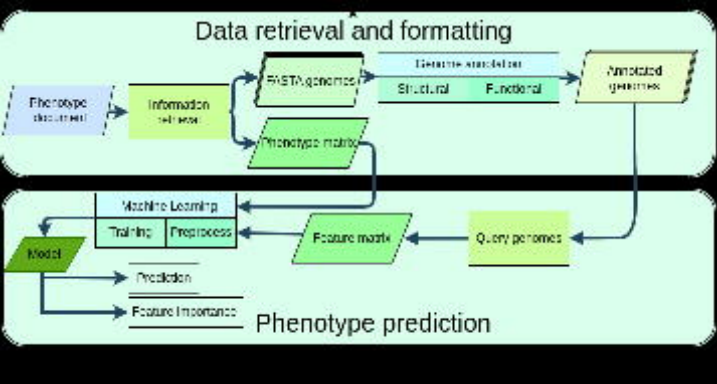
30. *Conserved hypothetical ATP binding protein (PF03029) - Pfam entry - InterPro* https://www.ebi.ac.uk/interpro/entry/pfam/PF03029/ (2025).

31. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158–D169. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkw1099 (2023) (Jan. 2017).

32. Poncheewin, W., van Diepeningen, A. D., van der Lee, T. A. J., Suarez-Diez, M. & Schaap, P. J. Classification of the plant-associated lifestyle of Pseudomonas strains using genome properties and machine learning. *Scientific Reports* **12,** 10857. ISSN: 2045-2322. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9237127/ (2023) (June 2022).

33. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. eng. *Nature Biotechnology* **35,** 725–731. ISSN: 1546-1696 (Aug. 2017).

34. Molnar, C. *9.5 Shapley Values — Interpretable Machine Learning* https://christophm.github.io/interpretable-ml-book/shapley.html#shapley (2024) ().


Data retrieval and formatting
Phenotype
Information
FASTA genomes
Phenotype matrix
Genome annotation
Structural
Functional
Annotated
Machine Learning
Training
Preprocess
Model
Prediction
Feature importance
Feature matrix
Query genomes
Phenotype prediction