Journal Title Here, YEAR, , 1–8
DOI added during production
Published: Date added during production
Paper

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 & Research and Delft University of Technology, The Netherlands
*Corresponding author. mike.loomans@wur.nl

Abstract

The availability of microbial genomic data and the development of machine learning methods have created a unique opportunity to establish associations between genetic information and phenotypes. Here, we introduce a computational workflow for Microbial Genome Prospecting (MiGenPro) that combines phenotypic and genomic information. MiGenPro serves as a workflow for the training of machine learning models that predict microbial traits from genomes that have been annotated. Microbial genomes have been consistently annotated and features were stored in a semantic framework that is easy to query using SPARQL. 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, a hyper parameter halving grid search was used to determine optimal parameter settings followed by a five-fold cross-validation which 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 easy to use interoperable workflow to build and validate models to predict phenotypes from microbes based on their annotated genome.

Keywords Bioinformatics, Machine Learning, Genomics (microbiology), Visualization, Python library, Linked data, Genotype-phenotype association

Abbreviations AUC: Area Under the Curve; CWL: Common Workflow Language; DT: Decision Tree; FAIR: Findable, Accessible, Interoperable, Reusable; GB: Gradient Boosting; GBOL: Genome Biology Ontology Language; HDT: Header Dictionary Triples; JSON: JavaScript Object Notation; JSON-LD: JSON for Linking Data; MCC: Matthews correlation coefficient; ML: Machine Learning; MiGenPro: Microbial Genome Prospecting; NCBI: National Center for Biotechnology Information; REST API: Representational State Transfer Application Programming Interface; RDF: Resource Description Framework; RF: Random Forest; SAPP: Semantic Annotation Platform with Provenance; SMOTEN: Synthetic Minority Over-sampling Technique for Nominal; SPARQL: SPARQL Protocol RDF Query Language.

Graphical Abstract

Key Messages

- MiGenPro merges phenotypic and genomic data for microbial trait prediction using machine learning.
- MiGenPro mitigates phenotype prediction model constraints by leveraging linked data technologies.
- MiGenPro's FAIR design enables adaptation for various phenotypes with available training data.

Received: Date Month Year. **Revised received:** Date Month Year. **Accepted:** Date Month Year
© The Author(s) YEAR. Copyright and licence statements to be updated by the publisher during production.

Introduction

Advances in experimental and computational techniques for DNA extraction, sequencing, assembly, and structural and functional annotation have increased the availability of microbial genotypes. 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 genetic sequences and the result of annotations 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 sequences still lacks behind as dedicated experiments are often required to characterise each aspect such as optimal temperature range, substrate utilisation potential or potential for sporulation among others.

Yet the exploitation of microbes depends on the microbial phenotype. Consider the global biotechnology sector that is expanding in the coming years. The expansion of this industry is coupled with developments in strain optimization and the discovery of strains with industrial potential, . The selection of industrial strains often considers an increased performance of the cell factory and relevant traits include along with desired metabolic pathways, thermotolerance, osmotolerance, and general robustness. For example, the production of polyhydroxyalkanoates in *Schlegelella thermodepolymerans* benefits from high temperatures as it reduces the risk of contamination. Therefore, exploring these phenotypes can accelerate biotechnological progress. Similarly, bioremediation strategies could benefit 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, , , . 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. It also highlights the need for uniform data annotation and preserving the association between genotype and phenotype data in the training datasets. Here, we demonstrate 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 standardised workflows for uniform genome data annotation, . 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. The predictions are complimented with the addition of methods for feature selection that extract biological knowledge from the developed models.

Methods

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

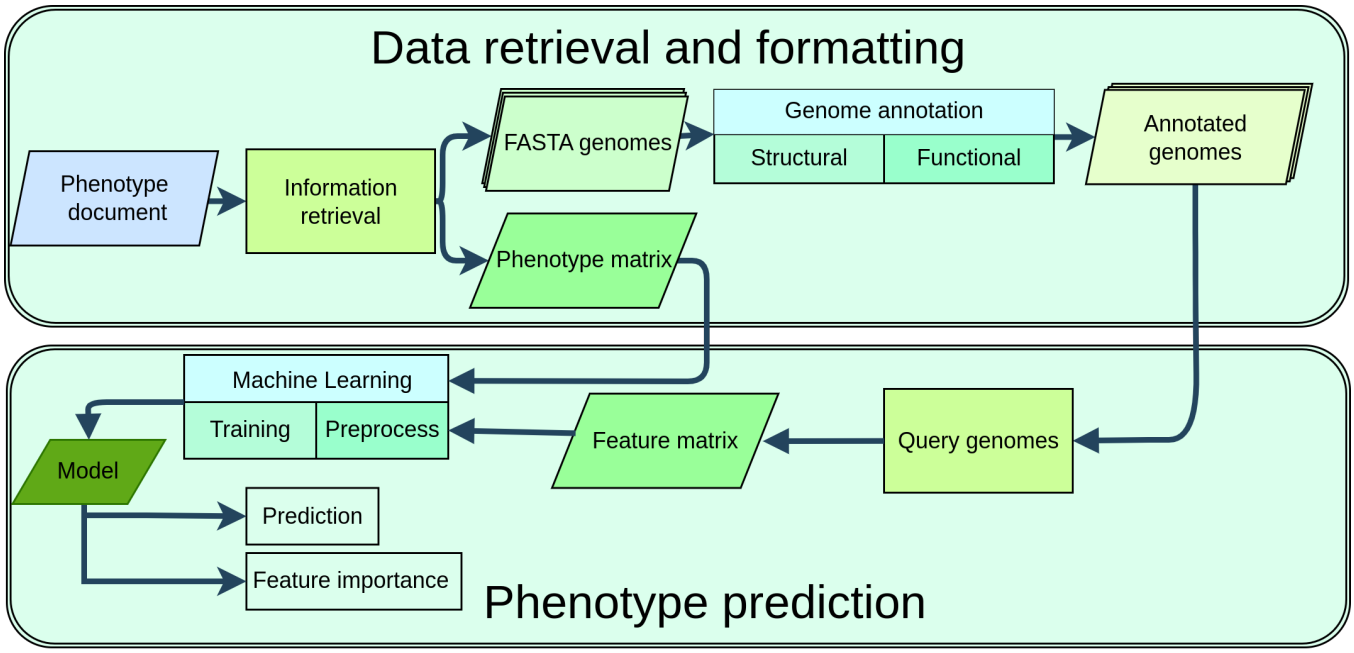


Figure 1 The workflow starts with by loading a phenotype graph, which is queried using SPARQL to generate a list of genome IDs and a matrix containing the phenotypes. These IDs are used to obtain corresponding FASTA files and the phenotype matrix links each genome ID to its respective phenotype. The 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.

Retrieval of phenotype graph

The term phenotype graph refers to a document that describes 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 with 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).

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. The 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, the ete4 (v.3.1.3) library was used for its NCBI taxonomy tree to ensure $\leq$10 randomly chosen genomes per species to avoid over-representation of model organisms. The genome data in FASTA format was retrieved using the NCBI Datasets conda package.

Genome annotation

The genome data were then structurally and functionally annotated using a standardised workflow written in the Common Workflow Language, . 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 , , . 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.

Phenotype prediction

Query genomes and feature matrix construction

The 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 mutual_info_classif function was used to calculate the mutual information of protein domains. By subsequently selecting the top 50% of mutual information ordered protein domains redundant features were filtered out, addressing the multi-dimensional nature of the

dataset. The selected 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 an 80% train and 20% test with the scikit-learn train_test_split function. 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) . This technique was used for categorical oversampling with a maximum oversampling factor of two to balance the difference in frequency amongst phenotypes.

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	6539
negative	4253
positive	2286
Motility	6048
no	3191
yes	2857
Oxygen	10826
aerobe	5324
anaerobe	3127
facultative anaerobe	1447
microaerophile	928
Spore	4383
no	3211
yes	1172
Temperature	18315
mesophilic	16298
psychrophilic	1100
thermophilic	917

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 .

Parameter optimisation was performed per phenotype and classifier using the HalvingGridSearchCV function with the grid space for decision tree models being the maximum tree depth (3, 5, 10, 15). Random forest models tuned the number of trees (50, 100, 200, 500, 1000, 2000), maximum depth (5, 10, 20, 30), minimum samples per leaf (1, 5, 10), and minimum samples per split (2, 5, 10). The Gradient Boosting grid explored estimator counts (50, 100, 200, 500, 1000) and learning rates (0.01, 0.1, 0.2), alongside constraints on maximum depth (3, 5, 7) and minimum samples per split (2, 5, 10).

The parameter max_resources was set to the number of samples and min_resources was set to exhaust with the Matthews coefficient as the target metric. A successive halving grid search first trains all parameter combinations with few resources, in this scenario a small subset of data; it then selects the best performing parameter combinations and repeats the process with more resources until the optimal combination has been found.

Decision tree classifiers used a maximum depth of 15 for spore formation, oxygen requirements, Gram stain prediction, and optimal temperature range, and 10 for motility prediction. Random forest classifiers were trained using a maximum depth of 30 for all phenotypes . 1000 trees were used to train random forest classifiers that predicted motility, oxygen requirements, Gram stains, and optimal temperature range, with spore formation models trained with 500 trees. A minimum sample split of 5 was used for the prediction of spore formation and motility, while a split of 2 was used for oxygen requirements, Gram stains, and temperature. Gradient boosting classifiers used a learning rate of 0.2 and a maximum tree depth of 5 for spore formation, motility, and optimal temperature range, while oxygen requirements and Gram stain prediction used a depth of 3. Minimum sample splits were set to 10 for spore formation and oxygen requirements, 5 for Gram stains and temperature, and 2 for motility. The number of estimators was 1000 for motility, 500 for spore formation and oxygen requirements, 200 for Gram stain prediction, and 50 for optimal temperature range .

Model assessment and feature importance

The models were evaluated in accuracy, Matthew coefficient, F1 score, and AUC on a held-out test set which consisted of 20% fo the 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 of 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.

Results

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 such as BacDive. 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 models is shown in Table 2. Comparison to different studies shows that models reach similar bottlenecks in terms of performance due to the datasets being used. It should 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. The differences between datasets used is vast as Feldbauer, et al. 2015 used 97 genomes for training, Weimann, et al. 2016 used 234, Lingner et al. 2010 used 795 finished and 237 unfinished genomes, and Koblitz, et al. 2025 used between 3000 and 12000 genomes for training depending on the phenotype.

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 and values were rounded to two decimals. The mean and standard deviation are displayed in the table.

Source	Accuracy
Gram	
Feldbauer (2015)	0.99 ± 0.010
Koblitz (2025)	0.96 ± 0.000
Lingner (2010)	0.97 ± 0.000
This study decision tree (2026)	0.96 ± 0.006
This study gradient boosting (2026)	0.98 ± 0.003
This study random forest (2026)	0.98 ± 0.000
Weimann (2016)	0.99 ± 0.010
Motility	
Feldbauer (2015)	0.87 ± 0.060
Koblitz (2025)	0.98 ± 0.000
Lingner (2010)	0.93 ± 0.000
This study decision tree (2026)	0.79 ± 0.020
This study gradient boosting (2026)	0.86 ± 0.008
This study random forest (2026)	0.81 ± 0.003
Weimann (2016)	0.95 ± 0.020
Oxygen requirement	
Feldbauer (2015)	0.90 ± 0.050
Koblitz (2025)	0.94 ± 0.000
Lingner (2010)	0.97 ± 0.000
This study decision tree (2026)	0.79 ± 0.002
This study gradient boosting (2026)	0.87 ± 0.012
This study random forest (2026)	0.86 ± 0.006
Weimann (2016)	0.97 ± 0.010
Spore formation	
Feldbauer (2015)	0.93 ± 0.060
Koblitz (2025)	0.96 ± 0.000
Lingner (2010)	0.95 ± 0.000
This study decision tree (2026)	0.94 ± 0.009
This study gradient boosting (2026)	0.97 ± 0.004
This study random forest (2026)	0.97 ± 0.006
Weimann (2016)	0.95 ± 0.050
Temperature	
Feldbauer (2015)	-
Koblitz (2025)	0.99 ± 0.000
Lingner (2010)	-
This study decision tree (2026)	0.90 ± 0.001
This study gradient boosting (2026)	0.92 ± 0.002
This study random forest (2026)	0.91 ± 0.002
Weimann (2016)	-

Table 3. Classifier performance for each phenotype using the F1 score, receiver operating characteristic (AUC), 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). The mean and standard deviation of five randomised iterations are displayed in the table.

Classifier	F1 score	Matthews	AUC
Gram			
DT	0.96±0.01	0.92±0.00	0.97±0.00
GB	0.98±0.00	0.96±0.01	0.99±0.00
RF	0.98±0.00	0.96±0.00	1.00±0.00
Motility			
DT	0.79±0.02	0.57±0.04	0.82±0.01
GB	0.86±0.01	0.72±0.02	0.93±0.01
RF	0.82±0.00	0.63±0.01	0.90±0.00
Oxygen requirement			
DT	0.79±0.00	0.67±0.01	0.82±0.01
GB	0.87±0.01	0.80±0.02	0.95±0.00
RF	0.86±0.01	0.79±0.01	0.96±0.00
Spore formation			
DT	0.94±0.01	0.85±0.02	0.88±0.02
GB	0.97±0.00	0.91±0.01	0.99±0.01
RF	0.97±0.01	0.93±0.02	0.99±0.01
Temperature			
DT	0.90±0.00	0.34±0.01	0.70±0.01
GB	0.92±0.01	0.51±0.02	0.86±0.01
RF	0.91±0.00	0.43±0.02	0.87±0.01

An overview of the model performance scores can be found in Table 3. Performance indicators for all methods and all physiological characteristics considered are high, indicating the suitability of this 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.

Feature analysis and biological interpretation

Although the performance difference between random forests and gradient-boosting machines is dependent on the variables used. 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 5 most relevant protein domains for predicting the motile potential are provided in Table 4).

The rank sum method aggregates feature importance across five iterations by assigning inverse ranks (with 5 indicating highest importance) to each feature and computing their cumulative sum. Lower total scores indicate higher overall importance because features that are consistently ranked as critical across iterations accumulate smaller sums.

In the analysis of protein domains to predicting motility using a gradient boosting classifier, several domains consistently

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

Domain	Rank sum	Description	Reference
PF02120	5	C-terminal domain of FliK; controls flagellar hook length	Shibata, et al. 2007
PF02203	10	Ligand-binding domain in methyl-accepting chemotaxis receptors	Mise 2016
PF00672	15	HAMP domain common in chemoreceptors	Hulko, et al. 2006
PF02743	25	PAS-like domain found in chemoreceptors	Upadhyay, et al. 2016
PF00563	29	EAL domain found in signalling proteins	Galperin, et al. 2001

emerged with a low cumulative sum. Functional annotations were retrieved from Uniprot using the Uniprot API.

The PF02120 domain was identified as the most important domain, as shown by the rank sum of 5 over five iterations. This domain is located in the C-terminal of the FliK protein controls flagellar length. The domains PF02203, PF00672, and PF02743 emerged as the second, third, and fourth most important domains for the prediction of motility, , . These domains are associated with chemoreceptors and possibly chemotaxis, which determines the direction of motility. The fifth most important domain, PF00563, is associated with signal proteins in the EAL gene that have been associated with regulation of motility, .

Discussion

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, , , . The application of this workflow requires genomes combined with a database or matrix with phenotype information of the corresponding genomes.

Consistent annotation of genome data prior to comparative genomic studies has been regarded as essential to prevent bias and artifacts due to differences in the tools used for genome annotation . MiGenPro enables automation of these tasks that facilitate comparative genomics. Here we have illustrated the use of MiGenPro using phenotype data from BacDive, and due to its modular design can be extended to other data resources using Linked Data formats. The starting point is a phenotype graph file that associates physiological measurements to standardised genome identifiers such as the ones in NCBI or an internal system. Once this starting point is available, the workflow provides the tools to retrieve and canonicalise the data and enable the development of machine learning models. By standardising genomic data as Linked Data, a high level of interoperability is obtained. This interoperability allows for efficient development of phenotype prediction learning models. The quality of the genomes was not compared to model performance; however, it stands to reason that genomes of higher-quality would yield more accurate results, as machine learning methods could use the absence of a protein domain to infer a phenotype.

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

Model performance is not artificially inflated

Phenotypic traits often relate to the species and genus of microbes however, exceptions are known where traits occur only in strains of specific species within a species or genus. When training classifiers for the prediction of phenotypes, model performance should not base predictions on taxa-specific features. The results from five-fold cross-validation show no significant deviations in feature importances across iterations, with importance remaining consistent throughout the process, as 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 with regard to the observed standard deviation in Table 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 are biologically relevant (See Table 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 as the user could use other methods for the evaluation of feature importance, such as Shapley additive explanation values.

Model performance matches previous phenotype prediction methods

Cross-study comparison of machine learning approaches (Decision Trees, Random Forests, Gradient Boosting) in Table 2 reveals marginal performance variations (Δ accuracy $< 7\%$) barring the motility, suggesting algorithm choice is less critical than data quality for phenotype prediction. From Table 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 the last 15 years, accuracy was selected as one of the few performance metrics available, , , . 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, . These motility subgroups are not evenly distributed and have varying levels of complexity, resulting in lower than expected performance metrics.

Impact of data quality on performance

BacDive data lacks or under-represent certain phenotypes (e.g., temperature ranges of extremophiles). Low-performance phenotypes such as motility that is derived from complex multi-gene regulation and have multiple methods for achieving the same result. The rapid development of AI technology represents a promising future area of research where phenotype analysis is based on gene sequences combined with protein structures to predict these more complex phenotypes. In addition, AI technology can also more easily process literature for the identification of historical records of phenotypes that are currently not incorporated into dedicated repositories such as BacDive.

Data availability and description

The source code of MiGenPro is available free of use under the MIT license. In addition, all data generated in this project, the annotated genomes are available as compressed and pre-indexed HDT files combined with the phenotype matrix at 16995284.

Supplementary file 1

SPARQL queries to retrieve microbial data from the BacDive database, extracting NCBI genome accessions, taxonomic identifiers, and five key phenotypic traits: Gram stain status, motility, oxygen requirements, spore formation, and optimal growth temperature.

Supplementary file 2

SPARQL query used to query annotated genomes in the GBOL format for protein domain frequencies.

Supplementary file 3

Halving grid search space used for hyper parameter tuning of decision tree, random forest, and gradient boosting classifiers.

References

Sparql 1.2 query language. Technical report, World Wide Web Consortium (W3C), 08 2025. W3C Working Draft. Former editors: Steve Harris, Andy Seaborne, Eric Prud'hommeaux. Latest published version: https://www.w3.org/TR/sparql12-query/. Latest editor's draft: https://w3c.github.io/sparql-query/spec/. History: https://www.w3.org/standards/history/sparql12-query/. Test suite: https://w3c.github.io/rdf-tests/. Latest Recommendation: https://www.w3.org/TR/2013/REC-sparql11-query-20130321.

Eric Abbate, Jennifer Andrion, Amanda Apel, Matthew Biggs, Julie Chaves, Kristi Cheung, Anthony Ciesla, Alia Clark-ElSayed, Michael Clay, Riarose Contridas, Richard Fox, Glenn Hein, Dan Held, Andrew Horwitz, Stefan Jenkins, Karolina Kalbarczyk, Nandini Krishnamurthy, Mona Mirsiaghi, Katherine Noon, Mike Rowe, Tyson Shepherd, Katia Tarasava, Theodore M Tarasow, Drew Thacker, Gladys Villa, and Krishna Yerramsetty. Optimizing the strain engineering process for industrial-scale production of bio-based molecules. *Journal of Industrial Microbiology & Biotechnology*, 50(1):kuad025, September 2023.

Matthias Blum, Hsin-Yu Chang, Sara Chuguransky, Tiago Grego, Swaathi Kandasaamy, Alex Mitchell, Gift Nuka, Typhaine Paysan-Lafosse, Matloob Qureshi, Shriya Raj, Lorna Richardson, Gustavo A Salazar, Lowri Williams, Peer Bork, Alan Bridge, Julian Gough, Daniel H Haft, Ivica Letunic, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, and Robert D Finn. The InterPro protein families and domains database: 20 years on. *Nucleic acids research*, 49:D344–D354, 01 2021.

Robert M. Bowers, Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, Susannah G. Tringe, Natalia N. Ivanova, Alex Copeland, Alicia Clum, Eric D. Becraft, Rex R. Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M. Weinstock, George M. Garrity, Jeremy A. Dodsworth, Shibu Yooseph, Granger Sutton, Frank O. Glöckner, Jack A. Gilbert, William C. Nelson, Steven J. Hallam, Sean P. Jungbluth, Thijs J. G. Ettema, Scott Tighe, Konstantinos T. Konstantinidis, Wen-Tso Liu, Brett J. Baker, Thomas Rattei, Jonathan A. Eisen, Brian Hedlund, Katherine D. McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W. Tyson, Christian Rinke, Genome Standards Consortium, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H. Parks, A. M. Eren, Lynn Schriml, Jillian F. Banfield, Philip Hugenholtz, and Tanja Woyke. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, 35(8):725–731, 08 2017.

Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, October 2001.

Rida Chaudhary, Ali Nawaz, Mireille Fouillaud, Laurent Dufossé, Ikram ul Haq, and Hamid Mukhtar. Microbial Cell Factories: Biodiversity, Pathway Construction, Robustness, and Industrial Applicability. *Microbiology Research*, 15(1):247–272, March 2024. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 06 2002.

Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, and The CWL Community. Methods included: standardizing computational reuse and portability with the common workflow language. *Commun. ACM*, 65(6):54–63, 05 2022.

Jesse C. J. van Dam, Koehorst Jasper J., Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez. Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining. Pages: 184747 Section: New Results, 09 2017.

Youssef El Mouali, Hyunhee Kim, Irfan Ahmad, Annelie Brauner, Ying Liu, Mikael Skurnik, Michael Y. Galperin, and Ute Römling. Stand-Alone EAL Domain Proteins Form a Distinct Subclass of EAL Proteins Involved in Regulation of Cell Motility and Biofilm Formation in Enterobacteria. *Journal of Bacteriology*, 199(18):10.1128/jb.00179–17, August 2017. Publisher: American Society for Microbiology.

Roman Feldbauer, Frederik Schulz, Matthias Horn, and Thomas Rattei. Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics*, 16(14):S1, 10 2015.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001. Publisher: Institute of Mathematical Statistics.

M Y Galperin, A N Nikolskaya, and E V Koonin. Novel domains of the prokaryotic two-component signal transduction systems. *FEMS microbiology letters*, 203(1):11–21, September 2001.

Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638, 06 2016.

Michael Hulko, Franziska Berndt, Markus Gruber, Jürgen U. Linder, Vincent Truffault, Anita Schultz, Jörg Martin, Joachim E. Schultz, Andrei N. Lupas, and Murray Coles. The HAMP Domain Structure Implies Helix Rotation in Transmembrane Signaling. *Cell*, 126(5):929–940, September 2006.

Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 03 2010.

Julia Koblitz, Lorenz Christian Reimer, Rüdiger Pukall, and Jörg Overmann. Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets. *Communications Biology*, 8(1):897, 06 2025. Publisher: Nature Publishing Group.

Jasper J. Koehorst and Bart Nijsse. Microbial (meta-) genome annotation. Technical report, Wageningen University & Research, 10 2024.

Jasper J. Koehorst, Jesse C. J. van Dam, Edoardo Saccenti, Vitor A. P. Martins Dos Santos, Maria Suarez-Diez, and Peter J. Schaap. SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. *Bioinformatics (Oxford, England)*, 34(8):1401–1403, 04 2018.

Xenie Kourilova, Iva Pernicova, Karel Sedlar, Jana Musilova, Petr Sedlacek, Michal Kalina, Martin Koller, and Stanislav Obruca. Production of polyhydroxyalkanoates (PHA) by a thermophilic strain of *Schlegelella thermodepolymerans* from xylose rich substrates. *Bioresource Technology*, 315:123885, 11 2020.

Vicky Levy. 2024 global life sciences sector outlook. Technical report, Deloitte, May 2024.

Thomas Lingner, Stefanie Mühlhausen, Toni Gabaldón, Cedric Notredame, and Peter Meinicke. Predicting phenotypic traits of prokaryotes from protein domain frequencies. *BMC Bioinformatics*, 11:481, 09 2010.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4768 – 4777. Curran Associates, Inc., 2017.

Takeshi Mise. Structural Analysis of the Ligand-Binding Domain of the Aspartate Receptor Tar from Escherichia coli. *Biochemistry*, 55(26):3708–3713, July 2016. Publisher: American Chemical Society.

Said Nawar and Abdul M. Mouazen. 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. *Sensors*, 17(10):2428, 10 2017. Publisher: Multidisciplinary Digital Publishing Institute.

Lisbeth Olsson, Peter Rugbjerg, Luca Torello Pianale, and Cecilia Trivellin. Robustness: linking strain design to viable bioprocesses. *Trends in Biotechnology*, 40(8):918–931, 08 2022.

Nuala A. O'Leary, Eric Cox, J. Bradley Holmes, W. Ray Anderson, Robert Falk, Vichet Hem, Mirian T. N. Tsuchiya, Gregory D. Schuler, Xuan Zhang, John Torcivia, Anne Ketter, Laurie Breen, Jonathan Cothran, Hena Bajwa, Jovany Tinne, Peter A. Meric, Wratko Hlavina, and Valerie A. Schneider. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI datasets. *Scientific Data*, 11(1):732, 07 2024. Publisher: Nature Publishing Group.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Wasin Poncheewin, Anne D. van Diepeningen, Theo A. J. van der Lee, Maria Suarez-Diez, and Peter J. Schaap. Classification of the plant-associated lifestyle of pseudomonas strains using genome properties and machine learning. *Scientific Reports*, 12:10857, 06 2022.

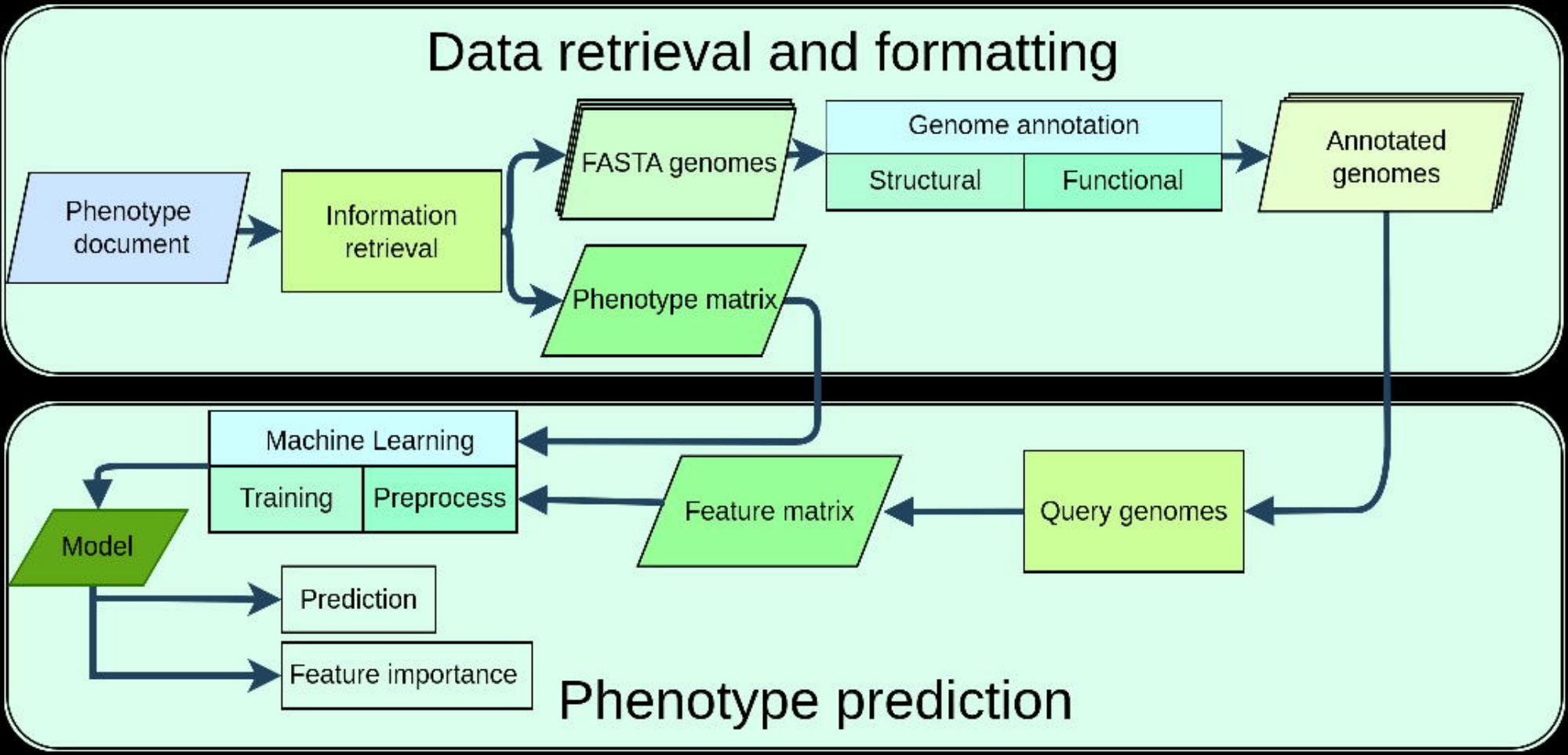
Isabel Schober, Julia Koblitz, Joaquim Sardà Carbasse, Christian Ebeling, Marvin Leon Schmidt, Adam Podstawka, Rohit Gupta, Vinodh Ilangovan, Javad Chamanara, Jörg Overmann, and Lorenz Christian Reimer. BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research*, 53:D748–D756, 01 2025.

Satoshi Shibata, Noriko Takahashi, Fabienne F V Chevance, Joyce E Karlinsey, Kelly T Hughes, and Shin-Ichi Aizawa. FliK regulates flagellar hook length as an internal ruler. *Molecular microbiology*, 64(5):1404–1415, June 2007.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45:D158–D169, 01 2017.

Amit A. Upadhyay, Aaron D. M. Fleetwood, Ogun Adebali, Robert D. Finn, and Igor B. Zhulin. Cache domains that are homologous to, but different from pas domains comprise the largest superfamily of extracellular sensors in prokaryotes. *PLoS computational biology*, 12(4):e1004862, April 2016.

Aaron Weimann, Kyra Mooren, Jeremy Frank, Phillip B. Pope, Andreas Bremges, and Alice C. McHardy. From genomes to phenotypes: Traitar, the microbial trait analyzer. *mSystems*, 1(6):e00101–16, 2016.


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