Final Report

Predicting plant-associated bacteria from genetic markers

Team 9
Alex Lukasiewicz, Muyoung Lee, Sarah Coleman

Introduction and Literature Review

As the global population continues to increase, agricultural productivity must increase by 70% to meet demand (Ke et al., 2021). Land for farming continues to become more scarce, so research has focused on increasing the efficiency of current crop systems with novel techniques. Once such technique involves exploring and exploiting beneficial relationships between microbes and plants to increase productivity, deemed microbiome engineering. Some microbes in their native state already have many plant growth-promoting properties, and genetic engineering of microbes to benefit agricultural crops for further improvement is an emergent field of research (Afridi et al., 2022). Some examples of these traits are biocontrol, biofertilization, and biostimulation (Ke et al., 2021). In order to select an appropriate microbial host for this synthetic biology engineering strategy, top-down approaches begin with the identification of native plant-associated bacteria. Once identified, top strains can be engineered with novel CRISPR-based technologies to incorporate specific genes that code for enzymes that may promote plant growth (Ke et al., 2021).

For this identification, recent publications exploring bacterial genomes in broad phyla (Levy et al., 2017; Martínez-García et al., 2016) and within the genera *Pseudomonas* (Saati-Santamaria et al., 2022) and *Xanthomonas* (te Molder et al., 2021), suggest that the presence of certain enzymes in bacterial genomes is related to plant association. Specifically, Levy et al. describe a relationship between carbohydrate-associated enzymes (CaZys) and plant association but do not develop a predictive classification model. While Martínez-García et al. do, they only work with a small dataset. Currently, plant association is inferred by fieldwork, which is time-consuming and unstandardized. An *in silico* method to classify bacteria from multiple genera as plant or non plant associated would aid in categorizing novel bacterial genomes. Identifying a core bacterial population or novel bacteria associated with plants could significantly accelerate metabolic engineering and synthetic biology approaches to engineer the plant rhizosphere (Afridi et al., 2022). Additionally, many bacteria have been uncovered in various soils with an unknown or unclear plant association. Inferences based on these bacterial genomes may assist with the prediction of plant (or root) association.

We hypothesized that machine learning and bioinformatics tools could be used to predict if a new bacteria is plant or non-plant associated (PA, NPA, respectively). This is a binary classification problem. To the best of the author's knowledge has only been reported in literature once, with a supervised Random Forest machine learning model yielding a precision of 93% (Martínez-García et al., 2016). For feature generation, the authors of Martínez-García et al. used PIFAR (searches for protein sequences and PFAM domains) as well as an annotation tool

developed by the authors, T346Hunter, which identifies genomic clusters identified in various types of secretion systems. The authors used a dataset containing 354 data points for model training and 77 for testing.

The features used to train the model described herein include the presence of CaZys and COGs (Clusters of Orthologous Genes) in the available genome sequences as well as the horizontal gene transfer of each genome. This totals approximately 4,500 features. The total number of genomes is approximately 3,700, about an order of magnitude greater than existing literature.

Data and Exploratory Analysis

We used a large dataset of 3,837 well-curated bacterial genomes identified in Levy et al., 2017. Metadata was downloaded from their website, including taxon IDs, classifications (PA, NPA, and soil), taxonomic groups, and the abundance of COGs. We excluded 65 genomes that were no longer hosted on the DOE Joint Genome Institute (JGI) website as there were no total gene counts available. Therefore, 3,772 of the 3,837 genomes described by Levy et al., 2017 were used in this project. In addition to the total gene count, we extracted an additional feature, horizontal gene transfer (as a percentage) from the JGI website.

These genomes' amino acid fasta files were downloaded from the same website by Levy et al., 2017, and CAZy annotation was performed with the most updated version of dbCAN, dbCAN4 (Zhang et al., 2018). Following the guidelines listed on the example results tab of the author's website, we filtered out annotations found by less than two methods provided. Genes with multiple annotations were repeatedly counted for each annotation.

Bacterial genomes can vary widely in size. Therefore the counts of our aligned CAZy families and COG counts reported by Levy et. al., 2017 were normalized by each genome's total gene count. After normalization, we excluded soil bacteria from the model training since they were not classified as PA or NPA. We further normalized all numerical features using the SciKit-learn StandardScaler and removed any features with zero variance. We split the data into the train (67%) and test (33%) sets stratified on the PA/NPA classification. The same random state (42) was used for this and all future steps for reproducibility.

We then visualized the data by two methods: classification, to understand the relative abundance of our bacterial genome samples, and principal component analysis (PCA) to identify how our train/test split, classification, and different bacterial orders comprised the PCA linear transformation space. These visualizations are shown in **Fig. 1** and **Fig. 2,** respectively. It is clear from **Fig. 1** that NPA is the largest classification in the dataset, and the largest bacterial phylum present in the dataset is *Proteobacteria*. In **Fig. 2a**, we see a little clear separation between the PA and NPA classes, which is also true for the soil class (**Fig. 2d**). In **Fig. 2b** and **e**, there is an apparent reconstruction of bacterial order in both PCAs. The test/train split does not unequally represent any components in the linear PCA transformation, as shown in **Fig. 2c**.

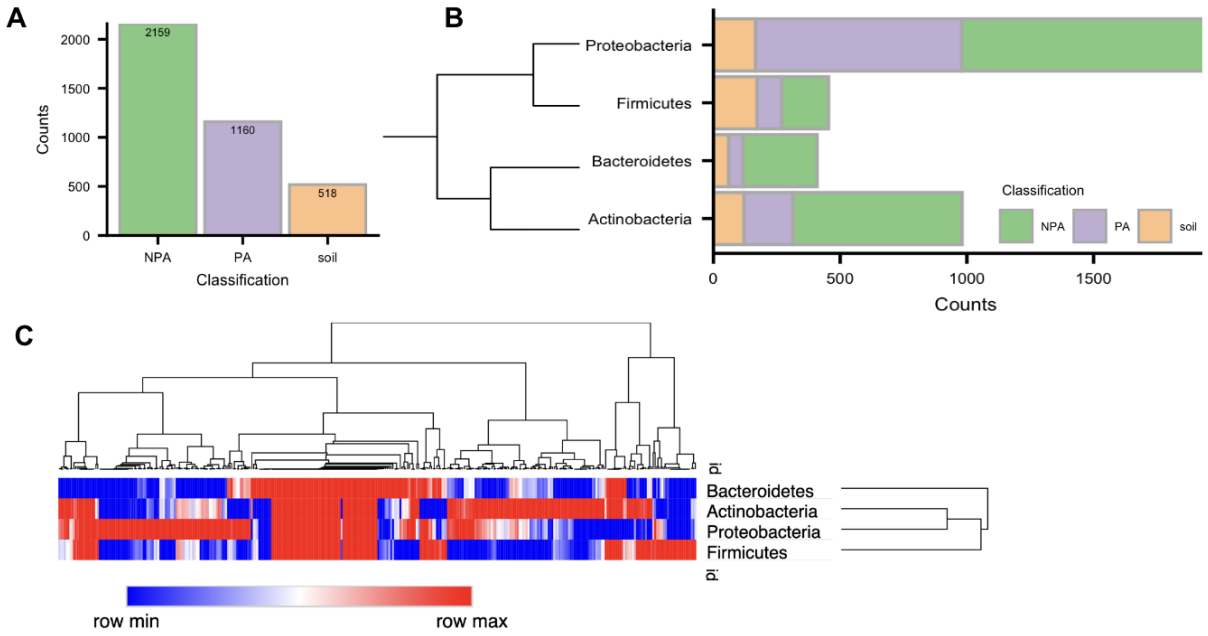

Fig. 1. (A) Distribution of counts within the dataset. (B) Classification by bacterial phylum, with the phylogenetic relationship between phylum shown. (C) Feature clustering and visualization of CAZy families by phylum classification.

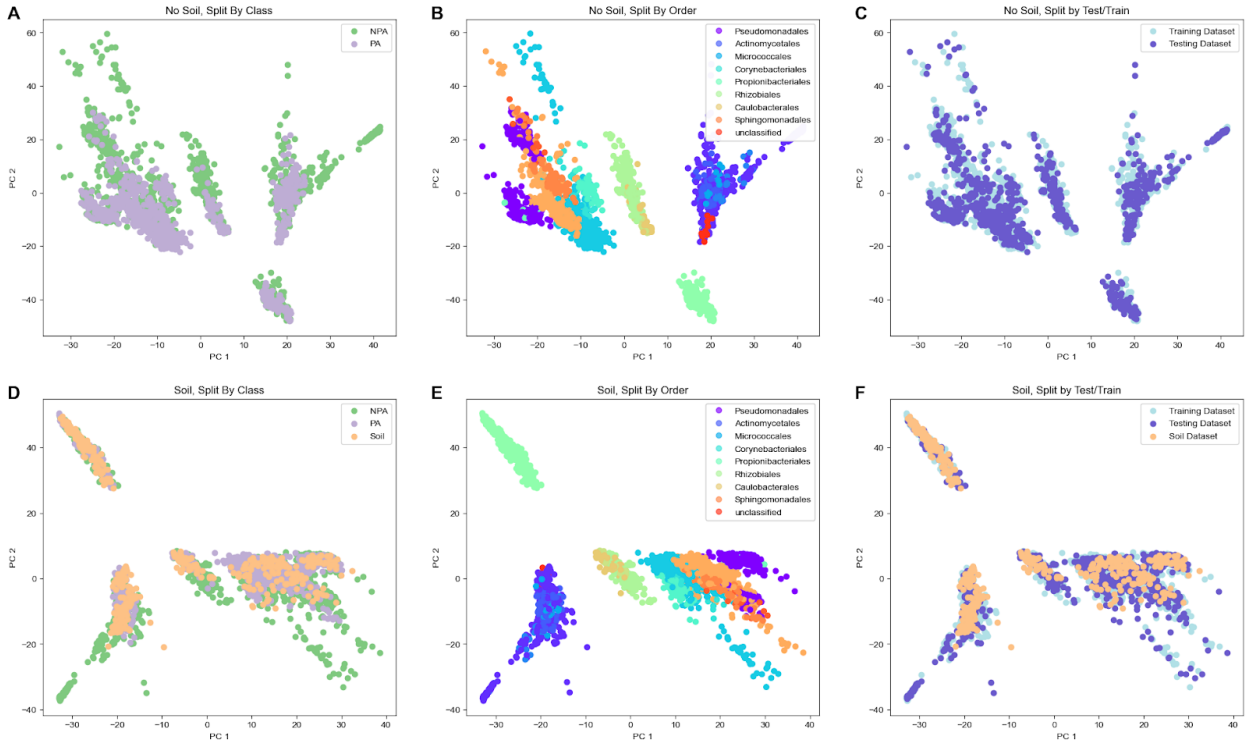

Fig. 2. Principal component analysis (PCA) of the dataset without soil, split by classification (A), order (B), and test/train (C), and a separate PCA of the dataset containing soil, split by classification (D), order (E) and test/train (F).

Feature Selection

Because many COG and CAZy families exist within the genome, the number of features in our dataset is larger than the data points we have. To avoid the situation where data is overfitted to the model, we performed feature selection with a variety of established methods, such as recursive feature elimination with a Random Forest classifier, GridSearchCV with an L1 penalty Logistic Regression function, and SelectKBest with a K-NN Classifier. We hypothesized that the union of features identified by different methods would be able to accurately predict the binary classification without model overfitting. Before feature selection, we removed the soil dataset (as classification is unknown) and split the dataset into test/train. Feature selection was only performed on the train dataset. The feature selection flowchart shown in **Fig. 3** details these steps, and **Fig. 4** shows the counts of the features selected by each method. 440 (= 43 + 62 + 300 + 35) features were selected for the final model. **Fig. 5** details the enrichment in COG families between each of the three different feature selection methods.

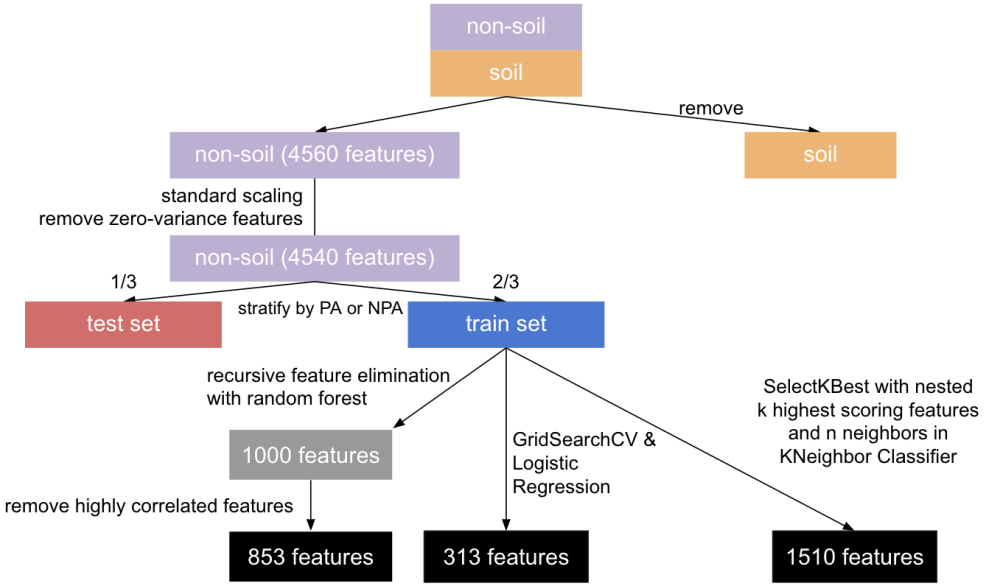

Fig. 3. Feature selection flowchart for this analysis. After the soil genomes were removed, the data was scaled, and zero variance features were removed. Next, the test/train split separated out ⅓ of the data, and feature selection was independently performed on the remaining train set by three different methods.

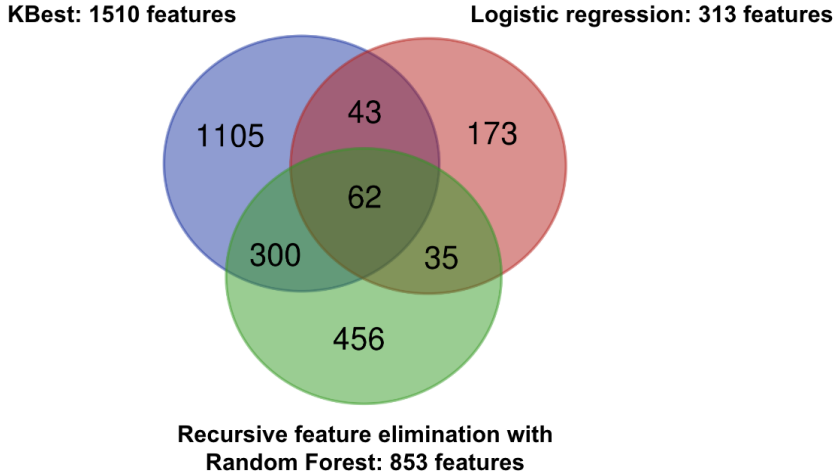

Fig. 4. Overlap of features identified by three independent feature selection methods. All were performed on the train dataset. The union of 440 features identified by these methods (at least ⅔) was used for all further analysis.

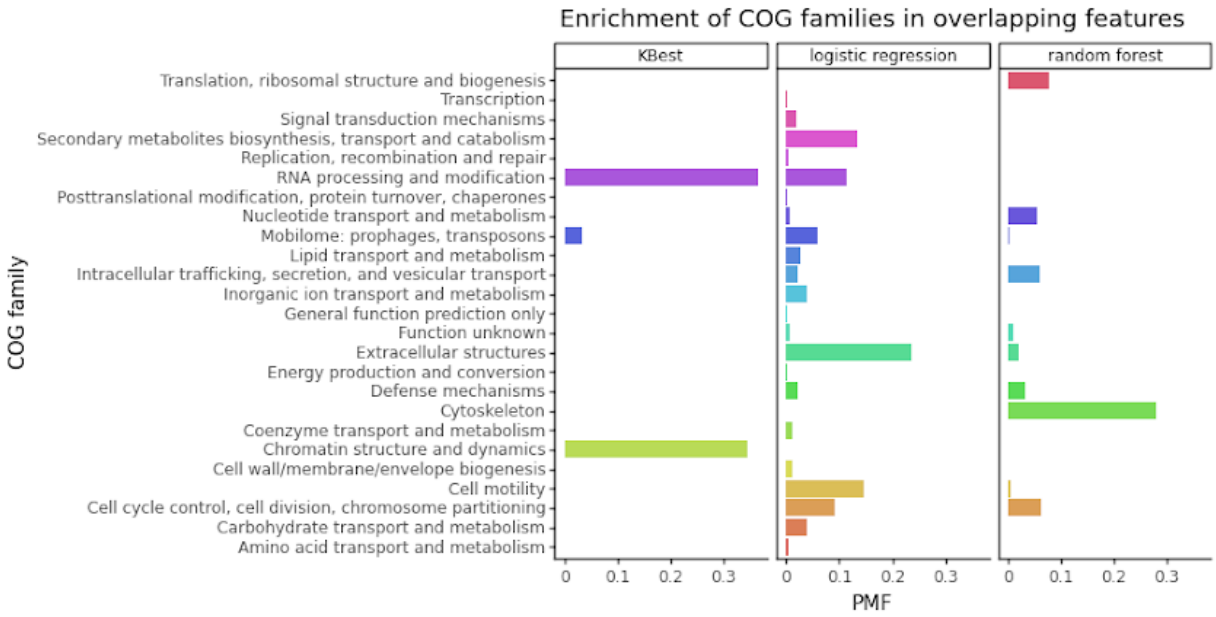

Fig. 5. Enrichment of COG Families in overlapping features. Across the three methods of feature selection used (KBest, logistic regression, and random forest), we identified more feature unity (as shown as similar PMF) between logistic regression and random forest than either model with KBest.

To calculate the level of enrichment of different COG families in the overlap and within each of our feature selection methods, we used SciPy.stats hypergeom function to calculate the probability mass function (PMF) of each family for each selection method. The PMF is a measure of whether we see specific COG families at a higher rate in our feature set than if they were randomly selected from the entire population and interpret this value as a level of enrichment. Each feature selection approach displayed different levels of enrichment and breadth of representation for each family (**Fig. 5**). The SelectKBest method enriched 2 families as features that were important for model performance: RNA processing and modification and chromatin structure and dynamics. Logistic regression had enrichment in a broader range of families but showed the highest enrichment for COG counts belonging to the extracellular structures. The Random Forest aided feature selection also showed enrichment across broad families, but the highest represented group were annotations belonging to cytoskeletal processes. Interestingly, Levy et. al., 2017 identified these COG families as either being significantly enriched or depleted in plant-associated bacteria relative to non-plant associated genera. RNA processing and modification, chromatin structure and dynamics, and cytoskeletal processes represent families with the most extreme enrichment or depletion given the authors' PhyloGLM test. This observation led us to the conclusion that the union of selected features would be able to differentiate between PA and NPA bacteria.

Modeling and Validation

After preprocessing and feature selection, we trained our data with the same selected features on the following models: Decision Tree and Random Forest, XGBoost, Logistic Regression, and K-nearest neighbors. Additional models tested included SVC and a Naive Bayes Classifier, but

they are eliminated from this report due to clarity and poor model performance. For each model, we report (if applicable) hyperparameter tuning results. Precision-Recall (PR) curves for all models are shown in **Fig. 6**. Specific model methodology is described in the following section.

Decision Tree, Random Forest, and XGBoost

Decision tree is a non-parametric supervised learning method to create a model by learning simple decision rules inferred from the data features. It is easy to understand, and the model can be visualized, but serious disadvantages include overfitting, errors due to bias, and variance. To solve this problem, a random forest generates many trees with different samples in parallel and averages out their solutions. However, more trees mean more processing time. On the contrary, XGBoost (Extreme Gradient Boosting) is based on boosting, a sequential model. In boosting, each model in a series trains upon its predecessor's mistakes to correct them. Gradient boosting optimizes the loss as the inputs go through the sequence of models. It calculates the loss and adds the next model that reduces the loss. XGBoost is a specific implementation of gradient boosting with more accurate approximations and speedy processing.

All hyperparameter tunings were done based on accuracy scores from the test set. Details are below.

- Tested hyperparameters and their ranges:
  - Decision tree: max_depth: 1 - 20, min_sample_leaf: 1 - 20
  - Random forest: n_estimators: 50, 100, 150; max_depth: 1 - 20; min_samples_leaf: 1- 20
  - XGBoost: n_esimators: 50, 100, 150; max_depth: 1 - 20

- Selected hyperparameters and models' accuracies on the test set
  - Decision tree
    max_depth: 11, min_sample_leaf: 5, accuracy: 0.848
  - Random forest
    n_estimators: 50, max_depth: 18, min_samples_leaf: 1, accuracy: 0.889
  - XGBoost
    n_esimators: 150; max_depth: 8, accuracy: 0.900

PR curves for these three models are shown in **Fig. 6**. PR curves from the Random Forest and XGboost models showed similar patterns, but the PR curve from the decision tree was atypical. It would be because the curve was from only one decision tree different from the other two.

Logistic Regression

Logistic Regression is one of the most interpretable models, which would be an advantage for the downstream applications of this project. For the logistic regression model, K-fold hyperparameter tuning of C (the inverse of regularization strength) is shown in **Fig. 7**. While the default value of C is 1, a value of 0.029471 yielded the highest accuracy with the lowest variance. For model convergence, if it was not reached (as indicated by an error message), max_iter was intuitively increased until convergence was found. To explore the best Logistic Regression fit for the data in this paper, different solver methods and penalty functions were

explored. The following three specifications (only specifications that are different from default are reported) were made to test different Logistic Regression models:

- Model A: penalty = 'l1', solver = 'saga'
- Model B: penalty = 'l2', solver = 'lbfgs'
- Model C: penalty = 'elasticnet', solver = 'saga', l1_ratio = 0.5.

While all had a very similar PR curve, Model A was selected as the best Logistic Regression performer for feature reduction and interpretability purposes (data not shown). The L1 penalty has the advantage of penalizing when parameters are close to (but not) zero. While the L1 penalty is often associated with worse runtimes, which was observed when training the model on Google Colab, our dataset was small enough that this did not affect runtime.

K-Nearest Neighbors

The K-Nearest Neighbor algorithm was also trained on the stratified training dataset. The 440 features that overlap between the three feature importance methods (**Fig. 4**) were used to subset the training and test data. In an effort to maximize accuracy while minimizing variance during the parameter search, a stratified K-fold cross-validation was performed. The mean and standard deviation of this accuracy score were calculated for the n number of neighbors between 1 and 100 (**Fig. 8**). This revealed that the number of neighbors that maximized accuracy (0.89) and minimized variance (0.01) was n = 3. This parameter was then selected to fit the model for further evaluation.

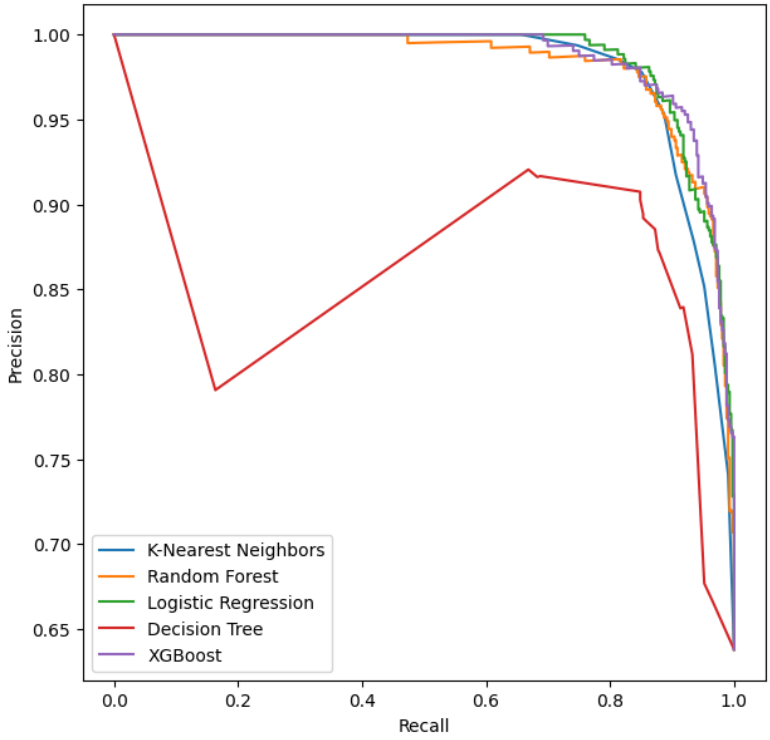

Fig. 6. Precision-Recall curves of the tested models in this study. PR curves were generated on the test dataset, with the 440 union features and optimal hyperparameters for each model. XGBoost generally appears to be the best model with this metric.

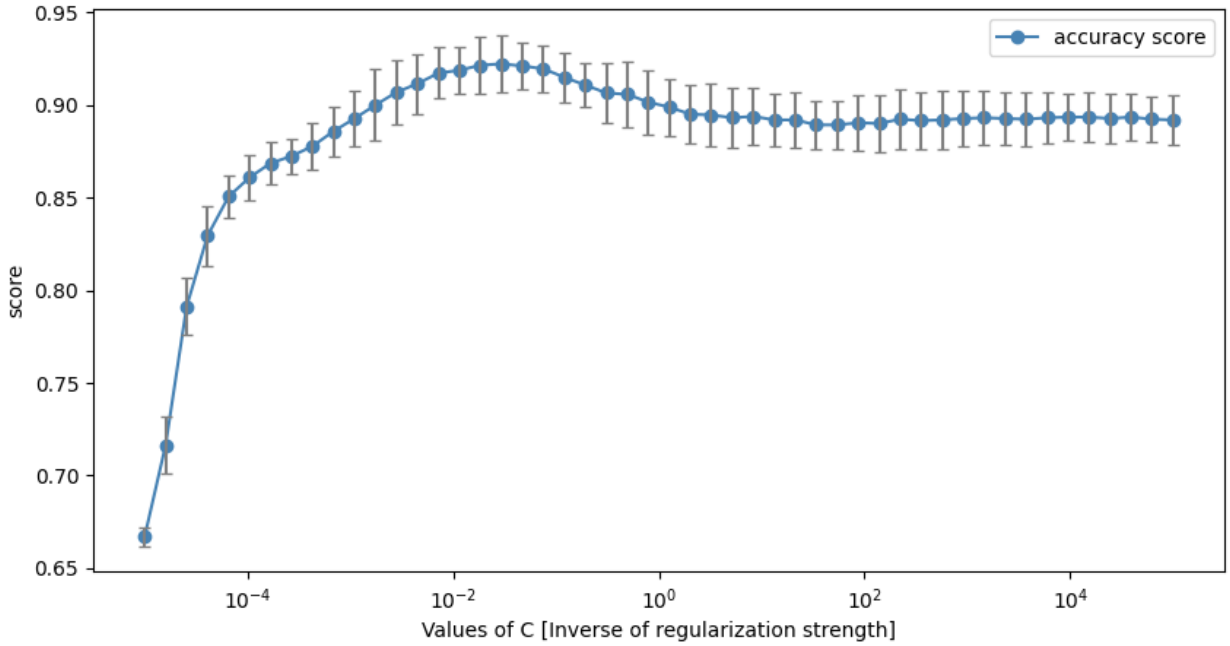

Fig. 7. K-Fold hyperparameter tuning of the Logistic Regression Classifier reveals that a C value of 0.029471 yields the highest accuracy score and lowest variance. Note that the x-axis is log scale.

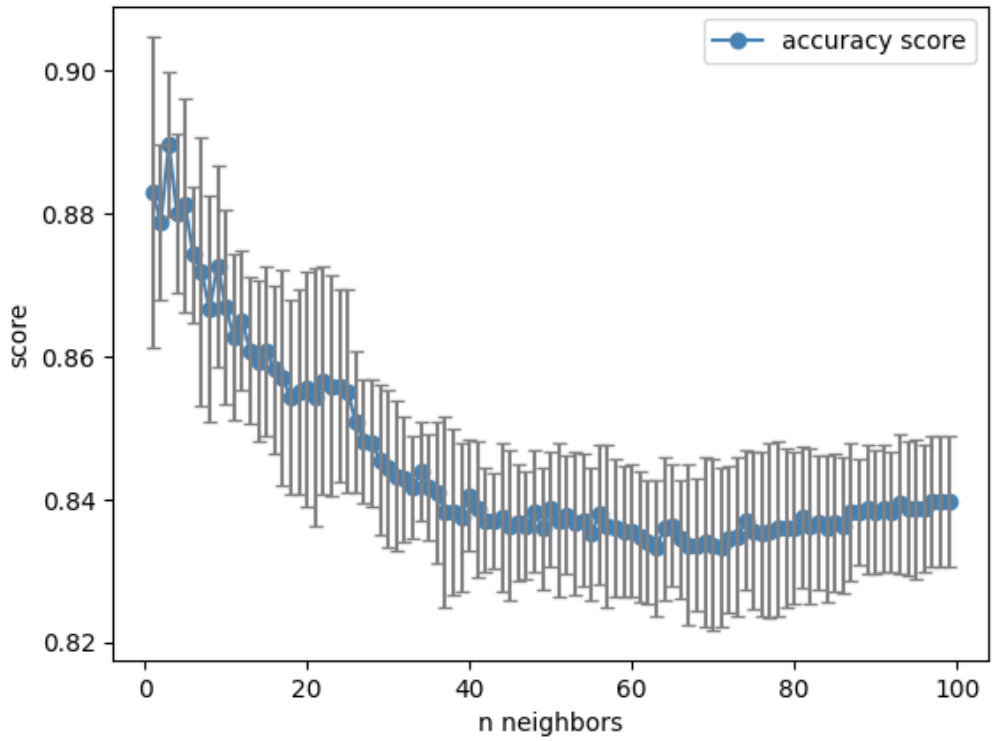

Fig. 8. K-Fold hyperparameter tuning of the K-Nearest Neighbor Classifier reveals that an n of 3 yields the highest accuracy score and the lowest variance.

Results

We next performed K-fold (K=5) cross-validation with our five models on the test dataset, as shown in **Table 1** and **Fig. 6**. The best-performing models are XGBoost and Logistic Regression in terms of the highest evaluation metrics. However, XGBoost also showed the smallest variation in most of the evaluation metrics across all k-fold splits of the dataset, which indicates its generalizability to unseen data. Therefore, we decided to move ahead with this model.

Table 1. Comparing the five-fold cross-validation results across the five main models applied in this study.

Model	Accuracy	F1	Precision	Recall
Decision Tree	0.847 ± 0.006	0.787 ± 0.012	0.771 ± 0.028	0.805 ± 0.013
KNeighbors	0.895 ± 0.012	0.854 ± 0.02	0.831 ± 0.032	**0.878 ± 0.018**
Logistic Regression	0.894 ± 0.012	0.847 ± 0.021	0.856 ± 0.038	0.839 ± 0.035
Random Forest	0.900 ± 0.013	0.857 ± 0.023	0.857 ± 0.035	0.858 ± 0.015
XGBoost	**0.911 ± 0.006**	**0.873 ± 0.012**	**0.872 ± 0.020**	0.875 ± 0.011

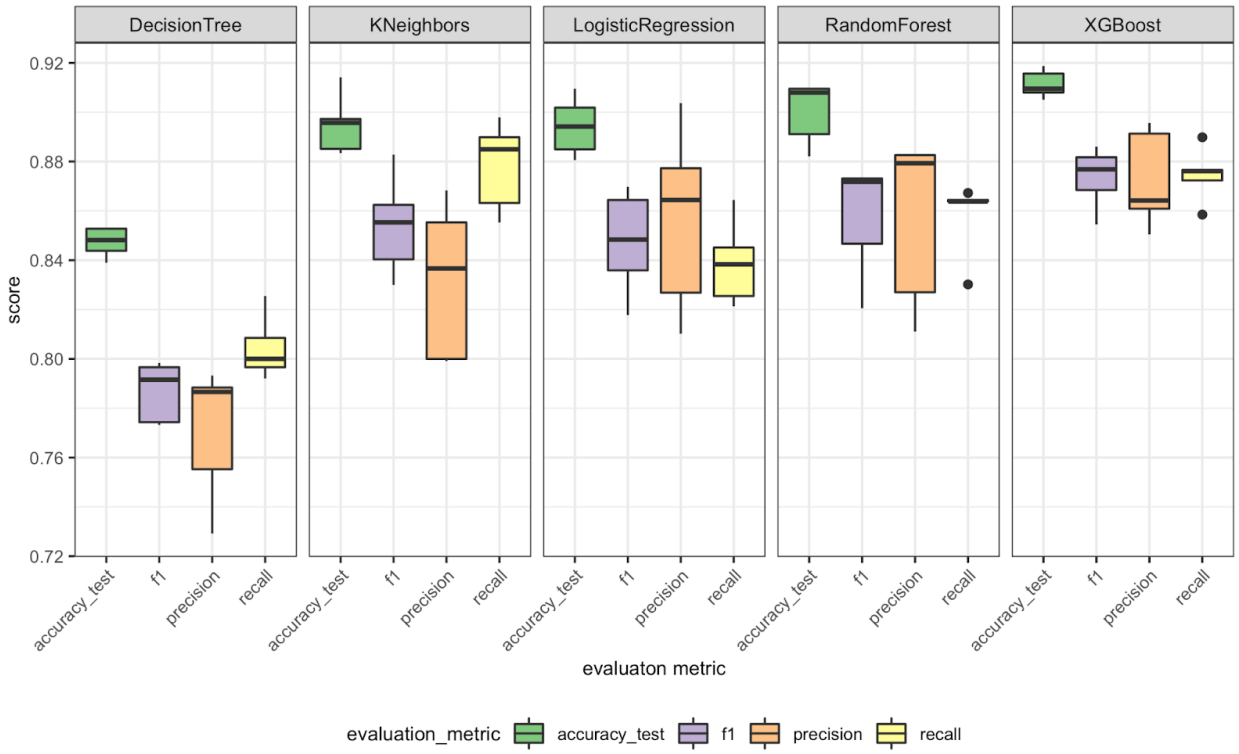

Fig. 8. Visualization of different evaluation metrics for the Decision Tree, K Neighbors, Logistic Regression, Random Forest, and XGBoost models. Box plots were derived from K-fold (K=5) cross-validation.

The top ten most important features identified in the XGBoost model are shown in **Table 2**. All

of the ten features are COG annotations. We hypothesized that the CAZy features predicted by dbCAN4 would be influential to PA/NPA status, and while they may be, they are not captured by the feature importance of our top-performing model. However, we do see that the top feature is associated with a sugar lactonase, supporting the hypothesis that carbohydrate associated enzymes are influential in determining plant association.

Table 2. Top 10 features important to the performance of the XGBoost classifier and information about their function.

Feature	Feature Importance	Description
COG3386	0.1189	Sugar lactone lactonase YvrE
COG3569	0.0797	DNA topoisomerase IB
COG3131	0.0597	Periplasmic glucan biosynthesis protein OpgG
COG0667	0.0591	Pyridoxal reductase PdxI or related oxidoreductase, aldo/keto reductase family
COG1609	0.0449	DNA-binding transcriptional regulator, LacI/PurR family
COG0798	0.0374	Arsenite efflux pump ArsB, ACR3 family
COG1744	0.0366	Lipoprotein Med, regulator of KinD/Spo0A, PBP1-ABC superfamily, includes NupN
COG0514	0.0257	Superfamily II DNA helicase RecQ
COG0297	0.0248	Glycogen synthase
COG1349	0.0230	DNA-binding transcriptional regulator of sugar metabolism, DeoR/GlpR family

Next, we drew partial dependence (PDP) plots (**Fig. 9**) and individual conditional expectation (ICE) plots (**Fig. 10**) to investigate the relationship between prediction results and the most important features. Our model seems to have binary dependencies for features, dependent on whether a feature is valued above or below a specific threshold. Considering XGBoost is an applied version of the Decision Tree model, this reasonably explains the relationship. For the ICE plots shown in **Fig. 10**, we observe again how the decision threshold modifies specific partial dependencies, but they do not otherwise change as a function of parameter values.

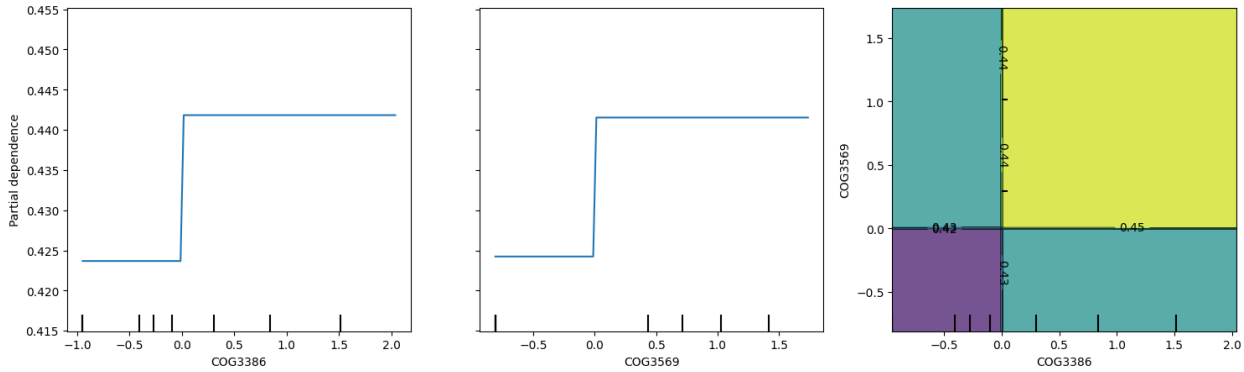

Fig. 9. Partial dependence plots (PDP) of the top 2 most important features of the XGBoost model. Feature names are shown on the x-axis.

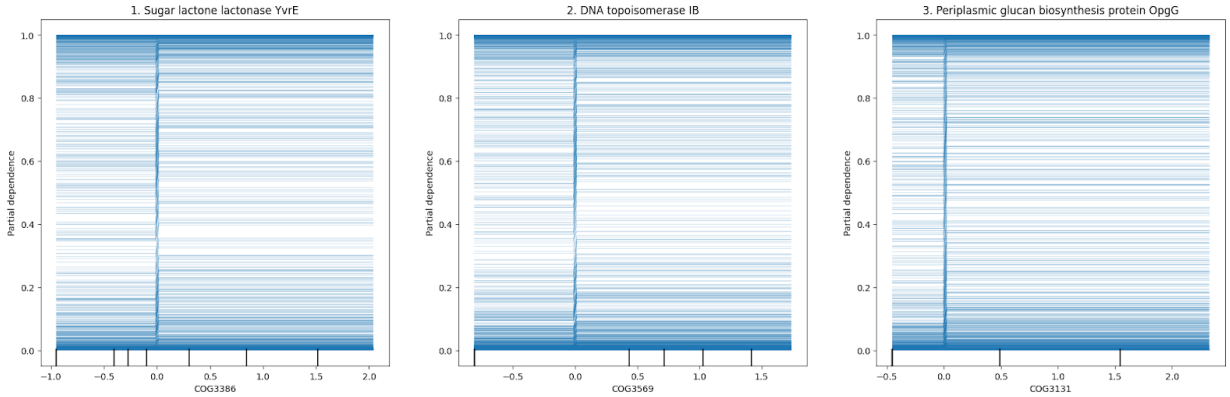

Fig. 10. Individual conditional expectation (ICE) plots of the top 3 most important features of the XGBoost model. Feature names are shown on the x-axis.

Prediction on soil bacteria data

Aftering re-training the model with whole data (both of train and test sets), we made predictions of soil bacteria, which do not have PA/NPA annotations in our dataset (**Fig. 11**). Interestingly, there was a slightly higher percentage of NPA classifications (71% NPA) in the soil population than there were in the rest of the dataset (65% NPA). We were curious whether trends were similar across bacterial orders. It appears that for some taxa (such as *Pseudomonas*) there is a much lower percentage of plant association relative to the total dataset. Only one taxon has the opposite trend; *Actinobacteria*, where there is a soil bacteria enrichment relative to the entire dataset.

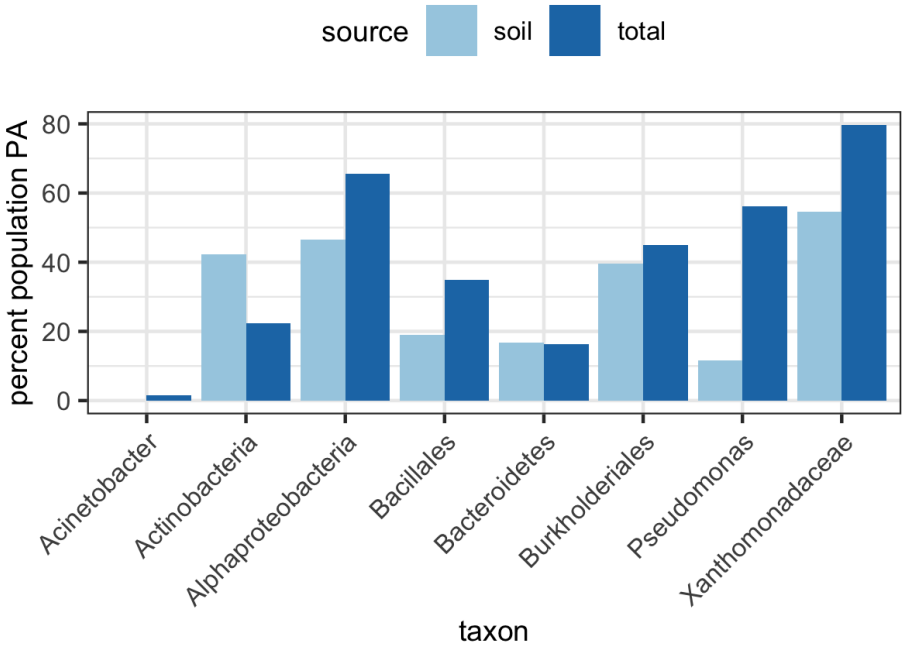

Fig. 11. Percentages of PA-classified bacteria in each phylum represented in the dataset. Counts in light blue are the percentage of PA classifications in the soil dataset. Dark blue bars represent the percentage of PA in the total dataset.

Discussion

There has been a burgeoning interest in developing machine learning classifiers to identify bacteria that are associated with or may be capable of being beneficial to plants, with applications in agriculture and microbiome engineering. To date, classifiers have only been built for members of specific bacterial genera (te Molder et al., 2021; Saati-Santamaria et al., 2022), or only small datasets that may not be comprehensive (n = 431) (Martínez-García et al., 2016). Herein, multiple supervised learning models were trained on a dataset containing almost ten times the number of bacterial genomes as Martínez-García et al. (n = 3,772) of known plant or non-plant association. Because each data point had 4,540 features (after preprocessing), we were concerned about overfitting the data. Therefore, we performed feature selection by selecting the union of features identified by different methods, and reduced the number of features approximately tenfold to 440 (see Feature Selection section and **Fig. 3**, **Fig. 4**, **Fig. 5**).

For our supervised binary classifier, we tested the performance of multiple models trained on the training dataset on the test dataset. As shown in **Fig. 6**, **Fig. 8,** and **Table 1**, XGBoost was the best-performing model across most metrics, with an accuracy of 0.911 $\pm$ 0.006 and low variance across other metrics during K-fold cross validation**.** Therefore, XGBoost was selected out of all the models tested for further evaluation and understanding of feature importance. The highest-performing model observed by (Martínez-García et al., 2016) was a random forest. While our random forest model was not too much worse than the XGBoost, the XGBoost model has a similar architecture to a random forest, which may explain this result.

We then made predictions on soil bacteria to classify them as PA or NPA based on our trained models. Interestingly, a majority of these bacteria were classified as NPA. This is surprising given their proximity to plants at the isolation sources. Upon further investigation of classifications within the genus *Pseudomonas*, NPA-classified genomes were isolated from contaminated soil, or from soil sampled beside a highway whereas PA-classified samples originated from a potato field, agricultural soil, and rice paddies. Indeed, the classified members of the *Pseudomonas* were easily distinguishable via PCA (**Fig. 12**) and may therefore have specific genetic features that promote plant association, or association with other hosts. While further experimental validation is necessary, this suggests that our developed model may be able to discern between PA and NPA classified bacteria present in soil, which is, to the best of our knowledge, a novel discovery.

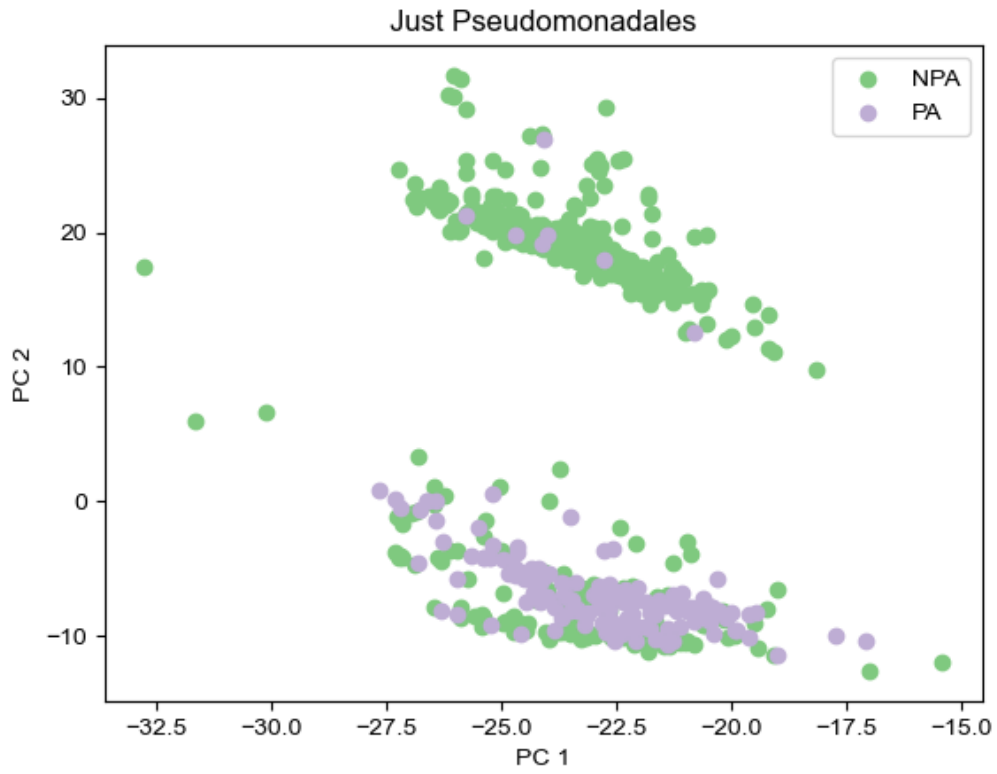

Fig. 12. PCA of the bacterial genomes belonging to the order *Pseudomonales*, colored by classification in full dataset. This PCA is simply a replot of only *Pseudomonales* genomes within the PCA shown in **Fig. 1 A**, **B**, **C**.

We also made predictions on the soil dataset with other trained models (data not provided). We found there are some discrepancies between the predictions, which makes sense, since each model assigns different importance to features. It would be interesting to see which features contributed to these inconsistencies. However, it should be noted that *in silico* analyses cannot confirm or deny PA/NPA status. Since there is no data to validate predictions, field studies are required to determine whether they are found in plants or rhizospheres, or simply present due to error (observational, transcribing, contamination, or other).

Biases and limitations of classifiers

Although our accuracy metrics were high, there are several limitations to both our ability to train and interpret these models. The main limitation to our dataset is that the classification of plant association is entirely dependent on the reported location that the sample was isolated from. Therefore this variable is subject to human error in sample handling, data entry, and classification. An additional bias that we may be introducing in our model stems from repeated species that are present in the dataset. Although the initial publication asserts that the ~3800 genomes are divergent enough to count as separate entries, there is still a possibility that the high accuracy may stem from different strains of the same species being present in the train and test splits of our dataset.

In addition, even after feature selection, our dataset contained a high number of features which made it difficult to discern which feature had the highest impact on model performance. In the

future, we may attempt to reduce the feature list further, or convert our numeric features (abundance of gene) to categorical (presence of gene).

Conclusion

In this study, we showed that genomic features (clusters of orthologous genes, COGs and carbohydrate-active enzymes, CaZys) could successfully classify bacterial genomes as PA or NPA. Although all models tested showed high accuracy metrics, our best-performing model was the XGBoost classifier. This model is highly accurate (0.91) in classifying existing data, can prioritize features essential for the classification, and can be used to predict unlabeled data. To the author's knowledge, this is the largest classification study on plant-associated bacterial genomes to-date, and has similar accuracy scores to the only other multi-order classifier study in literature. This study may help elucidate which genes or enzymes are significant for plant association, which will uncover novel host organisms for microbiome engineering to increase agricultural crop yield. Additionally, this study may help infer a PA or NPA classification from bacteria uncovered in soil.

Acknowledgment

Name (*alphabetical order*)	**Contribution**
Alex Lukasiewicz	100
Muyoung Lee	100
Sarah Coleman	100

Bibliography

Afridi, M. S. *et al.* New opportunities in plant microbiome engineering for increasing agricultural sustainability under stressful conditions. *Front. Plant Sci.* **13**, 1–22 (2022).

Ke, J., Wang, B. & Yoshikuni, Y. Microbiome Engineering: Synthetic Biology of Plant-Associated Microbiomes in Sustainable Agriculture. *Trends in Biotechnology* **39**, 244–261 (2021).

Levy, A. *et al.* Genomic features of bacterial adaptation to plants. *Nat. Genet.* **50**, 138–150 (2017).

Martínez-García, P. M., López-Solanilla, E., Ramos, C. & Rodríguez-Palenzuela, P. Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environ. Microbiol.* **18**, 4847–4861 (2016).

te Molder, D., Poncheewin, W., Schaap, P. J. & Koehorst, J. J. Machine learning approaches to predict the Plant-associated phenotype of *Xanthomonas* strains. *BMC Genomics* **22**, 848 (2021).

Saati-Santamaría, Z., Baroncelli, R., Rivas, R. & García-Fraile, P. Comparative Genomics of the Genus *Pseudomonas* Reveals Host- and Environment-Specific Evolution. *Microbiol. Spectr.* **10**, 1–15 (2022).

Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

URLs

1. Data website by Levy et al., 2017**:**

http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/index.html

http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/faa_trees_metadata.html

2. This study's GitHub repository**:**

https://github.com/ajlukasiewicz/Team_9_MacLearn

3. JGI Genome Database

https://genome.jgi.doe.gov/portal/