Michigan Technological University

Michigan Technological University
Digital Commons @ Michigan Tech

Dissertations, Master's Theses and Master's Reports

2021

APPLICATIONS OF MACHINE LEARNING IN MICROBIAL FORENSICS

Ryan B. Ghannam
Michigan Technological University, rghannam@mtu.edu

Copyright 2021 Ryan B. Ghannam

Recommended Citation

Ghannam, Ryan B., "APPLICATIONS OF MACHINE LEARNING IN MICROBIAL FORENSICS", Open Access Dissertation, Michigan Technological University, 2021.
https://doi.org/10.37099/mtu.dc.etdr/1196

Follow this and additional works at: https://digitalcommons.mtu.edu/etdr

 Part of the Biochemistry Commons, Bioinformatics Commons, Biotechnology Commons, Computational Biology Commons, Genomics Commons, Molecular Biology Commons, and the Other Ecology and Evolutionary Biology Commons

APPLICATIONS OF MACHINE LEARNING IN MICROBIAL FORENSICS

By

Ryan B. Ghannam

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Biochemistry and Molecular Biology

MICHIGAN TECHNOLOGICAL UNIVERSITY

2021

© 2021 Ryan B. Ghannam

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Biochemistry and Molecular Biology.

Department of Biological Sciences

Dissertation Advisor:	*Stephen M. Techtmann*
Committee Member:	*Paul D. Goetsch*
Committee Member:	*Carsten Külheim*
Committee Member:	*Ebenezer Tumban*
Department Chair:	*Chandrashekhar P. Joshi*

Table of Contents

List of Figures

List of Tables

List of Algorithms

Author Contribution Statement

A statement of contribution as part of the original work presented in this dissertation:

R.B.G., Ryan B. Ghannam. S.M.T., Stephen M. Techtmann., T.M.B., Timothy M. Butler., L.G.S., Laura Grace Schraer.

R.B.G. and S.M.T. designed the overall studies. R.B.G., S.M.T., and T.M.B. undertook field work. R.B.G. and L.G.S. processed the samples. R.B.G. performed molecular methods for DNA library preparation and DNA sequencing. R.B.G. developed theory and performed computation for machine learning, statistical and other quantitative data analysis. R.B.G. and S.M.T. wrote the manuscripts as presented here with assistance from L.G.S. and T.M.B. S.M.T. aided in interpreting results and oversaw the research.

Work from part of this dissertation come from the following publications:

Chapter 1

Machine learning Applications in Microbial Ecology, Human Microbiome Studies, and Environmental Monitoring

Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal. 2021.

Credit authorship and contribution statement
R.B.G.: Conceptualization, Writing - original draft, Writing - review & editing, Visualization. S.M.T.: Conceptualization, Writing - original draft, Writing - review & editing, Funding acquisition.

Chapter 2

Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities

Ghannam RB, Schaerer LG, Butler TM, Techtmann SM. Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities. Msphere. 2020;5(1).

Credit authorship and contribution statement
S.M.T. and R.B.G. designed the study. R.B.G., S.M.T., and T.M.B. undertook field work. R.B.G. and L.G.S. processed the samples. R.B.G. performed DNA sequencing, data analysis, machine learning, and statistical analysis and wrote the manuscript with assistance from S.M.T., L.G.S., and T.M.B. S.M.T. oversaw the research.

Acknowledgements

This dissertation would have not come to successful conclusion without the encouragement, help, support and trust of colleagues, friends and family.

Foremost, I would like to sincerely thank my advisor Stephen M. Techtmann for his mentorship and freedom I was granted to explore science. I am grateful to all committee members for assistance in evaluating and critiquing my work and this dissertation.

I also thank the Bioconductor team and all its contributors, along with Max Kuhn for developments in machine learning. Open-source developments heavily contributed to this work and helped shape my vision of approaching complex biological questions at scale.

This work was sponsored by DARPA Young Faculty award D16AP00146. I would also like to thank the those who helped coordinate the field work of this project in a large way. In particular, Jamey Anderson and Christopher Pinnow, Thorsten Brinkhoff, Gian Marco Luna, Stanley Lau, and Sukhwan Yoon.

List of Abbreviations

ML	Machine Learning
USML	Unsupervised Machine Learning
SML	Supervised Machine Learning
tSNE	t-distributed Stochastic Neighbor Embedding
PCoA	Principal Coordinate Analysis
ASV	Amplicon Sequence Variant
RF	Random Forest
SVM	Support Vector Machine
GB	Gradient Boosting
ANN	Artificial Neural Network
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
$\log_{loss}$	Logistic Loss
Y	Class Label
$\hat{Y}$	Predicted Label
N	Sample/Observation
X	Feature
*R*U	Interaction Impact
Ri	Constituent Impact
Fi	Constituent Taxon
*F*U	Full Feature Space

Data and Code Availability

The National Center for Biotechnology Information (NCBI) Sequence Read Archive has archived the raw sequencing data and associated metadata used in this study under the accession numbers PRJNA542890 and PRJNA542685 (for *Chapter 2*). All code used for statistical analysis, machine learning, figures, and other relevant data necessary for these proposed workflows or that support the findings in dissertation (e.g., *Supplementary Data*) are available from me or corresponding authors upon reasonable request, and through GitHub (https://github.com/rghannam).

Abstract

Microbial ecosystems are complex, with hundreds of members interacting with each other and the environment. The intricate and hidden behaviors underlying these interactions make research questions challenging – but can be better understood through machine learning. However, most machine learning that is used in microbiome work is a black box form of investigation, where accurate predictions can be made, but the inner logic behind what is driving prediction is hidden behind nontransparent layers of complexity.

Accordingly, the goal of this dissertation is to provide an interpretable and in-depth machine learning approach to investigate microbial biogeography and to use micro-organisms as novel tools to detect geospatial location and object provenance (previous known origin). These contributions follow with a framework that allows extraction of interpretable metrics and actionable insights from microbiome-based machine learning models. The first part of this work provides an overview of machine learning in the context of microbial ecology, human microbiome studies and environmental monitoring – outlining common practice and shortcomings. The second part of this work demonstrates a field study to demonstrate how machine learning can be used to characterize patterns in microbial biogeography globally – using microbes from ports located around the world. The third part of this work studies the persistence and stability of natural microbial communities from the environment that have colonized objects (vessels) and stay attached as they travel through the water. Finally, the last part of this dissertation provides a robust framework for investigating the microbiome. This framework provides a reasonable understanding of the data being used in microbiome-based machine learning and allows researchers to better apprehend and interpret results.

Together, these extensive experiments assist an understanding of how to carry an in-silico design that characterizes candidate microbial biomarkers from real world settings to a rapid, field deployable diagnostic assay. The work presented here provides evidence for the use of microbial forensics as a toolkit to expand our basic understanding of microbial biogeography, microbial community stability and persistence in complex systems, and the ability of machine learning to be applied to downstream molecular detection platforms for rapid and accurate detection.

1 Machine learning Applications in Microbial Ecology, Human Microbiome Studies, and Environmental Monitoring

Preface

It is common for microbiome researchers to employ machine learning to investigate their research questions. This section provides an overview of many of these studies as they relate to our health and built environments. We discuss a variety of machine learning algorithms that are common to microbiota analysis – along with a comparison to traditional multivariate statistics. Additionally, this section compares open-source machine learning toolkits that can be used to investigate microbial data. Following this, we present an argument for the need for reporting interpretable metrics in microbiome-based machine learning studies. As reproducibility is a concern in studies that employ machine learning, more interpretive open-source software should be acknowledged as an integral part of the modern workflows of investigating microbiota.

This section aimed to review machine learning in the context of microbes as they relate to our health and built environments. Provided was an in-depth overview of microbiome studies employing a variety of machine learning algorithms – from microbial ecology to the human microbiome and environmental monitoring. We then proceeded to compare the machine learning algorithms (supervised and unsupervised) used in these studies, with brief mention of advantages over traditional multivariate statistics. Additionally, provided was a thorough comparison of open-source toolkits that can be used for predictive and exploratory machine learning modeling of microbial datasets. Our review followed with mentions of shortcomings of common machine learning practice in the experimental microbiome literature, and how machine learning interpretation could be improved and reported. As reproducibility is a concern in studies that employ machine learning, more interpretive open-source software should be acknowledged as an integral part of the modern workflows of investigating microbiota.

Abstract

Advances in nucleic acid sequencing technology have enabled expansion of our ability to profile microbial diversity. These large datasets of taxonomic and functional diversity are key to better understanding microbial ecology. Machine learning has proven to be a useful approach for analyzing microbial community data and making predictions about outcomes including human and environmental health. Machine learning applied to microbial community profiles has been used to predict disease states in human health, environmental quality and presence of contamination in the environment, and as trace evidence in

forensics. Machine learning has appeal as a powerful tool that can provide deep insights into microbial communities and identify patterns in microbial community data. However, often machine learning models can be used as black boxes to predict a specific outcome, with little understanding of how the models arrived at predictions. Complex machine learning algorithms often may value higher accuracy and performance at the sacrifice of interpretability. In order to leverage machine learning into more translational research related to the microbiome and strengthen our ability to extract meaningful biological information, it is important for models to be interpretable. Here we review current trends in machine learning applications in microbial ecology as well as some of the important challenges and opportunities for more broad application of machine learning to understanding microbial communities.

1.1 Introduction

Environmental microbial communities are extremely diverse and play a role in driving many biogeochemical cycles and regulating human health. These environmental communities also have various applications in biotechnology. The ability to probe microbial diversity has been enabled through the increasing availability of high throughput sequencing (HTS) technologies. The microbial diversity of the human microbiome as well as soil and ocean microbial communities has been expanded through large-scale collaborative sequencing efforts such as the Human Microbiome Project [1] and the Earth Microbiome Project [2] as well as the TARA OCEANS project [3]. These large-scale efforts have provided baseline data for the microbial communities found in diverse settings. The low cost of sequencing now allows for large scale studies of systems and the generation of microbial community profiles for hundreds and thousands of samples. This scale of data necessitates methods capable of extracting meaningful information from these large datasets. Natural microbial communities have the potential to provide key insights into environmental phenomena and may be useful in predicting environmental phenomena. Machine learning (ML) has been employed to find patterns in data that can be predictive of various phenomena. In recent years machine learning has been applied to microbial community data to classify samples and predict various outcomes [4], [5], [6]. There is potential for expansion of the use of ML for microbial ecology studies. In this review, we seek to provide an overview of ML applications in microbial ecology and present some challenges and opportunities for the expansion of ML applications in the study of microbial communities.

1.2 Next-generation sequencing methods in microbial ecology

Molecular methods have been used in microbial ecology for decades employing sequencing of ribosomal RNA genes to profile microbial diversity in settings

ranging from soil and aquatic environments to hydrothermal vents and the built environment [7], [8], [9]. With the expansion of high throughput sequencing, the ability to generate thousands of sequences from hundreds of samples in a single sequencing run is possible [10], [11]. The application of next generation sequencing in microbial systems follows a pipeline that includes both wet lab and computational methods (**Fig. 1.1**). A goal of molecular profiling of microbial communities is to obtain a comprehensive assessment of the taxonomic and functional diversity of a community. In order to obtain this assessment, there are a number of important considerations that must be addressed during the analysis pipeline. The pipeline starts with wet lab methods for molecular profiling of microbial communities, which involves sample collection, extraction of nucleic acids from the environment or host and library preparation for sequencing **Fig. 1.1**. There are a number of biases that can be introduced with the wet lab portion of the methods [12]. In particular, extraction of DNA with different methods can result in differential extraction efficiencies for different taxa and thus has the potential to skew diversity assessments. Often sequencing of DNA for microbial community profiling can take the form of marker gene surveys which profile the diversity of either taxa (small subunit rRNA such as the 16S rRNA for bacteria and archaea and 18S rRNA for eukaryotes) or of a particular functional gene. The choice of sequencing primers for marker gene surveys can also introduce bias as degenerate primers are not truly universal and may miss key microbial groups. These biases are important to consider in planning study design. Alternatively, shotgun metagenomic methods can be employed to profile the complement of genes that are present in a sample.

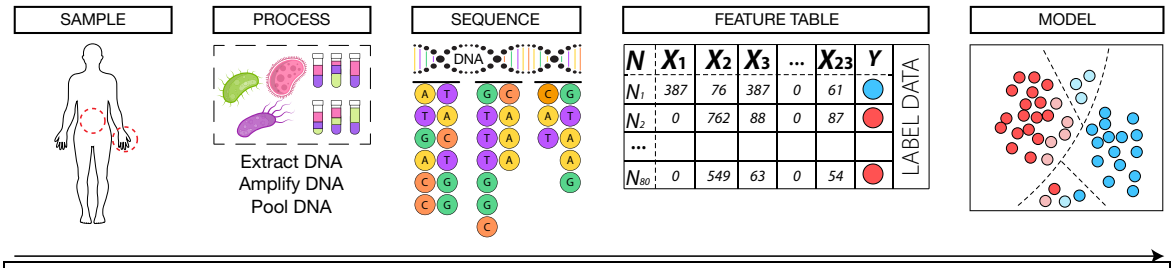


Figure 1. 1

Illustrative pipeline for the investigation of microbial communities using metagenomics.

Sequencing depth is another key consideration in the process of profiling microbial communities. Sequencing is a sampling-based approach. Therefore, with increased sequencing depth, the diversity of reads is more completely sampled and thus diversity estimates are more reflective of the natural system. After sequencing, the primary analyses are computational. While the computational portion of the pipeline greatly depends on the goals of the study, often this portion is divided into sequence processing to generate a table of

samples and taxonomic features in each sample followed by analysis methods to assess and link microbial diversity with various outcomes. A lot of work has been done related to processing of sequencing reads into meaningful data tables. This has included methods for binning marker genes into operational taxonomic units (OTUs). These OTUs are features that are representative of some biologically meaningful categories. Methods such as UCLUST [13] (often implemented in QIIME1 [14]) and mother [15] bin 16S rRNA reads based on percent identity to other reads in the dataset into OTUs. More recent methods have sought to cluster sequences into identical groups rather than cluster by some fixed percent identity. These approaches such as DADA2 [16] have employed denoising algorithms to correct sequencing errors and then dereplicate sequences into bins of identical sequences known as Amplicon Sequencing Variants (ASVs) or Exact Sequence Variants (ESVs). Each of these methods has advantages and limitations. They are all similar in that they are approaches for grouping marker gene sequences into biologically meaningful bins that result in feature tables for downstream analysis.

Much of the downstream analyses of feature tables generated from microbial community data has focused on the application of commonly used ecological measures for processing microbial community data. Methods such as alpha diversity assessments as well as beta diversity and multivariate statistics have been commonly used. A number of issues have been identified related to the application of methods designed for datasets with tens of features to highly dimensional datasets with thousands of features [17]. For example, specific diversity metrics have been shown to be highly impacted by the dimensionality and scale of the data, while others are less prone to errors resulting from highly dimensional data. Additional metrics, such as UniFrac distances, were developed that allowed researchers to more fully extract meaningful information from these marker gene surveys [18]. However, many of these methods seek to understand the data through decreasing dimensionality of the data and often can lose some of the important information that is contained within the rich datasets of microbial community profiles. For example, principal coordinate analysis (PCoA) is commonly used to assess overall differences in diversity. PCoA analysis is performed using distance or dissimilarity matrices of the microbial community profiles using metrics such as UniFrac distance or Bray-Curtis dissimilarity. While useful, these methods collapse the highly dimensional datasets and assess overall similarity or dissimilarity. This process can often lose important information and bias observations to highly abundant or highly prevalent features.

OTU-level analyses have also been important for analyzing the relationship between particular features in microbial data and specific outcomes. Indicator Species Analysis has been important for environmental monitoring. In Indicator Species Analysis, the prevalence of a species or OTU is linked to particular treatments or environmental states. Each OTU is given an indicator value (IndVal)

which details how indicative that species is of the particular outcome. Additionally, differential abundance analysis has often been used to better understand differences between samples, categories, or particular outcomes on the level of particular features. Methods such as DESeq2 [19] and metagenomeSeq [20] have been used to identify which features are differentially abundant between different categories. DESeq in particular was originally developed to understand differentially expressed genes in RNASeq datasets. One advantage of the use of these methods for differential abundance analysis is that these approaches have been designed to work well with sequencing datasets and use normalization approaches tailored specifically to sequencing data. One of the limitations of differential abundance analysis is the ability to understand the importance of multiple features or the interaction of features in a particular outcome. Differential abundance analyses treat features as independent and it can be difficult to glean how increased or decreased abundance of groups of features may be a hallmark of a particular sample type or treatment category.

In addition to the methods described above, advanced computational methods are being employed to assist in analyzing the increasing amounts of data. In particular ML is being used with increasing frequency to use microbial communities to predict different outcomes. ML has advantages in that it is able to more fully appreciate the depth of data generated in microbiome studies as well as build predictive models for outcomes based on microbial community data. In the following sections we will provide an overview of commonly used ML methods, discuss key steps to be considered in the ML process and provide examples of how ML can be used in microbiome studies in the human microbiome, environmental monitoring, and forensics.

1.3 Machine learning and microbial community data analysis

In the context of microbial ecology, applied machine learning involves creating and evaluating models that use algorithms capable of recognizing, classifying and predicting specific outcomes from data. ML approaches take various forms including unsupervised, semi-supervised, reinforced, or supervised learning [21]. For example, often the goal of supervised machine learning (SML) applied to microbial community data is to construct a decision rule (i.e. a *model*) from a set of collected observations (i.e. samples) to predict the condition (i.e. *response label* (Y); such as an assigned category or value to each observation that have meaning to the model-operator) of an unlabeled sample using a set of measurements from next generation sequencing instruments. In microbiome studies this input data takes the form of a frequency count matrix of the observed microbial taxa from a sample (i.e., *input variables* (X) and their assigned values). Input variables are often referred to as *features* and samples as *observations* and will be used interchangeably throughout this review.

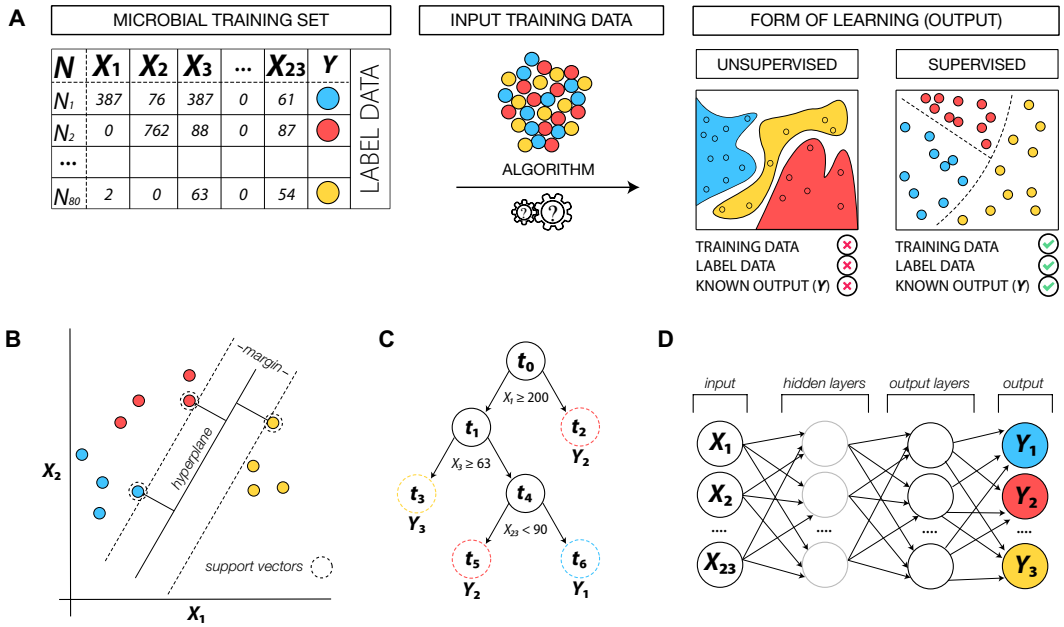


Figure 1. 2

Schematic representation of unsupervised and supervised forms of learning and several ML methods predicting three conditional response labels (blue/red/yellow). (A) Depicts a common microbial frequency matrix containing observations or samples (*N*), features (*X1, …, X23*) and multiple class labels (*Y*). Input data are algorithmized and processed to either predict which cluster *Y* belongs to (unsupervised) or to find a best fit decision boundary between *X* and *Y* (supervised). (B) Linear SVM classifier demonstrating separation between class labels where the hyperplane maximizes the distance (margin) between the nearest data training points. Support vectors refer to the three position vectors drawn from the origin of the sample positions (dashed circle) with the goal of maximizing the distance between the optimal hyperplane and the support vectors (max-margin) so that a decision boundary can be drawn. (C) A decision tree constructed for the classification of samples into *Y* based on input feature values. Trees start from a root node (t_0) and are grown to various leaf nodes (closed circle) to end at a terminal node (dashed circle) so that bootstrap aggregated predictions across terminal nodes are averaged across k-trees for best predictions of $\hat{Y}$. (D) A neural network displaying the structure of successive layers. Input values of *X* are transmitted to the proceeding hidden layer which passes weighted connections to the output layer for predictions of $\hat{Y}$.

While there are other forms of learning that have been used on microbial community data, this manuscript discusses unsupervised techniques and supervised machine learning methods commonly applied to microbial datasets. The principal distinction between unsupervised (USML) and supervised machine learning (SML) is that in USML samples are segregated using features without any

reference to response labels and the prediction is to which cluster a response may belong to, whereas the SML finds a best fit decision boundary between features and response labels [22] (**Fig. 1.2A**). A more precise overview of these methods is introduced below.

1.3.1 Unsupervised multivariate analysis common to marker-gene analysis

Unsupervised techniques are often employed for initial exploratory analysis of high-dimensional metagenomic data and for generating hypotheses for subsequent analysis as they aid in visualization and can search for structure in data that do not have predefined response labels assigned to observations. These methods operate with the goal of identifying homogenous subgroups by clustering data (hierarchical or centroid) or to detect anomalies by finding patterns through dimensionality reduction (DR) techniques. An example of DR by some unsupervised techniques (Principal Coordinate Analysis: PCoA and t-Distributed Stochastic Neighbor Embedding: t-SNE) is to take data points from a high-dimensional feature set and project them in low-dimensions to encapsulate the largest amount of statistical variance in a set of observations while preserving structure and minimizing information loss [23]. In USML, input data is either continuous data of features for each observation or a distance matrix of similarities between community composition. Observations that cluster together in USML represent microbial communities from samples that are more similar in composition.

1.3.1.1 K-means clustering (centroid)

The objective of K-means [24] is to cluster samples into a specified number of (k) non-overlapping subgroups (clusters) using distances calculated between features so each data point belongs to only one group. This technique assigns data points to a cluster such that the sum of squared distance between data points and the centroid (average of all data points represented by the geometric center of the cluster) is minimized. By reducing intra-cluster variation, data points are arranged to construct a cluster that assumes a spherical shape surrounding the centroid and allows different subgroups of data to remain as far apart as possible. A drawback of K-means is that it cannot construct clusters well on data points that are distanced to a more complex geometric shape. An additional constraint is that a pre-defined number of clusters is required, which necessitates assumptions to be placed on the structure of data prior to analysis.

1.3.1.2 Principal Coordinate Analysis (PCoA)

In PCoA analysis, data are decomposed into components to maximize the linear correlation between data points in a dissimilarity matrix, such as microbial taxa as input features [25]. Through a "coordinate transform", x number of data points are

replaced to newly derived y coordinates, thus reducing the dimensionality of a dataset by discarding the coordinates that may not capture a threshold of variance in the microbial community data. This technique preserves the global structure of the data while projecting it to low dimension. By mapping nearby points to each other and faraway points to each other, linear variance in the global relationships of the data are maximized to retain a faithful representation of the actual distances between original data points [23]. This method works from distance matrices calculated from biologically meaningful metrics such as UniFrac and is commonly employed in microbial analysis [26].

1.3.1.3 t-Distributed stochastic neighbor embedding (t-SNE)

In t-SNE analysis, data points are transformed and assigned a probability based on similarity to define relationships in high-dimension, guided by a Students t-distribution to help reduce crowding of data points during visual projection [27]. As the name entails, this method tries to identify close neighbors (samples with similar measurements) and tries to arrange these points in a low-dimension projection such that the close neighbors remain close and distant points remain distant. In contrast to PCoA that tends to preserve long distances for global retention of the original data, t-SNE tries to represent local relationships in the data, thus capturing non-linear variance and is not as faithful to the original state of the data [23]. Certain fields that use high-dimensional data currently benefit from the local non-linear structure of this method, such as single cell RNA-seq [28]. T-SNE is not commonly employed for metagenomic data despite its utility as a promising exploratory technique for the analysis of microbial communities [29], [30].

1.3.2 Supervised machine learning methods common to microbiome study

Supervised machine learning (SML) is a more elaborate form of exploring marker-gene datasets since unlike unsupervised methods, response labels (Y) are assigned to each sample in the dataset, grouping them into meaningful categories. A more targeted investigation of data can be achieved since the model is being trained to learn the structure of features (X) (*training set*) to create rules where they can serve as predictors of phenomena or outcome. In other words, which feature (X) maps to the response label (Y). Once trained, this model can intake new unlabeled samples with similar features (*testing set*) and predict their output (Y) based on what it has learned from the training set. SML can be used with continuous numerical outputs (regression: $Y = \mathbb{R}$.; continuous traits such age, blood pressure, concentration of contaminant) or categorical outputs (classification: $X \rightarrow Y$; binary or symbolize grouped conditions such as 'diseased' or 'healthy'). The following section seeks to provide an overview of some of the most common SML algorithms for microbiome-based prediction tasks (outlined as implemented methods

in **Table 1.1**). We have focused our overview to primarily classification approaches, although this collection of methods could apply to regression as well.

Table 1. 1

Summary of ML techniques used for microbiome-based prediction tasks. This table briefly summarizes each technique, provides the source of the software, noteworthy ML implementations and interpretation of its result with reference to either the source study or specific studies that have applied these techniques for microbiome profiling. This table is not exhaustive but mentions current and commonly employed ML and ML related pipelines tailored to the characteristics of microbiome data or that are domain agnostic but relevant to research questions relating to the microbiome.

Software name	Summary	Source	Example Impleme-ntation	Remarks	URL
SIAMCAT (*)	**S**tatistical **I**nference of **A**ssociations between **M**icrobial **C**ommunities **A**nd host pheno**T**ypes	R package 'SIAMCAT' https://siamcat.embl.de/	FS, ML, INTERP, VIS	Confounder analysis Enables cross-study comparison Advances visualization	https://www.biorxiv.org/content/10.1101/2020.02.06.931808v2
DeepMicro (*)	Deep representation learning for disease prediction based on microbiome data	Python: https://github.com/minoh0201/DeepMicro	DR, ML	Deep representation learning using autoencoders to handle high-dimensional data Accelerates model training and hyperparameter optimization	https://www.nature.com/articles/s41598-020-63159-5
MetAML (*)	**Met**agenomic prediction **A**n alysis based on **Ma**chine **L** **e**arning	Python: https://github.com/segatalab/metaml	FS, ML, INTERP, VIS	Enables cross study comparison of models on single cohorts, across stages of same the same study and across different studies.	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977
mAML (*)	An automated machine learning pipeline with a microbiome repository for human disease classification	Python: https://github.com/yangfenglong/mAML1.0 Web: http://lab.malab.cn/soft/mAML/	FS, ML, INTERP, VIS	Automates optimized, interpretable and reproducible models Deployed on a user-friendly web-based platform Advanced visuals	https://pubmed.ncbi.nlm.nih.gov/32588040/
BiomMiner (*)	An advanced exploratory microbiome analysis and visualization pipeline	Docker: https://mbac.gmu.edu/mbac_wp/biomminer-readme/	FS, DR, ML, INTERP, VIS	Automatically tunes optimal hyper-parameters Tailored to clinical datasets Generates web-enabled visuals	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234860
MIPMLP (*)	Microbiome Preprocessing Machine Learning Pipeline	Python: https://github.com/louzounlab/microbiome/tree/master/Preprocess Web: http://mip-mlp.math.biu.ac.il/Home	FS, DR, ML, INTERP, VIS	Approaches for standardized ML preprocessing Consensus methods for optimal performance	https://www.biorxiv.org/content/10.1101/2020.11.24.397174v1.full#ref-12
MicrobiomeAnalystR (*)	**C**omprehensive statistical, functional,	R package 'MicrobiomeAnalystR'	FS, DR, ML, INTERP, VIS	Comprehensive analysis reporting	https://www.nature.com/artic

	and meta-analysis of microbiome data	Web: https://www.microbiomeanalyst.ca/		Real time feedback and recommendations Visual comparison with a public dataset	les/s41596-019-0264-1
Meta-Signer (*)	**Meta**genomic **Sign**ature Identifi**er** based on Rank Aggregation of Features	Python: https://github.com/YDaiLab/Meta-Signer/tree/master/src	FS, ML, INTERP	Ensemble learning for feature ranking Identifies a robust set of highly informative taxa	https://www.biorxiv.org/content/10.1101/2020.05.09.085993v1
QIIME2 (*)	**Q**uantitative **I**nsights **I**nto **M**icrobial **E**cology	https://qiime2.org/	FS, DR, ML, INTERP, VIS	Automatic tracking of data provenance Multiple user interfaces Plugin support	https://www.nature.com/articles/s41587-019-0209-9
mothur (*)	Microbial community analysis pipeline	http://mothur.org/	FS, DR, ML, INTERP, VIS	Can handle data from multiple sequencing platforms Encapsulates large elements of the pipeline in single command	https://aem.asm.org/content/75/23/7537
scikit-learn	Simple and efficient tools for predictive data analysis	Python: https://scikit-learn.org/stable/	FS, DR, ML, INTERP, VIS	Robust machine learning library and support system Supports end-to-end projects with extensive documentation	https://arxiv.org/abs/1201.0490
Keras	Simple deep learning API	R package 'keras' Python: https://pypi.org/project/Keras/	FS, DR, ML, INTERP, VIS	High-level learning API that limits the number of user actions Multiple deployment capabilities Provides clear and actionable error messages	https://link.springer.com/chapter/10.1007/978-1-4842-2766-4_7
caret	**C**lassification **A**nd **RE**gression **T**raining	R package 'caret'	FS, DR, ML, INTERP, VIS	Streamlines complex predictive tasks Large library of available models	http://topepo.github.io/caret/index.html
mlr	Machine learning in R	R package 'mlr3'	FS, DR, ML, INTERP, VIS	Modern and extensible ML framework for	https://joss.theoj.org/papers/10.21105/joss.01903

		https://mlr3.mlr-org.com/		developers and practitioners Provides a unified interface to many learners	
H2O.ai	Fast scalable ML API	R package 'h2o' Python: http://h2o-release.s3.amazonaws.com/h2o/rel-zermelo/3/index.html	FS, ML, DR, INTERP, VIS	End-to-end engine specialized for big data Parallel distributed ML algorithms Automatic ML interface	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238648
iml	Interpretable machine learning	R package 'iml'	FS, ML, INTERP, VIS	Feature effects on the influence of predictions	https://joss.theoj.org/papers/10.21105/joss.00786
LIME	Local interpretable model-agnostic explanations	R package 'lime' Python: https://github.com/marcotcr/lime	FS, ML, INTERP, VIS	Explains individual predictions of a black box ML model Model-agnostic	https://arxiv.org/abs/1602.04938
inTrees	Interpretable tree ensembles	R package 'inTrees'	FS, ML, INTERP	Extracts, measures, prunes, selects and summarizes rules from a tree ensemble Specific to decision trees	https://link.springer.com/article/10.1007/s41060-018-0144-8
dtreeviz	Decision Tree Visualization	Python: https://github.com/parrt/dtreeviz	FS, ML, INTERP	Advanced visualizations Provides user-friendly interpretations of prediction paths Specific to decision trees	https://explained.ai/decision-tree-viz/index.html
ranger	**RAN**dom forest **GE**ne**R**ator	R package 'ranger'	FS, ML, INTERP	Fast implementations of random forests optimized for high-dimensional data Has advanced and convenient functions for decision trees	https://arxiv.org/abs/1508.04409
partykit	A toolkit for recursive partitioning	R package 'partykit'	FS, ML, INTERP	Can coerce tree models from different sources into a unified infrastructure Contains a variety of novel decision	https://dl.acm.org/doi/10.5555/2789272.2912120

				tree implementations Parameterization requires expertise	

FS, Feature Selection; DR, Dimensionality Reduction; ML, Machine Learning; INTERP, Interpretation Measures; VIS, Visualization Outputs. (*) Denotes whether the software Is microbiome-specific (as opposed to domain agnostic).

1.3.2.1 Random Forests (RF)

Random forests [31] have been extensively deployed to solve a variety of problems in microbiota analysis. This method constructs multiple forests composed of decision trees by using the information contained in input features (abundance of microbial taxa, for example) to successively split samples based on their assigned (Y) values. The forests are guided by bootstrapping (drawing a random subset of samples with replacement, to be drawn multiple times) and a node splitting criterion that uses the information contained in a random subset of features to decide how to split each node in each tree (**Fig. 1.2C**), where the best split is selected based on a node impurity estimate (the likelihood of misclassifying new samples as a classifier) or the prediction squared error (as a regressor) [32]. The fact that hundreds or thousands of decision trees are being constructed in each forest using a subset of both samples and features allows an aggregate average of the predictions made at each terminal node (**Fig. 1.2C**). The combination of bootstrapping and then aggregating is jointly known as bagging (bootstrap aggregating) and frames RF as an ensemble learning method, where multiple forests are leveraged to obtain better performance than any single decision tree alone [33]. By this effect, RF are an ideal framework for consistently identifying "true effects" in complex and heterogenous data (multiple feature types; numerical or categorical). Additional factors that make RF appealing in practice is that they are an off the shelf, computationally tractable and top performing classifier that are robust to outliers, inherently noisy and non-linear data (such as metagenomic), and errors in manually curated response labels [34], [4]. Using SML with highly dimensional data with limited numbers of observations, such as microbial community data, can lead to overfitting. The RF method is less prone to overfitting than other SML methods, which contributes to its appeal in microbial community analysis [35]. Lastly, decision tree models in general are considered interpretable in their evaluation as they aid in extracting meaningful information from RF models [36] (**Fig. 1.3**).

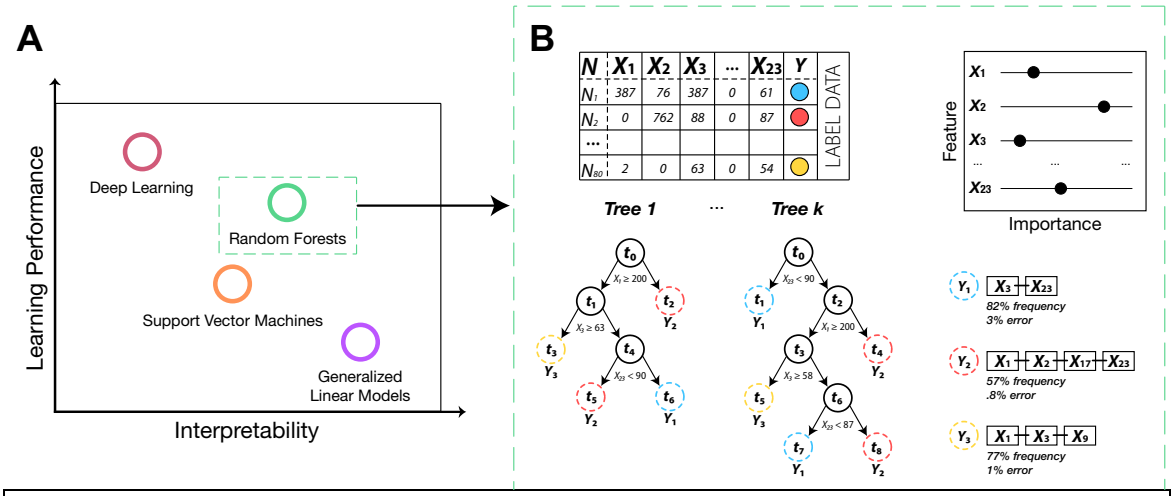


Figure 1. 3

Depiction of performance-interpretability trade-off and random forests interpretation. Note that these figures are fictional and are not based on experimental quantification (the axes in this figure lack meaning). (A) Performance-interpretability tradeoff of commonly deployed algorithms in microbiome research. However, in practice, the models characterized here tend to varying degrees of accuracy and interpretability based on experimental procedure. Had a plot been generated from experiment, model choice and complexity could vary such that inconsistent illustrations could arise. By way of example: tuning models to become more accurate could result in the belief that more accurate models are less interpretable and may not respect whether the model infrastructure supports inherently easier interpretation. (B) Hypothetical extraction of 'association' rules that measure frequent microbial community member interactions from fictional decision tree ensembles (***tree1, ..., tree k***) for low error predictions of $\hat{Y}$. Additionally diagramed is a feature 'importance' schematic that scores each feature on its relative importance in making predictions of $\hat{Y}$.

1.3.2.2 Gradient Boosting (GB)

Gradient boosting [37], when used for decision trees, is an ensemble method that uses a process called boosting to combine individual learning algorithms (decision trees) successively to arrive at a strong learner. Gradient boosted trees contrast RF as an ensemble learner in that each decision tree is constructed in series in attempt to reduce the errors of the preceding tree, rather than in parallel. In addition, each tree built in GB is a fixed size and is fit on the original data, instead of bootstrapping samples as done in RF. Similar to RF, both numerical and categorical features can be used, but may be harder in practice to find optimal tuning parameters for a good model fit, such as the number of tree estimators. This method in particular is sensitive to outliers but efficient for both classification and regression, with reports of achieving similar or better accuracy to RF [38].

1.3.2.3 Support Vector Machines (SVM)

The goal of an SVM [39] is to find the best generalized line separation of response labels (Y) through a hyper-plane that maximizes the margin between different values of Y (or each class) in the label data. A decision boundary is drawn such that each class is separated while keeping maximum distance from the closest samples as possible (called support vectors that dictate this decision boundary) (**Fig. 1.2B**). SVMs are in the category of linear discriminant SML techniques. Although, if a hyperplane cannot justify the separation between classes with a clear margin of separation (in the case of non-linear metagenomic data), a so called "kernel trick" for a nonparametric form of SVM can be introduced to transform the data and satisfy a non-linear separation [40]. Several factors contribute to the success of SVM in microbial community analysis in that it is effective in the high-dimensional nature of the data, where $X > N$ (or the number of members of a microbial community as features is larger than the sample set) and that it is also computationally tractable since the decision function only uses a subset of the data. These models can handle various feature types but can be inherently hard to interpret as they do not directly provide probability estimates in their evaluation.

1.3.2.4 L2 regularized logistic regression

Regularization is a technique used to reduce overfitting. For example, if a model is parameterized to learn every small bit of information in the structure of the microbial community composition under a given set of labels in training, it may not generalize well to make predictions on samples collected and processed outside of the training set and is considered overfit. Ridge (L2) regression [41] satisfies a model that reduces variance without increasing bias and is achieved by placing restrictions on the complexity of parameters (i.e. where to ultimately draw the decision boundary to separate response labels). This technique adds information to features used in training the model and by adding a penalty term to a loss function (estimation of how wrong the relationship is between X and Y), enables a constraint on parameter complexity so as to not capture every specific detail of the training data. Ridge regression can be used for both classification and regression but can be computationally expensive in the case of large input feature space.

1.3.2.5 Neural Networks

Neural networks [42] use a hierarchical model building architecture where multiple structured networks of interconnected nodes (neurons) are constructed with weights attached at each edge of the network to facilitate mapping inputs of X to responses Y (weights being parameters to define strength of connection, for example) (**Fig. 1.2D**). Networks are interconnected through a feed-forward propagation mechanism, where each neuron receives input from preceding

neurons. The network starts from input layers (microbial taxa feature set; *X*1, *X*2, …, *X*i), that are linked to each neuron in the one or many hidden layers that use a backpropagation algorithm to maximize the weights placed at each neuron to improve predictive power. This process is iterative, where the last hidden layer is met by an output layer to produce a predicted response output (*Y*) (**Fig. 1.2D**). Neural networks are very dynamic in their ability to identify intricate structure in very high-dimensional and complex datasets, making them a tractable technique to investigate the role of microbes in complex settings [43]. Neural nets are often referred to as "black box" methods as it can be difficult to interpret how decisions are made.

1.3.2.6 Deep vs. shallow learning

Deep learning is a family of both unsupervised and supervised techniques that belong to the class of neural networks (**Fig. 1.2D**). Despite shallow networks (dependent on number of layers), all non-deep learning methods such as those summarized above can be qualified as shallow learners. Whereas deep learning methods automatically alter raw input features by successively extracting abstractions of the data to be used as more discriminative features to the learning process, shallow learners are more of a manual process that depend on domain knowledge for a reduced selection of features that would serve as good inputs for a model to make accurate predictions with (i.e., which microbial taxa are differentially abundant between response labels).

Shallow learners can also benefit from feature engineering, where new features are handcrafted as composites, or abstract representations of multiple raw features using heuristics of the domain problem (i.e., agglomerating multiple high-resolution taxonomic features (ASVs) into a single lower resolution feature (Phyla)).

Although deep learning has shown to create models that are more accurate compared with shallow learning methods for microbiome-based prediction tasks [44], the models often sacrifice interpretability or understanding of the inner logic behind the predictions, which, for microbial-based applications can be rewarding in addition to predictive accuracy. An example of learning performance-interpretability trade-off is displayed in **Fig. 1.3A**.

1.4 Advantages of machine learning vs. classical statistics for microbial community data

Microbial ecology has for long relied on traditional statistical analyses to summarize data, test hypotheses, and to interpret interactions between features and responses on microbial datasets [45]. However, researchers and developers

are starting to realize the enormous potential for machine learning in the microbial realm. ML methods have some advantages over standard statistical methods. A principal distinction between statistical models and ML is that the goal of the former is to describe and infer the relationships between variables, whereas the latter is designed to optimize the ability to predict an outcome on an external dataset. For example, typically SML will use a training set (supplied labels) to learn patterns associated with an outcome and a test set (hidden labels) to determine the performance of the model. On the other hand, statistical models are primarily interested in determining the relationship of the values to the outcome and unlike many studies that use ML, most do not require partitioning the data to measure performance.

Classical statistical analysis presents microbial ecologists with a two major issues: (1) the assumption that features of metagenomic data are independent and identically distributed is often harmed through molecular methods of sample processing and sequencing; and (2) that data with NGS imposed characteristics [46] such as being high-dimensional (number of data points is large), sparse (contain a lot of zeros), and compositional (feature set of microbial taxa may be co-abundant and are a part of a unit sum) often cannot be met by specific assumptions in classical statistics [47]. Many machine learning methods, such as the ones summarized above, can accommodate these dynamics of marker-gene data for a robust interrogation of the complex association patterns in microbial communities.

Some of the benefits ML has over classical statistics is that it is particularly effective in identifying subtle variation in microbial community structure and can identify specific bacterial taxa that underlie prediction of a conditional outcome. Another strength of ML is its ability to model a non-linear combination of bacterial count data and environmental parameters (a feature space resembling the real-world system) that do not need to assume complex transformations or preprocessing, which are challenging to molecular data.

However, since ML can operate without explicit user-instruction, is highly configurable, and requires a considerable amount of data, the tendency of these methods to overfit data are often overlooked. ML interpretation is also model-specific, meaning that some ML algorithms have easily understandable metrics that can be used to evaluate how the model arrived at the prediction (random forests), while some only provide vague accuracy statistics (neural networks). A consequence of these less interpretable 'black box' machine learning methods is that they may leave the user without the utility to uncover associations that underlie predictions, or to access probability thresholds of why certain observations were grouped to a particular response output.

We urge that there is no best use scenario when it comes to ML, and that individual researchers should select methodologies that are consistent with the specific domain problem, the questions being asked, and based on available data. If the goal of the research is to build a predictive understanding of an outcome based on microbial community data, SML has appeal since these algorithms are tailored to optimize predictive accuracy. However, if the goal is to relate specific microbial groups to particular outcome, classical statistical models have utility as well. Statistical models have strengths in microbial community analysis, but SML can provide a research strategy that can be based on less *a priori* assumptions of the data (such as formulating decisions based on predefined significance level) and more emphasis can be placed on ML to identify intricate associations and confounding variables that may be hard to detect but are often responsible for cause-effect.

In practice, ML can perform surprisingly well on datasets that are sampled from and represent messy real-world systems, such as the human body, soil, and water [48], [49], [50] and demonstrates superiority over traditional multivariate statistics in analyzing metagenomic data. In addition to these benchmarks, there is an increase in the development of microbiome-specific 'pipelines' that have user-friendly ML implementation and can be accessed through web-interfaces, the statistical compute language R [51], or Python [52]. A collection of methodologies is described in **Table 1.1** and although not exhaustive, mentions microbiome-specific or domain-agnostic procedural extensions of predictive data analysis, such as interpreting and visualizing model outputs, as will be described moving forward.

1.5 Optimizing model construction and evaluation

In most domains, input features can be challenging and economically expensive to obtain. In the case of marker-gene analysis there is often an overabundance of features as a result of how high-throughput sequencing platforms capture genetic diversity within samples. It is therefore the goal of those using machine learning on microbiome data to consider feature selection methods to identify and remove non-informative, noisy or redundant features. As opposed to using every available feature in training a model, carefully selecting features may lower the cost of computation, reduce the complexity of the model for easier interpretation, and in some cases improve generalized predictive performance of the model.

In most cases, it then becomes tractable to understand microbial community data at a deeper and more targeted level, since feature selection allows for easier evaluation of the relationship between each input feature (i.e., as a microbial taxa) to a response label, or whether any features are used together to drive predictions. In addition to its predictive capabilities, ML can be used as a powerful data mining

tool and to access a translational component of data, such as assessing whether feature-response label linkages in a model correspond to similar conditions in the real-world system after which the model is constructed. A noteworthy caveat of using SML in translational research is that it would require subsequent testing and hypothesis validation independent of the modeling procedure to conclude such relationship, since this initial interpretation is at the level of the model only.

The use of ML in microbiome research is motivated by a range of research questions and expected outcomes of modeling. This makes ML a very dynamic approach to predictive and exploratory modeling with many user defined parameters to be considered for each objective. Many of the 'pipelines' described in **Table 1.1** enforce optimal parameter tuning of ML and associated *post-hoc* analysis that enables more of an understanding of microbiome-specific research questions; however, it should be noted that the more informed a researcher is of how parameterization benefits their domain problem and research questions, the better. Likewise, as 'pipelines' offer more customization to allow more user-defined decision making, there calls for an increase in knowledge of the broadly applicable methodologies for predictive data analysis.

Accessibility of evaluation metrics that aid this interpretation may depend on which learning method is used. It is integral to consider this at the point of model selection in order to optimize ML for microbial community analysis. The remainder of this section will describe various techniques for feature selection, preferred model evaluation metrics and *post-hoc* model interpretations, with consideration of why particular methods may be better for certain problems.

1.5.1 Exploring feature selection methods

It is often the case that features in microbiome data greatly exceed the number of samples, which can lead to a model overfitting, provides overoptimistic model evaluations, and may limit cross-study comparison [53]. Feature selection methods generally dictate how well a model generalizes to novel input data by allowing for fewer and more discriminative features that maximize performance. This section discusses three main categories of feature selection: filter methods, wrapper methods, and embedded methods.

Filter methods are typically a pre-processing step performed outside of the modeling procedure that statistically measure and score correlations (i.e., univariate or multivariate: Spearman's rank correlation [54], MANOVA [55]) between input features so that only those passing some relevant criteria can be considered for downstream modeling. Although filter methods are advantageous in that they are easy to parameterize, computationally inexpensive and scalable, they can be challenging for the following reasons: (1) choosing a specific method

assumes prior assumptions about the relationships in the input feature space (2) filter methods become challenging when trying to satisfy a specific research question and account for potential feature heterogeneity or the multicollinearity and complex covariance structure of microbial community data and (3) since filter methods are done prior to modeling, they place no consideration on whether a specific ML model would maximize performance using the reduced set of features. Wrapper methods repeatedly construct models (e.g., classifiers) by iteratively adding (forward selection), removing (backwards elimination) and ranking (recursive elimination) features to search for an optimized combination that improves or marginalizes performance of preceding models. Since wrapper methods are a repeated learning process that can exhaust through features, it is not as ideal as filter methods because it becomes computationally expensive with the high-dimensional structure of metagenomic data.

Embedded methods are a more computationally tractable approach to feature selection by relying on the algorithm itself to inform a 'useful' feature. As discussed earlier, decision tree algorithms GB and RF satisfy the objective of modeling a problem and inherently have a built-in feature selection method that operates during model training. Importantly, this provides embedded methods the ability to search the full feature space, that is, if the algorithm infrastructure is in place to handle such high-dimensional data. To this extent, many feature-response associations have the potential to be discovered that would otherwise have been disregarded had data been pre-processed with restrictive assumptions prior to modeling with a filter method, or if certain potentially important features were left out of a resulting wrapper method if not considered a part of the 'optimal feature subset'. For these reasons, and on the basis of computational tractability, embedded methods are an ideal practical feature selection method for optimizing microbial-based ML models.

Despite not being as extensively reported in studies that profile the microbiome, new feature selection regimes that are more biologically motivated, such as taxonomy-aware hierarchical feature engineering (HFE) [53] are starting to gain traction and may be ideal for when embedded methods struggle with using the full search space when using very high-dimensional datasets.

1.5.2 Evaluating and interpreting estimator performance

For binary classification tasks (assigning samples to one of two response labels), receiver operating characteristic (ROC) [56] curves can be used to assess performance of the model at various decision thresholds by plotting TPR (true positive rate – sensitivity) as a function of the FPR (false positive rate – 1-specificity). By extension, computing the area under the ROC curve (AUC) [57] can provide a measure of how well the model could discriminate $\hat{Y}$. AUC can range from 0.5 (separation of $\hat{Y}$ was no better than random chance) to 1.0

(perfect separation of $\hat{Y}$), assumes that the cost of misclassifying each response label is equal and is sensitive to when response labels are skewed.

For multiclass classification (assigning samples to more than two responses), we advocate that logistic loss ($\log_{loss}$), also known as cross-entropy loss be used, as it measures the quality of predictions using the probabilistic confidence of sample separation into respective Y labels and penalizes incorrect or uncertain predictions [58]. A low $\log_{loss}$ is preferred and reflects the distribution of the certainty of predictions and like AUC, is also sensitive to when response labels are skewed.

When predicting continuous labels in regression, mean squared error (MSE) is a preferred metric that averages the squared difference of the known continuous Y value and the predicted value of $\hat{Y}$. This metric is desired because it is differentiable, which can be optimized better. A lower MSE is favorable as it measures how close a fitted line is to the data points.

Often in practice, these metrics are computed for predictions on a single cross-validated model rather than on separate models from splitting the same dataset into a training and testing set. Cross-validation is a method that holds out samples which are later used to validate prediction accuracy during the learning process and generally leads to models that are less biased and not as overoptimistic as compared to train/test splitting [36].

While accuracy measures as described above are useful, they cannot be used to explain why a model made a certain prediction. Typically, many algorithms have *ad-hoc* implementations for model interpretation, such as measuring the 'importance' of each feature or multiple features to response labels. In RF, for instance, this is usually done by permuting, or re-arranging the values of input features during the learning process, such that, if a feature is 'important', changing its values will lead to increased error rates in aggregated predictions. This process, also called variable importance, is often guided by model-specific information, such as the correlation structure between predictors, and usually scales features to have a maximum value of 100 to indicate the relative importance (**Fig. 1.3B**).

1.5.3 A use case summary of current software implementations

Table 1.1 describes recently developed and commonly employed toolkits designed to assist researchers through the steep learning curves of predictive data analysis. For instance, SIAMCAT [59] and BiomMiner [60] are comprehensive ML 'pipelines' tailored to clinical microbiome datasets. These pipelines include the ability to perform cross-study comparison, automatic tuning of optimal parameters for dimensionality reduction, feature selection and predictive

modeling, provide *post-hoc* interpretable measures of feature 'importance', and can demonstrate the influence of different parameter choices on resulting classification accuracy.

Another variety includes web-based tools such as MicrobiomeAnalystR [61], which is an ML-toolkit deployed through a web-interface to assist users who may lack computational expertise or resources. MicrobiomeAnalystR provides real-time comprehensive analysis reporting, recommendations, and visual comparisons of an implemented model to public datasets. Moreover, commonly used analysis pipelines such as QIIME2 [62] and mother [15] include implementations of SML algorithms such as RF and SVMs.

Another implementation of ML is DeepMicro[ref], which has been shown to perform well when using the microbiome to predict various diseases through deep representation learning. This method uses autoencoders to transform high-dimensional microbiome data into low-dimensional representations, then applies classification algorithms on the various learned representations. This method accelerates model training and parameter tuning by significantly reducing dimensionality of the microbiome profiles.

Many re-implementations of the original RF, namely cforest [63] and ranger [64], include novel resampling schemes for more unbiased estimates of prediction accuracy, measures of feature importance, and for computational efficiency on high-dimensional data. By extension, tools like inTrees [65] and dtreeviz can be used for *ad-hoc* knowledge discovery, such as to interpret predictions of black box models. These systems are designed for extracting, measuring and summarizing rules that govern splitting criteria in decision tree ensembles. A brief schematic illustration of this process is displayed in **Fig. 1.3B**.

Other software such as LIME [66] and iml [67] seek to offer robust, model-agnostic explanations. These include measuring feature effects on the influence of predictions, and in the case of decision tree algorithms, approximating black box predictions by constructing less complex 'surrogate' trees that provide accessible interpolations.

As comprehensive as some of the ML-toolkits described above may seem, they are still limited in their customization and cross-platform implementation. Given these constraints, more advanced users may consider domain agnostic end-to-end ML platforms with parallelized implementations for predictive data analysis, such as scikit-learn [68], keras [69], caret [70], and H2O.ai [71]. These 'pipelines' enable more customization for parameter tuning and parameter choices, allow multiple models to be built from scratch and ensembled using the same re-sampling parameters and provide more access to raw model contents (i.e., indexed predicted probabilities during cross-validation, as opposed to just an accuracy

metric). Although less intuitive, these methods allow more in-depth analysis than the more automated, user-friendly microbiome-specific platforms that are built for execution efficiency on smaller ML workloads, rather than for scale.

Nevertheless, the domain specific tools described in **Table 1.1** are useful for putting into context the biological relevance of the domain problem, allow fast and easy exploration, and serve as a good starting point for microbiome-based predictive data analysis. These 'pipelines' are also beneficial for those with a more advanced understanding of ML. While often the choice of pipeline comes down to optimization and comfortability as well as if visual outputs are necessary for data reporting, it is best practice to choose methodologies handle the characteristics of microbiome data and are interpretable, especially if the goal is to translate the research into diagnostics.

Aside from software implementations, it is worth mentioning that there are a few public repositories for curated microbiome datasets and related metadata from some of the most cited studies in the field of microbial ecology: GMRepo [72], MLrepo [73], curatedMetagenomicsData [74] and MicrobiomeHD. These public repositories can be used to practice ML, benchmark new approaches, and for cross-study comparison.

1.5.4 Machine learning for classification of human disease from microbiome data

Microbiome data has been used to link microbial community composition and disease state [75]. Diseases such as Inflammatory Bowel Disease, metabolic syndrome, obesity, hypertension, cancer, neurological diseases, among others have been linked to the human microbiome [76]. Many studies have sought to statistically link diversity metrics such as alpha diversity or abundance of particular taxonomic groupings with disease states [77]. However, as sample numbers have increased, these broad level relationships often do not hold up. For example, in studying obesity, it had been proposed that some taxonomic markers (*Firmicutes* and *Bacteriodetes*) [78] as well as decreased alpha diversity [79] were indicators of obesity. Reanalysis of this data, aggregating data across studies, demonstrated that some of these coarse measures for the microbiome did not adequately predict obesity across larger datasets [80]. The complexity and interpersonal variation within the microbiome of humans has complicated the use of the broad level metrics.

SML has been proposed as an alternative to other methods for associating microbiome with an outcome as SML may be a more robust analysis tool for predicting disease state based on microbial community profiles. **Table 1.2** summarizes key studies employing SML to link microbial community data to

a specific outcome to illustrate how SML has been previously used and highlight some considerations in employing SML to study microbial communities. One recent study used fecal microbial community profiles to predict the presence of colonic neoplasia [76]. The use of SML allows for models optimized for prediction of disease to be trained and validated on out of training set data that will enable more robust determination of the link between microbial communities and health states. This study explored multiple SML methods for this classification problem including L2 Regularized Linear Regression, RF, and SVM. This study found that many methods resulted in highly performing models, with RF performing the best (AUROC curve 0.695). Other models such as L2-regularized logistic regression, XGBoost, L2-regularized SVM with linear and radial basis function kernel all performed similarly with AUROC between 0.668 and 0.680. Interestingly, they found that while RF performed the best out of the tested models, some more interpretable approaches, such as L2-regularized logistic regression, had similarly high accuracies. These authors proposed that while more complicated models such as RF may result in higher accuracies, interpretability is an important factor in considering study design and the application of SML.

Table 1. 2

Studies using Machine learning in microbial ecology and microbiome studies.

System	Classification	Input data	Number of samples	Method	Training and Validation	Reference
Human	Colonic screen relevant neoplasias	16S rRNA	172 patients with normal colonoscopies, 198 with adenomas, and 120 with carcinomas	L2-regularized logistic regression, L1- and L2-regularized SVM with linear and radial basis function kernels, a decision tree, RF, and gradient boosted trees	80% Training 20% Validation 20% Test Five-fold cross validation	Topçuoğlu et al 2020 [78]
Human	Personalized postprandial glycemic response	16S rRNA	900 samples 800 in training 100 in validation	Gradient boosted trees	800 samples used and validated with a leave one out cross validation scheme 100 Sample validation cohort	Zeevi et al 2015 [84]
Environmental	Crop Productivity	Shotgun metagenomic	12 samples	RF	10 samples as training set 2 samples as validation set (all	Chang et al 2017 [91]

					combinations of the 12 samples	
Environmental	DOC level	16S rRNA	302 samples	feed-forward neural network regression and RF	257 samples as training set and 51 as test set	Thompson et al 2019 [92]
Environmental	Environmental quality status associated with salmon farms	SSU RNA (bacteria and ciliates)	152 across seven salmon farms	RF and SVM	Models trained on six of the salmon farms and tested with the seventh	Cordier et al 2018 [93]
Environmental	Environmental impacts of marine aquaculture	SSU RNA (five marker genes – one bacterial, one foraminiferal, and three universal eukaryote)	144 Sediment samples	RF	Models trained on four of the salmon farms and tested with the other farm	Frühe et al 2020 [94]
Environmental	Environmental quality status associated with salmon farms	Bacterial 16S rRNA	12 sediment samples collected from six sites	RF	12 samples validated with a leave one out cross validation scheme	Dully et al 2020 [95]
Environmental	Contamination state (uranium, nitrate, oil)	16S rRNA	93 samples for ground water contamination. 42 samples for oil contamination	RF	Performance metrics were determined from a confusion matrix based on out-of-bag predictions	Smith et al 2015 [89]
Environmental	Glyphosate presence	16S rRNA	32 16S rRNA gene samples and 32 16S rRNA samples	ANN and RF	32 samples used and validated with a leave one out cross validation scheme	Janßen et al 2019 [88]
Forensic	Postmortem Interval	16S rRNA	144 sample swabs were taken from a total of 21 cadavers	SVR, K-neighbor Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, RF regression, Bayesian Ridge Regression.	80% of samples for training set and 20% of samples for validation set	Johnson et al 2016 [102]
Forensic	Postmortem Interval	16S rRNA	176 samples	RF, SVM, ANN	70% for training and 30% for testing. Accuracy determined by mean absolute	Liu et al 2020 [103]

					error and goodness of fit of 15 models.	
Forensic	Geospatial location (port of origin)	16S rRNA	1,218 samples	RF	repeated k-fold cross validation (k 10 with 3 repeats)	Ghannam et al 2020[51]

The use of the microbiome to personalize treatment was further investigated in another study examining the interpersonal variation in the changes in blood glucose observed following meals (postprandial glycemic response (PPGR)). Previous studies have shown high interpersonal variability in PPGR in response to the same food [81]. This suggests that some foods might result in a high PPGR in some patients and low PPGR response in others. This finding coupled with the high interpersonal difference in microbiomes, led Zeevi et al (2015) to develop a classifier that could relate foods with the microbiome and other physiological data to accurately predict the PPGR of patients [82]. The authors used GB for regression to relate patient data, information about the meal, and the patient's microbiome to predict the PPGR. This study revealed that the SML models that incorporated microbiome data were able to more accurately predict PPGR than meal carbohydrates or meal calories alone. The combined microbiome and patient data model's prediction of PPGR was correlated with the measured PPGR with a Pearson correlation of 0.68. The model was trained using a cohort of 800 individuals and validated on a different 100 individuals. This type of analysis using ML with patient and microbiome information allows for a more tailored treatment that accounts for the high interpersonal variation that is often observed with human disease [83].

1.5.5 Machine learning for classification in environmental monitoring

In addition to prediction or disease state in the human system, coupling SML and microbial community profiling of microbial communities in the environment shows promise for the purpose of environmental monitoring [84]. Just like in the human environment, microbes in soil, water, or air can rapidly respond to changes in their environment. These changes in microbial community composition can often occur in a predictable manner. SML has been used in both natural and industrial settings to use microbial information to aid in predicting environmental quality [85], contamination state [86], [87] as well as rates of various processes including copper bioleaching [88], Previous studies have used microbial biomarkers as indicators of particular environmental processes or outcomes. Indicator species analysis has been used to identify taxa that are related to particular phenomena or treatments that could be used as biomarkers for that phenomena. However, like differential abundance analysis, indicator species is performed by analyzing the prevalence and abundance of individual features in

different categories and is not able to identify complex interactions between microbes in the dataset and the possibility that large groups of microbes may respond to the treatment.

ML is gaining popularity in predicting environmental phenomena from environmental microbial community data to develop and predict environmental health indices. One such study used SML to relate the microbial community present in agricultural soil with crop productivity [89]. In this study the authors coupled SML with metagenome wide association studies to identify potential differences in the microbial communities that were related to crop productivity. RF models built from metagenomic data were able to predict the crop productivity with an accuracy of 0.79. Another study sought to relate dissolved organic carbon (DOC) with microbial community composition [90]. RF and artificial neural networks (ANN) were used to construct models to predict the DOC concentrations of leaf litter based on microbial community composition. The models from this study were reasonably accurate with the ability to predict DOC correlating with observed DOC with a Pearson correlation coefficient of 0.636 and 0.676 for the feed-forward ANN and the RF models respectively. Interestingly, these researchers compared the important features identified through SML with indicator species identified in indicator species analysis. While they found some overlap in the features identified in both methods, only about 30% of the features were shared between indicator species analysis and RF. This suggests that ML often uses distinct features for classification than what would be identified through differential abundance or indicator species analysis and may be able to more sensitively identify groups of features that are related to an outcome.

Biotic indices have been used to assess environmental health as biotic organisms are impacted by the overall ecological quality status of an environment and may be more sensitive than measurement of abiotic factors. Therefore, various organisms have been proposed as indicators of environmental health. SML can be used to associate environmental genomic profiles with environmental quality status, which is commonly used by regulators to guide decision making in restoration and environmental monitoring [85]. A number of studies have provided a framework for the use of SML to identify patterns in microbial eukaryote and bacterial communities to predict biotic indices and environmental quality status using salmon farming as a test case. Cordier et al. (2018) [91] used various marker genes targeting the small subunit rRNA for bacteria, ciliates, and universal eukaryotes to compare the performance of SML to predict the environmental quality status and biotic indices compared to using environmental DNA to measure known indicator taxa. They found that SML outperformed the use of metazoan-assigned OTUs. The predictions obtained from metazoan-assigned OTUs had kappa values between 0.211 and 0.569, whereas the SML models had kappa values raging from 0.755 to 0.881. Following on from this study,

Frühe et al (2020) [92] compared the performance of SML with standard IndVal approach for prediction of environmental status. The indicator species approach directly from OTUs/ASVs has appeal due to there not being a need to assign taxonomy to the OTUs/ASVs like in the metazoan-assigned OTUs approach. This study found that SML outperformed the IndVal approach for prediction of environmental status. Furthermore, bacterial communities were better able to predict environmental quality status salmon farming compared to ciliates. These studies illustrated the utility of SML for environmental biomonitoring. However, these studies used training and validation data generated from the same lab. In order for this type of ML-coupled to molecular analysis approach to be used for environmental monitoring, there needs to be high replicability and generalization of the models. Therefore, Dully et al 2020 [93] performed an inter-lab validation study for the prediction of biological indices. In this study two series of samples were collected and split into technical replicates. From each site, biological replicates were also sampled. The authors of this study found that there was greater variability in diversity between biological replicates compared to technical replicates processed in each lab, which suggests that molecular methods can be standardized and have good replicability. Furthermore, SML models constructed from the two labs produced highly correlated data. These studies combine to demonstrate the promise, generalizability, and robustness of linking SML with environmental genomic data to assess environmental health status, which can guide decisions related to environmental health.

The prediction of environmental contamination is another growing area of interest in the application of SML and microbial ecology. Often contamination is identified in the environment through direct measurement of the contaminant of interest. While measuring of the contaminant is the gold standard for contaminant detection, often the contaminant may be present transiently. In these cases, the contaminant may not be detectable at the time of sampling. Smith et al (2015) demonstrated that RF could be used to predict the presence of uranium and nitrate contamination in groundwater [87]. This study demonstrates that a single set of microbial community profiles can be used to predict any number of response variables. Further, RF models were able to predict the presence of oil in the ocean with near perfect accuracy (F1 score of 0.98). Notably, RF could classify samples into no-oil, oil, and past oil contamination based on the microbial community alone. The past oil category contained samples that at one point in time had detectable levels of oil, but at the time of sampling, there was no detectable oil. This finding indicates that ML methods can identify patterns in the microbial community that are indicative of current and past contamination. The ability of the RF models to identify past contamination could be indicative of ecological resiliency and stability that allows microbial communities to maintain the signature of oil after the oil was no longer present.

The ability to predict contamination in the environment has been expanded to other systems including prediction of the herbicide glyphosate in the Baltic Sea [86]. In the Janßen et al (2019) study, the authors employed artificial neural networks and RF to predict the presence of glyphosate. Expanding on the previous work showing the ability of SML to predict contamination, Janßen et al (2019) identified important features through constructing a series of models leaving out individual features and monitoring the changes in accuracy of the models. This type of approach can be used to interrogate some of the more complex and less transparent approaches such as Artificial Neural Networks (ANNs). Another novel aspect of the work of Janßen et al (2019) is the use of a random forests proximity matrix as the dissimilarity measure in PCoA. This approach resulted in clearer separation of samples on the PCoA analysis compared to using a Bray Curtis dissimilarity matrix.

While these studies demonstrate the ability of ML to predict the presence of a specific contaminant in an environmental sample, other work has been used to predict more general properties such as environmental impacts of hydraulic fracturing as well as location and residence time of ballast water [94], [95], [96]. In both of these cases the specific relationship between the features and the output variable may not be clear. In other words, when predicting the presence of a specific contaminant, a single feature may increase or decrease in abundance in direct response to the contaminant due to the toxicity of the contaminant or ability of the contaminant to stimulate growth of the microorganism. These more generic phenomena may result in indirect impacts on the microbial community that are detectable using ML. These studies demonstrate that it is possible to detect and classify contamination both specifically and more generically using ML. Interrogation of the important features used in these classifiers may provide insights into specific biomarkers of contamination that could be used as tools for environmental monitoring.

In addition to the applied outcomes described above, SML has potential to be used to better understand the ecology of microorganism in the environment. Smith et al 2015 [87] demonstrated that microbial community composition as determined by 16S rRNA can be used to predict a diverse set of geochemical factors including pH, manganese and aluminum. Alneberg et al (2020) [97] also highlight the application of SML to predict the ecological niche of microbial groups with a focus on microbial communities from the Baltic Sea. The authors of this study use metagenomic binning to obtain 1962 metagenome assembled genomes (MAGs) representing the majority of prokaryotic diversity in in the Baltic Sea. These prokaryotic clusters demonstrated distinct ecological preferences along the various environmental gradients observed. Ridge Regression, RF, and GB were used to predict the niche gradient of the prokaryotic cluster based on the functional profile of genes found in each cluster. The authors of this study found

that the predicted niche gradient agreed with the observed niche gradient with a Spearman's rank correlation of 0.70 – 0.81. These studies highlight the fact that SML can be useful in identifying patterns in natural microbial communities and predicting the niche of an organism.

1.5.6 Microbial communities and machine learning for forensics

Microbes have been used for forensic applications for a long time. Normally microbial forensics is used to identify the source of particular organisms related to bioterrorism, disease, or contamination. However, it is possible to use microbial community composition as a tool for trace evidence [98], [99]. Previous work has shown the utility of microbial community composition in determining postmortem intervals (PMI). Various studies have examined the ability of the soil and skin microbiome to serve as a molecular clock for postmortem intervals. Other studies have used ML models constructed from skin microbiota to assess the PMI [100], [101]. Soil evidence has also been used as forensic information. In the same way that pollen can be used to identify the source of a particular soil sample, the microbial community in a soil sample may provide information about where that soil was derived. Metagenomic information from soils has been used to differentiate soil from different locations [102], [103]. These studies demonstrated that information contained within the microbial community from the soil sample could be used to identify the source of the soil. These studies used hierarchical clustering and non-metric multidimensional scaling (NMDS) to differentiate groups. More recently, SML has been applied to determining the geographic source of an ocean water sample based on the microbial community [50]. Ghannam et al (2020) [50] demonstrated that RF could be used to accurately differentiate the location of sampling of water from 20 different locations. This study is important in that it shows that SML can be used to identify important trace signals in the microbial community of water that can accurately distinguish between 20 diverse locations from around the world as well as specifically identify the location of collection within locations close in proximity to each other.

1.6 Summary and outlook

This review has sought to provide an overview of how ML has progressed the field of microbial ecology. Despite the unprecedented sophistication and promise of ML algorithms, there exist several outstanding issues that should be considered when applying ML to marker-gene datasets. Although ML models can be consistently constructed to produce high accuracy metrics on complex data, the underpinning decision support systems can often be largely black box methods of investigation where the rational and logic behind predictions are hidden behind layers that are challenging to interpret [104].

The large majority of studies using ML to investigate microbiome datasets gauge and validate hypotheses and report findings through performance and may apply *post-hoc* procedures to identify important biological taxa using variable importance metrics. However, due to the complexity of some modeling methods, inferring biological importance from feature importance could be problematic. Therefore, there is a need for increased interpretability in ML models used in microbial ecology studies. Often the learning algorithms applied to marker-gene datasets are developed and implemented for improved performance, rather than for model interpretation [104], [105]. In order to glean biologically meaningful data from these ML methods, it may be important to consider the choice of model with preference toward more interpretable algorithms as well as novel methods for interpreting models such as permutational approaches. Microbial ecology studies that demonstrate model transparency are limited to reporting single feature to response interaction or are overburdened by investigating feature contributions to each observation for accumulated local explanations of modeling procedures [48], [76], [106], [107].

There have been major improvements for model specific and model agnostic approaches for model interpretation [66], [108], [109], [110], [111], some were described in this review. However, these methods often cannot account for hidden heterogenous effects of the full feature space, which can reduce model fidelity and mislead researchers depending on algorithm selection.

Here, we argue that while methods for inferencing how single microbial community members influence single predictions are beneficial (local interpretations), appreciating the inner workings of multiple microbial community members and how they generally discern a group of the same response label is more robust and generalizable (global interpretation). In the context of microbial ecology, the lack of global interpretation techniques makes it challenging to inference on the basis of the full feature space and to identify all potential features that are interacting to most frequently to predict response labels with the least error. Often a condition is not attributable to a single feature, but multiple features. One of the strengths of ML is the ability to appreciate these groups of features in making a prediction. However, in interpreting a model, a focus on the importance of a single feature may limit the applicability to the real-world system that is being modeled (i.e., appreciating the full microbial community rather than subsets).

In high-risk domains like human health and biology, the ability to interpret and generalize a model has many downstream benefits, such as identifying biological relevance that support hypotheses of the system being investigated and the ability to extract actionable insights about the community of study. Many of the implementations described in this review seek to extract actional information from microbiome datasets that can be used in the clinic, environmental monitoring

applications, and forensics. It is important that in implementing the use of ML-identified biomarkers in diagnostic application that there is a need for common acceptance and trust of the algorithms employed which lead to critical decisions relating to the microbiome [112], [113], [114], [115], [116].

While other disciplines of biology such as single-cell RNA seq, drug discovery and development, and neuroscience have attempted to bring interpretation to black box ML models [117], [118], [119], [120], [121], investigation into microbial ecology applying ML on marker-gene datasets is lagging behind. This is surprising since there has been a rapid expansion of microbiome related research that will continue to expand. With a lack of interpretation of ML in this field, fundamental dynamics of a microbial system will be left unreported.

It is notable to mention that as a result of the structure of marker-gene datasets from HTS platforms, Classification and Regression Trees (CART) algorithms continue to dominate the field of microbial ecology. However, deep learning is a promising approach to revolutionize how we investigate microbial communities. Considerations should be placed on whether deep learning is necessary to investigate metagenomic datasets, since, although the inner workings of neural networks are the focus of ongoing research [122], [123], they are some of the most notorious black box methods that lack interpretability. In some cases, it may be better to choose a model that can be more easily interpreted over a more complex model that has a higher performance metric.

There are still a number of open questions and considerations that need to be taken into account when considering the use of employing SML for monitoring and diagnostics. One of the first considerations is the need for sufficient replication in experimental design. Human microbiome studies have paved the way with high replication with hundreds of samples used in training algorithms. However, for environmental monitoring, sample collection is often costly, which can limit replication. In cases where sample replication is limited, some test and validation approaches may be more useful. For example, a leave-one-out validation strategy could be useful when replication is low and the splitting of samples into a training and test set would result in even less replication. Another consideration is sampling depth. As was discussed earlier, diversity estimates from sequencing data highly depends on sequencing depth. Therefore, it is important to ensure the diversity of the samples have been sufficient covered in constructing models to be used in SML. This is an example of how an understanding of ecological diversity measures and coverage estimates (e.g. Good's coverage) may be an important first step in determining if the obtained data is sufficient for development of SML models. Another important question that must be addressed is the level of accuracy that a model must obtain to be useful for its purpose. This question is a little more difficult to answer and depends highly on the domain problem. In certain domains higher accuracy may be required for a model to be of use. While

100% accuracy may not be achievable in noisy real-life environments, it is important to consider the level of accuracy that is needed. This may vary between medical diagnostics, forensics, and environmental monitoring applications.

If we are to move toward a translational framework for microbiome analysis where features extracted from ML models are used to inform development of particular treatments or monitoring approaches, it is important to have a thorough understanding of the interpretability of the models. It is also important to ensure that ML is used to complement other approaches for profiling microbial communities that confirm the choice of selected biomarkers. Overall, it is important to consider how ML models are interpreted and reported in situations where actionable insight can be extracted from modeling procedures and used to construct downstream molecular applications such as in health and environmental diagnostics.

1.7 References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804-10.
2. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457-63.
3. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359.
4. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343-59.
5. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods. 2012;9(6):621-+.
6. Zhou YH, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. 2019;10.
7. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid-Determination of 16s Ribosomal-Rna Sequences for Phylogenetic Analyses. P Natl Acad Sci USA. 1985;82(20):6955-9.
8. Stahl DA, Lane DJ, Olsen GJ, Pace NR. Analysis of Hydrothermal Vent-Associated Symbionts by Ribosomal-Rna Sequences. Science. 1984;224(4647):409-11.
9. Norman R. Pace, David A. Stahl, David J. Lane, Gary J. Olsen. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In: K.C. M, editor. Advances in Microbial Ecology Advances in Microbial Ecology. vol 9. Boston, MA: Springer; 1986.
10. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". P Natl Acad Sci USA. 2006;103(32):12115-20.
11. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. Isme J. 2012;6(8):1621-4.
12. Hazen TC, Rocha AM, Techtmann SM. Advances in monitoring environmental microbes. Curr Opin Biotech. 2013;24(3):526-33.
13. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.
14. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6.
15. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microb. 2009;75(23):7537-41.
16. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581.
17. Preheim SP, Perrotta AR, Friedman J, Smilie C, Brito I, Smith MB, et al. Computational Methods for High-Throughput Comparative Analyses of Natural Microbial Communities. Method Enzymol. 2013;531:353-70.
18. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microb. 2005;71(12):8228-35.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12).
20. Paulson JN. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor package. 2014;1(0).
21. Sathya R, Abraham A. Comparison of supervised and unsupervised learning algorithms for pattern classification. International Journal of Advanced Research in Artificial Intelligence. 2013;2(2):34-8.

22. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction: Springer Science & Business Media; 2009.
23. Silva V, Tenenbaum J. Global versus local methods in nonlinear dimensionality reduction. Advances in neural information processing systems. 2002;15:721-8.
24. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics. 1965;21:768-9.
25. Ramette A. Multivariate analyses in microbial ecology. FEMS microbiology ecology. 2007;62(2):142-60.
26. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5(2):169-72.
27. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.
28. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature communications. 2019;10(1):1-14.
29. Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature communications. 2019;10(1):1-12.
30. Xu X, Xie Z, Yang Z, Li D, Xu X. A t-SNE Based Classification Approach to Compositional Microbiome Data. Front Genet. 2020;11:1633.
31. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
32. Biau G, Scornet E. A random forest guided tour. Test. 2016;25(2):197-227.
33. Louppe G. Understanding random forests: From theory to practice. arXiv preprint arXiv:14077502. 2014.
34. Mentch L, Zhou S. Randomization as regularization: A degrees of freedom explanation for random forest success. arXiv preprint arXiv:191100190. 2019.
35. Breiman L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA. 2002;1:58.
36. Probst P, Boulesteix A-L, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. J Mach Learn Res. 2019;20(53):1-32.
37. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
38. Wang X-W, Liu Y-Y. Comparative study of classifiers for human microbiome data. Medicine in Microecology. 2020:100013.
39. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters. 1999;9(3):293-300.
40. Soman K, Loganathan R, Ajay V. Machine learning with SVM and other kernel methods: PHI Learning Pvt. Ltd.; 2009.
41. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.
42. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.
43. Fiannaca A, La Paglia L, La Rosa M, Renda G, Rizzo R, Gaglio S, et al. Deep learning models for bacteria taxonomic classification of metagenomic data. BMC bioinformatics. 2018;19(7):198.
44. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. Front Microbiol. 2019;10:827.
45. Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS microbiology ecology. 2014;90(3):543-50.
46. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
47. Økland RH. Wise use of statistical tools in ecological field studies. Folia Geobotanica. 2007;42(2):123.

48. Aasmets O, Lüll K, Lang JM, Pan C, Kuusisto J, Fischer K, et al. Machine learning reveals time-varying microbial predictors with complex effects on glucose regulation. bioRxiv. 2020.
49. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, et al. Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes. 2018;9(2):104.
50. Ghannam RB, Schaerer LG, Butler TM, Techtmann SM. Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities. Msphere. 2020;5(1).
51. Team RC. R: A language and environment for statistical computing. Vienna, Austria; 2013.
52. Van Rossum G, Drake Jr FL. Python tutorial: Centrum voor Wiskunde en Informatica Amsterdam; 1995.
53. Oudah M, Henschel A. Taxonomy-aware feature engineering for microbiome classification. BMC bioinformatics. 2018;19(1):1-13.
54. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal. 2012;24(3):69-71.
55. O'Brien RG, Kaiser MK. MANOVA method for analyzing repeated measures designs: an extensive primer. Psychological bulletin. 1985;97(2):316.
56. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997;30(7):1145-59.
57. Ling CX, Huang J, Zhang H, editors. AUC: a statistically consistent and more discriminating measure than accuracy. Ijcai; 2003.
58. Bishop CM. Pattern recognition and machine learning: springer; 2006.
59. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine-learning toolbox. bioRxiv. 2020.
60. Shamsaddini A, Dadkhah K, Gillevet PM. BiomMiner: An advanced exploratory microbiome analysis and visualization pipeline. PloS one. 2020;15(6):e0234860.
61. Chong J, Liu P, Zhou G, Xia J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nature Protocols. 2020;15(3):799-821.
62. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology. 2019;37(8):852-7.
63. Hothorn T, Zeileis A. partykit: A modular toolkit for recursive partytioning in R. The Journal of Machine Learning Research. 2015;16(1):3905-9.
64. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:150804409. 2015.
65. Deng H. Interpreting tree ensembles with intrees. International Journal of Data Science and Analytics. 2019;7(4):277-87.
66. Ribeiro MT, Singh S, Guestrin C, editors. " Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
67. Molnar C, Casalicchio G, Bischl B. iml: An R package for interpretable machine learning. Journal of Open Source Software. 2018;3(26):786.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
69. Gulli A, Pal S. Deep learning with Keras: Packt Publishing Ltd; 2017.
70. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008;28(5):1-26.
71. Candel A, Parmar V, LeDell E, Arora A. Deep learning with H2O. H2O ai Inc. 2016.
72. Wu S, Sun C, Li Y, Wang T, Jia L, Lai S, et al. GMrepo: a database of curated and consistently annotated human gut metagenomes. Nucleic acids research. 2020;48(D1):D545-D53.

73. Vangay P, Hillmann BM, Knights D. Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. GigaScience. 2019;8(5):giz042.
74. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14(11):1023.
75. Durack J, Lynch SV. The gut microbiome: Relationships with disease and opportunities for therapy. J Exp Med. 2019;216(1):20-40.
76. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. Mbio. 2020;11(3).
77. Reese AT, Dunn RR. Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. Mbio. 2018;9(4).
78. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. P Natl Acad Sci USA. 2005;102(31):11070-5.
79. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480-U7.
80. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. Mbio. 2016;7(4).
81. Vrolix R, Mensink RP. Variability of the glycemic response to single food products in healthy subjects. Contemporary clinical trials. 2010;31(1):5-11.
82. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94.
83. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. Plos Comput Biol. 2016;12(7).
84. Techtmann SM, Hazen TC. Metagenomic applications in environmental monitoring and bioremediation. Journal of industrial microbiology & biotechnology. 2016;43(10):1345-54.
85. Cordier T, Lanzen A, Apotheloz-Perret-Gentil L, Stoeck T, Pawlowski J. Embracing Environmental Genomics and Machine Learning for Routine Biomonitoring. Trends Microbiol. 2019;27(5):387-97.
86. Janßen R, Zabel J, von Lukas U, Labrenz M. An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. Mar Pollut Bull. 2019;149.
87. Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu LY, et al. Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. Mbio. 2015;6(3).
88. Demergasso C, Véliz R, Galleguillos P, Marín S, Acosta M, Zepeda V, et al. Decision support system for bioleaching processes. Hydrometallurgy. 2018;181:113-22.
89. Chang HX, Haudenshield JS, Bowen CR, Hartman GL. Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. Front Microbiol. 2017;8.
90. Thompson J, Johansen R, Dunbar J, Munsky B. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PloS one. 2019;14(7):e0215502.
91. Cordier T, Forster D, Dufresne Y, Martins CI, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Molecular Ecology Resources. 2018;18(6):1381-91.
92. Frühe L, Cordier T, Dully V, Breiner HW, Lentendu G, Pawlowski J, et al. Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. Molecular Ecology. 2020.
93. Dully V, Balliet H, Frühe L, Däumer M, Thielen A, Gallie S, et al. Robustness, sensitivity and reproducibility of eDNA metabarcoding as an environmental biomonitoring tool in coastal salmon aquaculture–An inter-laboratory study. Ecological Indicators.121:107049.
94. Ulrich N, Kirchner V, Drucker R, Wright JR, McLimans CJ, Hazen TC, et al. Response of Aquatic Bacterial Communities to Hydraulic Fracturing in Northwestern Pennsylvania: A Five-Year Study. Sci Rep-Uk. 2018;8.

95. See JRC, Ulrich N, Nwanosike H, McLimans CJ, Tokarev V, Wright JR, et al. Bacterial Biomarkers of Marcellus Shale Activity in Pennsylvania. Front Microbiol. 2018;9.
96. Gerhard WA, Gunsch CK. Microbiome composition and implications for ballast water classification using machine learning. Sci Total Environ. 2019;691:810-8.
97. Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. Commun Biol. 2020;3(1):119.
98. Metcalf JL, Xu ZJZ, Bouslimani A, Dorrestein P, Carter DO, Knight R. Microbiome Tools for Forensic Science. Trends Biotechnol. 2017;35(9):814-23.
99. Hampton-Marcell JT, Lopez JV, Gilbert JA. The human microbiome: an emerging tool in forensics. Microb Biotechnol. 2017;10(2):228-30.
100. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, et al. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. Plos One. 2016;11(12).
101. Liu RN, Gu YX, Shen MW, Li H, Zhang K, Wang Q, et al. Predicting postmortem interval based on microbial community sequences and machine learning algorithms. Environ Microbiol. 2020;22(6):2273-91.
102. Khodakova AS, Smith RJ, Burgoyne L, Abarno D, Linacre A. Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils. Plos One. 2014;9(8).
103. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-Gonzalez A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. Science. 2018;359(6373):320-+.
104. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;8(8):832.
105. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research. 2019;20(177):1-81.
106. Bogart E, Creswell R, Gerber GK. MITRE: inferring features from microbiota time-series data linked to host status. Genome biology. 2019;20(1):1-15.
107. Richardson M, Gottel N, Gilbert JA, Lax S. Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures. MBio. 2019;10(4):e01054-19.
108. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.
109. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics. 2015;24(1):44-65.
110. Zhao Q, Hastie T. Causal interpretations of black-box models. Journal of Business & Economic Statistics. 2019:1-10.
111. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:161208468. 2016.
112. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data & Society. 2016;3(2):2053951716679679.
113. Bathaee Y. The artificial intelligence black box and the failure of intent and causation. Harv JL & Tech. 2017;31:889.
114. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
115. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608. 2017.
116. Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: is there a double standard? Philosophy & Technology. 2019;32(4):661-83.

117. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nature Reviews Nephrology. 2020:1-14.
118. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-77.
119. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019;50:71-91.
120. Netzer M, Hackl WO, Schaller M, Alber L, Marksteiner J, Ammenwerth E. Evaluating Performance and Interpretability of Machine Learning Methods for Predicting Delirium in Gerontopsychiatric Patients. Stud Health Technol Inform. 2020;271:121-8.
121. Fellous J-M, Sapiro G, Rossi A, Mayberg HS, Ferrante M. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. Frontiers in Neuroscience. 2019;13:1346.
122. Singla S, Wallace E, Feng S, Feizi S. Understanding impacts of high-order loss approximations and features in deep learning interpretation. arXiv preprint arXiv:190200407. 2019.
123. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018;73:1-15.

2 Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities

Preface

This section attempts to determine if natural microbial communities in aquatic environments can be used to characterize geospatial location on a global scale. This approach is leveraging the fact that microbes are ubiquitously dispersed and abundant throughout most environments, and how the types of microbes can vary in each location. We needed to collect real-world data, so we chartered research vessels and sampled the water at a variety of different locations near busy ports. Here we characterize the relation of microbes to geospatial location, which is possible by identifying differences in genetic material from the microbes collected (DNA) at each location. This allows an accurate and quantifiable fingerprint of one location in relation to another – and can be used to probe microbial biogeography. Below, a detailed experimental design of the sampling regime along with detailed methods used to assess these research questions are explored.

Abstract

We conducted a global characterization of the microbial communities of shipping ports to serve as a novel system to investigate microbial biogeography. The community structures of port microbes from marine and freshwater habitats house relatively similar phyla, despite spanning large spatial scales. As part of this project, we collected 1,218 surface water samples from 604 locations across eight countries and three continents to catalogue a total of 20 shipping ports distributed across the East and West Coast of the United States, Europe, and Asia to represent the largest study of port-associated microbial communities to date. Here, we demonstrated the utility of machine learning to leverage this robust system to characterize microbial biogeography by identifying trends in biodiversity across broad spatial scales. We found that for geographic locations sharing similar environmental conditions, subpopulations from the dominant phyla of these habitats (*Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, and *Proteobacteria*) can be used to differentiate 20 geographic locations distributed globally. These results suggest that despite the overwhelming diversity within microbial communities, members of the most abundant and ubiquitous microbial groups in the system can be used to differentiate a geospatial location across global spatial scales. Our study provides insight into how microbes are dispersed spatially and robust methods whereby we can interrogate microbial biogeography.

2.1 Introduction

There is increasing knowledge of the vast diversity and the abundance of microbes on our planet. However, we are only beginning to understand microbial dispersal and the potential for microbes to exhibit distinct biogeographic patterns. It has been proposed that the selection of microbes in certain locations occurs through various processes such as the environmental conditions (temperature, salinity, pH, etc.), ecological drift, diversification, and dispersal limitation (1–4). Numerous studies have outlined the relative influences of these proposed ecological drivers, which vary drastically across ecosystem type (terrestrial; in soil and sediments, marine and human) (5–9). This has resulted in a lack of consensus as to the seemingly stochastic nature of diversity observed within microbial communities and their geographic distribution.

Previous studies have applied high-throughput sequencing as a means of characterizing the microbial community composition and their underlying global spatial relationships (10–12). It is apparent that under similar environmental conditions, microbial communities can have distinct compositions across space and time (13–15). These efforts, however, have primarily been studied between local or unique habitats (such as extreme environments) (9, 16–19). Presently, the extent of variation within microbial communities on both local and regional spatial scales sharing similar environmental conditions is understudied, despite being an important component to understanding microbial biogeography. Microbial assemblages from aquatic communities surrounding shipping ports are a novel system for microbial ecologists to query biogeography in part because of the similar physiochemical conditions found between both local and regional scales in these ports.

Interfacing this unique, global data set with machine learning (ML) has allowed us to identify stark contrasts in the microbial community composition across a broad geographic range. We were able to observe subpopulations of the highly abundant and ubiquitous microbes of the same phyla that dominate these communities. Portions of the community belonging to the "rare biosphere" have been suggested to constitute much of the diversity across large spatial and temporal scales (18, 20–22) and are often attributed to the underlying distinction of a geographic location. As a result, observing variation in global biogeography through members of dominant taxa might be overlooked, and it may be possible to now explore this through certain machine learning applications. Applying machine learning to questions of biogeography may allow for resolution of fine-scale geographic differences by using a set of data that contains both microbial composition and class labels (geographic location to which the sample belongs) and learns from the relationship between these two to potentially find the microbial taxa which are most associated with a geospatial location (23). Leveraging the abilities of machine learning approaches, distinctions within

seemingly similar microbial communities across a global scale may allow for the future prediction or classification of a geospatial location based on a microbial community and could provide insights into the key microbial groups found in distinct geographic locations.

The coupling of cost-effective next-generation sequencing (NGS) technologies with well-established molecular techniques has allowed us to explore machine learning in the context of biology, ecology, and Earth science in unprecedented ways (24–27). Until now, the full potential of using machine learning to understand biogeography has yet to be achieved. This is largely a consequence of limited global microbial data sets with sufficient replication ranging across large spatial and physiochemical gradients that have been processed through standardized methodology. Here, we are seeking to combine high-resolution sequencing with machine learning to observe trends in biodiversity, investigate the potential for there to be biogeographic patterns in the microbial communities of ports, and determine the potential for machine learning to identify patterns in microbial community data not fully appreciated through the use of traditional statistical approaches used in ecology (27, 28).

Here, we investigate the global biogeography of microbial communities found to occupy shipping ports to determine whether there is a biogeographic signal to taxon distribution throughout this system. In determining the underlying distinctions in microbial community structures between these locations, we performed a community analysis of each microbial population from these ports through 16S rRNA amplicon sequencing. Amplicon sequence variants (ASVs) (29) were assigned to provide the highest resolution possible using this marker gene. As a result, we were able to investigate and identify taxon-spatial relationships across large spatial scales, with high resolution, using machine learning. We collected a total of 1,218 marine and freshwater samples from 604 geospatial locations spanning eight countries and three continents to catalogue 20 ports (each with metadata), initiating an expansive ecological study of port-associated microbes. Additionally, this data set provides a foundation for data mining and comparative ecology by accompanying the larger Tara Oceans Project (30) and Global Oceans Sampling Expedition (GOS) (31), with a focus on shipping ports. The aim of this project is to provide the framework to globally observe the process of microbial biogeography.

2.2 Results

2.3 Port sampling and microbial diversity profiling

To better understand how microbial community composition is influenced by geospatial location, we used 1,218 surface water samples from 604 locations surrounding ports spanning the Great Lakes, Pacific Ocean, Atlantic Ocean, North

Sea, Sea of Japan, South China Sea, Mediterranean Sea, and Adriatic Sea (**Fig. 2.1**). These samples were both from marine and freshwater settings and are representative of 20 globally important ports across a range of sizes and ship traffic levels, and they also vary environmentally by pH (5.67 to 9.33), temperature (3.1 to 30.8°C), and salinity (0.040 to 42.35 practical salinity units [psu]).

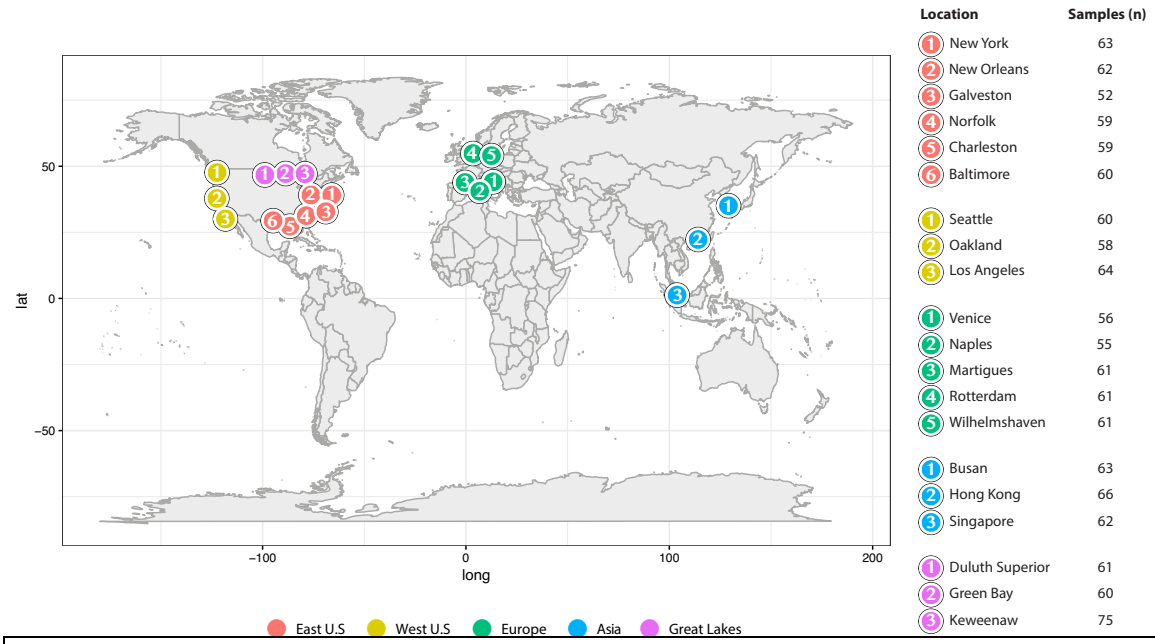


Figure 2. 1

Map of port sampling locations. Displayed on the map are the port locations from which samples were collected. All of the sampled ports are binned by the region in which each port is located (East Coast of the United States, West Coast of the United States, Asia, Europe, and the Great Lakes). Sampling depth is displayed by the number of surface water samples collected at each location (n = 1,218).

For these 1,218 samples, 86,411 amplicon sequence variants (ASVs) (29) were identified to assess diversity within microbial communities using the 16S rRNA marker gene. Instead of assigning traditional operational taxonomic units (OTUs), where sequencing reads are clustered by some fixed percent identity threshold, the raw sequence reads were denoised to account for the introduction of any DNA amplification and sequencing errors. By resolving these errors from our next-generation sequencing results, it is possible to dereplicate the reads and examine potentially meaningful information between biological sequences that differ by as little as one nucleotide. This single-nucleotide differentiation in the 16S rRNA marker gene of these bacteria, from all 1,218 samples, allows us to achieve a finer resolution of all the diversity within our data set.

2.4 Characteristics of the dominant microbial taxa of global port microbiomes

To investigate the distinct biogeographic patterns in the microbial communities of ports, we demonstrated that the taxonomic compositions from our sampling locations vary globally. There were four key bacterial phyla in our data set that dominated throughout all 20 port locations by being both highly prevalent (within 50% or more samples) and highly abundant (those with ≥10% of total 16S rRNA reads with taxonomic assignments at the phylum level). Collectively, these dominant phyla (*Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, and *Proteobacteria*) accounted for 92% of the total 16S rRNA reads across all samples and contained within them 84% of the total ASVs that were assigned throughout the data set.

The following six bacterial classes represented the majority of the variation of these four phyla in their assigned amplicon sequences (e.g., a bacterial class had ≥40% of its respective phylum's ASV content): *Acidimicrobiia, Actinobacteria, Bacteroidia, Oxyphotobacteria, Alphaproteobacteria,* and *Gammaproteobacteria. Proteobacteria* was the most abundant phylum overall (42% of total rRNA reads) across all 20 ports and included two of the six most dominant classes (*Alphaproteobacteria* and *Gammaproteobacteria*), which represented 21% and 20% of the total rRNA gene reads, respectively. Together, these six bacterial classes represent 91% of the total 16S rRNA reads and 81% of the total assigned ASVs in this global study and were sufficient to assess the majority of the diversity throughout our sampling locations (see **Table S1** in the supplemental material). These six classes were used to demonstrate data set-wide taxonomic composition throughout these globally distributed ports (**Fig. 2.2**). Despite such a high prevalence of these classes, there was substantial variation across all locations, with the highest range of variability belonging to the *Cyanobacteria*. For example, the *Oxyphotobacteria* dominated Galveston, TX, in the East Coast of the United States (44% average relative abundance) compared to the two port locations with the lowest abundances for this class, Rotterdam and Wilhelmshaven in Europe (0.01% and 1%, respectively). Additionally, the *Gammaproteobacteria* dominated Hong Kong in Asia (47%) and were least abundant in the East Coast in Galveston, TX, and New Orleans, LA (8%).

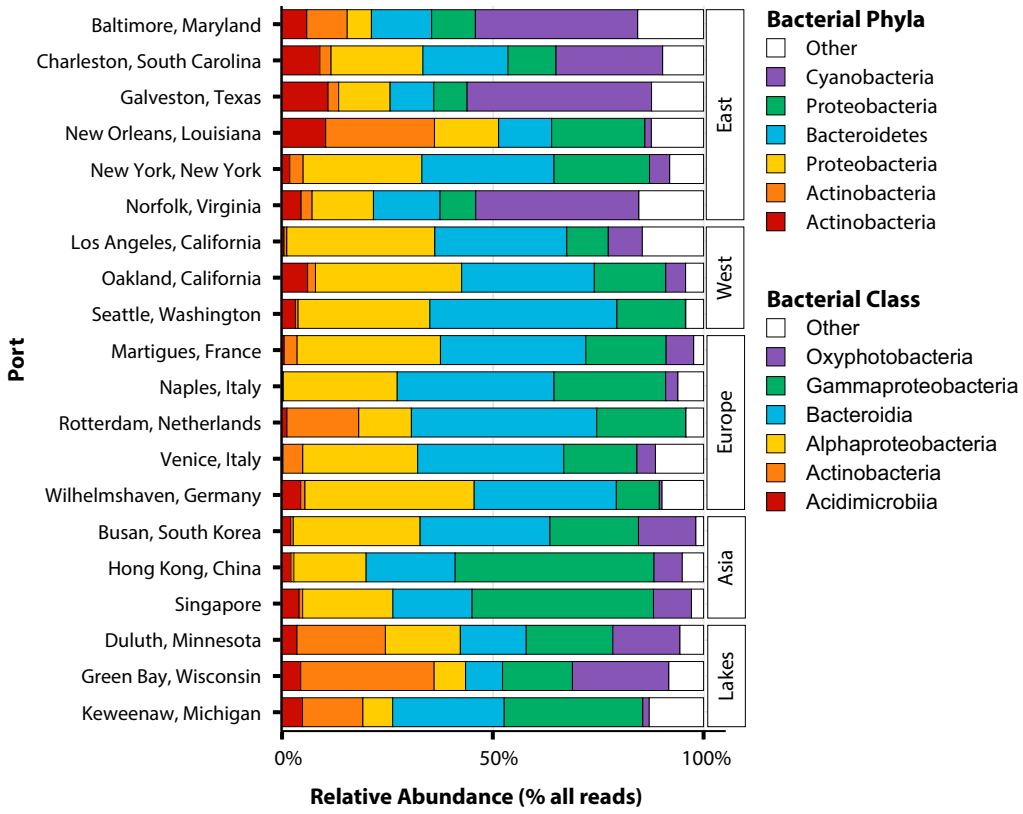


Figure 2. 2

Taxonomic composition and abundance of dominant bacteria across ports globally. Taxon plot of the composition and relative abundance of the top six dominant bacterial classes across each local port and the region to which they belong based on all 16S rRNA reads. Any bacterial class that did not comprise ≥ 40% of the ASVs belonging to the top four dominant phyla (Actinobacteria, Bacteroidetes, Cyanobacteria, and Proteobacteria) was categorized as "other."

Table 2. 1

16S rRNA read and ASV distribution of dominant taxa. Distribution of 16S rRNA reads after diversity profiling (ASV assignment) of dominant taxa. Bacterial classes comprising ≥40% of the ASV content of their respective phyla (assigned ASVs) were chosen as "dominant" classes for downstream analysis

Total 16S rRNA reads for dominant phyla			13,733,236
Total ASVs assigned during diversity profiling			3,214
Phylum	**% Total Reads**	**% Total ASVs**	**Assigned ASVs**
Actinobacteria	12.09	7.56	243
Bacteroidetes	26.07	29.37	944
Cyanobacteria	11.74	4.6	148
Proteobacteria	42.76	43.06	1384
Total	**92.66%**	**84.59%**	**2,719**
Class			
Acidimicrobiia	3.80	3.14	101
Actinobacteria	8.15	4.01	129
Bacteroidia	25.48	28.84	927
Oxyphotobacteria	11.74	4.60	148
Alphaproteobacteria	21.56	19.60	630
Gammaproteobacteria	20.91	20.87	671
Total	**91.64%**	**81.06%**	**2,606**

In addition to understanding fine-scale differences between each port, we also sought to determine broader spatial-scale patterns in biogeography of these microbial communities observed from the different regions. We analyzed the variability in the relative abundances of these dominant six classes after grouping each of the 20 port locations into one of the following five geographic regions: East Coast of the United States, West Coast of the United States, Europe, Asia, and the Great Lakes.

Our analysis of these regional taxon-spatial associations shows a substantial abundance of the *Alphaproteobacteria*, *Gammaproteobacteria*, *Bacteroidia*, and *Oxyphotobacteria* compared to the underrepresented *Acidimicrobiia* and *Actinobacteria* across all regions (**Fig. 2.3** and **Fig 2.4**). Notably, the Great Lakes have a much higher average relative abundance of *Actinobacteria* (19%) than do the other regions (average relative abundance, <10%). The *Alphaproteobacteria* predominate in the West United States (35%) and have the lowest representation in the Great Lakes (11%). The *Oxyphotobacteria* are more abundant in the samples from the East United States (median relative abundance, 17%) than the lowest median relative abundance belonging to samples from Europe (0.3%). Excluding *Actinobacteria* in the Great Lakes and *Oxyphotobacteria* in Europe and the West Coast of the United States, the six dominant classes had an average relative abundance of >10% across all regions (**Fig. 2.3**).

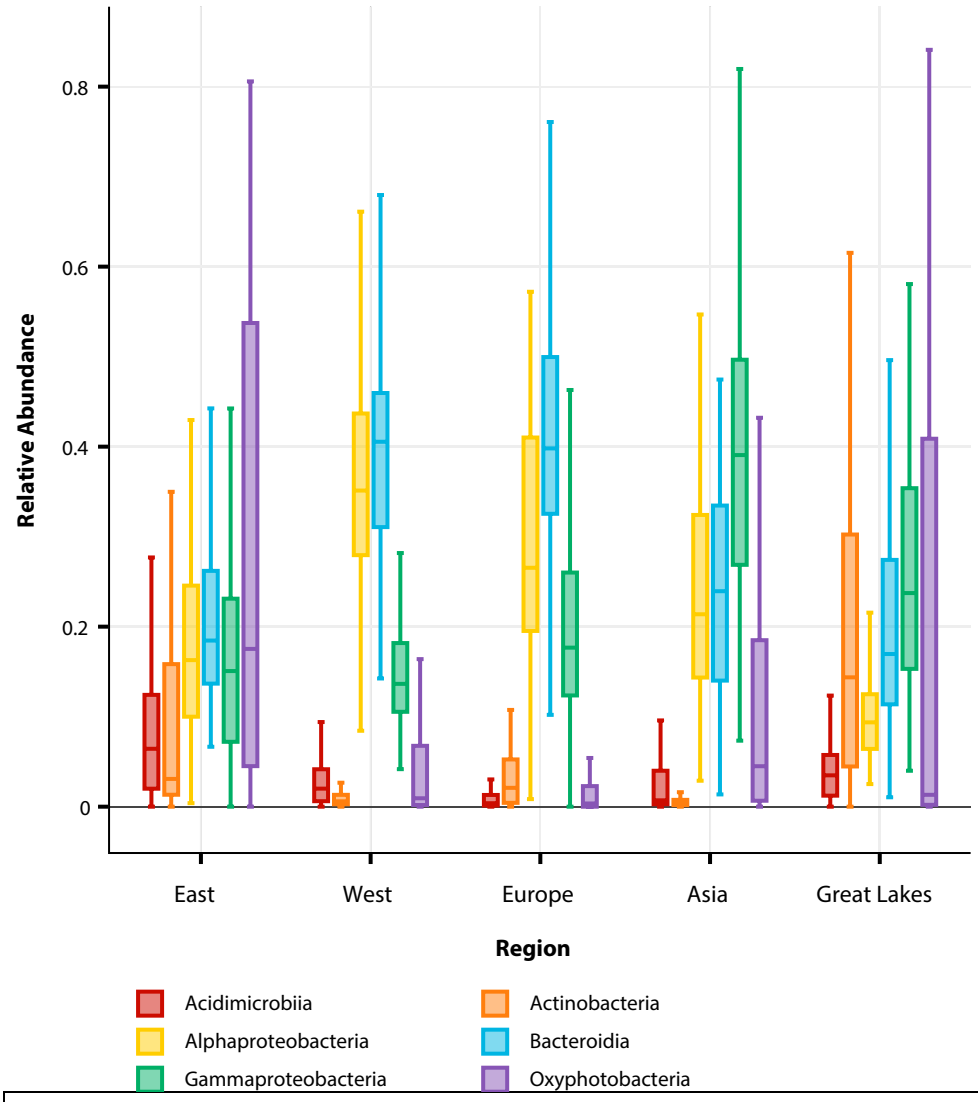


Figure 2. 3

Boxplot of dominant bacterial classes. Box plot displaying the differences in community composition of the top 10% most common (dominant) bacterial classes represented as a percentage of relative abundance. Each box represents the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively), and the median is represented by the vertical line inside the box. The lines protruding from either side of the box are the lowest and highest values within 1.5 times the IQR from the first and third quartiles, respectively. The relative abundances of all samples of these six dominant bacterial classes in each region are represented by density in **Fig 2.4**. The numbers of samples (*n*) of each region are as follows: East, 355; West, 182; Europe, 294; Asia, 191; and Great Lakes, 196.

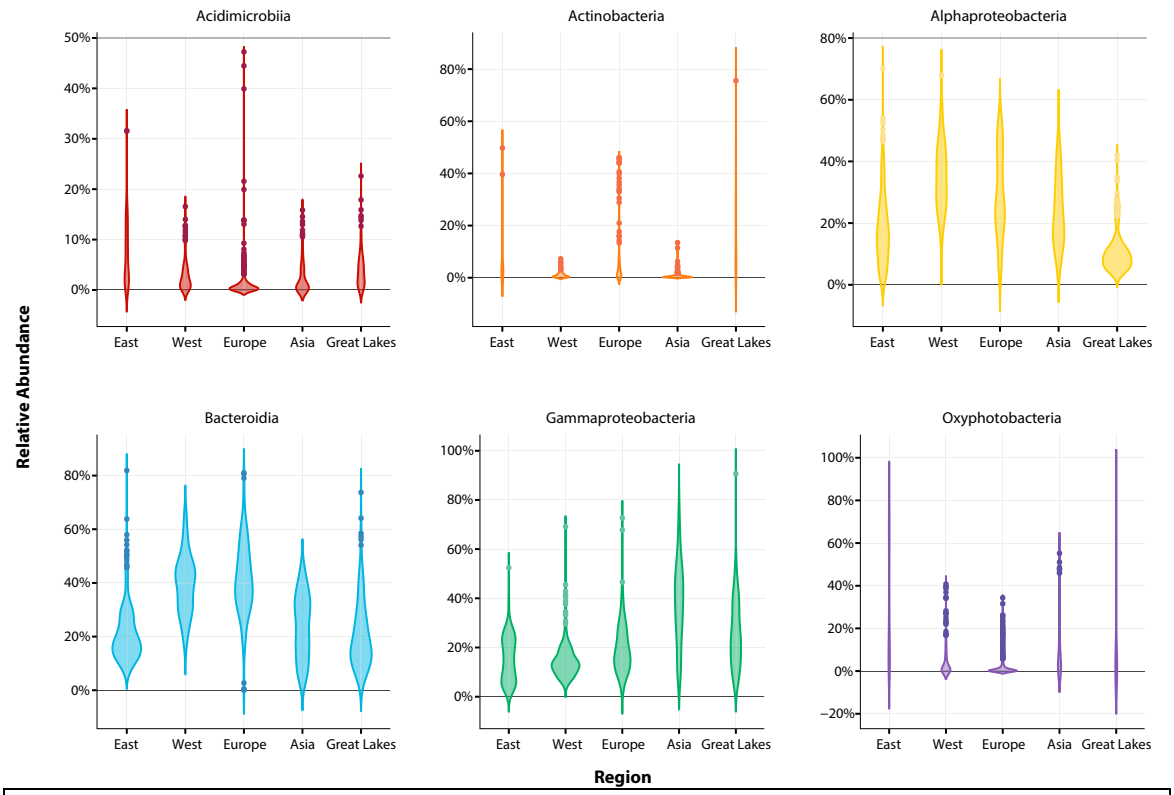


Figure 2. 4

Violin plot of relative abundance. Violin plot of relative abundances. Shown are the density and distribution by relative abundances of the six dominant bacterial classes within all samples ($n = 1{,}218$) and for each region. The wider the distribution, the more samples share similar relative abundances for that taxa. The shape is estimated via a kernel density estimation.

2.5 Machine learning uncovers the biogeographic component of microbial communities

Microbial data are known to be both highly dimensional and compositional (32, 33), and in many cases, the microbial features of the data set are shared between categories to which they belong (e.g., sample type). As a result, many machine learning techniques are often a good approach for understanding how microbial count features of a data set correlate to each other and to a dependent variable (outcome). Compared with the typical statistics used throughout ecology, biogeography, and Earth sciences (33–36), machine learning offers a robust, data-driven estimations of the taxon-spatial associations across globally distributed locations.

We first display the potential to differentiate spatial locations from microbial community data with a multivariate discriminant technique (analysis of similarity [ANOSIM]) applied to both local (all 20 ports) and regional (five regions) scales to

assess the ANOSIM in beta diversity. There were more similarities in the microbial communities between the five regions than between the 20 local locations (ANOSIM for regions, $|R| = 0.609$, $P < 0.001$; for local port locations, $|R| = 0.905$, $P < 0.001$ for Bray Curtis dissimilarity; **Fig. 2.5**), where a higher $|R|$ value suggests more dissimilarity between communities on the regional or local spatial scale. Similar performance was observed for additional distance metrics (**Table 2.2**).

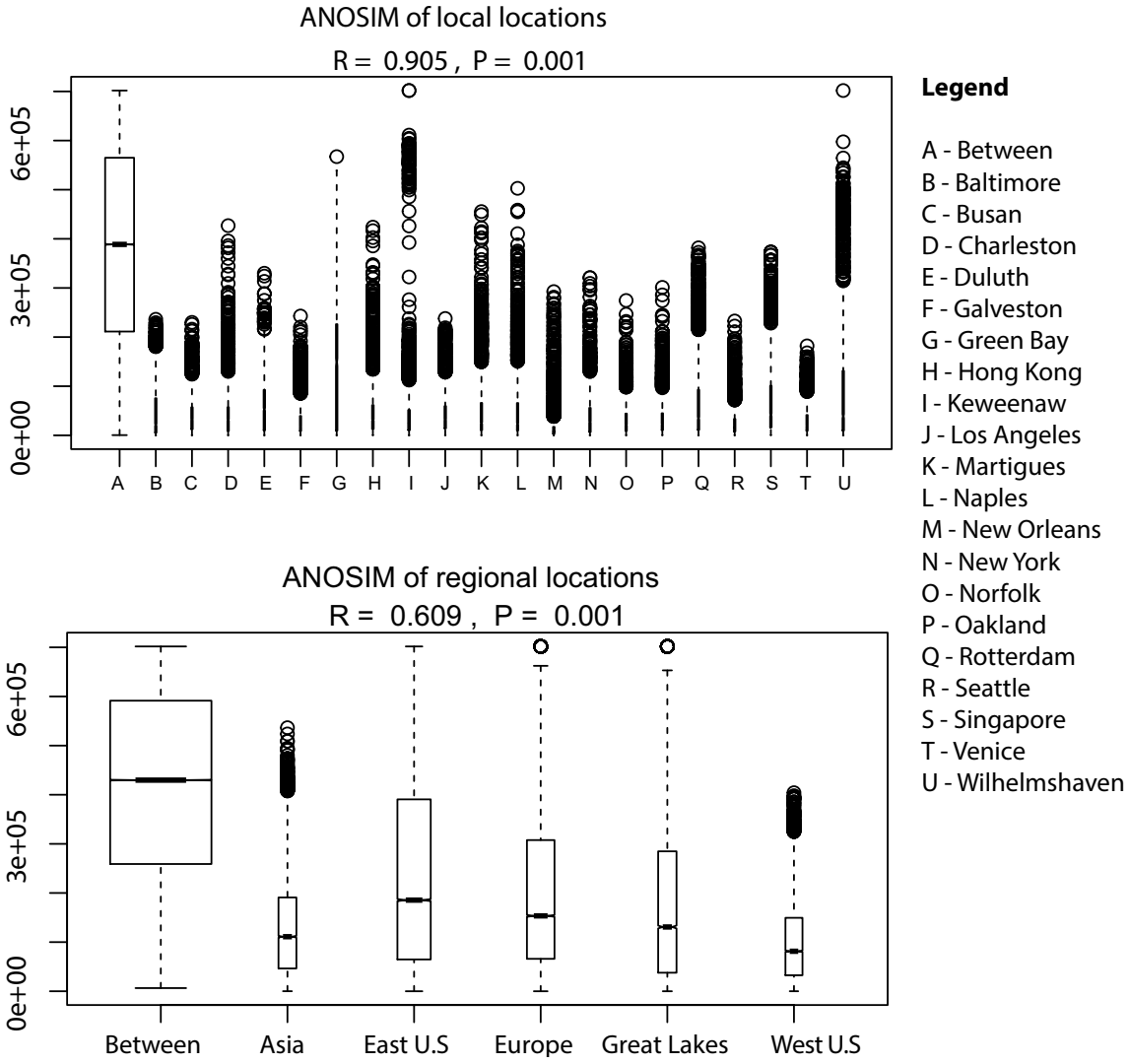


Figure 2. 5

Analysis of similarity of taxonomy to location. ANOSIM plot displaying the dissimilarity between and within local locations (ports) and regions to the microbial communities sampled from them (via Bray-Curtis dissimilarities with 999 permutations). The horizontal line in each box indicates the median; the bottom half of the box indicates the 25th percentile, and the top indicates the 75th

percentile. Whiskers project to the most extreme data point, and the width of each box is directly proportional to the sample size at each port or region. A higher $|R|$ value represents more dissimilarity between taxonomy and local port or regional scales and is based on the difference of mean ranks between and within groups.

Table 2. 2

Various dissimilarity indices through ANOSIM. Various dissimilarity indices through ANOSIM to complement the Bray-Curtis metric that was reported. All ANOSIM $|R|$ statistics displayed with 999 permutations and at 0.001 significance.

ANOSIM dissimilarity metrics		
Dissimilarity metric	Local	Region
	$\|R\|$	
Canberra	0.9607	0.6402
Manhattan	0.5653	0.2541
Jaccard	0.9048	0.6092
Jensen-Shannon	0.9325	0.6242

Additionally, we assessed the community composition through principal-coordinate analysis (PCoA; using Jaccard distances) to observe patterns in the microbial community composition at the regional scale. This form of unsupervised learning is able to simplify the complexity of high-dimensional data sets while retaining trends within bacterial features by transforming it to fewer dimensions. As expected, given how this is an oversimplification of the observed bacterial diversity, only 13.8% of the variation within these communities across each region could be explained by this technique (**Fig. 2.6**).

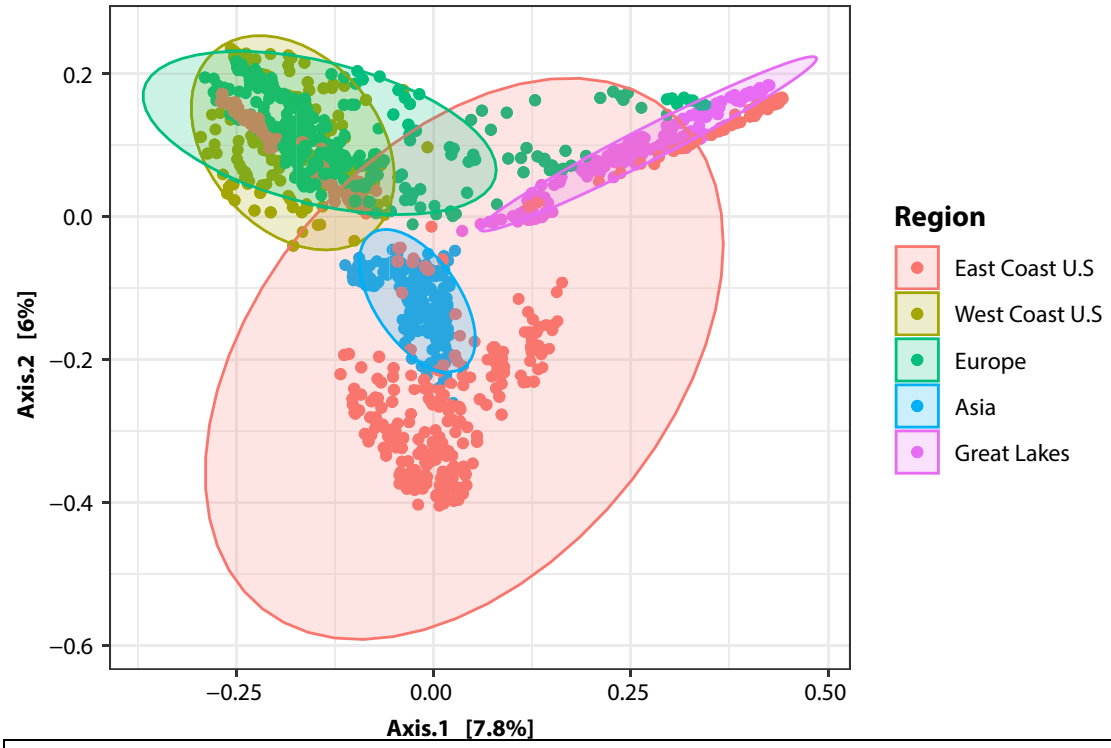


Figure 2. 6

Principal coordinate analysis (PCoA) plot of the global microbial community. (PCoA) plot of the global microbial community. This PCoA displays how much of the total sample variance can be explained in the community of all local port samples ($n = 1{,}218$) using 3,214 ASVs grouped by region (via Jaccard index). Ellipses were calculated assuming a multivariate *t*-distribution with a confidence level of 0.95. Coordinate points clustered closer to each other have more similar microbial communities. The axes indicate coordinate one (Axis.1) and coordinate two (Axis.2), where the percentages in parentheses explain the variation of the whole bacterial community from all of these regions.

Last, we assessed these taxon-spatial relationships through supervised machine learning. We were able to find distinctions in the bacterial community for each of the sampling locations locally (all 20 ports) and regionally (five regions) across our global data set. Using random forests (RF; a form of supervised learning) (37), two independent models were used to classify these local and regional geospatial locations (Y) from their microbial community alone. At both local ($Y = 20$) and regional ($Y = 5$) levels, all samples ($n = 1{,}218$, as observations) were able to be accurately binned into the respective geospatial location from which they were collected with high performance. However, these models had slightly more misclassifications while partitioning microbial communities on a local scale (logarithmic loss [$\log_{loss}$], 0.101; accuracy, 0.994) than on a regional scale ($\log_{loss}$, 0.045; accuracy, 0.995) (**Table 2.3**). Given how these models used the same microbial community structure (3,214 high-resolution bacterial predictors [p]; as

ASVs), the difference in performances between local and regional models suggests that while able to perform global spatial-scale classifications from microbial communities alone, there were more differences within the microbial communities between regions than there were locally between ports in the same region.

Table 2. 3

Model performance index. Machine learning classifier performance index of each taxonomic resolution at either the local port or regional scales. These metrics are reported as the macro averaged results of the ensemble of random forests tuned by the same hyperparameters.

Model resolution	Metric			Predictors
	Accuracy	log_{loss}	Precision/Recall	p
Local: Y = 20				
Phylum	0.84150114	0.58928161	0.85460672/0.83943254	24
Class	0.91242087	0.44398276	0.92566468/0.91223214	38
Order	0.9654911	0.28712688	0.97033929/0.96514087	114
Family	0.97759414	0.25411364	0.98079497/0.97707738	223
Genus	0.97806738	0.1352262	0.98106349/0.97784722	484
ASV	0.99478113	0.10128259	0.99547619/0.995	3,214
Region: Y = 5				
Phylum	0.90991625	0.33421786	0.91091799/0.9037573	24
Class	0.95531278	0.2656141	0.95568987/0.95368017	38
Order	0.98300733	0.16181949	0.98246165/0.98374711	114
Family	0.98277012	0.1286815	0.98300881/0.98234723	223
Genus	0.98930743	0.07437649	0.98949777/0.99056985	484
ASV	0.99535038	0.04514882	0.99497995/0.99539333	3,214

Here, classification performance is observed through a reduction in log_{loss} and its relation to increased accuracy. Model accuracy is the overall proportion of correctly classified samples to the local or regional scale to which they belong. Logarithmic loss (log_{loss}) measures the quality of predictions and is the probabilistic confidence of how each sample was classified to its local port or region (Y) and works by penalizing the incorrect or uncertain predictions. A low log_{loss} is preferred and reflects the distribution of predictions made on a sample toward the true location to which it belongs and how close each sample (observation) was to being misclassified to incorrect geospatial locations.

Interestingly, our ANOSIM results indicated more dissimilarities in the microbial community locally than between five regions, which is in contrast to the RF models which performed better when binning samples into their respective region rather

than their individual port. Multidimensional scaling through principal-coordinate analysis of the global port microbial community composition suggested fewer distinctions in microbial community composition than revealed by modeling through RF. Taken together, these results suggest that the ability for a microbial community to be differentiated on the basis of location is possible through a variety of metrics. However, modeling through RF achieved the highest accuracy in differentiating between samples, suggesting that this form of learning can identify differences in the microbial community better than do many of the standard methods for examining microbial community composition.

2.6 The most abundant microbial taxa can be used to discriminate geospatial locations

The focus of many studies in microbial biogeography has been toward "rare" indicator species as biomarkers for biogeography, as they are assumed to be present in one location and not another (38). Alternatively, highly abundant taxa are easier to detect, as they can differ from the rare biosphere by many orders of magnitude in abundance (18). Therefore, a more generalizable approach for studying biogeographic patterns of microbes may be to leverage the dominant taxa of a system (39).

By observing the overall importance of each bacterial ASV predictor used in both models (local and regional), we identified the microbial taxa responsible for the distinction of these globally distributed geospatial locations. There were 342 of 3,214 ASV predictors used in both models that were considered important (overall importance, ≥1 predictor; local, 250 predictors; regional, 92 predictors). Of these predictors, 68 were shared between the two models. These shared bacterial predictors were classified into eight bacterial classes. Notably, 91.17% of these 68 shared predictor ASVs belong to the six most dominant bacterial classes reported previously (**Fig. 2.2**), while the remaining two classes ("other") accounted for only 8.82% (**Fig. 2.7A**).

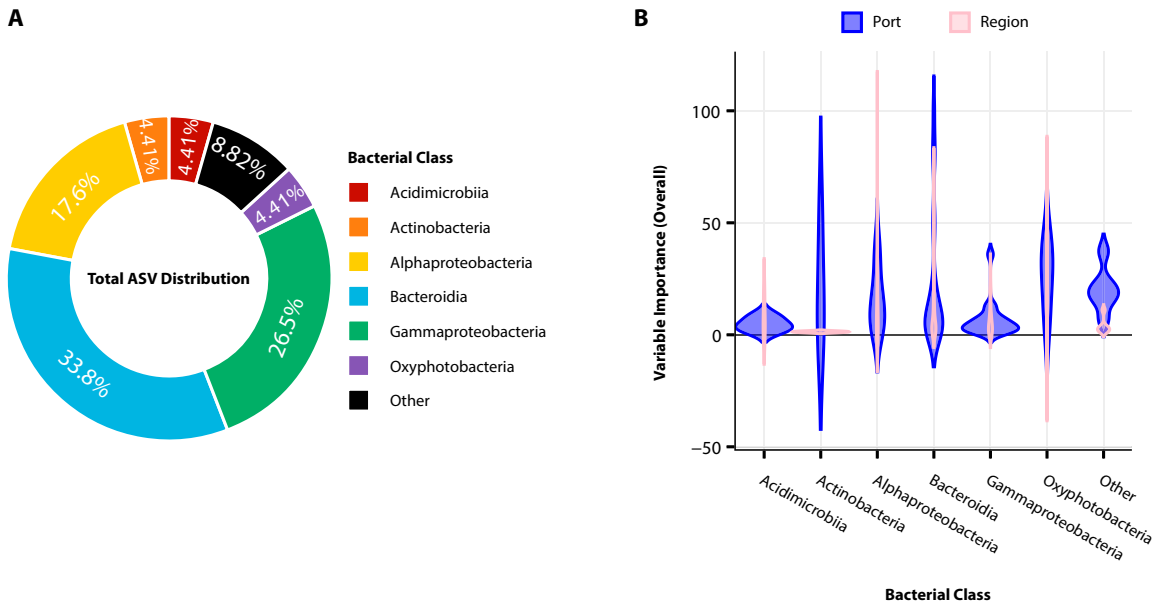


Figure 2. 7

Distributions of shared bacterial classes between machine learning models. Distributions of shared bacterial classes between machine learning models. (A) Donut chart showing percentages of the 68 shared ASV bacterial predictors after binning the ASVs into the dominant bacterial classes to which they belong. (B) Violin plot (shape via kernel density estimation) of the variable importance by distribution and density of the 68 shared predictor ASVs (overall variable importance, ≥1) and binned by the bacterial class to which the ASV belongs, displayed between both local (port) and regional machine learning models. Here, overall importance for each predictor is the scaled mean decrease in accuracy across all class labels (Y) (port, $Y = 20$; region, $Y = 5$). The wider the distribution means, the more similar the importance that ASV predictors belonging to the bacterial class share.

The 68 shared predictors were sorted by their overall importance to show how the dominant bacterial taxa are leveraged to make predictions on both local and regional scales (**Fig. 2.7B**). There were more ASVs considered important in the model used to classify a sample into individual ports than to regions, suggesting that more of the overall community was found to be important while identifying distinctions at the highest resolution of spatial scales. The majority of the predictors used in local classifications were distributed across wider ranges of importance, whereas predictors used to make regional classifications are weighted more similarly (**Fig. 2.7B** and **Fig. 2.8**).

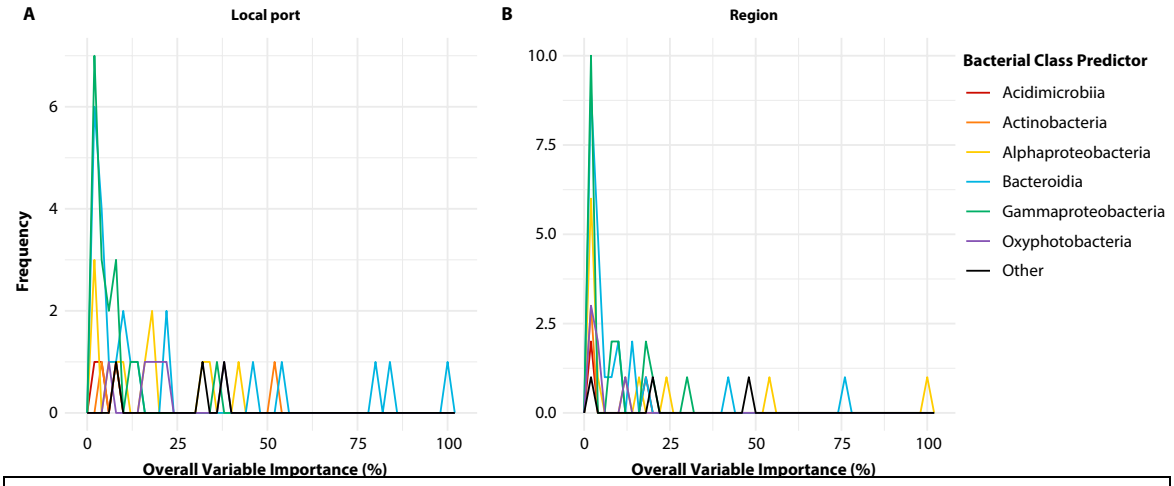


Figure 2. 8

Frequency polygon of shared bacterial predictors from local and regional models. Shared bacterial predictors from local and regional models. This displays the 68 shared predictors from our ASV models (from **Fig. 2.7**). The ASV predictors were binned into their respective taxonomic classes to show the frequency of features across the variable importance gradient. The *y* axis is the proportion of ASVs belonging to each of the bacterial classes, and the *x* axis is their overall importance to the model.

The local model leveraged predictors belonging mostly to the *Bacteroidia* to accurately classify samples, while the regional model used predictors from the *Alphaproteobacteria* (e.g., there is a higher density of predictors in higher overall importance for these classes). (Fig. 4B). *Bacteroidia* and *Alphaproteobacteria* accounted for a large proportion of shared predictors (33.8% and 17.6%, respectively). Between the two models, predictors belonging to the *Oxyphotobacteria* shared similar overall importance and only accounted for 4.41% of the shared predictors. Similarly, the *Acidimicrobiia* also accounted for 4.41% of the shared predictors and had nominal influence as an important predictor, with the highest overall importance of an Acidimicrobiia ASV being 8.25 in the local model and 18.56 regional model.

These results align with the distribution of relative abundances of these six dominant classes reported earlier (**Fig. 2.2** and **Fig 2.3**). The *Proteobacteria* and *Bacteroidetes* were the two most dominant phyla and accounted for the highest percentage of total sequencing reads, along with the two most dominant bacterial classes belonging to the phyla *Alphaproteobacteria* and *Bacteroidia* (Table S1). *Oxyphotobacteria* had the widest range of variability in relative abundance across all samples. Further, there is a correlation between how these models utilize members of the *Alphaproteobacteria* and *Bacteroidia* and their relative abundances on a local or regional scale. Sequence variants of *Alphaproteobacteria* were considered the most important to the regional model, while variants from

Bacteroidia were most important to the local model. The choice of members of these classes as being the most important to these models is consistent with the increased differences observed between relative abundances of *Alphaproteobacteria* observed between regions and of *Bacteroidia* observed between local ports (**Fig. 2.2** and **Fig. 2.3**).

These models used information about members of the most dominant and ubiquitous classes of microbes to make accurate classifications. This suggests that subpopulations in dominant, globally dispersed species are best at explaining geographic patterns in microbial populations. More so, the use of high-resolution ASVs in this study allow for the dissection of fine-scale differences that may represent distinct species, or potentially subspecies, in these populations. These fine-scale differences are able to discriminate between geography at both local and regional spatial scales with high accuracy through machine learning. Observing these regional geographic patterns through abundant taxa has been a challenge largely due to a lack of sufficient sampling density and uniformity in sampling and processing methodology on large spatial scales (9).

2.7 Environmental conditions do not fully explain microbial-spatial diversity on a global scale

How microbial community composition differs between geospatial locations could be attributed to differences in environmental conditions. It has been suggested that the observed composition of abundant taxa in marine environments is likely a reflection of both historical and current environmental influences (18). A number of environmental variables were measured at the time of sample collection, including conductivity, optical dissolved oxygen (ODO) content, pH, salinity, total dissolved solids (TDS) content, and temperature. The distribution of these six physiochemical variables and their association with each region were analyzed. Each region displayed distinctions between each other for each physiochemical condition other than pH (assessed through analysis of variance [ANOVA], $P < 0.001$). (**Fig. 2.9**). Further, we correlated the abundance of each bacterial class with these same physiochemical variables for each region (**Fig. 2.10**). There are many taxa that are strongly correlated with these measured environmental variables. These findings follow previous work that has shown that the environment plays a key role in selecting for the microbial taxa present in a location in marine environments (40, 41).

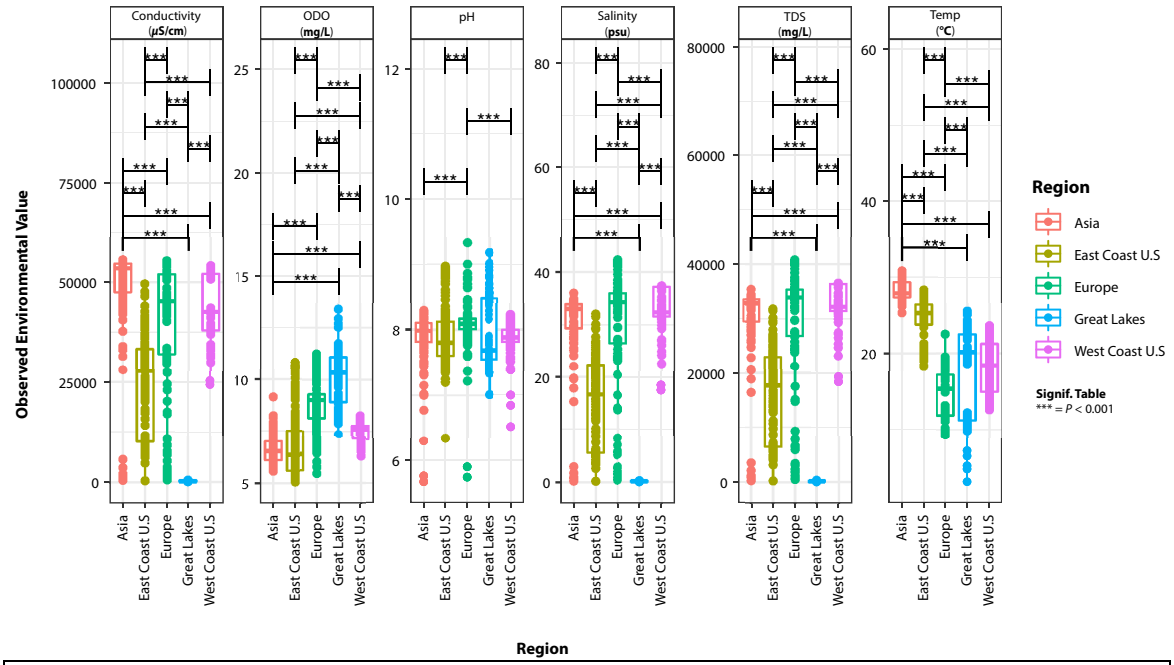


Figure 2. 9

Comparative ANOVA of physiochemistry across each region. Annotated ANOVA of physiochemistry across each region. This annotated ANOVA displays the association of each geographic region and six of the measured environmental variables. All comparisons displayed are considered significantly different between each region ($P < 0.001$).

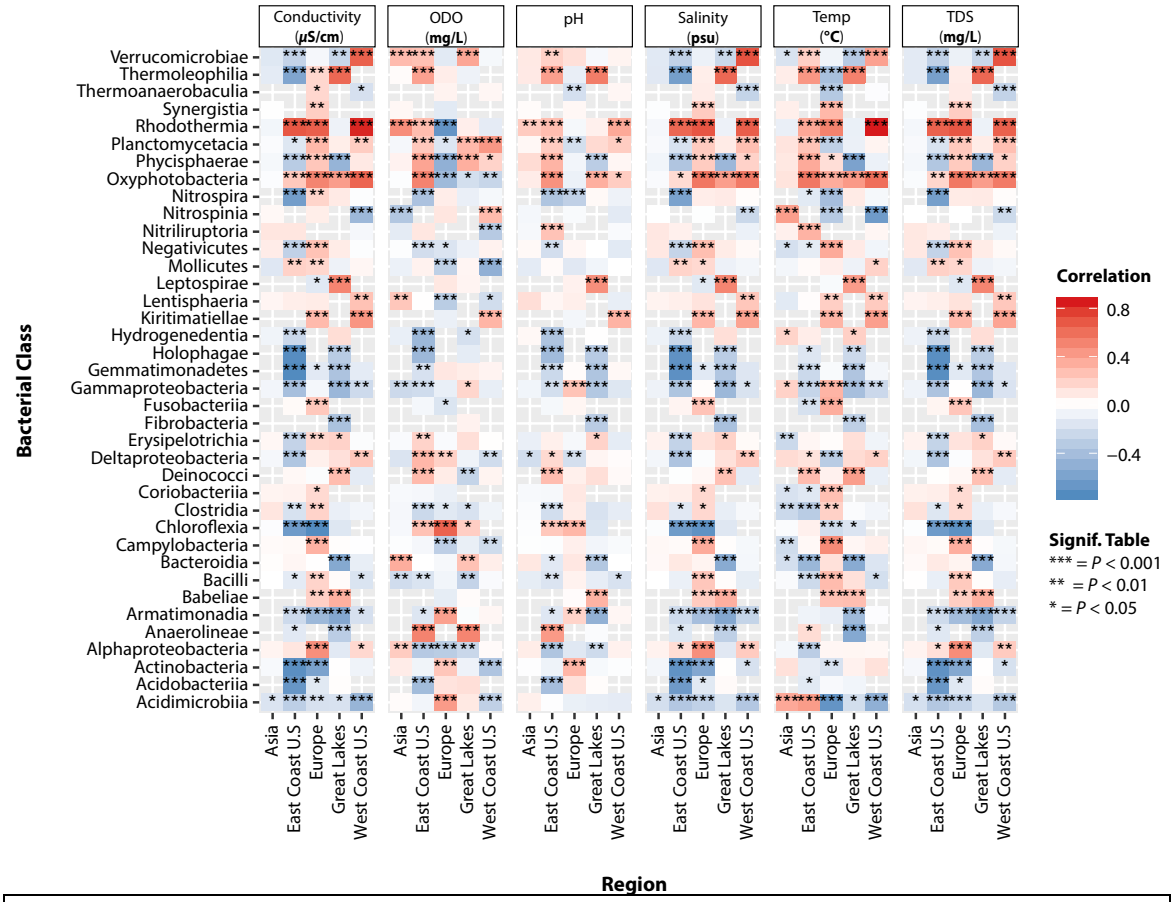


Figure 2. 10

Taxonomic association with region and physiochemistry. Taxonomic association with region and physiochemistry. Displayed are the correlations between the taxonomic abundances of 38 bacterial classes (after agglomerating all 3,214 ASVs used in our models) and environmental variables at each geographic region. A correlation test (Pearson coefficient $|r|$) was performed, and associated *P* values were adjusted for multiple comparisons for environmental variables (Benjamini-Hochberg).

Our classification models were able to accurately discriminate between all 20 ports and five regions by modeling only relative abundances of microbes from the sampled community. To further understand the relationship between environmental conditions and the biogeographic diversity of port microbes, we sought to quantify the amount of variance in the microbial community explained by these measured environmental variables. Across all samples from the 20 port locations, these six physiochemical parameters and their corresponding microbial community composition were used to perform a permutational multivariate analysis of variance (PERMANOVA) (42). This analysis was performed to find the significant conditions that could explain the observed diversity. Conductivity, salinity, and TDS content displayed significant contribution as environmental

factors [adonis, Pr(>F) = 0.001, R^2 = 0.833; 0.002 and 0.002, respectively), which cumulatively explains 83% of the variation in microbial diversity within all 20 port locations as one global community. While these environmental variables were considered significant across our data set, the majority of the significance from conductivity could arise from the range of variability in this environmental parameter across samples, for example, since our samples used in analysis come from environments that are either marine water, brackish water (East Coast United States), or freshwater (Great Lakes) (**Fig. 2.9**). A constrained analysis of principal coordinates (CAP; Bray-Curtis) was subsequently performed on all six of the physiochemical parameters and microbial community data from the 20 geospatial locations. As expected, given the dimensions of the data set, these six environmental conditions could only explain 22.2% of the observed diversity within this global study (**Fig. 2.11**).

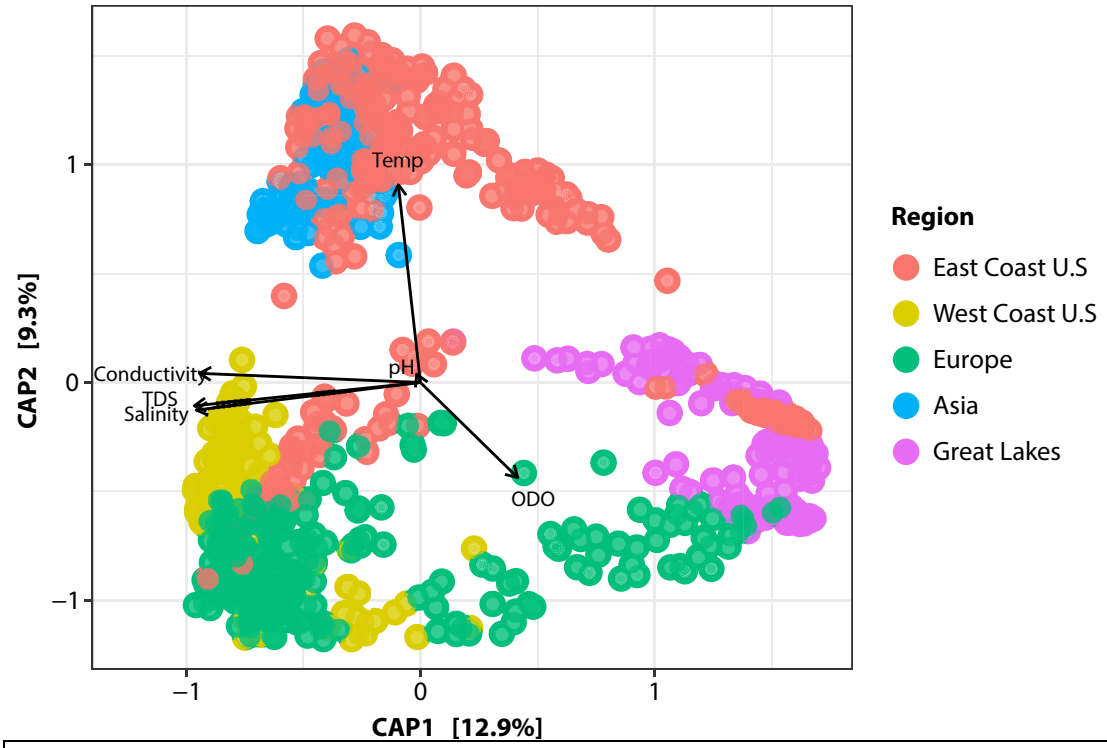


Figure 2. 11

Constrained Analysis of Principal Coordinates (CAP) of beta diversity and physiochemistry. (CAP) plot of beta diversity and physiochemistry. CAP plot (Bray-Curtis distances) displaying the measured environmental variables and their association to sample variance within the microbial community (grouped by region) of all 3,214 ASVs used in modeling. ANOVA on constrained axis used in this ordination, $F = 94.66$, $P < 0.001$.

These findings, along with how our ML models perform independent of any physiochemical parameters supplied, show that although the microbial

community may be influenced by its environment, the measured environmental information alone is not sufficient to explain the observed biogeographic separation in the microbial community composition.

2.8 Differentially enriched taxa lose discriminant ability across large spatial scales

To better understand the microbial groups that explain the observed differences in the microbial communities between locations, we employed pairwise differential abundance (DA) analysis. This approach is commonly used in microbial ecology to identify taxa that are overrepresented in a particular sample (43). After assigning all 3,214 ASVs used in this study into 38 bacterial classes, pairwise DA analysis was done comparing each location against all other locations for these 38 bacterial classes in all of the 20 ports in a one-versus-all manner, resulting in 7,220 pairwise comparisons. Our analyses indicated a complement of microbes that are differentially present in these ports around the world.

A large proportion of these bacterial classes (30/38 [78%]) displayed positive enrichment (log fold change [logFC], ≥ 2; adjusted *P* value [false-discovery rate{FDR}], ≤0.05) in one location over at least one other location. We have termed this the enrichment factor (EF). For example, a bacterial class with an EF of 12 for a location means that the bacterial class has a greater abundance (or is enriched) in that location than in 12 other locations. Our results indicate that each location is composed of a unique consortium of enriched taxa. By assigning an EF, we can ascribe a single bacterial class to a geospatial location that can discriminate it from others. Of the most dominant bacterial classes previously described, only four of the six (*Acidimicrobiia*, *Actinobacteria*, *Gammaproteobacteria*, and *Oxyphotobacteria*) were differentially abundant with an EF of ≥1. *Alphaproteobacteria* and *Bacteroidia* were not considered differentially enriched in any one location more than another (EF, 0), as the relative abundance across each location is too similar to differentiate geospatial location. Of the 30 bacterial classes that were differentially enriched, 28 unique bacterial classes had an EF of ≥10 throughout all 20 locations (**Fig. 2.12**). The distribution of how the 28 unique bacterial classes predominated these locations regionally are as follows: East Coast of the United States, 21 classes; Asia, 9 classes; Great Lakes, 14 classes; Europe, 18 classes; and West Coast of the United States, 5 classes (**Fig. 2.12**). The reported enrichments of the dominant classes at EF of ≥10 were congruent with the relative abundances reported earlier and allow for better differentiation than with abundances alone (**Fig. 2.2** and **2.3**).

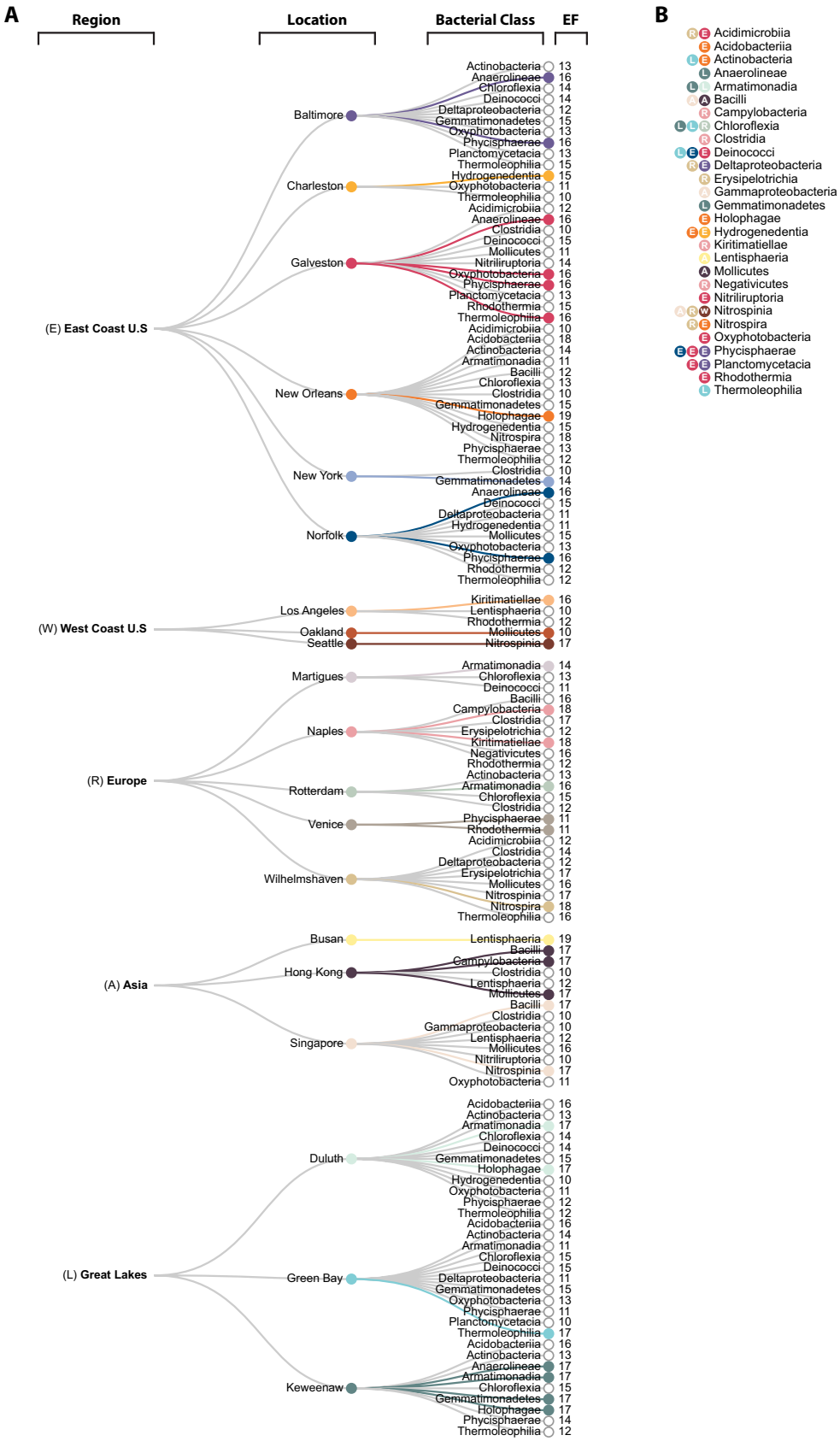

A
Region
Location
Bacterial Class
EF
(E) East Coast U.S
(W) West Coast U.S
(R) Europe
(A) Asia
(L) Great Lakes
Baltimore
Charleston
Galveston
New Orleans
New York
Norfolk
Los Angeles
Oakland
Seattle
Martigues
Naples
Rotterdam
Venice
Wilhelmshaven
Busan
Hong Kong
Singapore
Duluth
Green Bay
Keweenaw
Actinobacteria 13
Anaerolineae 16
Chloroflexia 14
Deinococci 14
Deltaproteobacteria 12
Gemmatimonadetes 15
Oxyphotobacteria 13
Phycisphaerae 16
Planctomycetacia 13
Thermoleophilia 15
Hydrogenedentia 15
Oxyphotobacteria 11
Thermoleophilia 10
Acidimicrobiia 12
Anaerolineae 16
Clostridia 10
Deinococci 15
Mollicutes 11
Nitriliruptoria 14
Oxyphotobacteria 16
Phycisphaerae 16
Planctomycetacia 13
Rhodothermia 15
Thermoleophilia 16
Acidimicrobiia 10
Acidobacteriia 18
Actinobacteria 14
Armatimonadia 11
Bacilli 12
Chloroflexia 13
Clostridia 10
Gemmatimonadetes 15
Holophagae 19
Hydrogenedentia 15
Nitrospira 18
Phycisphaerae 13
Thermoleophilia 12
Clostridia 10
Gemmatimonadetes 14
Anaerolineae 16
Deinococci 15
Deltaproteobacteria 11
Hydrogenedentia 11
Mollicutes 15
Oxyphotobacteria 13
Phycisphaerae 16
Rhodothermia 12
Thermoleophilia 12
Kiritimatiellae 16
Lentisphaeria 10
Rhodothermia 12
Mollicutes 10
Nitrospinia 17
Armatimonadia 14
Chloroflexia 13
Deinococci 11
Bacilli 16
Campylobacteria 18
Clostridia 17
Erysipelotrichia 12
Kiritimatiellae 18
Negativicutes 16
Rhodothermia 12
Actinobacteria 13
Armatimonadia 16
Chloroflexia 15
Clostridia 12
Phycisphaerae 11
Rhodothermia 11
Acidimicrobiia 12
Clostridia 14
Deltaproteobacteria 12
Erysipelotrichia 17
Mollicutes 16
Nitrospinia 17
Nitrospira 18
Thermoleophilia 16
Lentisphaeria 19
Bacilli 17
Campylobacteria 17
Clostridia 10
Lentisphaeria 12
Mollicutes 17
Bacilli 17
Clostridia 10
Gammaproteobacteria 10
Lentisphaeria 12
Mollicutes 16
Nitriliruptoria 10
Nitrospinia 17
Oxyphotobacteria 11
Acidobacteriia 16
Actinobacteria 13
Armatimonadia 17
Chloroflexia 14
Deinococci 14
Gemmatimonadetes 15
Holophagae 17
Hydrogenedentia 10
Oxyphotobacteria 11
Phycisphaerae 12
Thermoleophilia 12
Acidobacteriia 16
Actinobacteria 14
Armatimonadia 11
Chloroflexia 15
Deinococci 15
Deltaproteobacteria 11
Gemmatimonadetes 15
Oxyphotobacteria 13
Phycisphaerae 11
Planctomycetacia 10
Thermoleophilia 17
Acidobacteriia 16
Actinobacteria 13
Anaerolineae 17
Armatimonadia 17
Chloroflexia 15
Gemmatimonadetes 17
Holophagae 17
Phycisphaerae 14
Thermoleophilia 12
B
Acidimicrobiia
Acidobacteriia
Actinobacteria
Anaerolineae
Armatimonadia
Bacilli
Campylobacteria
Chloroflexia
Clostridia
Deinococci
Deltaproteobacteria
Erysipelotrichia
Gammaproteobacteria
Gemmatimonadetes
Holophagae
Hydrogenedentia
Kiritimatiellae
Lentisphaeria
Mollicutes
Negativicutes
Nitriliruptoria
Nitrospinia
Nitrospira
Oxyphotobacteria
Phycisphaerae
Planctomycetacia
Rhodothermia
Thermoleophilia

Figure 2. 12

Cluster Dendrogram of pairwise comparisons from differential abundance analysis. Cluster dendrogram of pairwise comparisons from differential abundance analysis. (A) Dendrogram displaying the 28 unique bacterial classes across all locations with an EF of ≥10. The colored line projecting from each location indicates which class(es) had the highest EF in that location. (B) Displayed for each of the 28 bacterial classes is which location (indicated by color) had the highest total EF for that class along with which region in which it is located (indicated by letter).

Although DA analysis could identify the dominating bacteria in different ports, we observe that for multiple bacterial classes, the same EF was observed at multiple locations (**Fig. 2.12B**). Collectively, the use of EF profiles could only differentiate 15 different geospatial locations using 24 bacterial classes, while our machine learning models found 68 subpopulations belonging to eight bacterial classes adequate enough to differentiate all 20 port locations (**Fig. 2.7B** and **2.12B**). Machine learning approaches are able to integrate the interaction of multiple features for classification, which is not possible when considering each microbial class independent of each other as DA analysis does. This outlines another strength of the use of machine learning approaches for understanding microbial diversity and biogeography.

The use of enrichment factor and DA analyses did not pick up on some of the most abundant and prevalent taxa that were found to be important for the machine learning classification (*Alphaproteobacteria* and *Bacteroidia*). Instead, low-abundance and low-prevalence taxa were used as discriminators of geospatial location. This observed limitation of DA analysis is consistent with the more generalizable approach of leveraging the highly abundant and ubiquitous taxa for discriminating globally distributed geospatial locations. In the case of ML, as shown with our modeling, accurate classifications are achieved by incorporating the entire community, despite using either all high-abundance taxa, low-abundance taxa, or a mixture of these taxa. This discrepancy between DA analysis and ML may be that the ML models were constructed using ASVs and that the DA analysis was done using an agglomerated table at the taxonomic class level. The use of the class taxon table in the DA analysis was out of the necessity to limit the number of comparisons needed. However, some resolution in the data was lost by agglomerating ASVs into a single class category. Therefore, ML allows for an appreciation of high-resolution microbial count data to observe biogeography.

2.9 The ability to discriminate patterns of biogeography is apparent at the phylum level

Our previous machine learning models performed very well at the highest level of resolution (ASVs). Therefore, we wanted to determine the ability lower levels of resolution of the microbial community to discriminate geographic location. To decrease resolution, all 3,214 raw sequence variant features (ASVs) from our amplicon reads were binned into their respective taxonomic level (phylum, class, order, family, and genus) and modeled through RF to predict local and regional spatial scales from our samples (n = 1,218). Interestingly, the ability for machine learning to establish contrasts in geospatial diversity is apparent at lower taxonomic resolution than expected (**Fig. 2.13**).

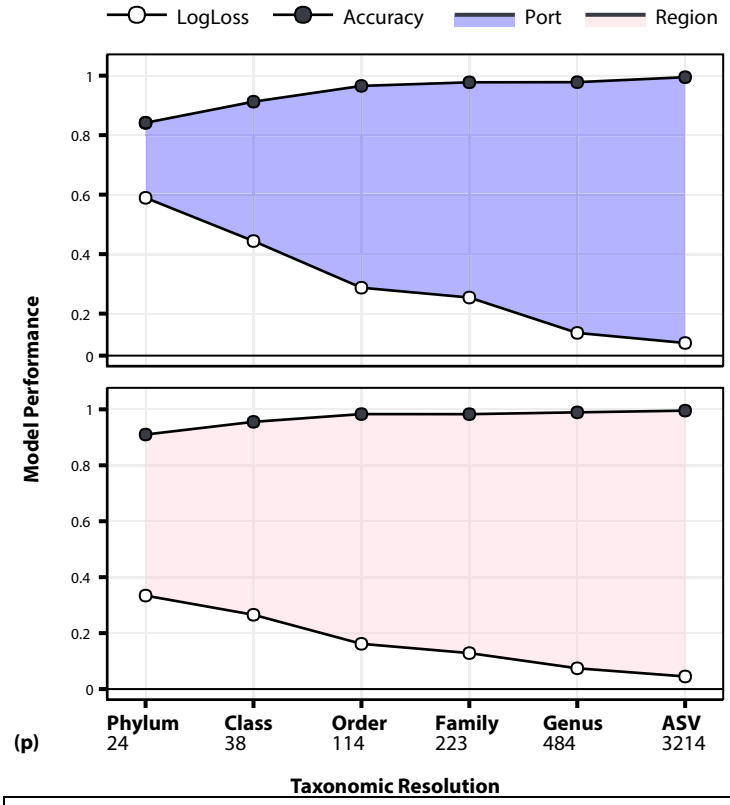


Figure 2. 13

Machine learning model performance at each taxonomic resolution. Filled line plot displaying overall logarithmic loss (log_{loss}) and accuracy in our machine learning models at each level of taxonomic resolution. Taxonomic resolution is in increasing order on the x axis along with the number of predictors (p) used in each model. These models vary in their feature space or number of predictors and class labels (Y) (port, Y = 20; region, Y = 5). All of these multiclass classification models were transformed to 20 one-versus-all or 5 one-versus-all binary classification tasks based on Y. The performance metrics log_{loss} and accuracy are expressed as

the respective models' macro averaged results of the ensemble of random forests tuned by the same hyperparameters.

There were considerable improvements in our performance metrics (log_{loss}/accuracy) between spatial scales (local or regional) with models built from the lowest to highest levels of taxonomic resolution (phylum to genus) (**Table 2.3**). As taxonomic resolution increased, there was a consistent increase in accuracy and decrease in log_{loss}, indicating that our models performed better with increasing taxonomic resolution. Overall, the regional models outperformed the local port models, supporting our earlier findings that learning the biogeography of each sample becomes more challenging as the number of potential geographic locations (Y) it could have come from increases (**Fig. 2.13**).

Even at the lowest taxonomic resolution of phylum, our models were quite accurate in differentiating geospatial locations locally (log_{loss}, 0.58; accuracy, 0.84) and regionally (log_{loss}, 0.33; accuracy, 0.90). These accuracies are well above what would be expected for random classifications taking place in our models (based on model kappa, local, 0.83; region, 0.88). The highest reduction of log_{loss} was observed between class-order resolution in both the local and region models (local, 0.16; regional, 0.1) (**Fig. 2.13** and **Table 2.3**).

It is notable that of the ASV models which are composed of all ASVs, 3,214 performed better than all lower levels of taxonomy (phylum to genus), where the features arise from agglomerating all 3,214 ASVs into their respective taxonomic levels. This observed trend in increased resolution (e.g., increased predictors [p]) to model performance can be explained by how lower-taxonomic resolutions offer a lower bacterial feature space for which models learn. This finding likely suggests that ML model performance is a result of how much of the microbial community it has available to make data-driven spatial distinctions. Although we observe this resolution-performance scaling, an interesting finding is that at the phylum level, enough differences in the community were observed to bin all samples into their respective port and region with relatively high accuracy. Additionally, we display the ability to agglomerate taxa, which reduces the dimensionality of the data by more than an order of magnitude and results in only a marginal decrease in classification performance (**Fig. 2.13**).

To determine how these models leverage what we know about the underlying structure of the microbial community at these spatial locations, we assessed the regional model at the taxonomic class-level resolution (log_{loss}, 0.26; accuracy, 0.95) (**Table 2.3**). In this model, the top 10 important bacterial classes and their overall importance across each region were assessed. This reflects how well these bacterial classes could be leveraged by the ML model to help differentiate samples from all 20 ports or five regions. We found that five of the 10 important predictors were

among the most dominant classes in this data set, as reported previously, each with an overall variable importance of >50% (**Fig. 2.14**). *Acidimicrobiia*, *Bacteroidia*, and *Oxyphotobacteria* were considered most important for samples from Europe (overall importance, 100%, 97.88%, and 51.94%, respectively), while the importance of *Actinobacteria* and *Gammaproteobacteria* was highest for samples from Asia (71.17% and 59.14%, respectively). The overall importance of these taxa in differentiating each region through ML is not directly proportional to the average relative abundance reported for these regions.

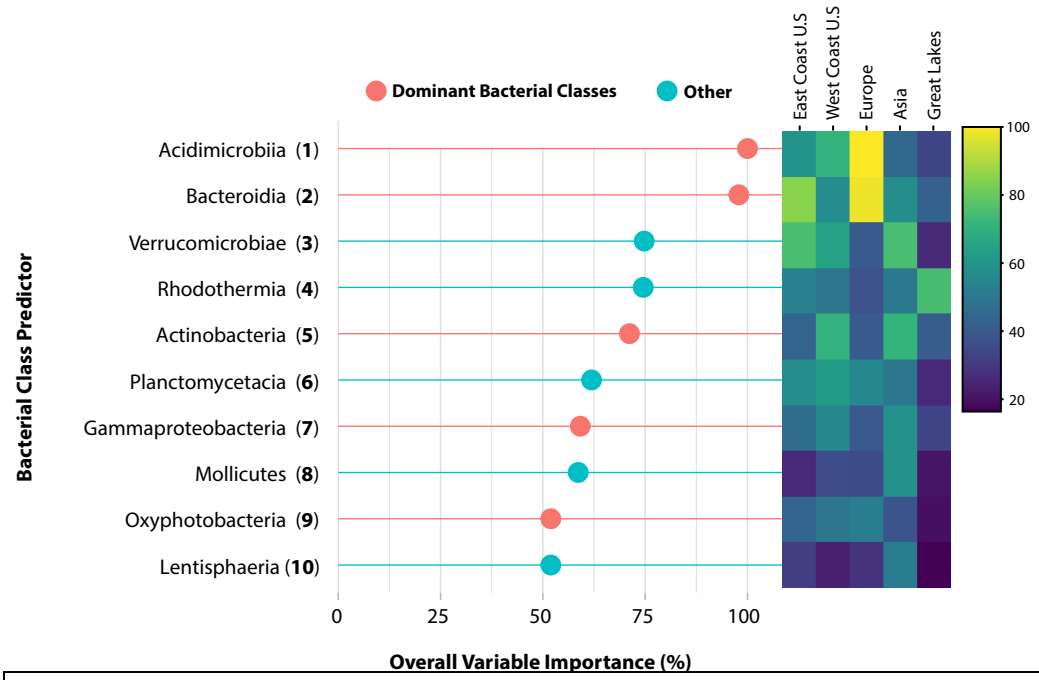


Figure 2. 14

Distribution of taxonomic importance by region. Displayed are the important predictor variables identified by the regional model at taxonomic class-level resolution (**Fig. 2.13**) and are taxa that are best at differentiating geospatial location. The red lines indicate that these taxa were among the top six dominant bacterial classes. The overall variable importance is the scaled mean decrease in accuracy for that predictor across all regions ($Y = 5$) and for the ensemble of random forest classifications (these predictors were consistently important across the decision trees in the model). The heat map to the right displays the distribution of overall importance across each region to show the relationship between these bacterial taxa and how they were leveraged by the model to classify samples into each geographic region.

It is notable that during these taxon-spatial assessments through ML, Europe has the lowest average relative abundance for *Acidimicrobiia* and the highest for *Bacteroidia* despite the two taxa having the highest variable importance in this

region (**Fig. 2.3** and **2.14**). In differentiating regions employing DA analysis through enrichment, we observe the opposite behavior. This could be indicative of these ML models making classifications off a common trend in the microbial abundance (low abundance in one location over others). This finding suggests that caution must be used while inferring associations of a microbial community based on the interpreted importance of taxa in a machine learning model. As such, the variable importance of a taxon is not a direct representation of its biological enrichment in a particular location.

Alphaproteobacteria was the only dominant bacterial class that was not considered an important predictor in the bacterial class-level resolution regional model. Interestingly, the absence of this class as part of the top 10 important predictors is consistent with DA analysis results, where *Alphaproteobacteria* could not be considered differentially enriched in any one location more than another. Despite how *Alphaproteobacteria* seemed negligible when observed from both a lower resolution (ML model, class) and DA analysis (**Fig. 2.12** and **Fig. 2.14**), the ML model utilizing the highest-resolution predictors (ASVs) found *Alphaproteobacteria* to be quite a significant predictor. Sequence variants of *Alphaproteobacteria* were given the highest overall importance in our ASV models regionally (100%), while the same variant was given an overall importance of 42.67% locally (**Fig. 2.7** and **Fig. 2.8**). The combination of these findings suggests that computationally, these ML models are using different microbial community information at each level of taxonomic resolution to make their predictions and to maintain high accuracy. Biologically, this suggests that biogeographic patterns exist in the presence of distinct ASVs within ubiquitous classes which are present at similar abundances throughout these locations (e.g., ASVs can differentiate location, but the total abundance of the bacterial group to which the ASV belongs is not observably different between locations).

Gammaproteobacteria had relatively similar average relative abundance across all regions (**Fig. 2.3**). Our machine learning model assigned an overall importance to *Gammaproteobacteria* commensurate to how useful it was to the model for making spatial distinctions across all regions (33.12% to 59.14%) (**Fig. 2.14**). This could provide insight into how bacterial taxa with low variability in abundance between locations contribute to machine learning model performance. Similar and notable distinctions between the ML overall importance and DA analysis enrichment metrics were found for the two dominant classes that were not considered differentially enriched (*Bacteroidia* and *Alphaproteobacteria*) yet were assigned an overall variable importance of 100% and 0%, respectively (**Fig. 2.14** and **Fig. 2.15**).

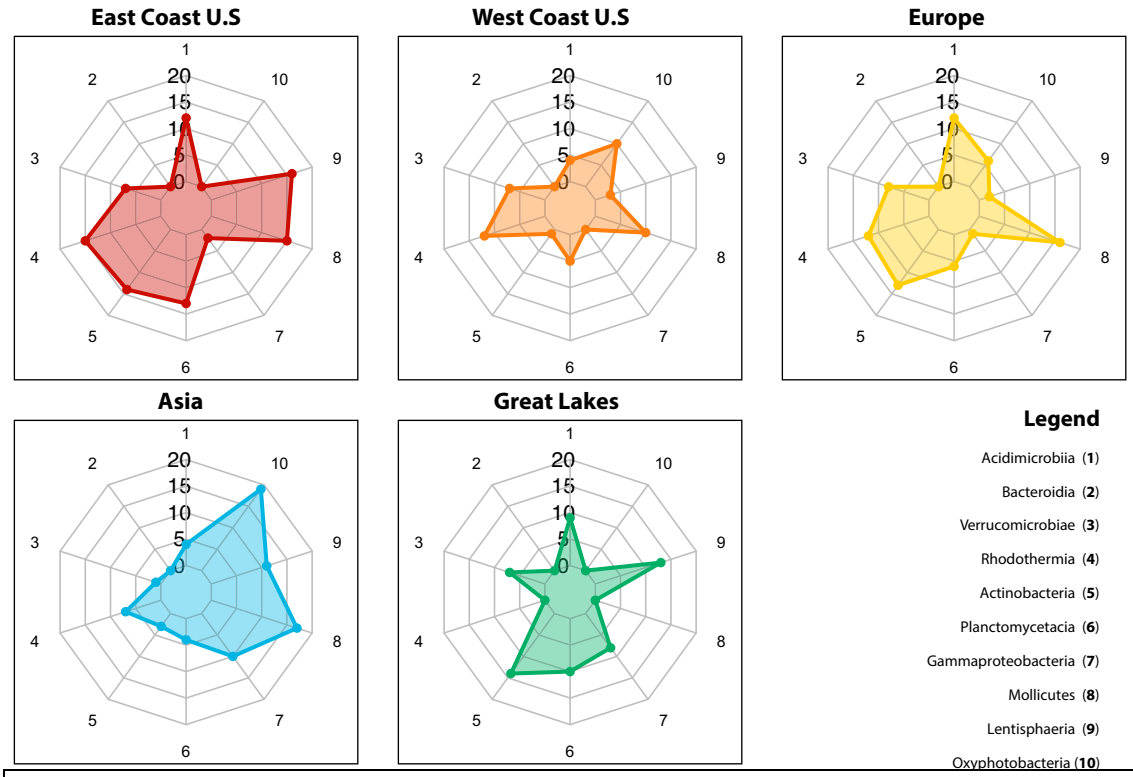


Figure 2. 15

Radar chart of enrichment factors of important predictor taxa. Radar chart of enrichment factors of important predictor taxa. These plots show the enrichment factor (EF) of the top 10 important predictor taxa (**Fig. 2.14**) assigned during DA analysis. The vertical axis represents an EF scale of 1 to 20 (as there are 20 local ports). The numbers around the radar charts correspond to the taxa in the legend and indicate those considered most important in their ability to differentiate these geographical regions.

2.10 Summary and outlook

The ability for us to accurately differentiate between locations using microbial abundance information at high taxonomic levels (albeit low resolution compared to ASVs) suggests that broad differences exist in these microbial communities globally. However, these ML models were slightly more accurate with higher-resolution data, which signifies the importance of geographically distinct subpopulations of the dominant and ubiquitous groups.

This study reports the microbial biogeography of 604 locations belonging to 20 shipping ports distributed globally. We provide a comprehensive data set for the largest study of port-associated microbial communities to date that permits the robust analysis of microbial biogeography across global spatial scales and physiochemical gradients. Accompanying the larger Tara Oceans Project (30) and

Global Oceans Sampling Expedition (GOS) (**31**), this work expands our ability to understand the biogeography of microorganisms in our world's marine and freshwater aquatic ecosystems.

We identified how much of the complex microbial community structure could be explained in these locations by enrichment through differential abundance analysis and machine learning. Our machine learning models could detect biogeographical patterns in the presence of distinct ASVs within the most ubiquitous and abundant groups (*Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, and *Proteobacteria*), despite these groups having seemingly relatively equal abundances throughout each location. Distinctions in the microbial community for all 20 ports and five regions into which they group were observable at the lowest level of taxonomic resolution (phylum) and became more granular as we increased to the highest resolution (ASVs) both locally (for phylum, $\log_{loss}$, >0.58; accuracy, 0.84; for ASV, $\log_{loss}$, 0.10; accuracy, 0.99) and regionally (for phylum, $\log_{loss}$, 0.33; accuracy, 0.90; for ASV, $\log_{loss}$, 0.04; accuracy, 0.99).

Machine learning could discern how each location contained a distinct composition of sequence variants belonging to these highly abundant taxa better than could commonly used multivariate discriminant techniques and differential abundance analysis. This strongly suggests that between machine learning, commonly used multivariate discriminant techniques, and differential abundance analysis, ML is an optimal approach to uncover biogeographic patterns. Our ML models could appreciate the nature of microbial count data in how both high- and low-abundance bacterial features of the community are distributed across samples and therefore across geospatial locations. As such, these ML models provide a way of finding patterns in diversity and gauging the relative importance of taxa in the overall microbial community at each location on a global scale. Notably, we observed biogeographic patterns in the microbial community composition at a regional scale, where this has previously been a challenge in microbial biogeography across large sampling densities and spatial scales (9).

The work presented here only included samples from a single time point, all during the summer. Therefore, we were unable to address the impact of seasonal changes and/or severe weather events on the observed biogeographic patterns. Since microbial communities are known to vary by season and in response to episodic weather events, we expect there be seasonal impacts on the observed patterns. Analysis of the microbial diversity across two seasons, fall and summer, in the Great Lakes stations used in this study (Duluth, MN; Green Bay, WI; and Keweenaw, MI) shows that the microbial community composition in these locations maintained geospatial taxonomic indicators through these two seasons (44). Future work could include investigation into the temporal dynamics of the observed microbial biogeography of this system. It has been shown previously

that community composition shifts in response to seasonal changes can be detected at the level of major taxa (41). We expect that despite the changes in community composition, the dominant and ubiquitous groups would remain throughout seasonal changes. In contrast, taxa that are less abundant and considered rare are seldom retrieved by common molecular techniques that we use on large-scale sampling expeditions (45). Our observation that members of abundant and ubiquitous groups are indicators of geospatial location suggest that these biogeographic patterns may be robust to seasonal changes. Despite longitudinal research showing how dominant bacteria of a system persist throughout the year (40, 41, 45–47), more work is needed to observe exactly how abundant taxa may proportionally stabilize their community composition across large spatial scales and after seasonal changes.

Additionally, severe weather events may perturb the system and may result in transient excursions in microbial community composition. Future studies could investigate the ability of the machine learning classifiers developed in this study to accurately classify samples from a location before, during, and after severe weather events to clarify the persistence of biogeographic patterns despite perturbations. While our study demonstrates the utility of random forests machine learning in modeling and identifying biogeographic patterns, additional work is required to more fully appreciate and model the impact of temporal variation, both seasonal and short term, on biogeographic patterns in microbial communities. Furthermore, while our results suggest that random forests machine learning can be used to more fully appreciate biogeographic patterns, more work could be performed that characterizes the potential for random forests to be applied for modeling of temporal variation in microbial communities.

Although we observed that several existing methods were able to provide insights into our global microbial data set, machine learning appears to provide to deepest insights. This in part may be due to the high-dimensional, highly compositional, and naturally sparse (e.g., contains a lot of zeros) nature of microbial community data (**32**). There still, however, remains a challenge in ecology to accurately infer associations between microbial communities (48) and, further, their association between geographic locations (39). Despite observing clear trends in biogeography through this robust system, this outlines the urgency to develop statistical methods that are biologically motivated enough to understand the complex taxon-spatial relationships in microbial count data.

2.11 Materials and methods

2.11.1 Port Selection

Twenty ports were selected to cover globally important ports that varied across a range of environmental conditions, ship traffic, and traffic type (cargo or passenger) and covered multiple continents and various bodies of water. Samples were collected from the following ports: in the Great Lakes at Duluth, Green Bay, and Keweenaw; in the East Coast of the United States at New York (NY), New Orleans (LA), Galveston (TX), Norfolk (VA), Charleston (SC), and Baltimore (MD); in the West Coast of the United States at Seattle (WA) and Oakland and Long Beach (CA); in Europe at Venice and Naples (Italy), Martigues (France), Rotterdam (the Netherlands), and Wilhelmshaven (Germany); and in Asia at Busan (South Korea), Hong Kong, and Singapore.

2.11.2 Sampling

The samples used in this study (n = 1,218) were collected from 604 locations across eight countries and three continents at a total of 20 ports spanning the Great Lakes, Pacific Ocean, Atlantic Ocean, North Sea, Sea of Japan, South China Sea, Mediterranean Sea, and Adriatic Sea. All samples were collected between May and August 2017. Between 27 and 38 sampling stations were chosen in each port to provide sufficient replication and adequate representation of the range of conditions found within that port. At each station, surface water samples (1 liter) were taken from various locations within that port, each with metadata. Samples were subsequently filtered through a glass fiber prefilter with a 1.6-µm pore size (47-mm diameter) and a 0.2-µm pore-size polyethersulfone (PES) membrane postfilter (47-mm diameter) (Sterlitech Corporation) using a Cole-Parmer Masterflex E/S 115 VAC portable sampler. Filters were placed in 2-ml Eppendorf tubes with 500 µl RNA/DNA shield (ZymoBIOMICS) and stored at ambient temperatures until transported back to the laboratory to be stored at –80°C. Multiparameter data of water quality (conductivity, ODO, pH, salinity, TDS content, temperature, and dissolved oxygen content) along with global positioning system (GPS) coordinates of each sampling site were recorded *in situ* with a YSI ProDSS digital sampling system that was calibrated before each sampling trip.

2.11.3 DNA extractions

DNA was extracted from each filter using the ZymoBIOMICS DNA microprep D4305 kit (Zymo Research, Irvine, CA, USA), and for each sample, both the prefilter (1.6-µm pore size, 47-mm diameter) and postfilter (0.2-µm, 47 mm

diameter) were cut in half, where one half was to be used in the DNA extraction and the other half stored as a contingency.

2.11.4 DNA sequencing

First-stage amplification PCRs were carried out in 25-µl mixtures consisting of 12.5 µl Phusion high-fidelity PCR master mix (Thermo Fisher Scientific, Waltham, MA, USA) containing deoxynucleoside triphosphates (dNTPs) at a concentration of 200 mM each, optimized reaction buffer, 1.5 mM $MgCl_2$, and 1 U high-fidelity polymerase per reaction in 96-well VWR polypropylene plates. The primer pair 515f and 926r was used at a concentration of 0.4 µM to amplify a construct that spans the variable regions 4 and 5 (V4–V5) of the 16S rRNA gene (49). The PCR thermal cycler settings were as follows: 95°C for 3 min; 25 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and 72°C for 5 min. PCR cleanup was performed after first-stage amplification PCR to remove residual primers and excess reagents from PCR mixtures. For this cleanup, we followed the MiSeq library preparation guide (Illumina, San Diego, CA) and deviated from the standard protocol by using AxyPrep Mag PCR cleanup beads (Corning, Big Flags, NY, USA), using 10 mM Tris at a pH of 8 (down from 8.5) and by using 28 µl AxyPrep beads in the second-stage cleanup since the PCR volume was 25 µl (down from 50 µl). Second-stage indexing PCRs took place under the same mixture conditions as first-stage amplification PCR and with primers that contained a unique index sequence for each sample and the Illumina sequencing adaptors. An additional PCR cleanup was done after second-stage PCR, eluting to a final volume of 50 µl. Library preparation and sample pooling were performed according to the MiSeq 16S sequencing library preparation guide (Illumina). The products from the second-stage indexing PCR and subsequent cleanup stages were pooled into a library for sequencing at an equimolar concentration of 10 nM after ensuring that primer contamination was absent or at a minimum using a 2100 Bioanalyzer (Agilent, Santa Clara, CA). Denaturation and dilution of the pooled 16S rRNA gene library were performed according to the MiSeq 600-cycle V3 reagent kit guide (Illumina) to produce a 2 × 300-bp paired-end run. These samples were sequenced over three separate sequencing runs containing 672, 480, and 396 samples, respectively.

2.11.5 Computational analysis and visualization

All statistical analysis, machine learning models, and visualization were conducted on a local server (Red Hat Enterprise Linux server 7.3 [Maipo]; 256 Gb of random-access memory [RAM]) and on R environment version 3.5.0 (50) using the following packages and associated dependencies: DADA2 (51), phyloseq (52), DESeq2 (53), hpgltools (54), microbiome (55), microbiomeSeq, vegan (56), caret (57), caretEnsemble (58), and randomForest (59), the visualization packages ggplot2 (60) and plotly, and through rawgraphs.io.

2.11.6 ASV identification and taxonomic profiling

Raw 16S rRNA sequencing reads were demultiplexed using the Illumina MiSeq platform. Through the divisive amplicon denoising algorithm (DADA2 package) (51) in R, primer nucleotides were removed, and overlapping paired-end reads were merged, quality filtered, and cleansed of internal standard phiX; to distinguish amplification and sequencing errors from true biological variation in our collected samples, amplicon sequence variants (ASVs) were inferred. To account for learning the inherently different error rates in each of the three separate sequencing runs, samples (672, 480, and 396) from each run were inferred independently (from >100 million bases) so as not to bias the true sequence diversity contained in the final data set of the combined samples. The three independent ASV count tables were merged and then used to resolve and remove chimeric artifacts with higher accuracy as a result of the resolution of ASVs. Traditionally with OTU picking, chimeric sequences are removed in a conservative manner, as closely related sequences are later merged into the same OTU. While using ASVs, a more sensitive removal is accomplished by performing a Needleman-Wunsch global alignment of each sequence, finding bimeras (two-parent chimeras) and localizing combinations from a left and right parent chimera that overlaps the child sequence exactly. From 52,316,084 paired-end input reads, a total of 23,235,684 nonchimeric reads passed our filtering parameters and were used in ASV identification and analysis in this study. We obtained a count table analogous to the generally used OTU table; similarly, our features in this table are composed of the uniquely inferred ASVs that map how many of these amplicon variants were observed in each sample. Taxonomy of ASVs was assigned through DADA2 (51) with a reimplementation of a rapid assignment naive Bayesian classifier that compares our biological sequence variants to a training set of previously accurately classified sequences using the SILVA v132 training set (61, 62).

2.11.7 Dimensionality reduction and normalization of data

A series of filtering criteria were applied to the final sequencing count table of 1,514 samples and 117,397 ASVs. Initially, only samples only from open water and those that had >1,000 16S rRNA reads were chosen to be in our data set for microbial community analysis. Additionally, every ASV that was not under the kingdom *Bacteria* was removed, along with a prevalence filtering step to only keep ASVs that were within ≥15 samples (e.g., an amplicon sequence variant had to be present in 15 or more samples from 1,218 total samples). Subsequently, singleton ASVs that either had a quantity of one in any sample or were only present in one sample along with ASVs that summed to zero across all samples were removed, resulting in a data set of 1,218 open-water samples and 3,214 ASV features. The absolute ASV read counts were logarithmized with the standard $\log_{10}(x + 1)$ using the transform function in the microbiome package in R (55); this count table was

used for all downstream statistical analysis and machine learning. To simplify downstream visualization, supply count tables with reduced feature dimensions, and compare lower-taxonomic-level model performance against high-resolution ASVs both locally and regionally, phyloseq (52) was used to agglomerate all 3,214 ASVs into their respective levels of taxonomy (phylum to genus).

2.11.8 Annotation of environmental conditions

All 3,214 ASVs were used to identify which environmental conditions were considered significant in explaining beta diversity in our microbial community across spatial scales. PERMANOVA (42) was conducted using distance matrices (Bray-Curtis) with 999 permutations in vegan (56), and significance ($P < 0.001$) was assessed through F testing based on the sequential sums of squares between the physiochemical parameters chosen and the five geographic regions to which the local ports were assigned. To account for the trends in environmental conditions and their correlation to each region, these same physiochemical parameters were used to annotate an ANOVA of each condition across all regions ($P < 0.001$). In order to detect the biotic relationships of the taxa and their association to the six physiochemical parameters, we used our ASVs to identify correlations using Pearson coefficient $|r|$ (63), and associated P values were adjusted for multiple comparisons for environmental variables (Benjamini-Hochberg). Finally, to define how well these six physiochemical parameters could explain the total sample variance in the microbial community, a constrained analysis of principal coordinates (CAP; Bray-Curtis) was applied to all 3,214 ASVs using vegan (56).

2.11.9 Analysis of similarity and ordinations

To show whether the microbial community structures of the 3,214 ASVs were significantly different between local ports and regional ports, ANOSIM ($|R|$) was performed on absolute ASV counts using a Bray-Curtis dissimilarity matrix with 999 permutations. To visualize differences within this community, a principal-coordinate analysis (PCoA) was generated using phyloseq (52) using the ordination function (Jaccard index) and visualized through the plot_ordination function, where ellipses were calculated assuming a multivariate t-distribution with a confidence level of 0.95.

2.11.10 Differential abundance analysis and identification of enrichment factors

We used the count table that was agglomerated to the class level as a sufficient level of taxonomic resolution to detect differentially abundant taxa between all ports. These data were used to create an experimental design model with hpgltools (54) so that a pairwise contrast could be made for each of 20 locations against the other and across all features (38 bacterial classes), with n biological replicates supplied as n samples per location, ranging from 52 to 75, with a total of 1,218

samples (**Fig. 2.1**; samples [*n*]). These counts were normalized assuming a negative binomial distribution, and a parametric gamma-family generalized linear model fitting scheme was applied over taxon-wise dispersion estimates using DESeq2 (53). Of these 7,220 pairwise comparisons, taxa were only considered differentially enriched and were assigned an enrichment factor (EF) if they satisfied the following conditions: had a logFC of ≥2, had an adjusted *P* value (FDR) of ≤0.05, and were in one location over at least one other location.

2.11.11 Machine learning

Our normalized ASV and agglomerated genus, family, order, class, and phylum count matrices were used as input data from which to learn. The same hyperparameters were chosen to ensemble the random forests in caret (57) and caretEnsemble (58) as follows: repeated *k*-fold cross-validation ($k = 10$ with 3 repeats) so as to estimate the generalization performance of the models, ntree = 501 (number of trees grown), and a random search for best mtry (the number of predictors sampled at each node); last, input data were centered by removing the mean value of each feature and scaled by dividing nonconstant features by their standard deviation. All models were trained with a multiclass summary function so that macroaveraged results of the ensemble of all random forests tuned by these same hyperparameters could be reported. As these are multiclass classifications, depending on the model type (local, $Y = 20$; regional, $Y = 5$), each model was transformed to either 20 one-versus-all or 5 one-versus-all binary classification tasks. Each model in the ensemble was fit with the same resampling indexes across each *k*-fold.

2.12 References

1. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci U S A 105:7774-8.
2. Fierer N, Lennon JT. 2011. The generation and maintenance of diversity in microbial communities. Am J Bot 98:439-48.
3. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB. 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat Rev Microbiol 10:497-506.
4. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S. 2013. Patterns and processes of microbial community assembly. Microbiol Mol Biol Rev 77:342-56.
5. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111-20.
6. Delgado-Baquerizo M, Reich PB, Khachane AN, Campbell CD, Thomas N, Freitag TE, Abu Al-Soud W, Sorensen S, Bardgett RD, Singh BK. 2017. It is elemental: soil nutrient stoichiometry drives bacterial diversity. Environ Microbiol 19:1176-1188.
7. Hernando-Morales V, Ameneiro J, Teira E. 2017. Water mass mixing shapes bacterial biogeography in a highly hydrodynamic region of the Southern Ocean. Environ Microbiol 19:1017-1029.
8. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-7.
9. Power JF, Carere CR, Lee CK, Wakerley GLJ, Evans DW, Button M, White D, Climo MD, Hinze AM, Morgan XC, McDonald IR, Cary SC, Stott MB. 2018. Microbial biogeography of 925 geothermal springs in New Zealand. Nat Commun 9:2876.
10. Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, Fierer N, Townsend AR, Cleveland CC, Stanish L, Knight R. 2011. Global patterns in the biogeography of bacterial taxa. Environ Microbiol 13:135-144.
11. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences 108:4516-4522.
12. Ghiglione JF, Galand PE, Pommier T, Pedros-Alio C, Maas EW, Bakker K, Bertilson S, Kirchman DL, Lovejoy C, Yager PL, Murray AE. 2012. Pole-to-pole biogeography of surface and deep marine bacterial communities. Proceedings of the National Academy of Sciences 109:17633-17638.
13. Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreas L, Reysenbach AL, Smith VH, Staley JT. 2006. Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4:102-12.
14. Gonzalez A, King A, Robeson Ii MS, Song S, Shade A, Metcalf JL, Knight R. 2012. Characterizing microbial communities through space and time. Current Opinion in Biotechnology 23:431-436.
15. Gibbons SM, Gilbert JA. 2015. Microbial diversity--exploration of natural ecosystems and microbiomes. Curr Opin Genet Dev 35:66-72.
16. Mandakovic D, Rojas C, Maldonado J, Latorre M, Travisany D, Delage E, Bihouee A, Jean G, Diaz FP, Fernandez-Gomez B, Cabrera P, Gaete A, Latorre C, Gutierrez RA, Maass A, Cambiazo V, Navarrete SA, Eveillard D, Gonzalez M. 2018. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Sci Rep 8:5875.

17. Raes EJ, Bodrossy L, van de Kamp J, Bissett A, Ostrowski M, Brown MV, Sow SLS, Sloyan B, Waite AM. 2018. Oceanographic boundaries constrain microbial diversity gradients in the South Pacific Ocean. Proc Natl Acad Sci U S A 115:E8266-E8275.
18. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci U S A 106:22427-32.
19. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-Gonzalez A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. Science 359:320-325.
20. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115-20.
21. Vergin K, Done B, Carlson C, Giovannoni S. 2013. Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. Aquatic Microbial Ecology 71:1-13.
22. Szabó K, Itor P, Bertilsson S, Tranvik L, Eiler A. 2007. Importance of rare and abundant populations for the structure and functional potential of freshwater bacterial communities. Aquatic Microbial Ecology 47:1-10.
23. Mohri M, Rostamizadeh A, Talwalkar A. 2018. Foundations of machine learning, Second edition. ed. The MIT Press, Cambridge, Massachusetts.
24. Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu L, Campbell JH, Fortney JL, Mehlhorn TL, Lowe KA, Earles JE, Phillips J, Techtmann SM, Joyner DC, Elias DA, Bailey KL, Hurt RA, Jr., Preheim SP, Sanders MC, Yang J, Mueller MA, Brooks S, Watson DB, Zhang P, He Z, Dubinsky EA, Adams PD, Arkin AP, Fields MW, Zhou J, Alm EJ, Hazen TC. 2015. Natural bacterial communities serve as quantitative geochemical biosensors. MBio 6:e00326-15.
25. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343-59.
26. Roguet A, Eren AM, Newton RJ, McLellan SL. 2018. Fecal source identification using random forest. Microbiome 6:185.
27. Thessen A. 2016. Adoption of Machine Learning Techniques in Ecology and Earth Science. One Ecosystem 1:e8621.
28. Paliy O, Shankar V. 2016. Application of multivariate statistical techniques in microbial ecology. Mol Ecol 25:1032-57.
29. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639-2643.
30. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans c, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, et al. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359.
31. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16.
32. Lo C, Marculescu R. 2017. MPLasso: Inferring microbial association networks using prior microbial knowledge. PLoS Comput Biol 13:e1005915.

33. Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random forests for classification in ecology. Ecology 88:2783-92.
34. Olden JD, Lawler JJ, Poff NL. 2008. Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83:171-93.
35. Prasad AM, Iverson LR, Liaw A. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9:181-199.
36. Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E. Zimmermann N. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.
37. Breiman, L. 2001. Random Forests. Machine Learning 45, 5-32, doi:10.1023/a:1010933404324.
38. Siddig AAH, Ellison AM, Ochs A, Villar-Leeman C, Lau MK. 2016. How do ecologists select and use indicator species to monitor ecological change? Insights from 14 years of publication in Ecological Indicators. Ecological Indicators 60:223-230.
39. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A 113:5970-5.
40. Ward CS, Yung CM, Davis KM, Blinebry SK, Williams TC, Johnson ZI, Hunt DE. 2017. Annual community patterns are driven by seasonal switching between closely related marine bacteria. ISME J 11:2637.
41. Bunse C, Pinhassi J. 2017. Marine Bacterioplankton Seasonal Succession Dynamics. Trends Microbiol 25:494-505.
42. Schaerer LG, Ghannam RB, Butler TM, Techtmann SM. 2019. Global Comparison of the Bacterial Communities of Bilge Water, Boat Surfaces, and External Port Water. Appl Environ Microbiol 85.
43. Pedros-Alio C. 2012. The rare bacterial biosphere. Ann Rev Mar Sci 4:449-66.
44. Liao J, Cao X, Wang J, Zhao L, Sun J, Jiang D, Huang Y. 2017. Similar community assembly mechanisms underlie similar biogeography of rare and abundant bacteria in lakes on Yungui Plateau, China. Limnology and Oceanography 62:723-735.
45. Logares R, Lindström ES, Langenheder S, Logue JB, Paterson H, Laybourn-Parry J, Rengefors K, Tranvik L, Bertilsson S. 2013. Biogeography of bacterial communities exposed to progressive long-term environmental change. The ISME Journal 7:937-948.
46. Anderson MJ. 2005. Permutational multivariate analysis of variance. Department of Statistics, University of Auckland, Auckland 26:32-46.
47. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200-2.
48. Cai W, Lesnik KL, Wade MJ, Heidrich ES, Wang Y, Liu H. 2019. Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells. Biosens Bioelectron 133:64-71.
49. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403-14.
50. R Development Core Team. 2017. R: A Language and environment for statistical computing.
51. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-3.
52. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.

53. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

54. Belew, A., Hughitt, K. 2018. hpgltools: A pile of (hopefully) useful R functions. R package version 2018.03.
55. Lahti, L., Shetty, S. 2017. microbiome: Utilities for microbiome analysis. R package version 2.4-3.
56. J. Oksanen, R. Kindt, P. Legendre, B. O'Hara, G. Simpson, P. Solymos, M. Stevens, H. Wagner. 2017. vegan: Community ecology package. R package version 2.4-3.
57. Kuhn, M. 2018. caret: Classification and Regression Training. R package version 6.0-80.
58. Deane-Mayer, Z., Knowles, J. 2016. caretEnsemble: Ensembles of Caret Models. R package version 2.0.0.
59. Liaw, A. & Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3),18--22.
60. Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag, New York)
61. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-6.
62. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261-7.
63. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27-46.

3 Persistence and stability of aquatic microbial communities on surface objects: A longitudinal experimental design for object provenance

Preface

The work described previously in *Chapter 2* demonstrates the ability to accurately classify a location a sample was taken from using the microbial community profile alone. In order to extend this application into a real-world detection scheme – we designed a longitudinal field experiment to determine if microbes can serve as indicators of object provenance (previous known origin). More specifically, we sought to detect provenance of vessels in marine and freshwater settings as they travel from one location to another. This form of detection is made possible by understanding the microbial community persistence and stability dynamics of this system in the aggregate and with a subset of the most robust candidate taxa in the community.

Abstract

The work presented here provides a systematic way of investigating microbial communities in the Chesapeake Bay — a real world, complex ecosystem that is abundant with different microbes. We aim to understand the microbial communities of this system by transecting from the home port of Baltimore to the destination port of Norfolk, sampling along the way using two research vessels for two independent voyages (296km each). By understanding the microbes from the open water in this system, we can gain insight into how an object (a vessel) is effectively seeded from and adopts the signatures of the open water in the system. We investigate this because if a microbe from the open water colonizes an object that is moving, and persists on the object across space and time, the microbial signature is likely a good indicator of that object's geospatial history. The work presented here provides a better understanding of how microbes in aquatic systems can be leveraged as utility for object biosensors.

3.1 Introduction

The oceans and other aquatic systems have been important in globalization, enabling around eighty percent of global trade by volume/value, military mobilization, leisure and personal travel, etc. (1, 2). Maritime Domain Awareness (MDA) surveillance systems such as Automated Identification Systems (AIS) estimate that an average of 200,000 vessels transit the oceans each day(3). The requirement for AIS transponders only applies for large vessels (greater than 300 gross tonnage) or those used as passenger ships. This means that many smaller craft are able to subvert surveillance technologies described above and are responsible for a range of illicit maritime crimes and other unlawful acts, including

transport of illegal goods (arms and drugs), human trafficking, smuggling of migrants, unregulated fishing and piracy and is responsible for a large percentage of transnational crime globally(1, 3, 4).

A measurable understanding of the origin of objects passing through exclusive economic zones (EEZs), the high seas and other navigable waterways can be beneficial to law enforcement and national security – as noted by the National Strategy for Maritime Security, a plan to achieve Maritime Domain Awareness/MDA through the National Security Presidential Directive-41/Homeland Security Presidential Directive-13 (NSPD-41/HSPD-13) (Maritime Security Policy, December 21, 2004)(5).

Natural microbial communities have enormous potential as toolkits for forensic science, and possibly serve as "fingerprints" to establish object provenance (previous known origin). A wide variety of microbes are ubiquitously and abundantly dispersed through our planet – each with their own unique DNA(6). As objects pass through our natural and built environments, they will adopt the resident microbes of each geospatial location – and this gradually forms a colonization fingerprint that can be used to determine where the object has been through space-time(6, 7).

As mentioned throughout this dissertation, applications like this are made possible by a timely coupling of less expensive and accessible molecular platforms and computation: high-throughput next-generation sequencing (NGS) and machine learning (ML). This complement gives the ability to harness the power of microbes as sensors of particular phenomena that we are interested in and often in a non-invasive way. For example, machine learning coupled to microbiome data has enabled identifying individuals from their personalized microbiome(8), estimating post-mortem intervals(9), identifying clandestine grave sites(10) and tracking food sources(11). There however has been a focus primarily on human microbiomes. More recently, scientists have started to examine the potential for machine learning to use microbial communities for environmental diagnostics(12, 13). However, little work has been done to characterize microbiomes of many objects in the built environment – which can provide insights into the history of that object. Since objects possess a dynamic and unique microbiome that changes over time in response to surrounding environment (7), this approach has great potential such as in our application.

Our previous work has shown that microbes can be used to discriminate geospatial location at global scales of which we sampled(6), suggesting microbes can be used as sensors of object provenance. Other work from our group has shown that vessels from around the world are home to a diverse community of bacteria that are in some ways shaped by the waters in which they come in

contact(14, 15). By way of real-world experiment, we leverage that in a 1ml of surface seawater there can be approximately 10^5-10^6 microbes present(16). These microorganisms are able to be picked up by moving objects and carried from one destination to another. In this study we followed the microbial community (using the 16S rRNA gene-marker) of two vessels and the surrounding water through which they passed over a transect down the Chesapeake Bay from Baltimore to Norfolk and back. This work advocates an enhanced and innovative perspective of using natural microbial communities in complex environments as forensic tools and provides a framework for the integration of field sampling, data acquisition, and computational techniques necessary to approach such a problem. Lastly, the data generated as part of the experimental field design complements public datasets accessible to other researchers to validate and benchmark hypotheses relating to the study of natural microbial communities in complex settings.

3.1.1 Field design and sampling

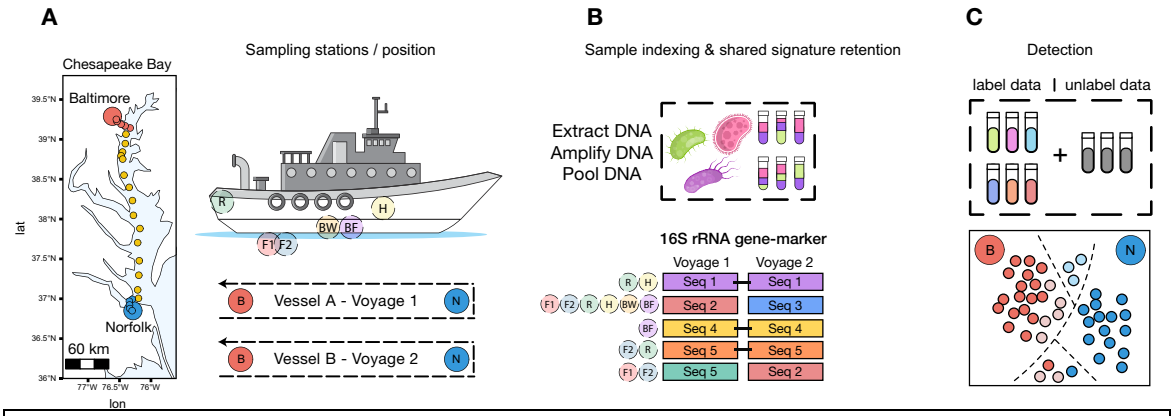


Figure 3. 1

Illustration of experimental design: field work, molecular and computational workflow. (**A**) Depicted is a map of 25 sampling stations (colored circles on map) (coordinates supplied in *Supplementary Data*) along the Chesapeake Bay spanning a transect between the port of Baltimore, MD (B: Red) and Norfolk, VA (N: Blue). For the analyses in this *Chapter*, samples along the transect were divided into three groups (1) Port of Baltimore (Red), (2) Chesapeake Bay (Yellow) and (3) Port of Norfolk (Blue). A number of samples were collected from the *vessel* and surrounding *open water* at each sampling station as follows: R, rear; F1, open water filtered through a pre-filter (see methods); F2, open water filtered through a post-filter (see methods); BW, bilge water inside of the engine compartment; BF, bilge biofilm inside of the engine compartment; H, hull (side). Sample positions R, BW, BF, H were swab samples taken from the surface of the vessel – and F1, F2 samples were collected as 1L of surrounding open water and subsequently filtered to collect biomass through a pre and a post filter that have different pore sizes. Two independent voyages were chartered to transit from Baltimore to Norfolk and

back, sampling both directions for a total of 616 samples to be used in a downstream in silico detection scheme and to test for system robustness. (**B**) Illustrates the molecular techniques used to prepare sequencing libraries of the 16S rRNA gene marker and identify shared sequences between both *voyages*. (**C**) A detection scheme using machine learning that identifies object provenance from learning the *vessel's* home port resident microbial signatures (Baltimore) and intra-transit to destination signatures (mid Chesapeake Bay – Norfolk). In addition to collecting biological samples, multiparameter data of water quality (Conductivity, ODO, pH, Salinity, TDS, Temperature, Dissolved Oxygen, Total Dissolved Solids) along with the GPS coordinates of each sampling site was recorded in situ with a YSI ProDSS digital sampling system calibrated before each sampling trip. Nutrient data (phosphate, silicate, nitrate, ammonia), total organic carbon (TOC) and chlorophyll content was recorded for each station on each transit. Additional information regarding this experimental field design can be found in materials and methods.

Previous work indicated that microbes in port water exhibit some sort of biogeography, suggesting that natural microbial communities can serve as sensors for object provenance(6). The aim of this experimental design is to extend this information to potentially use microbes as quantitative indicators of the transmission of a vessel through space and time. For this test case, we have designed a field experiment to acquire the datasets necessary to ultimately train and validate a machine learning model objectivizing previous known origin of an object, provided a sample has been taken from the object. More specifically, we aim to investigate if upon sampling a vessel that leaves from the port of B and arrives at the port of N, what is the retention and level of detection of the microbial signatures of the home port B, and can these signatures geospatially quantify, with accuracy, the previous port the vessel was stationed in, or detect if it has traveled along a specific route (**Fig 3.1**).

We chartered two research vessels to transit from the port of Baltimore (B) to Norfolk (N) and back on independent voyages for a total of four transits along the Chesapeake Bay (C). Along departing and returning transits on both vessels, samples were collected at the same stations from the surrounding *open water*, the transom of the vessel (rear (R)), hull of the vessel (side) (H) and bilge compartment water (BW) and biofilms (BF) (engine room that collects water) to sample 25 stations on departing transits (B to N) and 9 stations on the returning transits (N to B) (**Fig 3.1A**) for a total of 616 samples including processing blanks and negative controls.

We chose to sample the boat at specific locations (R, H, BW, BF) because R and H are constantly being splashed by surrounding water (representative of F1/F2) – and also because surrounding water influxes into the bilge compartment (BW).

The bilge of a vessel is generally where the engine is held and is the lowest compartment which collects water during transit. Microbes that enter the bilge compartment through open water influx are effectively shielded from outside environmental perturbation and are given the opportunity to colonize to form biofilms (BF). This process of how microbes can colonize artificial surfaces to form biofilms has been well characterized in marine settings(17, 18). Since the formation of a biofilm provides advantages to microbes by supporting the stability of the community and various ecological and biogeochemical functions – this compartment may support the longest microbial signature retention over time (**Fig. 3.1B**). More so, biofilms can be more resistant to change from influx of new organisms into the bilge compartment relative to other positions on the vessel.

Although microbes have many ways of colonizing a vessel, we hypothesize that the microbial community of a boat will be shaped by two communities. The first is defined by the microbial signature acquired at the home port from sitting at the dock, and the second would be any new microbial assemblages picked up by the vessel from the water during its transit. While microorganisms can colonize the outside of the vessel (R and H), these boat sites are exposed to the environment and thus may be more variable than BW and BF.

Additional factors behind choosing these sampling positions and sample sizes at each station stem from previous studies including our global sampling expedition for geospatial signatures(6) and associated work(14, 15). Here, we show that taking low bio-mass samples (swabs) at the selected locations on the vessel (**Fig. 3.1A**) still provides enough genetic information to obtain sufficiently representative microbial community profiles for predictive algorithms to recognize biological differences between sampling stations (**Fig. 3.1C**). In addition to taking biological samples, multiparameter data of water quality (Conductivity, ODO, pH, Salinity, TDS, Temperature, Dissolved Oxygen, Total Dissolved Solids) along with the GPS coordinates of each sampling site were recorded in situ with a YSI ProDSS digital sampling system calibrated before each sampling trip. Further, nutrient data (phosphate, silicate, nitrate, ammonia), total organic carbon (TOC) and chlorophyll content was recorded for each station on each transit so that causal ecological influence on microbial signatures can be inferenced (environmental data supplied in *Supplementary Data*).

3.2 Results

3.2.1 General characterization of system-wide microbial community

Determining if and where there are apparent distinctions in the samples collected in our system in the aggregate is important for selecting a set of microbial candidates as the most robust signatures and for machine learning model building.

Initial analysis of all samples taken from the system either on the *vessel* or from the *open water* is displayed in **Fig. 3.2**. Here, similar patterns in sample composition were observed across both *vessels* and from the *open water* sampled from independent voyages (taken at different times), which suggests relative community stability across the system during the time period of sample collection and indicates that our experimental design is relatively robust (i.e., not limited to single type of vessel or very short distances). This clustering of the PCoA (**Fig. 3.2A**) suggests that samples collected from the *vessel* (BW/BF) are clearly separable from the *open water* (F1/F2), which we know seed these compartments(14). Notably, locations such as H/R are also known to be seeded by *open water*. The samples from these surface locations (external: exposed to environment) share more similar community compositions to *open water* (F1/F2) than do the compartmentalized (internal: shielded from environment) surface locations of the *vessel* (BW/BF) have to *open water* (**Fig. 3.2B**).

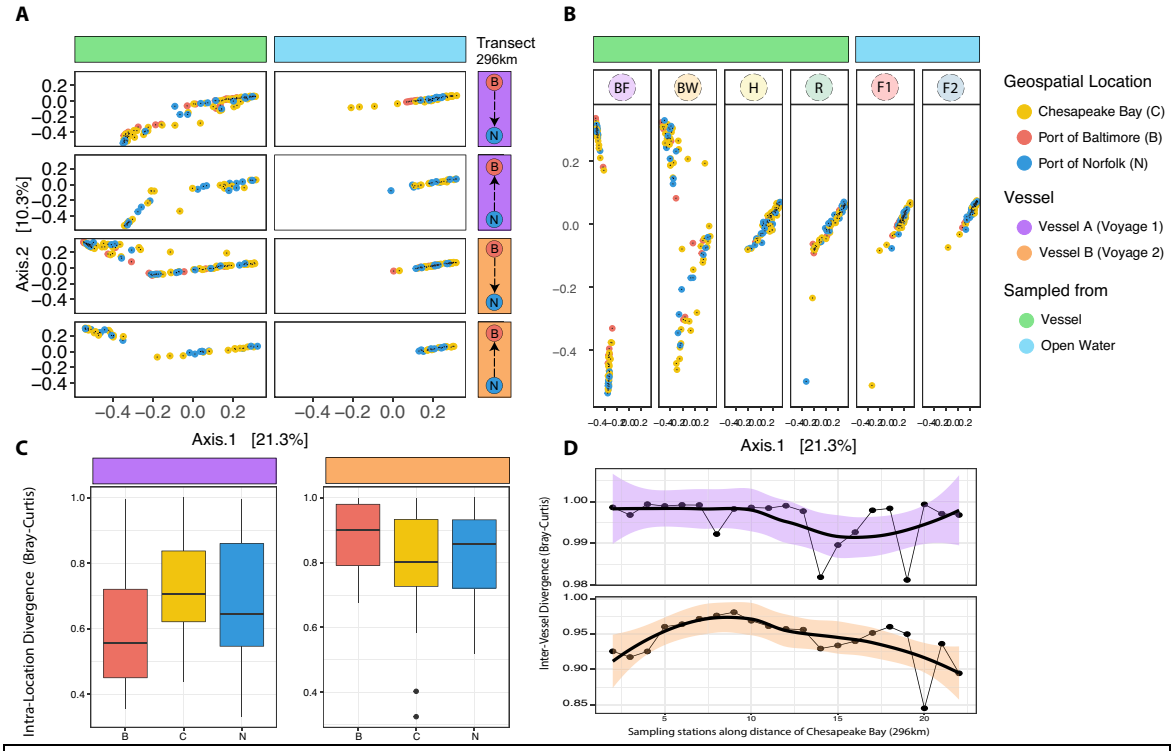


Figure 3. 2

Principal Component Analysis (PCoA) and distance metrics of variability in microbial communities. (**A**) PCoA plot comparing the entire microbial community composition of samples taken to display inter-boat variation between compartmentalized sampling positions of the *vessel* (BW/BF; shielded from environment) and external sampling positions (H/R; exposed to environment) as well as the *open water* (F1/F2). These ordinations are faceted by both *voyages* (*vessel* A and B) as illustrated in **Fig. 3.1A**. (**B**) PCoA plot displaying inter and intra-variation of each sampling position calculated using Bray-Curtis dissimilarity. (**C**)

Divergence plot showing the intra-location variation in community dissimilarity between each *voyage*. (**D**) Divergence plot showing the inter-*vessel* variation in community dissimilarity from all sample types (BF/BW/H/R/F1/F2) between each sampling station across the Chesapeake Bay (296km) for each *voyage*. A loess (locally weighted scatterplot smoothing) function with a y ~ x formula was used to better track divergence across the transect.

Intra-geospatial location (B/C/N) variability between each voyage (**Fig. 3.2C**) reveals that there are differences in microbial community composition between Baltimore, Chesapeake Bay and Norfolk, which supports an argument for resident microbes serving as signatures of detection across the transect. Looking at the inter-vessel divergence between each sampling station for each voyage (**Fig. 3.2D**), upon leaving home port (Baltimore), the open water microbial community diverges from the resident microbiome of the port of Baltimore – and continues to diverge as the vessel transits the middle of the Chesapeake Bay. Upon reaching the final destination (Norfolk), the microbial community diverges again – to what resembles more of a near-shore community profile, similar to earlier stations of Baltimore. To potentially explain observed divergence, when comparing divergence results to the PCoA plot (which clusters each sample in terms of similarity), it is observed that although each voyage is variable, the microbial community composition is still relatively similar in the aggregate.

These preliminary clustering results may suggest that compartmentalized positions of the vessel (BW/BF) that are shielded from the external environment (constant splashing, winds, environmental parameters) may influence a microbe's ability to colonize the surface of the compartment for longer retention. The *open water* is generally going to be more of a homogenous community composition than communities found in a compartment. Thus, the community composition observed between external sampling positions and *open water* support the hypothesis that as an object moves through the *open water*, the microbes present at each geospatial location along the transect (captured in F1/F2) will reflect the community composition on the external and compartmentalized surfaces of the *vessel* (H/R/BF/BW) (**Fig. 3.2AB**). However, investigating specific retention times of microbes in the water and the microbes at each sampling location on the vessel requires a more thorough analysis.

3.2.2 Generalizing candidate biomarkers

In order to develop a system for detection of past location, we wanted to focus on a subset of the community that would be the most informative for model construction and as detectable signatures. Moving forward with all characterizations of this system, we use a derived set of 9 'persistent' microbial

features (single nucleotide variants (ASVs, as taxa))(19) (**Fig. 3.3**). These features were chosen based on the fact that they are both detectable (≥0.001 relative abundance (RA)) and frequently observed (≥.5 prevalence/frequency) in the aggregate of samples along the transect from Baltimore (B) to Norfolk (N) across the Chesapeake Bay (C). The microbial ASVs are identified as a 'core' biome that are the most robust and persistent genetic signatures in the system as well as strong indicators of geospatial location. We chose to focus on abundant and prevalent ASVs because 'rare' or underrepresented microbial taxa may be harder to consistently detect due to variation and other biases introduced during sample processing. For example, we are considering a good microbial signature to one that is present in the system and retained through the majority of samples throughout the 296km voyage using *vessel A* on *day 1-5* and was also present and retained on the voyage using *vessel B* on *day 6-10*, as illustrated in **Fig. 3.3**.

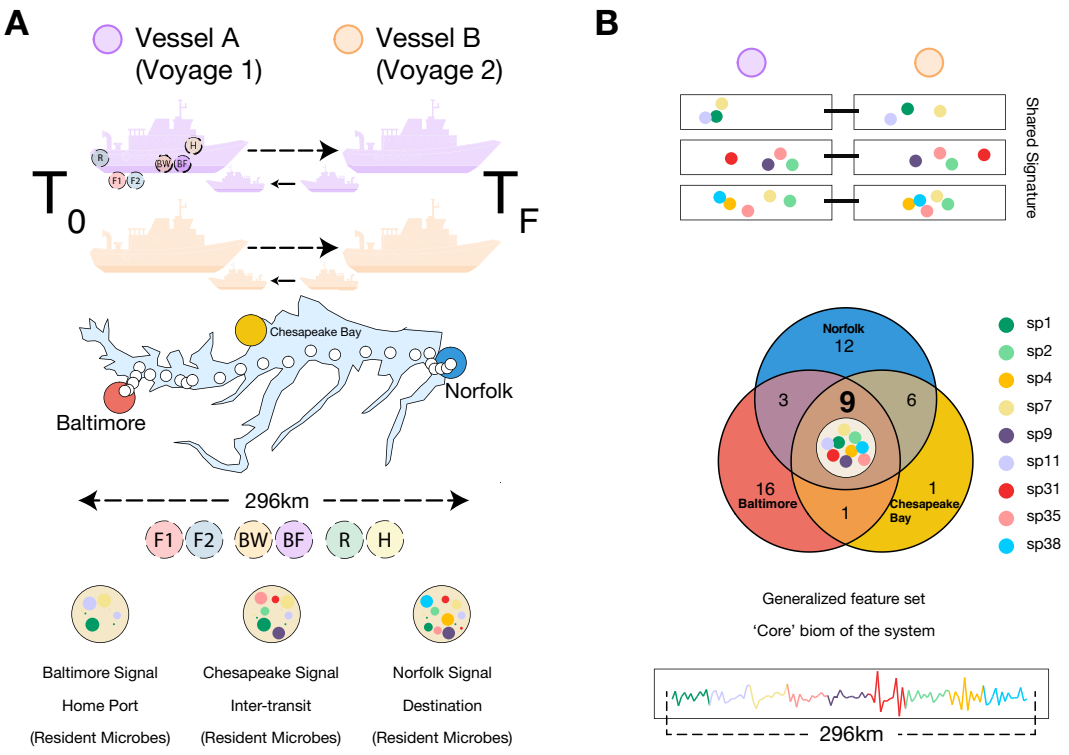


Figure 3. 3

Schematic illustration to identify a core biome of the system.

This heuristic threshold of detection/prevalence is chosen for computational and practical feasibility and for parsimonious model selection given that this model criterion can generalize to downstream molecular detection applications. Other thresholds may be useful to address other questions from this dataset. Examples

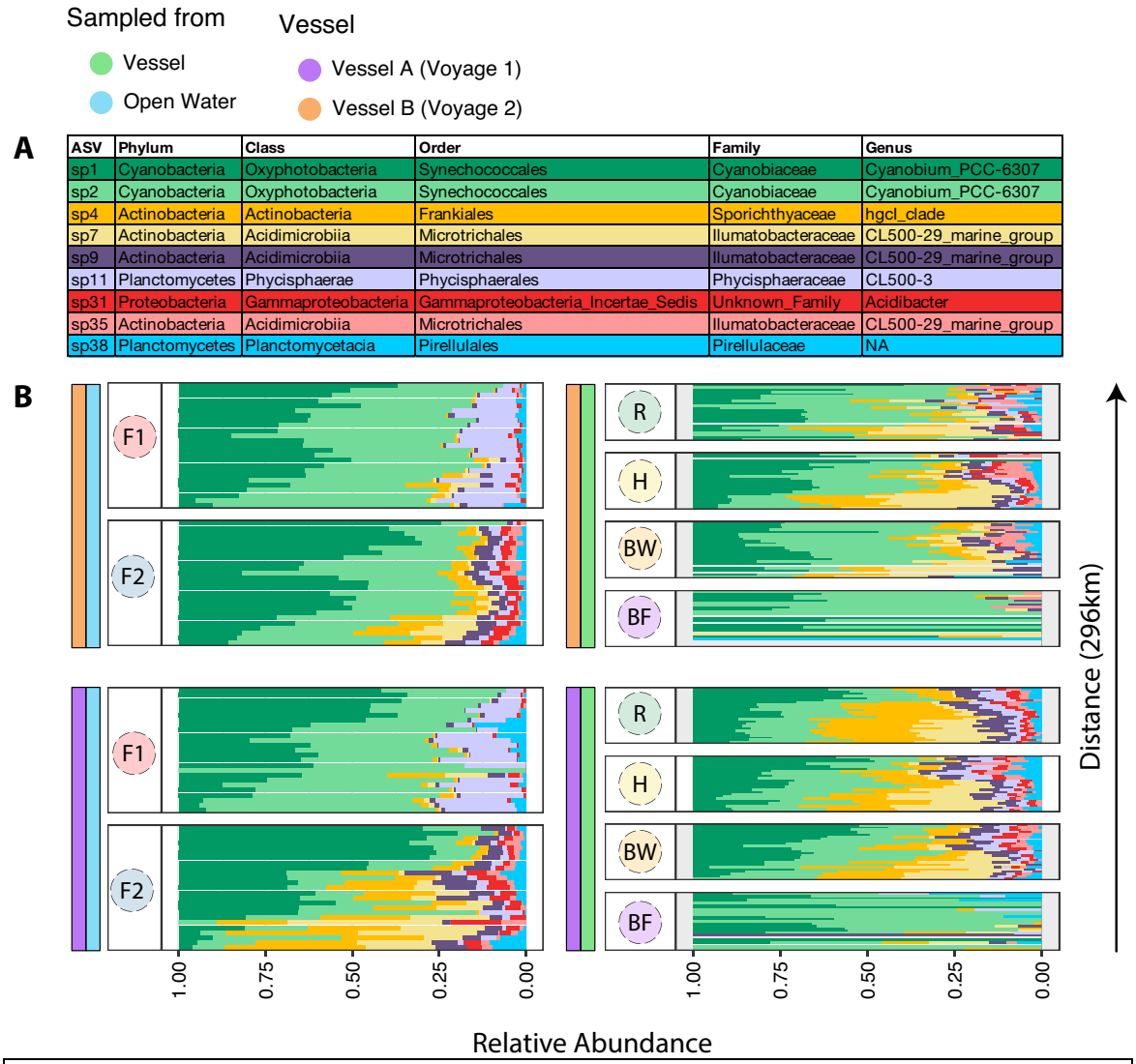
of translational applications from in silico modeling – such as the translational potential of this work is described extensively in the next chapter.

Certain ASVs may be 'persistent' resident members of the microbial community of *open water* (F1/F2) surrounding Baltimore (detectable at sampling stations near Baltimore) and are retained as the voyage continues to Norfolk (detectable at sampling stations near Norfolk). Using taxa that may be easily washed away (only retained from 0-100km, *hypothetical*) or are only present in certain sampling stations (**Figure 3AB**) does not allow for robust and generalizable biomarkers to be identified through the system. The reasons for loss of signature may result from many things in this dynamic system and is not limited to but can include the inability to rapidly become abundant by colonizing or adapting to environmental conditions (e.g., constant mixing, low nutrient availability, salinity, inability to adhere to vessel material, …).

Finally, the approach of using the core microbiome for detection is well suited for downstream modeling, such as a supervised machine learning algorithm that can use a feature space fully comprised of the microbial community, and not external metadata (environment, time, …). By selecting a core microbiome that is present on multiple vessels, we effectively reduce the inter-*vessel* and inter-*open water* sample variation such as in noisy and spurious taxa resulting from confounding variables that are not collected during sampling (e.g., nutrients on vessel: enhanced biofilm formation), while capturing the subtle intra-variation of the microbial communities across the system as 'signals' of detection (**Figure 3B**). In systems like the human microbiome, where there is high inter-personal variation in microbial community composition, this approach using the core microbiome to classify outcomes may allow for the construction of more generalizable models.

3.2.3 Taxonomic profiling

The generalized approach of 'core' microbiota described in ***Section 3.2.2*** and displayed in **Fig. 3.1** and **Fig. 3.3** is first demonstrated here by taxonomically profiling the community of our system (**Fig. 3.4**). Relative abundance profiles of the 9 ASVs (system's 'core' biome) reveal fine scale differences in microbial community composition between signatures sampled from each voyage. Of the 9 core ASVs, *sp1* and *sp2 (Cyanobacteria)* are observed system-wide at high proportions. This is consistent with what is expected to be observed – since our samples were derived from surface water, we expected that photosynthetic microbes that occupy the surface water during the day would be abundant in our surface water samples.



ASV	Phylum	Class	Order	Family	Genus
sp1	Cyanobacteria	Oxyphotobacteria	Synechococcales	Cyanobiaceae	Cyanobium_PCC-6307
sp2	Cyanobacteria	Oxyphotobacteria	Synechococcales	Cyanobiaceae	Cyanobium_PCC-6307
sp4	Actinobacteria	Actinobacteria	Frankiales	Sporichthyaceae	hgcI_clade
sp7	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29_marine_group
sp9	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29_marine_group
sp11	Planctomycetes	Phycisphaerae	Phycisphaerales	Phycisphaeraceae	CL500-3
sp31	Proteobacteria	Gammaproteobacteria	Gammaproteobacteria_Incertae_Sedis	Unknown_Family	Acidibacter
sp35	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	CL500-29_marine_group
sp38	Planctomycetes	Planctomycetacia	Pirellulales	Pirellulaceae	NA

Figure 3. 4

Taxonomic assignment and profiling of system-wide sampling regime. (A) Taxonomic assignment of each 'core' microbiota from highest resolution (ASV) to lowest resolution (Phylum). (**B**) Microbial community composition of samples taken from the system. The X axis is representative of relative abundance of each signature faceted by samples taken from the *vessel* (positions: BF/BW/H/R) and from the *open water* (positions: F1/F2) during both voyages across the distance of the Chesapeake Bay.

Although consistent in both voyages, samples from BF have a less diverse microbial community composition compared to all other samples, which is to be expected since the formation of a biofilm is usually dominated by few organisms(20). A closer look at *Cyanobacteria* specific ASVs (*sp1, sp2*) is outlined in *Supplemental File 1*, where we compared their relative abundance to the total community rather than just to the other signature taxa as presented in **Fig. 3.4**.

However, it is notable to mention that assuming *Cyanobacteria* in this system are phototrophic, these taxa could also survive as heterotrophs on the available nutrients after having been carried from *open water* (F1/F2) to the *bilge* compartment (BW) to ultimately form a biofilm (BF) in the absence of sunlight(21).

However, we would still expect there to be a lower representation of viable *Cyanobacteria* in a bilge tank. Similar work has explored this and found increased enrichment of *Cyanobacteria* in open water (F1/F2) relative to bilge (BW/BF)(14) (log_2-fold change of 3.1-5.5), and is consistent with what we show with *Cyanobacteria* signatures in this study (*sp1*, *sp2*). It is worth mentioning that *Cyanobacteria* analyzed as part of the mentioned in Schaerer et al may not map to similar taxa as in this study, since *sp1* and *sp2* here are derived from a different amplicon reads (ASVs). However, the fact that cyanobacteria are found in both surface water and in the bilge compartment is consistent across studies from diverse settings and varied vessels.

Analyzing other taxa, a large proportion of *sp11* is consistently present in F1 samples (filter: 1.6-*u*m pore size) relative to F2 samples (filter: 0.2-*u*m pore size), across both voyages. This enrichment in the larger pore size filters could indicate that *sp11* (Planctomycetes) could be particle attached. This is consistent with other studies that have shown that many members of the Planctomycetes often live associated with particles and are recovered on larger pore sized filters(22). Their presence at lower abundance on the smaller filter (F2) could potentially be explained by when water is filtered through the larger pores of F1 into F2 samples, increased abundance of the same, or a variety of different microbes can be liberated from particulate matter of the surface samples F1 (e.g., organic matter). This could also indicate that microbes could be then better captured by molecular detection techniques(23), such as *sp4*, *sp9* and *sp31* signatures as observed in F2. Together, these results suggest that a higher resolution of the distribution of signature taxa is reflected through analyzing biomass collected on F2 samples, and other compartments that resemble similar patterns, such as on the vessel itself (R, H, BW) (**Fig. 3.4B**).

In this system, we hypothesized that all sample types are primarily seeded by the surrounding *open water*, either through splashing or inflow, since work in similar systems showed that roughly 40% of the bilge communities (BW/BF) and 52% of the surface communities (R/H) are sourced from surrounding *open water*, or what is effectively captured in F1 and F2(14). Thus, the observed consistencies between each *voyage* type support that our experimental design and sampling regime is robust to distance (at least 296km), vessel type and to the particular seasonality of this system of when we sampled (environmental data along with season can be accessed in *Supplementary Data*).

Overall, profiles between samples from each position are consistent between both *voyages*. The microbial community diversity in BF is the least variable across both voyages, which could also support this sample type as the most stable for signature retention. Early-*voyage* sampling stations seem to have the most community variability as compared with mid-*voyage* and late-*voyage* samples, which could indicate that increased dynamics in community composition in coastal water could influence community turnover (e.g., taxa gain/loss) compared to the conditions that mid-bay communities are exposed to(14). Some work has been done to assess community turnover as a function of seasonality and severe weather events in aquatic systems(24, 25). After a perturbation event, microbial communities seem to stabilize back to a 'resident' community profile, which may support the work shown here that resident signatures are robust to season and time or will shift back to a steady state community.

It is notable to mention that these profiles in **Fig. 3.4B** are a result of DNA sequencing and observed taxa profiles could have come from either having sampled and sequenced free nucleic acid from 'core' microbiota that were part of influx through the voyage, rather than viable cells as part of the sampled biofilm (BF). More work would have to be done looking at the RNA level or using alternative methods for inferring viability of these signature taxa.

3.2.4 Enrichment profiles of signature taxa

Signature taxa that would be useful for determination of provenance would require the ability of a resident microbe to be picked up by an object and carried through a system as a persistent signature of geospatial location. This section more quantitatively analyzes these patterns of signature taxa using differential abundance profiling (**Fig. 3.5**). Different measures of association between signature taxa can be assessed as either being enriched in sample types from *bilge* (BW/BF) or *open water* (F1/F2). Focusing on the same ASVs as characterized in **Fig. 3.4** in order, *Cyanobacteria* (*sp1*, *sp2*) show different enrichment profiles (Abundance (log10-scale)) despite both being assigned to the same Genus. Additionally, these taxa show different patterns of enrichment with *sp1* being significantly enriched in the *open water* relative to *bilge* (logFC: 2.09). For *Actinobacteria* (*sp4, sp7, sp9, sp35*), excluding *sp4*, these taxa share the same taxonomic assignment through Genus and appear to have inconsistent enrichment profiles, with *sp7* slightly enriched in *bilge* (logFC: 1.18) and *sp9* and *sp35* enriched in *open water* (logFC: 3.23, 2.31 respectively). Although not significant, the profiles seem to be impacted by difference in prevalence. The *Planctomycetes* (*sp11*, *sp38*) deviate at Class and are both significantly enriched in *open water* relative to *bilge* (logFC: 5.47, 4.14, respectively). *Planctomycetes* appear to be significant by all metrics provided: (-log10(adj. p value)), generalized fold change and with respect to prevalence.

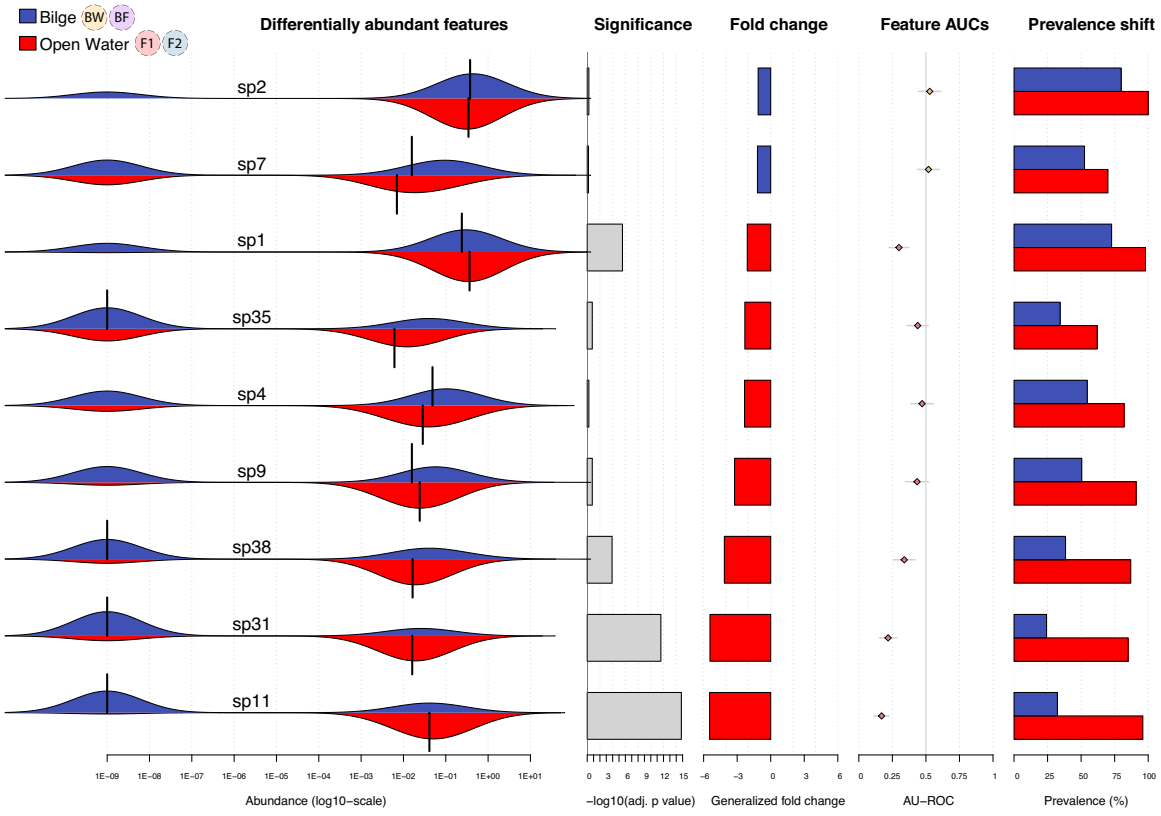


Figure 3. 5

Differential abundance analysis for signature taxa. This plot provides various measures of association between signature taxa and sample positions from *bilge* and *open water* samples from both voyages (*n* = 199; 100 *open water*, 99 *bilge*). All *axis* values are provided underneath each analysis. Differentially abundant features are represented as bean plots to show abundances (log10) of each signature in the sample distribution used from *bilge* (BW/BF) or *open water* (F1/F2) samples. Vertical lines in each bean plot represent the mean value for each taxon in the sample distribution. Significance is computed by a Wilcoxon test followed by multiple hypothesis testing correction and is meant to denote how 'important' enrichment is for each signature taxa in *bilge* relative to *open water*. Fold change is the generalized fold change in enrichment, or the geometric mean of differences between each signature taxa in *bilge* relative to *open water*. AU-ROC (Area Under the Receiver Operating Characteristic Curve) is computed as a non-parametric measure of enrichment (corresponding to the effect size of the Wilcoxon test). AU-ROC here measures the observed difference in read counts in the sample distribution for each signature (sample separability) where an AU-ROC from 0.5-1 is in favor of taxa being enriched in *bilge* and an AU-ROC from 0-0.5 is in favor of *open water*. Prevalence shift indicates the difference in prevalence (proportion of samples the taxa is in) of each signature taxa between *bilge* and *open water*.

The only signature taxa belonging to *Proteobacteria* (*sp31*) seems to be significantly enriched as well as substantially more prevalent in *open water* relative to *bilge*. This is consistent to the taxonomic profiles of *sp31* shown in **Fig. 3.4B**, where there is notably more abundance and prevalence for this taxon in samples from the *open water* relative to *bilge*.

In the aggregate, there seems to be a key role in prevalence leading observed enrichment, which is consistent to an intuitive perspective of this system, since a low prevalence of a certain taxa may impact its ability to be detected as a signature. A higher overall prevalence should indeed be detected in the seeding source of *open water* (F1/F2) – rather than the compartments it sinks into (BW/BF/R/H) (**Fig. 3.4B**). However, in ideal conditions (proper nutrients, …), it can be expected that certain taxa can be enriched in a *biofilm* (BF) relative to the source of the biofilm over an adequate amount of time/distance**.**

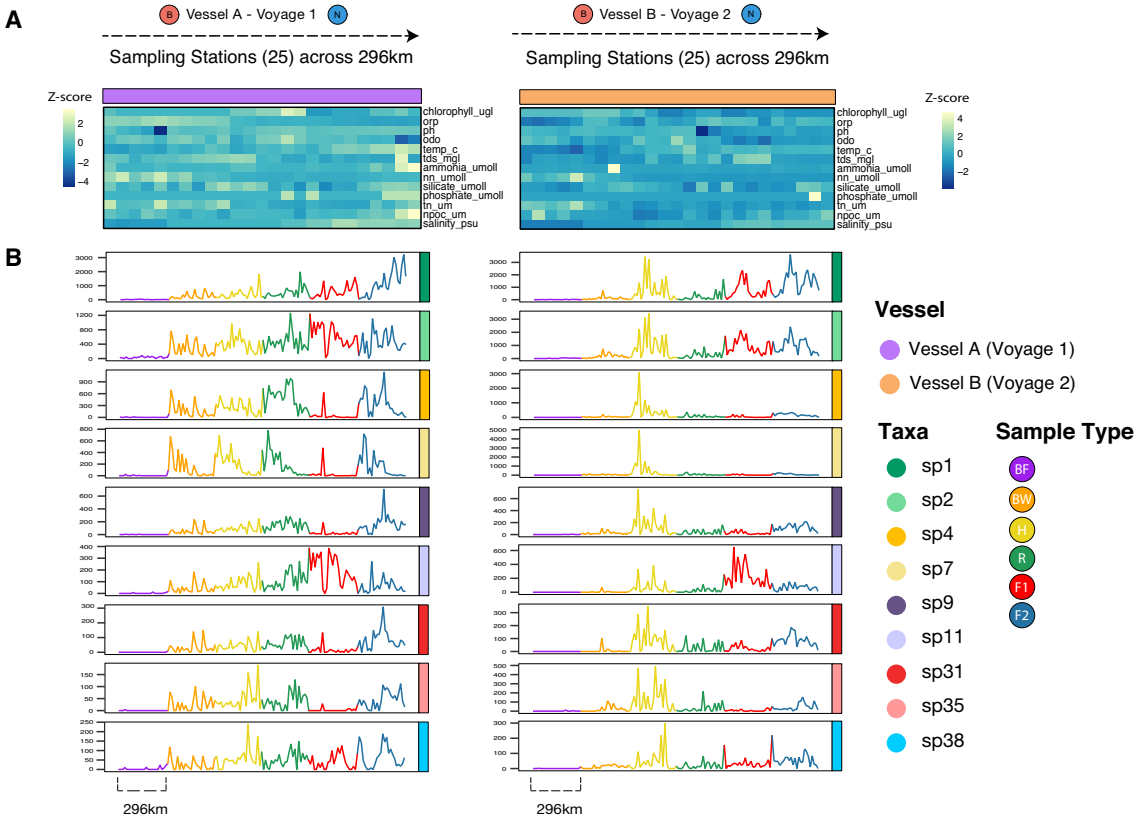
Although each signature taxa are more prevalent in *open water* as compared with *bilge*. It is notable that each taxon is at least present in bilge, which support previous results that influx from *open water* is effectively seeding this compartment. It is also worthwhile to mention that signatures enriched in *bilge* are not indicated to be significant, and that the largest generalized fold change in enrichment also have the largest shift in prevalence (**Fig. 3.5**).

The data presented in this section supports similar work(14) that show ASVs that map to similar taxa as our signatures are as well enriched in *open water* relative to *bilge* – and was previously described in the above *Section 3.2.3*. Together, these results suggest that system dynamics for signature retention are consistent in other similar systems, and that very few taxa are consistently found in *bilge* from different ports located in different regions around the world(14). This indicates that there are no single nucleotide variants of microbes that are globally found in the *bilge* compartment of *vessels* – and the resident signature in *bilge* is rather representative of the microbes that enter this compartment by being docked at home port or through influx along transit, further suggesting *bilge* could maintain the most persistent signature. As to what could explain the variation observed so far, more targeted approaches will be used to explore research questions in the rest of this chapter and moving forward.

3.2.5 Temporal and environmental dynamics of microbial persistence

In order for microbial communities to serve as biomarkers of provenance in this system, the microbial community of the *open water* and the *vessel* must have some stable state and the signature taxa must remain on the *vessel* for a period of time. To explore this, we move beyond studying the spatial variation of microbial

communities in this system and look toward the temporal component of our longitudinal sampling design – addressing questions related to the persistence of signature taxa on the vessel.



C

Environmental Parameter	Mantel r	Significance
chlorophyll_ugl	-0.02043	0.6942
orp	0.2996	0.0001
ph	0.07144	0.0468
odo	0.006432	0.4112
temp_c	0.2031	0.0001
tds_mgl	0.06741	0.0068
ammonia_umoll	0.03066	0.1178
nn_umoll	0.1246	0.0001
silicate_umoll	-0.02132	0.7633
phosphate_umoll	0.1002	0.0002
tn_um	0.06353	0.013
npoc_um	0.02423	0.269
salinity_psu	0.2654	0.0001

Environmental Parameter	Mantel r	Significance
chlorophyll_ugl	0.05996	0.0106
orp	0.09232	0.002
ph	0.004623	0.4312
odo	-0.01506	0.6322
temp_c	0.2011	0.0001
tds_mgl	0.000799	0.4797
ammonia_umoll	-0.01841	0.6237
nn_umoll	0.1809	0.0001
silicate_umoll	-0.01937	0.7586
phosphate_umoll	-0.0005372	0.4903
tn_um	0.2275	0.0001
npoc_um	-0.00245	0.5043
salinity_psu	0.2102	0.0001

Figure 3. 6

Environmental gradient and signal tracking profiles across the system. (**A**) Heatmap showing the environmental parameters across the Chesapeake Bay (296km^2). Each tile along the *x-axis* in the heatmap (25 total) represent sampling stations from Baltimore (B) to Norfolk (N), as described in **Fig 3.1** and **Fig. 3.3**. Environment is represented as Z-score values to show how each environmental parameter deviates from the mean. (**B**) Taxonomic 'signal' profiles (absolute abundance) of signature taxa faceted by sample type across each voyage along the

Chesapeake Bay. For ease of visualization, all sample types for each taxon were plotted across a continuous *x-axis*. A different color spanning each 296km^2 segment represents 'signal' across the full voyage (25 sampling stations), in order. Both plots in **A** and **B** represent a continuous voyage along the *x*-axis, covering all sampling stations. Raw values of each environmental parameter as well as time in between each sample station can be accessed in *Supplementary Data*. (**C**) Correlation between abiotic and biotic features of the system. Mantel statistic (**r**) measures the correlation between each parameter and absolute abundance (using dissimilarity matrices (Jaccard). Mantel statistic (**r**) can range from -1 (strong negative correlation) to 1 (strong positive correlation). Significance is calculated by permuting *N* rows and *N* columns of one of the dissimilarity matrices, because there are $N(N-1)/2$ entries for just *N* observations, thus cannot be accessed directly from the correlation. Raw values of each environmental parameter as well as time in between each sample station can be accessed in *Supplementary Data*.

This section examines the limits of signature detection by investigating how taxa can be traced through the environment along voyages, both on the object (*vessel*) and from the seed (*open water*). Here, we visually represent the environmental gradient probed at each sampling station and the temporal variability in microbial abundance across each voyage (**Fig. 3.6**). The environment and abundance visuals are from independent analyses and can be assessed in parallel to inference potential ecological interactions on observed community stability. Together, this will demonstrate whether patterns of microbial diversity within the 'core' microbiome are robust to distance, time, and whether taxonomic patterns reflect environment across the Chesapeake Bay, as demonstrated using two independent research vessels.

To explore these relationships, each taxon is tracked as a 'signal' across each voyage and for each sample type in the system (**Fig. 3.6B**). Here, 'detection' targets temporal variability along the voyage and will be discussed in terms of 'signal' peak, which represents absolute abundance (read counts in the sample distribution). It is worth mentioning that although a 'signal' may show a low read/abundance throughout the voyage for some sample types, such as in the bilge biofilm (BF), these taxa were still detectable by our methodology: sampling, filtering, molecular processing and sequencing (see methods). Subsequently, these analyses indicate that a sufficient number of taxa are present for detection by potential downstream molecular platforms. These include other sensitive and specific sequence-independent molecular techniques such as LAMP (Loop-mediated isothermal amplification)(26) and SHERLOCK (Specific high-sensitivity enzymatic reporter unlocking)(27). Although the qualitative aspect of this figure can be deceiving, it is important to consider the relation of absolute counts (*y-axis*) while comparing taxa at each sample position, as a seemingly similar 'signal' may

be orders of magnitude higher in some sample types for certain taxa. These taxa are detectable, albeit in low abundance in all sample types.

Since each voyage covered a distance of 296km from Baltimore (B) to Norfolk (B) – one can attempt to relate the environmental parameters to the observed microbial persistence in this system. This approach would further an understanding of gain/loss 'detection' of single nucleotide variants across distance, and potentially time. Some work has been done to characterize microbial persistence on objects in built environments(28, 29). Others have looked at the colonization of inanimate objects in oceans at lower depths(30). Microbial persistence however is highly dependent upon the system, and generalizations are constrained to microbial behavioral patterns in response to extraneous determinants and assumed perturbations of a particular system (e.g., confounders; resident microbial gene expression, mixing patterns, environmental parameters, etc.). In effect, inferences made here to the causal effect of environment on microbial persistence is strictly hypothetical – more so since the gene-marker resolution being used is 16S rRNA, and functional annotations of these signature taxa cannot be made. For example, metabolic capabilities such as substrate utilization at each sampling station for these taxa cannot be inferred from 16S rRNA sequencing alone. Although, long-term studies conducted in marine settings show that microbiota response to seasonal dynamics is highly predictable(25, 31), and although primary succession after a perturbation event is more variable, the microbial community has shown to stabilize. This response has been shown to be partially driven by predictable seasonal succession(24, 32). Community assembly is also known to be driven in part by environmental selection, biotic interactions, as well as dispersal(33). With that, the role of environment on ecological interactions can still provide insights for new research questions. This is explored below, first, with a high-level overview of trends observed from 'signal' in terms of signature retention, followed by relating these trends to environmental parameters.

3.2.6 Characterizing signal of detection

As a high-level overview, most signatures show gradual trends of decrease in abundance over time, such as *sp7*, while others show a gradual increase (*sp1, sp2, sp11*). Notably, there is a gradual increase in abundance for F2 samples for most taxa starting around the same sampling station in the 296km transect. Also observed is a trend toward increased abundance overall for many taxa for *voyage 1* relative to *voyage 2*, which could indicate that vessel material could influence colonization patterns (see methods). Consistent in both *voyages*, there seems to be a trend of increase and then decrease in abundance across the transect in sample type H. These results suggest that the method of microbial seeding may play a role in 'signal' variability. Additionally, since H (side of vessel) receives its microbial

signature input from the *open water* splash during transit, the rate at which the vessel is traveling could influence the strength of this signal.

Taxa that are most consistent between both voyages correspond to *Cyanobacteria* (*sp1*, *sp2*) and *Proteobacteria* (*sp31*) – which is the only phyla of its kind in the 'core' biome. Interestingly, *Cyanobacteria* appear to be the most abundant in the system (**Fig. 3.4**), contrasted by *Proteobacteria* – the only phyla with zero prevalence in any *bilge* biofilm (BF) samples on either voyage but shows a relatively consistent 'signal' in other sample types across the transect of both *voyages*. Between both voyages, *Cyanobacteria* (*sp1*, *sp2*) located in the *bilge* water (BF) and *sp9* (one of four signature variants belonging to *Actinobacteria*) in the *bilge* water (BW) are a good example of this (**Fig. 3.4B**). Interestingly, some taxa (e.g., *sp7*; *voyage A*) exhibit consistent trends in having increased 'signal' early on in the transit (from Baltimore) and then drop off later in the transit (approaching Norfolk) – suggesting that these taxa may be strongly indicative of home port (Baltimore).

In order for a microbe to liberate a 'signal', it should persist (i.e., is constantly prevalent at some threshold of detection) and have low variance in abundance across the transect (296km), which denotes how robust the 'signal' is. For example, a detectable 'signal' matters less about high abundance and more about prevalence across distance/time. In analyzing the longitudinal taxon distribution of each sample type through the system, inter-voyage variability could be a result of vessel type, as each vessel may vary by its ability to be seeded by the *open water* (e.g., splashing, rate of influx into bilge, etc.). Moreover, *bilge* biofilm (BF) is the least variable of the sampling positions in the aggregate, which suggests long-term geospatial history. This may be partly because this compartment can sustain a signature of resident microbes because it is less exposed to the outside environment, but still allows influx of microorganisms in from the *open water*. Each vessel was observed to carry a distinct 'signal' profile for each signature taxa (**Fig. 3.6B**) and may be reflective of the environmental parameters along the voyage (which needs to be further investigated) as well as specific selective forces unique to that vessel (cleaning history, vessel material, inboard vs. outboard motors, etc.).

One thing that is apparent are that the 'core' taxa are persistent in the system in the aggregate and at some point, are picked up by the vessel and carried to another destination – further supporting that microorganisms have the ability to be used as detectable sensors of objects passing through an environment.

3.2.7 Impact of environmental parameters on signal variability

In analyzing the longitudinal taxon distribution of each sample type through the system, there could be many extraneous determinants of the observed inter-voyage variability. To determine the causal influence of abiotic factors on

microbial abundance and therefore on 'signal' of detection, we collected a range of environmental parameters across the transect of Baltimore to Norfolk at each sampling station alongside sample collection (**Fig 3.6A**). Correlations between absolute abundance and environmental parameter across each transect are displayed by a Mantel statistic (**r**) and corresponding Significance value (**Fig. 3.6C**). The environmental factors that most strongly influence the microbial community compositional and stability on the *vessel* itself and in the *open water* are temperature (**r**: *voyage A; 0.2031, voyage B; 0.2011*), nn (nitrate) (**r**: *voyage A; 0.1246, voyage B; 0.1809*), tn (total nitrogen) for *voyage B* (**r**: *0.2275*), salinity (**r**: *voyage A; 0.2654, voyage B; 0.2102*).

This demonstrates that although we observe similar Mantel **r** for environmental parameters in each voyage, there is variability between abundance ('signal') – suggesting that the environment alone cannot fully explain the microbial persistence and stability observed in this system. For example, viewing the taxonomy profile (**Fig 3.4B**), *sp1* has a different abundance between *voyage A* and *voyage B*, and the correlation between significant parameters are relatively similar between these voyages. This is consistent to *Chapter 2*, **Fig 2.9**, **Fig 2.10**, **Fig 2.11**, where it was observed that similar environmental drivers are somewhat important but could only explain 22.2% of the observed microbial diversity and community stability variability in on a global scale.

In this study and in this particular system, if an environmental parameter had a large impact on the 'signal' of detection, we would expect there to be a higher mantel correlation. Although, it is noteworthy to mention that if we extended our sampling across season, there may have been notable differences between these abiotic correlations to microbial abundance.

3.3 Summary and outlook

This chapter demonstrates the microbial community structure and dynamics for the vessel-associated microbial community of two vessels along two independent voyages on an extended transit. Surprisingly, preliminary comparisons show similar microbial community composition of the *vessel* and *open water* despite sampling on independent *voyages* (e.g., different dates) (**Fig. 3.2**). This led us to probe for taxa responsible for observed similarities and uncover to what extent these taxa are robust to system dynamics and temporality.

By viewing the community from the perspective of a shared 'core' set of taxa, we could still thoroughly investigate research questions and better understand the system in the aggregate (**Fig. 3.3**). Additionally, we demonstrated that our derived signatures are robust to the complexity of this real-world system. For example, these signatures persist at least across a distance of 296km, and are present across

the two vessels tested despite the overall high variability in the microbial community between the two vessels.

This system-wide signature profiling supports our sampling regime and the use of a generalized feature set, since, after resolving the microbiota to a 'core' biome, there are still enough differences in community composition that given the proper quantitative analysis – large enough disparity in community composition of sample types can be quantified over the distance across the Chesapeake Bay (**Fig. 3.4**, **Fig. 3.6**).

To the extent of seeding, we demonstrated that the majority of signature taxa are enriched in *open water* relative to *bilge*. Hence, we conclude that microbes present in the bilge compartment arrive from influx seeded by *open water*. (**Fig. 3.5**). It is however notable to mention that other compartments such as hull (H) and rear (R) are also seeded by the *open water* microbial community. The results presented here also demonstrate which sampling position on the vessel has the most stable microbial community over time, supporting that *bilge* derived samples are most robust to complex dynamics of the system (**Fig. 3.6**).

Consistent to other work as part of an earlier project(6, 14), the variability between microbial communities in the *bilge* compartment from around the world suggests that there are no single nucleotide variants (ASVs) that are globally found in the bilge tank of all boats sampled but is rather representative of the microbes that enter this compartment as it is docked at home port and along its transit. This is further supported by analyses presented here (**Fig. 3.5**, **Fig. 3.6**).

Also, we demonstrate here that vessels can carry microbes from one place to another and that these microbial passengers can serve as fingerprints of geospatial locations of where the vessel has passed (**Fig. 3.6B**). From here, we concluded that both the *bilge* (BW/BF) and the boat surface (H/R) can be used for determining provenance, but that likely the *bilge* may be ideal for long-term microbial signature retention. This is partly because it is less exposed to the outside environment, but still has influx of microorganisms from the open water. Moreover, *bilge* biofilm (BF) is the least variable of the sampling positions and may be indicative of long-term geospatial history, as this compartment can sustain a signature of resident microbes. It is however worth mentioning that not all signature taxa could persist in the *bilge* tank, and that there was observed variability between vessel type (**Fig. 3.6B**). In particular there appears to be vessel-specific differences most markedly observed in the *bilge* biofilms (BF). The *bilge* compartment may lead to less turnover relative to microbes colonizing external positions of the vessel (H/R), although, less turnover also indicates most stable sample type. There is also a possibility that distinct nutrients supplied to microbes in the *bilge* compartment, such as hydrocarbons from products used in the engine compartment and lack of

sunlight, could contribute to turnover and selection of a particular community not related to geospatial location. Similarly, the rate at which influx of *open water* is seeded into the *bilge* compartment could influence signature retention and could potentially be a reason for observed variability between *voyages* and vessels. With that, each vessel was observed to carry a distinct 'signal' profile for each signature taxa (**Fig. 3.6**) and may be reflective of the environmental parameters along the voyage as well as specific selective forces unique to that vessel (cleaning history, vessel material, inboard vs. outboard motors, etc.).

3.3.1 A note on future work:

3.3.1.1 Confounders

Extraneous determinants of taxonomic signal variation could be partially explained by abiotic factors, such as environmental parameters (**Fig. 3.6**), or the likeness of a particular taxon to adhere to each vessel used in this system. However, these, and other system confounders still need to be explored, including the potential for other sources of microbes such as the seeding of noisy and spurious taxa from air, ran and/or snow biomes to impact resident signature taxa thus influencing 'signal' of detection. This concept has begun to be explored in similar settings(15). Additional confounding variables such as boat type (influence on engine compartment, rate of influx through travel velocity, ability for microbes to adhere to specific vessel material, hydrocarbon release), boat management (cleaning and maintenance) and whether or not a vessel has a home port could all impact the ability for microbes to serve as biosensors as explored in this study.

Similar work would have to be done to address some specific details for signature retention, such as assessing the delayed response of one taxon to another, or taxon to environmental parameter. Ecological interactions such as cooperation and competition in response to environment are one example that and may limit the ability of microbes to colonize a surface and establish themselves in boat microbiome – and have shown to be factors in characterizing the persistence of taxa on an object(34). A more expansive study that uses more vessels and transects longer distances would help to clarify if the approach of using a 'core' microbiome might control for many of these potential confounding factors.

3.3.1.2 Detection in alternative stable states

Much of the work presented here is seeking to understand microbial community stability in response to changing environmental conditions over distances and time. It is important when considering community stability to identify the taxonomic level at which community stability is measured. Detection of alternative stable states is something of interest. Contrasting the community composition at various levels of taxonomic resolution could result in varying

estimates of system stability. For example, observing the system by agglomerating ASVs into lower resolution taxonomic levels (Phylum – Genus) could allow more conservative estimates of community stability – but could tend toward less specificity in real-world translational applications.

The role of community stability could be better assessed through comparing high-frequency sampling for equidistant but shorter time points combined with low-frequency and longer time points, as well as with analysis at varying levels of taxonomic resolution. It is therefore appropriate to consider sample frequency in the specific system being studied – as well as with consideration to any downstream application that is being considered (e.g., molecular detection platforms). Hence, experimental design targeting real-world application should foremost consider what would be required to integrate findings into state-of-the-art molecular technologies in translation. Additionally, more work for steady-state and resident microbial communities and how they respond to a rare perturbation event would be useful in assessing how robust this approach is.

3.3.1.3 Inferring function

As previously mentioned, and despite presenting environmental parameters as a function of persistence, functional annotations of microbes cannot extensively be considered in this work. This limits a comprehensive assessment of microbial persistence in this system as it would help address a lot of questions of how environment could sustain specific signatures in the event of a predictable altered state (seasonal succession patterns, …). Whole genomic sequencing, at least on the 'core' taxa identified as persistent and detectable signature taxa would have been beneficial, and since all samples used in this study were stored as contingency, it could be pursued in later work. This work would also help to clarify the persistence of particular taxa and the appropriate level of resolution for considering microbial stability. Many microbial with nearly identical 16S rRNA genes have been shown to vary greatly in their genomic and functional content. Therefore, a better understanding of the functional genetic content of these core taxa would not only help with defining the impacts of the environment on community composition, but also allow for a better identification of how populations of microbes vary as a function of distance and persist on vessels.

3.3.1.4 Moving toward a more targeted analysis

Through our other work, we demonstrate that some methods are proven to be more promising in investigating microbial interactions through machine learning(35). Here, rather than traditional competition-cooperation models adopted from ecology, machine learning can find very small but meaningful patterns in very large and complex data.

Our long-term goal is to be able to identify a subset of taxa that are useful as biomarkers of provenance to be able to develop those into a rapid detection technology. Here we used the 'core' microbiome approach to narrow in on a subset of taxa that would be useful in predicting past location and be robust to inter-vessel variability. In order to understand the role of certain signature taxa and their weight on predicting object provenance, it is necessary to use more targeted approaches, such as exploratory and predictive machine learning. However, the use of machine learning in environmental systems – especially pertaining to the microbiome, is limited. The following *Chapter* demonstrates a comprehensive, ground-up in silico workflow to investigate this specific question of object provenance – and other questions relating to microbiome research. These new computational methods are essential for identifying the biomarkers of maritime provenance, interpreting machine learning algorithms, and better understanding how microbial interactions inform model prediction.

3.3.1.5 Conclusions

It is fair to state that microbial persistence can be highly subjective to the system, and generalizations are constrained to microbial behavioral patterns in response to extraneous determinants and assumed perturbations of a particular system (e.g., confounders; resident microbial gene expression, mixing patterns, environmental parameters, …). In effect, inferences made here to the causal effect of microbes as biosensors of object provenance and of microbial persistence patterns may be limited to this specific system. However, this work demonstrates that a subset of microbes can persist on ships over long distances and can provide information about the previous location of these vessels. Our approach of using a 'core' microbiome may be helpful in addressing systems with high variability.

3.3.1.6 Contributions

This work will hopefully provide insight into the interpretability and capabilities of this widely studied system and similar systems from a practical and theoretical point of view. Lastly, the data generated as part of this experimental field design complements public datasets accessible to other researchers to validate and benchmark hypotheses relating to the study of natural microbial communities in complex settings. To the best of our knowledge at the time of writing this dissertation – this is the only publicly known experimental design of its kind – set out to predict vessel provenance and provides a detailed study into the dynamics of boat-associated microbial communities during transits.

3.4 Methods

All data referred to in methods can be found in file: Supplementary Data

3.4.1 System selection and sampling

The Chesapeake Bay spanning from Baltimore, MD to Norfolk, VA was an ideal system to study vessel-associated microbes and satisfy sufficient replication and adequate representation of the range of conditions found within these systems and nearby ports. The samples used in this *Chapter* ($n = 616$) were collected from June 24, 2018 to July 3, 2018 (details of which samples were collected on specific date/time can be accessed in *Supplementary Data*). Sampling stations were chosen to represent an appropriate longitudinal distribution across the transect to collect samples necessary to construct machine learning models. Two independent research vessels were charactered in collaboration with Chestertown, MD as follows: ***Voyage A***: 'Callinectes': fiberglass hull with ablative marine hull paint, and ***Voyage B***: 'Lookdown': aluminum coated in Kevlar, with ablative marine hull paint. Samples were collected across 25 sampling stations (coordinates supplied in *Supplementary Data*) spanning 296km along the Chesapeake Bay. At each station, surface water samples (1 liter) were taken and subsequently filtered through a glass fiber prefilter with a 1.6-µm pore size (47-mm diameter) (F1) and a 0.2-µm pore-size polyethersulfone (PES) membrane postfilter (47-mm diameter) (F2) (Sterlitech Corporation) using a Cole-Parmer Masterflex E/S 115 VAC portable sampler. Sample positions R, rear; BW, bilge water; BF, bilge biofilm; H, hull, were swabbed samples from the surface of the vessel. Filters and swabs were placed in 2-ml Eppendorf tubes with 500 µl RNA/DNA shield (ZymoBIOMICS) and stored at ambient temperatures until transported back to the laboratory to be stored at –80°C. Multiparameter data of water quality (conductivity, ODO, pH, salinity, TDS content, temperature, and dissolved oxygen content) along with global positioning system (GPS) coordinates of each sampling site were recorded *in situ* with a YSI ProDSS digital sampling system that was calibrated before each sampling trip. Nutrient data (phosphate, silicate, nitrate, ammonia), total organic carbon (TOC) and chlorophyll content was recorded for each station on each transit. Detailed metadata as described here can be found in *Supplementary Data.*

3.4.2 DNA extractions

DNA was extracted from each filter (F1/F2) and from two swabs from each sampling station and sampling position of the vessel (using the ZymoBIOMICS DNA microprep D4305 kit (Zymo Research, Irvine, CA, USA. For each open water sample, both the prefilter (F1) (1.6-µm pore size, 47-mm diameter) and postfilter (F2) (0.2-µm, 47 mm diameter) were cut in half, where one half was to be used in the DNA extraction and the other half stored as a contingency. Since swabs contained low biomass, both swabs used in sample collection were treated as a single sample DNA extraction.

3.4.3 DNA sequencing

First-stage amplification PCRs were carried out in 25-µl mixtures consisting of 12.5 µl Phusion high-fidelity PCR master mix (Thermo Fisher Scientific, Waltham, MA, USA) containing deoxynucleoside triphosphates (dNTPs) at a concentration of 200 mM each, optimized reaction buffer, 1.5 mM $MgCl_2$, and 1 U high-fidelity polymerase per reaction in 96-well VWR polypropylene plates. The primer pair 515f and 926r was used at a concentration of 0.4 µM to amplify a construct that spans the variable regions 4 and 5 (V4–V5) of the 16S rRNA gene. The PCR thermal cycler settings were as follows: 95°C for 3 min; 25 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and 72°C for 5 min. PCR cleanup was performed after first-stage amplification PCR to remove residual primers and excess reagents from PCR mixtures. For this cleanup, we followed the MiSeq library preparation guide (Illumina, San Diego, CA) and deviated from the standard protocol by using AxyPrep Mag PCR cleanup beads (Corning, Big Flags, NY, USA), using 10 mM Tris at a pH of 8 (down from 8.5) and by using 28 µl AxyPrep beads in the second-stage cleanup since the PCR volume was 25 µl (down from 50 µl). Second-stage indexing PCRs took place under the same mixture conditions as first-stage amplification PCR and with primers that contained a unique index sequence for each sample and the Illumina sequencing adaptors. An additional PCR cleanup was done after second-stage PCR, eluting to a final volume of 50 µl. Library preparation and sample pooling were performed according to the MiSeq 16S sequencing library preparation guide (Illumina). The products from the second-stage indexing PCR and subsequent cleanup stages were pooled into a library for sequencing at an equimolar concentration of 10 nM after ensuring that primer contamination was absent or at a minimum using a 2100 Bioanalyzer (Agilent, Santa Clara, CA). Denaturation and dilution of the pooled 16S rRNA gene library were performed according to the MiSeq 600-cycle V3 reagent kit guide (Illumina) to produce a 2 × 300-bp paired-end run for 649 total samples including processing blanks.

3.4.4 Computational analysis and visualization

All statistical analysis, machine learning models, and visualization were conducted on a local server (Red Hat Enterprise Linux server 7.3 [Maipo]; 256 Gb of random-access memory [RAM]) and on R environment version 3.6.3 using the following packages and associated dependencies as follows (in no particular order): phyloseq, tidyverse, microbiome, eulerr, microbiomeutilities, pheatmap, Biostrings, MicrobeDS, dplyr, vegan, ggpubr, missForest, Hmisc, mi, grid, forecast, tseries, reshape2, randomForest, inTrees, caret, caretEnsemble, ggplot2.

3.4.5 ASV identification and taxonomic profiling (denoising)

Raw 16S rRNA sequencing reads were demultiplexed using the Illumina MiSeq platform. Through the divisive amplicon denoising algorithm (DADA2 package) in R, primer nucleotides were removed, and overlapping paired-end reads were merged, quality filtered, and cleansed of internal standard phiX; to distinguish amplification and sequencing errors from true biological variation in our collected samples, amplicon sequence variants (ASVs) were inferred. To account for learning the inherently different error rates from the sequencing run, all samples were used to infer errors (from >100 million bases) so as to identify true diversity contained in the final data set. The subsequent ASV count tables was merged and then used to resolve and remove chimeric artifacts with higher accuracy as a result of the resolution of ASVs. Traditionally with OTU picking, chimeric sequences are removed in a conservative manner, as closely related sequences are later merged into the same OTU. While using ASVs, a more sensitive removal is accomplished by performing a Needleman-Wunsch global alignment of each sequence, finding bimeras (two-parent chimeras) and localizing combinations from a left and right parent chimera that overlaps the child sequence exactly. From 234,064 paired-end input reads, a total of 89,521 nonchimeric reads passed our filtering parameters and were used in ASV identification and analysis in this study. We obtained a count table analogous to the generally used OTU table; similarly, our features in this table are composed of the uniquely inferred ASVs that map how many of these amplicon variants were observed in each sample. Taxonomy of ASVs was assigned through DADA2 with a reimplementation of a rapid assignment naive Bayesian classifier that compares our biological sequence variants to a training set of previously accurately classified sequences using the SILVA v132 training set.

3.5 References

1. McLaughlin R. Maritime Crime: A manual for Criminal Justice Practitioners. Maritime Crime: A Manual for Criminal Justice Practitioners. 2017.
2. Corbett JJ, Winebrake J, editors. The impacts of globalisation on international maritime transport activity. Global forum on transport and environment in a globalising world; 2008.
3. Brooks MR, Faust P. 50 Years of Review of Maritime Transport, 1968-2018: Reflecting on the Past, Exploring the Future. 2018.
4. Helmick JS. Port and maritime security: A research perspective. Journal of Transportation Security. 2008;1(1):15-28.
5. Guard UC. National Plan to achieve Maritime Domain Awareness. Washington, DC. 2005.
6. Ghannam RB, Schaerer LG, Butler TM, Techtmann SM. Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities. Msphere. 2020;5(1).
7. Qian J, Lu Z-x, Mancuso CP, Jhuang H-Y, del Carmen Barajas-Ornelas R, Boswell SA, et al. Barcoded microbial system for high-resolution object provenance. Science. 2020;368(6495):1135-40.

8. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, et al. Temporal variability is a personalized feature of the human microbiome. Genome biology. 2014;15(12):1-13.
9. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, et al. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PloS one. 2016;11(12):e0167370.
10. Metcalf JL, Xu ZZ, Bouslimani A, Dorrestein P, Carter DO, Knight R. Microbiome tools for forensic science. Trends in biotechnology. 2017;35(9):814-23.
11. Milan M, Maroso F, Dalla Rovere G, Carraro L, Ferraresso S, Patarnello T, et al. Tracing seafood at high spatial resolution using NGS-generated data and machine learning: Comparing microbiome versus SNPs. Food chemistry. 2019;286:413-20.
12. Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J. Embracing environmental genomics and machine learning for routine biomonitoring. Trends in Microbiology. 2019;27(5):387-97.
13. Chang H-X, Haudenshield JS, Bowen CR, Hartman GL. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. Frontiers in Microbiology. 2017;8:519.
14. Schaerer LG, Ghannam RB, Butler TM, Techtmann SM. Global comparison of the bacterial communities of bilge water, boat surfaces, and external port water. Applied and environmental microbiology. 2019;85(24).
15. Schaerer LG, Webb PN, Corazzola A, Christian WC, Techtmann SM. Impact of air, water, and dock microbial communities on boat microbial community composition. Journal of Applied Microbiology. 2020.
16. Jannasch HW, Jones GE. Bacterial Populations in Sea Water as Determined by Different Methods of Enumeration 1. Limnology and Oceanography. 1959;4(2):128-39.
17. Dang H, Lovell CR. Microbial surface colonization and biofilm development in marine environments. Microbiology and Molecular Biology Reviews. 2016;80(1):91-138.
18. Salta M, Wharton JA, Blache Y, Stokes KR, Briand JF. Marine biofilms on artificial surfaces: structure and dynamics. Environmental microbiology. 2013;15(11):2879-93.
19. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639-43.
20. Wimpenny J, Manz W, Szewzyk U. Heterogeneity in biofilms. FEMS microbiology reviews. 2000;24(5):661-71.
21. Meireles dos Santos A, Vieira KR, Basso Sartori R, Meireles dos Santos A, Queiroz MI, Queiroz Zepka L, et al. Heterotrophic cultivation of cyanobacteria: study of effect of exogenous sources of organic carbon, absolute amount of nutrients, and stirring speed on biomass and lipid productivity. Frontiers in bioengineering and biotechnology. 2017;5:12.
22. Fuchsman CA, Staley JT, Oakley BB, Kirkpatrick JB, Murray JW. Free-living and aggregate-associated Planctomycetes in the Black Sea. FEMS microbiology ecology. 2012;80(2):402-16.
23. Smith DP, Peay KG. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PloS one. 2014;9(2):e90234.
24. Ward CS, Yung C-M, Davis KM, Blinebry SK, Williams TC, Johnson ZI, et al. Annual community patterns are driven by seasonal switching between closely related marine bacteria. The ISME journal. 2017;11(6):1412-22.
25. Logares R, Lindström ES, Langenheder S, Logue JB, Paterson H, Laybourn-Parry J, et al. Biogeography of bacterial communities exposed to progressive long-term environmental change. The ISME journal. 2013;7(5):937-48.
26. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, et al. Loop-mediated isothermal amplification of DNA. Nucleic acids research. 2000;28(12):e63-e.
27. Kellner MJ, Koob JG, Gootenberg JS, Abudayyeh OO, Zhang F. SHERLOCK: nucleic acid detection with CRISPR nucleases. Nature protocols. 2019;14(10):2986-3012.

28. Kramer A, Assadian O. Survival of microorganisms on inanimate surfaces. Use of Biocidal Surfaces for Reduction of Healthcare Acquired Infections: Springer; 2014. p. 7-26.
29. Wilkins D, Leung MH, Lee PK. Microbiota fingerprints lose individually identifying features over time. Microbiome. 2017;5(1):1-9.
30. Krause S, Molari M, Gorb E, Gorb S, Kossel E, Haeckel M. Persistence of plastic debris and its colonization by bacterial communities after two decades on the abyssal seafloor. Scientific reports. 2020;10(1):1-15.
31. Faust K, Lahti L, Gonze D, De Vos WM, Raes J. Metagenomics meets time series analysis: unraveling microbial community dynamics. Current opinion in microbiology. 2015;25:56-66.
32. Faust K, Bauchinger F, Laroche B, De Buyl S, Lahti L, Washburne AD, et al. Signatures of ecological processes in microbial community time series. Microbiome. 2018;6(1):1-13.
33. Mittelbach GG, Schemske DW. Ecological and evolutionary perspectives on community assembly. Trends in ecology & evolution. 2015;30(5):241-7.
34. Clarke TH, Gomez A, Singh H, Nelson KE, Brinkac LM. Integrating the microbiome as a resource in the forensics toolkit. Forensic Science International: Genetics. 2017;30:141-7.
35. Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal. 2021.

4 A translational microbiome: Interpreting exploratory and predictive black-box machine learning models.

Preface

Demonstrated above in *Chapter 2* and *Chapter 3*, microbial communities can be used as tools for biosensing geospatial location, and microbes are shown to attach to objects in these environments, which suggests microbes as tools for object provenance. This chapter is concerned with the inferential validation of using the 'core' taxa identified in *Chapter 3* to predict the provenance of a vessel in aquatic systems. However, this approach largely relies on machine learning, which is a black box method of investigation. In order to properly validate whether predictions from machine learning models hold true to the biological system – a framework to extract biologically meaningful information from models that are otherwise uninterpretable must be developed and employed. The following work demonstrates a robust means of investigating microbiome datasets with a benchmarking dataset from the human microbiome, which links microbes to host subjects. This work is two-fold, to satisfy the specific research question of object provenance, while also providing a framework to more thoroughly investigate how patterns in microbial datasets can be linked to a conditional outcome through machine learning – a current gap in microbiome research at the time of writing this dissertation. The work presented here leads into the importance of forward thinking for in-silico models to translational molecular applications.

Abstract

Microbial ecosystems are rather complex, with hundreds of members interacting with each other and the environment. The intricate and hidden behaviors underlying these interactions make research questions challenging – but can be investigated using machine learning. By probing a microbial community using exploratory and predictive machine learning – we can begin to uncover system-wide patterns. This allows us to leverage this information and extend the application of microbes to be used as novel tools – such as biosensors of object provenance. This study explores how microbial community interactions that are linked to a system-state can be identified using some of the most widely used machine learning methods in microbiome analysis. This framework allows researchers to extract from these algorithms a reasonable biological understanding of the system being studied to better apprehend and interpret their results when using machine learning.

Introduction

Machine learning, as discussed throughout this dissertation – is fast becoming a routine tool for analyzing data and making predictions, with numerous health and environmental applications related to the microbiome(1-4). High-throughput

sequencing has enabled very targeted and inexpensive marker-gene studies with sufficient replication to benefit from machine learning in prediction of outcome(4). As such, the use of both human and environmental associated microbial community composition as predictive biomarkers of conditions of interest is now possible. This has led to the broad adoption of algorithmically informed decision support systems that catalyze our understanding of the microbiome in disease states, forensics, ecology and complex biochemical pathways(5-7). Discussed in *Chapter 1*, machine learning algorithms are exceptional at making predictions but have many limitations. A significant limitation is that often machine learning models are black box methods of investigation where the rationale behind decisions being made are hidden behind layers of non-transparent complexity(8, 9). Here, we address this issue, because black box modeling obstructs our understanding of the microbiome and potentially thwarts high-impact translational advances from the actionable insights that could be extracted from modeling procedures. These include, for example, constructing sensitive and robust molecular detection tools for complex biomarkers of disease states or environmental contaminants that are foreshadowed by microbial communities(10, 11).

Typically, machine learning generalization is defined as the model's ability to adapt to perform on new, unseen data, using the same set of predictive features used to create a model(12)**.** Here we argue that while using machine learning to make predictions using microbiome data, investigators should aim for both a model's ability to perform well on unseen data and that meaningful community member interactions can be measured, and actionable insight can be extracted. By inferencing machine learning models trained on microbial datasets using human interpretable outputs, researchers can expect to (1) provide transparency and gain a deeper understanding of the intra- and inter-microbial community interactions in various systems (2) determine the stability of the microbial community and mechanism(s) responsible for a transition from one state to another (e.g., from 'healthy' to 'diseased') and (3) identify robust, multi-microbe (taxa) biomarkers that can be used to more accurately predict an outcome relating to human health and outcomes in the built and natural environments.

Indeed, the ultimate goal of machine learning in microbial analysis is to train algorithms to recognize microbial behavioral patterns as they are in the nature of the system(13). In order to gain biologically meaningful insights from machine learning, it is important that the models we use be interpretable such that the information being used by the model to make a decision is clearly accessible. With a sizeable gap between domain expertise of microbiology and machine learning, there are few techniques that offer transparent outputs for machine learning models constructed using metagenomic sequencing data. Although several recent pipelines suggest improvements of machine learning interpretability in microbial

analysis(14-16), these developments, along with the majority of progress in biological machine learning, still focus on performance metrics and putative single microbe-response variable interactions and fail to appreciate the importance of microbial 'associations' predictive of an outcome. One of the strengths of machine learning is the ability to identify complex feature interactions that are predictive of a particular outcome. Previous methods that only determine single-feature importance are limited because in the real-world system after which machine learning models are constructed, the cause of a condition being studied is often attributed to multiple taxa, rather than single taxon-condition linkages as reported in most studies(4). This suggests the need to explore, develop, and apply biologically motivated algorithms to provide interpretations that measure relevant knowledge and properly describe what a model has learned from microbiome data.

To approach this issue, we explore the application of interfacing microbiome data with machine learning interpretability by (1) deriving 'association' rules from a fitted model (or models) and (2) provide descriptive accuracy by displaying how the derived decision rules impact model performance, and (3) present analyses in an intuitively comprehensible manner. Our framework discretizes a high-dimensional frequency count matrix into a list of interaction terms that objectively capture learned microbial relationships as simple, interpretive and generalizable plain-text conditional statements. The derived rules are meant to denote biologically meaningful co-abundant patterns, multicollinearity and the complex covariance structure of high-throughput next generation sequencing data.

We aim to produce models that are both accurate and transparent to microbiome researchers and offer the ability to debug and audit models – which allows integrating outputs into translational molecular applications at scale, with speed and with low overhead processing. Beyond applying these methods to continue to probe object provenance, we expect this work to serve as a meaningful perspective to further our understanding of the microbiome and for developments of interpretable, biologically motivated machine learning algorithms for marker-gene datasets.

4.1 An argument for interpretable machine learning in microbiome research

Both domain experts in microbiome research and machine learning developers struggle to properly verify, interpret and evaluate inner logic and reasoning behind the use of decision support systems(17, 18). By loose definition, interpretability is the degree to which a human can understand the cause of a decision or consistently predict a model's result(8). To this extent, we define interpretable machine learning as measuring relevant knowledge and the ability to properly describe what a model has learned from data. Thus, interpretations applied to the microbiome would allow a mechanistic view of the role of microbes in our health and environment that drive accurate decision making and prediction.

Interpretable machine learning is a domain-specific challenge and potential reason for why there are so few descriptive learning frameworks for microbial analysis, despite the many real-world implications that would result from it. For example: medical microbiology relies on Koch's postulates to determine microbial agents of disease – which requires a clear cause and effect determination between a specific microbe and disease state. However, microbiome studies often identify dysbiosis as a cause of disease despite how these health states are often less quantifiable. While machine learning allows for clarity to be brought to quantifying dysbiosis – without interpretable metrics – the ability to apply Koch's postulates and determine cause and effect of microbiome associated diseases remains implausible.

A noteworthy constraint to machine learning as it relates to domains that would greatly benefit from interpretability (not limited to the microbiome) is the social acceptance and trust that we can integrate machines and algorithms to drive critical reasoning. For wide-scale implementation, regulatory oversight and to ensure human acceptance and trust, the microbiome research community must consider the rigorous evaluation and scientifically verifiable interpretability of user-level computational reasoning that has significant impact on our everyday life – which is partly achievable through interpretable and descriptive machine learning(18, 19).

In view of recent, interpretable ML has been approached in biological disciplines including single-cell RNA seq, drug discovery and development, and neuroscience(20-24). In the case of microbiome research, the role of microbial community membership interactions in model prediction are often left unreported or left open to speculation. The traditional extent of model transparency is gauged by single feature (nucleotide sequence variants or OTU clusters) importance on the model as a whole and the association of that single feature to a response variable and does not account for which features of the community may be interacting with each other to explain decisions being made(25-27).

Historically, computational tools used to analyze taxa-distributions and co-occurrence patterns in multivariate microbiome datasets have been borrowed from other disciplines and tailored to accommodate marker-gene analysis(28). Novel machine learning techniques that can analyze microbiome datasets for actionable insight are no exception, since they can be and often are borrowed from disciplines that focus on developments that are not biologically motivated(17, 29, 30). Health and environmental domains such as the microbiome can benefit from the ability to debug and audit models. This could include leveraging domain specific knowledge to discern whether a model is incorporating rules consistent with experimentally determined biological phenomena to make predictions rather than predictions from phenomena resulting from data structure, preprocessing or model parameterization. Additionally, an interpretable microbiome-based model would be of interest to other disciplines that may be able to incorporate information from microbes to enhance existing toolkits (forensics, biomarker prediction, etc.)(31, 32).

Further, the ability to interpret these machine learning models has potential to drive translational research and provide the ability to satisfy Koch's postulates in linking microbiome function and human and ecosystem health. Further, the ability to identify the taxa and interactions that are related to a particular outcome will enable not only detection of disease states and environmental outcomes but may provide a framework for interventions that can restore the perturbed state to the healthy state. Within machine learning, there are various models that have been applied to microbiome data which vary in transparency. While advances have been made in explainable AI(4), sometimes the choice of a slightly lower performing model in favor of increased explainability would be justified to enable a more thorough understanding of the system and translation.

Despite the need for improvements on machine learning interpretability in marker-gene analysis, it is notable to mention that in our current state – machine learning continues to facilitate the progression of microbiome related research(4).

4.2 Study Framework: A scalable methodology for interpreting microbiome-based machine learning

As prefaced, before pursuing machine learning in the context of predicting object provenance, we first demonstrate the proposed framework for biologically interpretable machine learning in a use-case using a well-known microbiome study. The choice of this dataset extends the framework to provide other microbiome researchers additional tools for investigating the human microbiome. Further, to demonstrate that this framework is robust and that the methods used are domain agnostic (i.e., not limited to very specific domain problems such as vessel provenance) and can be applied to many problems in microbiome work.

The workflow developed in this study is first summarized using the 'moving pictures of the human microbiome' dataset(33) from two healthy host subjects, one male and one female sampled daily for 15 months and 6 months, respectively – at three body sites (gut (stool), sebum (left and right palms) and tongue (saliva)). This dataset is suited for this analysis as it is a dense microbiome study and has sufficient sample sizes to model microbe-host interactions from multiple body sites (**Fig. 4.1**). The taxonomic features used in modeling (ASVs (Amplicon Sequence Variants)) are not provided with biologically meaningful ranks as were in previous chapters but are supplied with full range of taxonomic resolution (Phylum - Genus) in ***Supplemental Data***. Using the experimental conditions described above, an initial black box model using random forests is fit and optimized to determine interaction effects and additional *post-hoc* explorative models are constructed to provide additional descriptions of the interaction effects. These analyses are demonstrated with binary classification but can be extended to multi-class classification and regression problems and using a variety of tree ensemble algorithms.

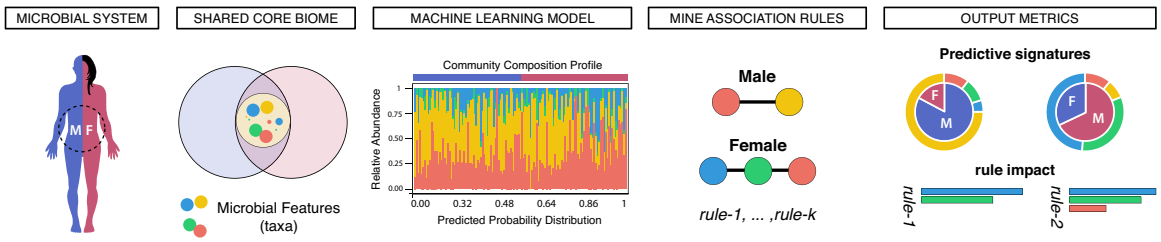


Figure 4. 1

Illustrative workflow for interpretive microbiome-based machine learning.

Model agnostic methods such as Local Interpretable Model-agnostic Explanations (LIME)(34) and iml(35) offer metrics for individual feature – response effects (local interpretation) but fail to indicate the marginal effects of features acting together to influence certain states in the system (global interpretation). The analyses presented here allow a comprehensive audit of the data by identifying a list of decision rules and provide constituent feature and whole rule 'impact' metrics. This allows researchers to weight interaction terms and parse out biologically meaningful contributions from microbial assemblages (e.g., co-occurring taxa) that are reasonably expected to occur within a study cohort or in the environment and are otherwise inextricably linked(25, 36). This explorative and predictive framework is described in the following two steps.

4.2.1 Identify generalized feature set

The previous *Chapter* deliberates the use of a generalized set of features. Here, we repeat this process. The features chosen to model are selected from a 'core' biome (10 ASVs – as taxa) shared between the two states/classes (male (M3)/female (F4)). How these core taxa were derived is similar to the previous chapter (*Section 2*), and also is described in this chapter (*see methods*). Using such a feature space comprised of highly prevalent and detectable taxa in both states is contrary to traditional biomarkers such as *Indicator Species* analysis(37), since the taxa associated with traditional biomarkers often use single taxa only present and prevalent in one state and not the other. Our approach aims to identify the hidden heterogenous effect of all taxa together and identify robust and generalizable multi-taxa biomarkers. In many systems, like the human microbiome, there is large inter-personal variation, which can make generalization difficult. The approach of using core microbiome effectively reduces the inter-sample variance while capturing subtle differences in intra-personal variation between taxa that are shared but are still strong predictors between each class/state in the system (**Fig. 4.1**). Modeling using taxonomic features that are not categorized as 'core' may identify taxa as biomarkers of inter-individual differences. These inter-individual biomarkers are noisy and often spuriously result from personal lifestyle (i.e., interaction, activity, diet, hygiene, which is not accounted for in the feature space) and is a less generalizable and robust approach as using a 'core' biome comprised of highly detectable and prevalent taxa throughout a population/cohort. Important earlier studies on the human microbiome project (NIH: HMP)(38), and other proposed strategies for comprehensive sampling(39), did not necessarily account for the future of powerful predictive models to be built around targeted sampling regimes, such as with object provenance.

Some work has been done to characterize a 'core' human microbiome using the same resolution gene-markers as in this dissertation (16S rRNA)(40). Using the stool biome as an example: if we consider how the microbial community

composition between male/female is variant to each sampling type (stool, skin, mouth, etc.), our approach of using a generalized feature set (as intra-personal 'core') of a sample population could lead to more accurate and generalizable models across the entire, or similar populations. Using such a feature set helps to normalize confounders from data points which cannot be accounted for all of the time (e.g., lifestyle metadata) and allow for models built solely on a search space indicating a microbial community (**Fig. 4.1**). Although this generalization is toward human microbiomes, the same concept can be applied to object microbiomes, as in the research question this work concerns.

4.2.2 Association rule mining

The principal methodology behind this approach is to identify significant feature interaction effects that occur in the aggregated voting process of many types of ensembles (e.g., random forests, regularized random forests, boosted trees). These types of models are appealing as they create an ensemble of decision trees that can be used in evaluating and identifying features consistently found in decision trees. A more thorough review of these processes is described in *Chapter 1*. Here, we describe feature interaction as derived from the model where two or more features (taxa) act together to influence and explain how predictions are made in a biological sense for samples/observations (**Fig. 4.1**).

This heuristic approach is three-fold. First, for practical and computational feasibility since a more complex model (i.e., a large feature space/increased parameters) would make the number of interaction terms containing noisy or non-informative but eligible features predominate each node. Second, a moderate number of predictors such as the shared 'core' biome between both states still generates a sufficient number of interaction terms from an interpretive and parsimonious standpoint. More so, exhaustively searching/screening a large feature space (many more features) could incur a penalty on account of spurious interactions overshadowing the small number of truly informative interactions, whereby a number of false positives could be identified. Third, all interaction terms are screened by a frequency and error threshold. A higher frequency indicates that interactions involving the same group of features are occurring at subsequent nodes and are more likely to be a global interaction across the observed range of samples and predictors in the cross validated ensemble (final model fit).

4.3 Issues with traditional machine learning metrics in microbiome research (use-case)

To demonstrate this framework sequentially and for context, a machine learning model was first constructed to accurately predict whether a stool sample could differentiate the host subject it was collected from (male (M3) or female (F4)) (**Fig. 4.2B**). Like many other machine learning models in the microbiome literature, this model was constructed to optimize performance, rather than for explanation. Hence, we are left largely with a black box model – where biologically meaningful metrics (as shown in **Fig.** 4.**1**) such as microbial interactions that drive the accurate decision boundary between male/female are hidden or inaccessible. Before demonstrating how interpretive metrics can be brought to the model – first outlined are some issues relating to traditional interpretation metrics (often in the form of performance) to put things into perspective.

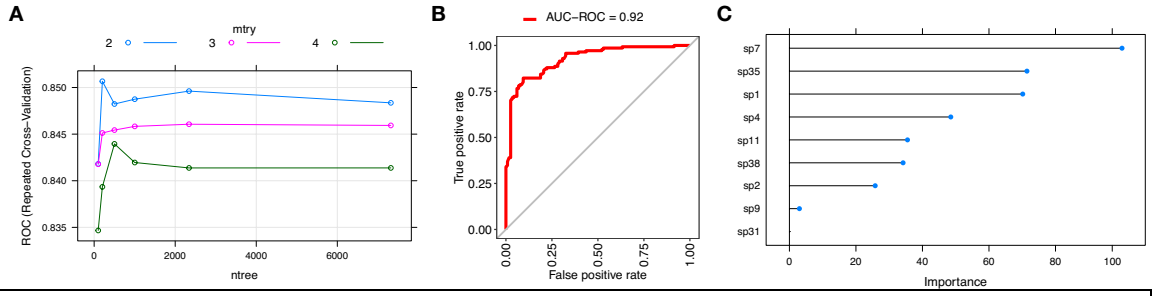


Figure 4. 2

Black box model of stool sample to host subject. Displayed are model assessment and metric outputs for a random forest model parameterized to make accurate predictions of host subject (male, M3 (128 samples); female, F4 (131 samples)) on the basis of stool samples containing 10 'core' taxa. (**A**) A grid search showing a threshold of optimal hyperparameters, **mtry**: number of predictor features (ASVS) to be randomly sampled as candidates at each split, and **ntree**: number of trees to grow in the cross-validated ensemble. (**B**) AUC-ROC (Area Under the Receiver Operating Characteristic) curve as a measure of the performance of repeated cross-validation classification at various thresholds. This is plotted as TPR (true positive rate – sensitivity) as a function of the FPR (false positive rate – 1-specificity). By extension, computing the area under the ROC curve (AUC) can provide a measure of how well the model could discriminate predictions of male vs. female samples. A higher AUC is preferred and can range from 0.5 (separation of samples was no better than random chance) to 1.0 (perfect separation of samples). (**C**) Variable importance plot describing which ASVs are 'important' in making accurate predictions. More information on how random forests work is thoroughly discussed in *Chapter 1*. All raw values for this model including predicted probabilities at each fold (*k*) in the cross-validation can be accessed in *Supplementary Data*.

Machine learning is a very dynamic process and is highly configurable to each domain problem. The complexities of model construction can, and often does, hinder a microbiome researchers' ability to understand the underlying biological complexity inside of an already complex modeling procedure. For example, often, it is widely acceptable to find an optimal model through a grid-search to test a gradient of hyperparameters together and localize a set yielding the highest performance metric (**Fig. 4.2A**). Accepting a model strictly on the basis of a performance metric is effectively cherry-picking the most 'accurate' model. In general, marginal increases in performance can correspond to learning more irrelevant noise in the data – with clear correlation between better performance and overfit models(41). As shown in **Fig. 4.2B**, this model was specifically tuned to reach a top performance AUC-ROC of 0.92, where 1.00 assumes perfect separation of samples into their respective class. Although AUC-ROC is well established metric – it does not mean that it an end-all metric, or that it needs to be fully optimized to achieve high performance. This is because the predicted probability distribution of samples a model is trained on is dynamic to the biological questions being asked. For example, often research questions can be satisfied with a model that has a lower performance metric but is generalizable and more interpretable, which has a better trade-off than optimizing performance to overfit a noisy and non-generalizable model. With variable importance (**Fig. 4.2C**), researchers continue to struggle with and outline the issues of using this metric in bioinformatics and related scientific fields – as this metric cannot reliably select accurate predictors in higher-dimensional data and is largely bias in most application(42, 43). This is because truly uninformative features (ASVs in this case) may be artificially preferred and selected in variable selection process while building trees – more so if a performance metric is optimized for.

Researchers are going to continue to use machine learning to help investigate biological problems. Moving forward, there must be a better solution to model more consistently to the real-world – as unknowingly modeling on noise is not actionable or valuable. Such a framework, as described through this *Chapter*, will show how interpretation can assist the black box modeling process and shine light on complex biological questions.

4.4 Modularity-based interpretation

There are some tools that allow insight into machine learning models – but there is not a widely accepted, robust and uniform machine learning framework to thoroughly investigate microbial associated outcomes(4). Often, microbiome researchers are tasked to come up with creative and novel methods to do so. As shown here, a good approach is to combine many different analytical tools that are borrowed from fields such as microbial ecology, or from machine learning development and theory. This approach as demonstrated here and is termed '*modularity-based interpretation*' – identifying the additive structure of a black box model(44). Modular approaches have been used in estimating biodiversity(45), and has been extensively studied for understanding community structures in complex systems with deep learning(46). This section will show how such an approach enables inference beyond what can be drawn from performance metrics (e.g., commonly reported: accuracy, AUC, etc.) alone and can increase the extent to which the black box is opened of the fitted model.

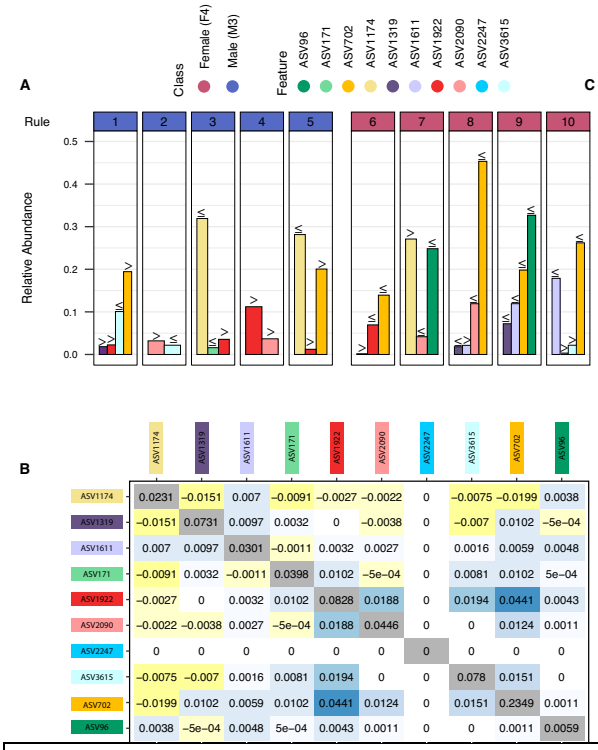
It is worth mentioning again that the techniques that will be described are not limited to microbiome studies – as they can be employed in any study using machine learning with frequency count matrices as the search space, such as with DNA or RNA sequencing, with expression profiles, and even domains outside of biology(47). In fact, the basis of association rule mining was first developed to understand rule-based learners in the context of large-scale transactions for point-of-sale (POS) systems in supermarkets(48). For example, to understand how people are likely to purchase certain grocery items that complement each other or based on product placement. To demonstrate the application framework, the prevailing question of object (*vessel*) provenance in marine settings using a rule-based learner is explored below.

4.4.1 Measuring microbial interactions

The initial model presented (**Fig. 4.2**) is only meaningful and interpretable to the extent that is described. To expand on this, when the same data are modeled and parameterized in pursuit of additional *post-hoc* analyses (e.g., mining recurrent patterns), researchers can better navigate convoluted microbial assemblages in complex real-world systems after which the machine learning model is constructed. This includes but is not limited to the relative impact of each association rule to stable/altered states of the system or to explain why certain microbes interact more frequently.

Below, we explore interpreting a black box model by (1) deriving 'association' rules from the fitted model and (2) by providing descriptive accuracy through displaying how the derived decision rules impact model performance, and (3) present analyses in an intuitively comprehensible manner. This framework

discretizes a high-dimensional frequency count matrix into a list of interaction terms that objectively capture learned microbial relationships as simple, interpretive and generalizable plain-text conditional statements. The derived rules are meant to denote biologically meaningful co-abundant patterns, multicollinearity and the complex covariance structure of high-throughput next generation sequencing data and are described in **Fig. 4.3**.



rule	feature interaction	reduced performance by feature	reduced performance by rule (rel. impact)	freq.	err.	complexity (length)	class
1	ASV702	0.235	0.38	0.263	0	4	
	ASV1319	0.060					
	ASV1922	0.072					
	ASV3615	0.075					
2	ASV2090	0.045	0.11	0.253	0	2	
	ASV3615	0.065					
3	ASV171	0.040	0.133	0.151	0	3	
	ASV1174	0.026					
	ASV1922	0.072					
4	ASV1922	0.083	0.111	0.075	0	2	
	ASV2090	0.047					
5	ASV702	0.235	0.317	0.059	0	3	
	ASV1174	0.026					
	ASV1922	0.072					
6	ASV171	0.040	0.28	0.043	0	3	
	ASV702	0.220					
	ASV1922	0.072					
7	ASV96	0.006	0.0726	0.054	0	3	
	ASV1174	0.026					
	ASV2090	0.040					
8	ASV702	0.235	0.352	0.022	0	4	
	ASV1319	0.060					
	ASV1611	0.030					
	ASV2090	0.046					
9	ASV96	0.006	0.291	0.022	0	4	
	ASV702	0.220					
	ASV1319	0.058					
	ASV1611	0.027					
10	ASV171	0.040	0.332	0.016	0	4	
	ASV702	0.220					
	ASV1611	0.030					
	ASV3615	0.075					

Figure 4. 3

Mined association rules and associated interpretation metrics. (**A**) Feature interaction terms (**Definition 1**: (*rule*) identified for each class (male/female). These rules meet a user-defined filtering criterion after being mined (*frequency, error, complexity*). Symbols above the bar plots represent boundaries of relative abundance for each 'core' community member satisfying each rule. (**B**) Feature interaction matrix (**Definition 1**: (*pairwise*) to examine the standalone and pairwise impact between any two features used in modeling. A higher value suggests that a constituent taxon or pairwise interaction is a strong influencer of predictions – and if we were to remove it, the model would become more error prone. (**C**) Model interpretation summary table displaying each *rule* (rule number), *feature interaction* (**Definition 1**: (*rule*), *reduced performance by rule (rel. impact)* (**Definition 3**), *freq* (**Definition 4**), *err.* (**Definition 5**), *complexity (length)* (**Definition 6**), *class*, the class membership the metrics belong to (male / female).

What will be described moving forward is a broad level overview of machine learning interpretation regarding human host subject predictions – since this section is meant to demonstrate how robust the framework is rather than to go into the biology of the human microbiome. Therefore, depictions will serve only as an introduction to the methods using an external microbiome dataset that is outside of the domain of object provenance. In the next section, a similar *modularity*-based approach will provide more detail of interpretation of results as they relate to the biological system of modeling object provenance in the Chesapeake Bay.

4.4.2 Approach toward microbial interaction metrics

For an interpretable machine learning model linking host subject – stool sample, a total of 9,834 interaction terms were fit to the final cross-validated model. These rules were subsequently pruned for a total of 10 relevant and non-redundant rules governing the majority of predictions throughout the training data (5 rules for each host subject: (male, M3 (85 samples); female, F4 (101 samples)) (**Fig. 4.3A**). The interaction rules derived from the microbial community composition could be used to predict either host with 87% accuracy. To test the how valid these interactions are as predictive multi-taxa biomarkers, the same 10 rules were used to construct a model to predict out-of-sample data (n=73) from the same hosts, achieving 79.5% accuracy (reproducible model can be found in *Supplementary Data*).

Next, we asked to what extent the contribution of each feature from each rule had on the fitted model. This was accomplished by fitting a rule-wise permutation (i.e., mixing of values of each constituent taxon feature in each interaction term) to provide two additional interpretive model metrics that provide weight to interactions. The first, for the reduction in performance by constituent taxon feature (*Ri)* (*constituent impact*) and second, for each rule as a whole (*R*U) (*interaction impact*), where a higher value for each of these metrics indicates a higher overall impact on the global model performance (measured through a performance metric). The third metric is *Fi*, denoting a constituent taxon that is present as a feature in the search space used to construct the model but was not present after pruning relevant and non-redundant interaction terms.

Notations
*R*U = all taxonomic features in interaction (*interaction impact*)
Ri = constituent taxon as feature present in interaction (*constituent impact*)
Fi = constituent taxon as feature present in full feature set (*e.g., not used as Ri*)
*F*U = full feature space (*Fi, Fj)*

Note: *all source code can be found under '**Data Code and Availability'; page xii**. All interpretation metrics are in terms of classification, although can be reimplemented for regression.*

Note: **Algorithms 1 & 2** were developed by Houtao Deng(29) and are adopted here to show detailed descriptions of the inner-workings of mining association rules from microbiome data. **Algorithm 1 & 2** descriptions are directly from Houtao Deng(29) to preserve original meaning.

Definition 1. ***Feature interaction (Figure 4.3A, B)***
In an interaction term (*rule*), ***Feature interaction*** is described as a conjugate variable-value pair that can be extracted from the root node to the leaf node in a tree. This was motivated by algorithms from original work on interpretable tree ensembles(29) and were reimplemented as a custom method for a tunable caret (classification and regression training) train function(49). Another form of feature interaction (*pairwise*) can be used to examine the standalone and pairwise impact of features used in modeling, measured as *(Err(Fi) + Err(Fj)) - Err({Fi,Fj})* after permuting a combination of any two features – motivated by algorithms for interaction measures for accuracy reduction(30).

Algorithm 1
ruleExtract(ruleSet, node, C): function to extract rules ***ruleSet*** from a decision tree. In the algorithm, let ***C*** denote the conjunction of variable-value pairs aggregated from the path from the root node to the current node, ***Cnode*** denote the variable-value pair used to split the current node, ***leafNode*** denote the flag whether the current node is a leaf node, and $\boldsymbol{pred_{node}}$ denote the prediction at a leaf node.

input : $ruleSet \leftarrow null, node \leftarrow rootNode, C \leftarrow null$
output: $ruleSet$

if $leafNode = true$ **then**
  $currentRule \leftarrow \{C \Rightarrow pred_{node}\}$
  $ruleSet \leftarrow \{ruleSet, currentRule\}$
  return $ruleSet$
end
for $child_i$ $=$ *every child of node* **do**
  $C \leftarrow C \wedge C_{node}$
  $ruleSet \leftarrow ruleExtract(ruleSet, child_i, C)$
end
return $ruleSet$

Other other than computational cost and complexity of the model (parameterization: *maxDepth*, …, and dimensions: (*X*) feature by (*N*) sample), there are not many constraints to how many rule conditions there can be. As such, this same work provides an efficient way to conditionally prune and extract rules

based on quality or by meeting some user-defined criteria to help address research questions, such as ***Relative impact***, ***Frequency***, ***Error***, ***Complexity*** (**Figure 4.3A**), or a combination of multiple metrics – as will be explored below.

Algorithm 2

condExtract(***condSet***, ***node***, ***C***, ***maxDepth***, ***curretDepth***): function to extract conditions ***condSet*** from a tree ensemble. In the algorithm, let ***C*** denote the conjunction of variable-value pairs aggregated in the path from the root node to the current node, C_{node} denote the variable-value pair used to split the current node, ***leafNode*** denote the flag whether the current node is a leaf node. Note one can set a maximum depth ***maxDepth*** of the tree where the conditions are extracted from. In a decision tree, most useful splits tend to happen in top levels of the trees (i.e., when *depth* is small), so setting a maximum depth can reduce computations, and may also avoid extracting overfitting rules. ***maxDepth***=-1 means there is no limitation on the depth.

input : $condSet \leftarrow null, node \leftarrow rootNode, C \leftarrow null,\ maxDepth \leftarrow -1, curretDepth \leftarrow 0$
output: $condSet$

$curretDepth = curretDepth + 1$
if $leafNode = true\ or\ curretDepth = maxDepth$ **then**
  $condSet \leftarrow \{condSet, currentCond\}$
  return $condSet$
end
for $child_i\ =\ every\ child\ of\ node$ **do**
  $C \leftarrow C \wedge C_{node}$
  $condSet \leftarrow condExtract(condSet, child_i, C, maxDepth, curretDepth)$
end
return $condSet$

Note: **Definitions 2 & 3** are adopted from the work of Sejong Oh(30) on feature interaction measures in terms of prediction accuracy, and were reimplemented as a custom method to extract interpretable metrics (notably, to permute {*ASVi*, *ASVj*} ⊂ *F*U and {*ASVi*, *ASVj*} ⊂ *R*U)) from a tunable caret (classification and regression training) train function(49).

Definition 2. ***Reduced performance by feature (feature impact) (Figure 4.3B, C)***
A reduction in performance of the model was measured after permuting each feature in the search space {*ASVi*, *ASVj*} ⊂ *F*U. A permuted feature is an effective measure of impact and has been studied extensively for determining feature importance from decision tree algorithms(43, 47, 50). By randomizing the values of a feature, the relationship between feature – response is broken and the corresponding loss function (generally an indication of performance) dictates model dependence on that feature for accurate predictions. This is especially important to note that in this method – permuting variables not only breaks the

relationship of feature to a response label, but in turn any interaction terms it that feature could be included in.

Definition 3. ***Reduced performance by rule (relative impact) (Figure 4.3C)***
The reduced performance of an interaction rule can be defined in the context of reduced performance by constituent feature. Instead of permuting each feature used in the search space as in **Definition 2** – the rules as a whole are permuted here (meaning all feature involved in a rule are permuted together), and reduction in performance is measured after each rule is permuted. Hence, a relative impact of each rule is reported.

Note**:** the motivation behind **Definitions 4, 5, 6** were developed by Houtao Deng(29) and are adopted here to determine rule importance and validity through filtering criteria – along with detailed descriptions of the inner-workings of mining association rules from microbiome data.

Definition 4. ***Frequency (Figure 4.3C)***
All interaction terms were screened by a ***Frequency*** metric to help determine the overall importance of a rule. Here, the ***Frequency*** of a rule is the proportion of samples that satisfy a rule condition (the number of times a rule could classify a sample into its proper class). A higher frequency indicates that interactions involving the same group of features are occurring at subsequent nodes and are more likely to be a global interaction across the observed range of samples and predictors in the cross validated ensemble (final model fit).

Definition 5. ***Error (Figure 4.3C)***
All interaction terms were screened by an ***Error*** metric to help determine the overall importance of a rule. Here, the ***Error*** of a rule is defined as the number of samples that were incorrectly classified governed by the rule divided by the total number of samples that were satisfied by the same rule.

Definition 6. ***Complexity (Figure 4.3C)***
Rule ***Complexity*** is measured as length of variable-value pairs in a rule (number of features). In this modeling regime, if two rules have similar ***Frequency*** and ***Error***, the rule of smaller length is chosen to be displayed since it is inherently more interpretable.

This work together suggested domain agnostic machine learning interpretation techniques could be wrapped in an accessible and intuitive format for microbiome researchers – providing graphical summary reports to help investigate the microbiome.

4.5 Predicted probability thresholds

In classification machine learning, each prediction of the class of an observation carries with it a predicted probability (ranging from 0.0-1.0). These predicted probabilities indicate for each sample the level the probability of a class membership (male or female) it belongs to – and is a good way to interpret the decision threshold between samples in a model (**Fig. 4.4A**). Often the class labels from predictions are accepted without consideration of predicted probabilities. Here, labels are not necessarily required, and instead, likelihood estimates for the class that each sample belongs to is assigned – and can later be interpreted in the context of the microbial community (**Fig. 4.4B**). It is notable to mention that this metric can satisfy multiple class memberships (*Yi,Yj*), and can help quantify microbial community members used as predictive features in machine learning.

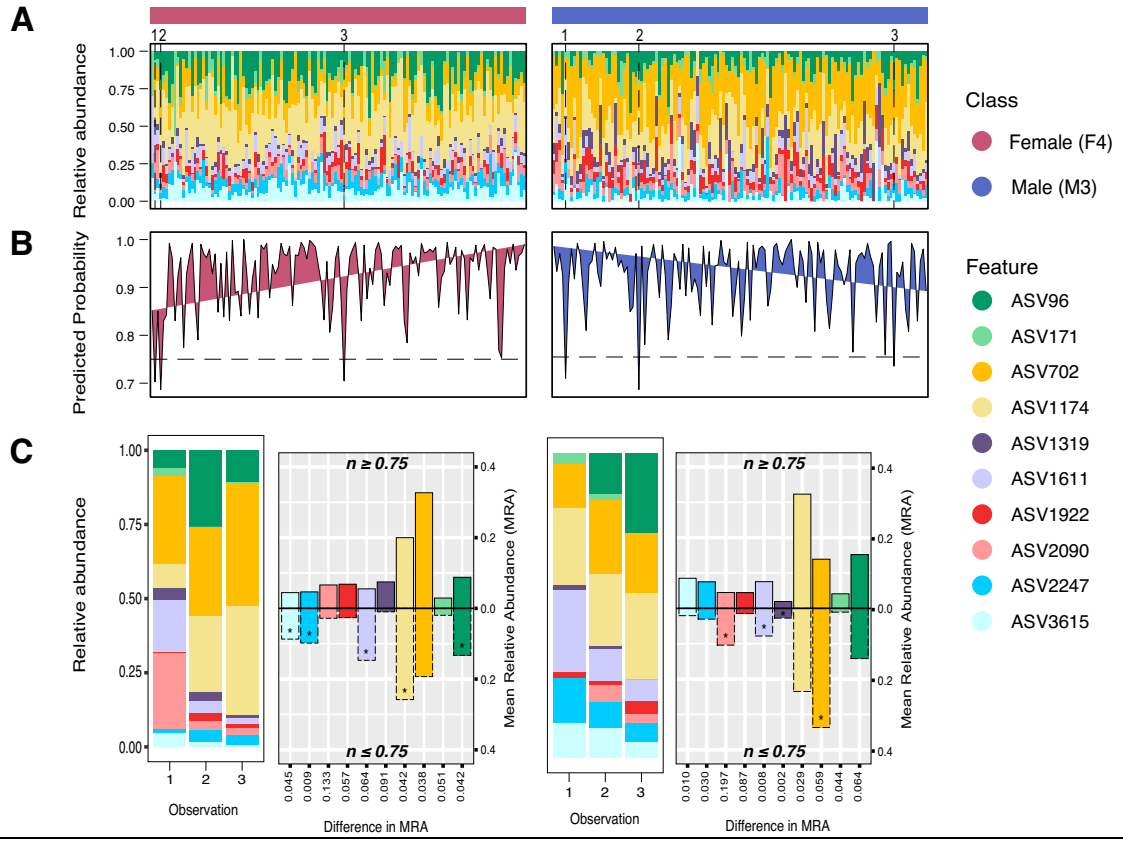


Figure 4. 4

Core biome community profile over sample-wise predicted probability distribution. (**A**) Relative abundance of shared 'core' microbial community profiles between male/female samples. Samples with predictions below user-defined threshold (predicted probability: 0.75) are denoted with a dashed line and are represented by an index corresponding to sample number in the training data. (**B**) Predicted probability distribution aligned to sample-wise community composition profiles to show the distribution of probabilities dynamic to

community composition. (**C**) Resolved samples that are 'misclassified', or that were predicted below the 0.75 probability threshold. Additionally, plotted are profiles of mean relative abundance (MRA) of all samples that are above the 0.75 predicted probability threshold (***n* ≥ *0.75***) (solid line) compared to samples predicted below the probability threshold (***n* ≤ *0.75***) (dashed line). A star (*) in a boxplot denotes that the feature had a higher MRA in the sample that was predicted below 0.75.

When a predicted probability assigned to a sample is below a user-defined threshold (≤0.75, for example), this likely suggests that the microbial community composition of that sample resembles more of the opposing, or another class (for more than two-class classification). **Fig. 4.4** provides a way to screen such samples, where three samples from each class were assigned ≤0.75 predicted probabilities for the actual class membership each of the samples belonged to (**Fig. 4.4A**). These 'misclassified' samples are further interrogated in **Fig. 4.4C**, where the relative abundance of 'core' microbes in those samples appear to be dominated by *ASV702* (female) and *ASV1174* (male). Interestingly, when comparing this to **Fig. 4.4A**, abundance of these same taxa in 'accurately' classified samples is the opposite – where *ASV702* was dominant in male samples and *ASV1174* was dominant in female samples in the aggregate.

Further, **Fig 4.4C** provides differences in mean relative abundance (MRA) for each taxon in each class membership. Here, female samples: (*ASV2090*, *A1922*, *ASV1319* and *ASV171*), and male samples: (*ASV3615*, *ASV2247*, *ASV1922*, *ASV1174*, *ASV171*), are all in lower proportion in the samples ≤0.75, compared with samples ≥0.75. It is also notable to mention taxon denoted with (*), as this demonstrates that for the three samples ≤0.75, the taxon has a greater difference in MRA than in samples predicted above the threshold ≥0.75.

To take a further look into predicted probabilities as a valuable metric, more in-depth profiles relating to the same model can be displayed to identify additional trends. **Fig. 4.5A** shows clear trends within each feature (ASV) along the class membership and corresponding predicted probability assigned to a sample (based on quantiles). Viewing trends in 'core' microbiota from a log10(x+1) abundance, *ASV96* seems to be more abundant in female samples along quantile 0-0.25. When viewed from **Fig. 4.5B**, the MRA seems to be relatively consistent across each quantile, other than the last quantile (>0.75), which shows that *ASV96* is less representative in male samples. Similar trends can be observed for *ASV702* and *ASV1174*. Some ASVs (notably *ASV171*) have a low overall abundance and thus a low MRA in each predicted probability quantile. This does not indicate that these taxa are unimportant drivers of prediction – but that their membership in the feature space holds value in the aggregate. Downstream, this could mean that an

interaction term containing ASV171 and another taxon, or multiple other taxa, could contribute to a high *feature impact* (**Fig. 4.3C** (*rule 3*)).

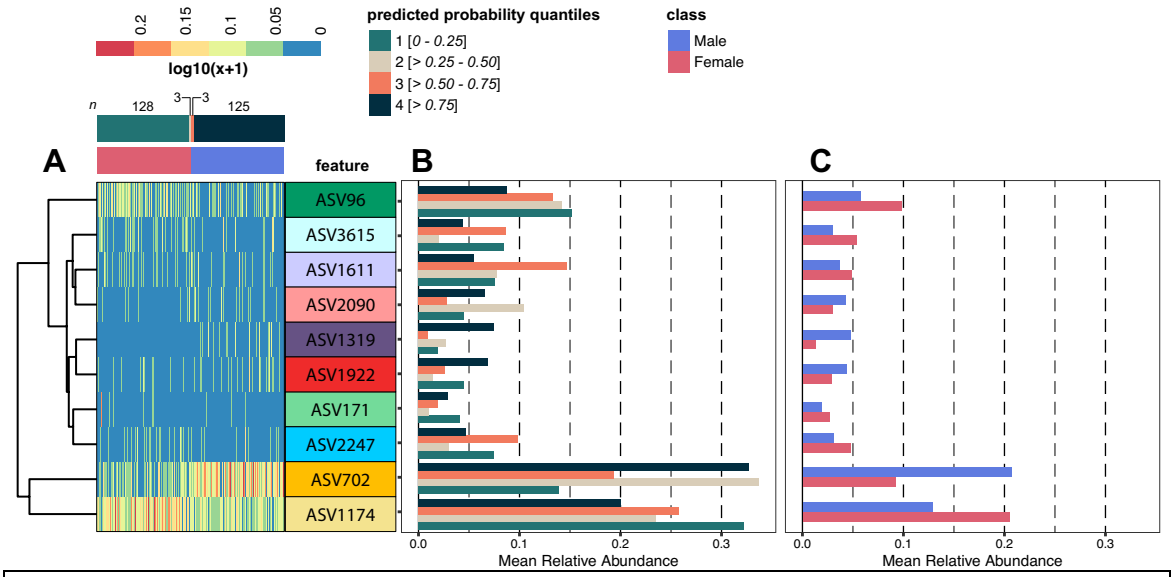


Figure 4. 5

Predicted probability associated community differences. (**A**) A heatmap showing microbial community abundance (log(10x+1)) profiles for each feature, distributed across each class (male/female) and across predicted probability quantiles (quarters: 1-4). (**B**) Displayed is the mean relative abundance (MRA) of each 'core' ASV in each predicted probability quantile. This represents how much of a single ASV is contained within each predicted probability from *1;0-0.25, 2;>0.25-0.50, 3;>0.5-0.75, 4;>0.75-1.00*. (**C**) Displayed is the mean relative abundance (MRA) of each 'core' ASV in each class membership (male/female).

Together, these approaches to interpretation outlines differences in microbial community composition between each class membership and between each taxon being used as a predictive feature. Not only could features be measured in terms of constituent impact, but in interaction with multiple other features and compared to the community as a whole. From this, one can apply domain knowledge to optimize a more generalizable or accurate model for targeted research questions or integrate such analyses as common interpretable report metrics for microbiome-based machine learning tasks – either through experimental literature, or for molecular workflows.

4.6 Summary and outlook

This work demonstrates a sound strategy for the interpretation of machine learning models built on microbial datasets and from the ecological context of microbial interactions. This framework provides methods that support commonly reported performance metrics to further validate research findings. From here, a variety of actionable insights can be extracted and linked back to the raw data used in model training and testing.

Although demonstrated here are interaction terms and corresponding metrics in the context of microbial community members, it is notable to mention that these methods are domain agnostic, and most data points being used in modeling procedures from other domains are satisfied by the limits of this approach.

While this was a proven and reproducible framework, work still needs to be done to wrap these functions into a user-end-ready package accessible to microbiome researchers (for both machine learning integration and for visual output). We plan to fine tune and publish these methods after ensuring that they are robust to variety of potential users.

We would also like to add additional features to this package, such as guiding researchers on labeling schemes for model training. For example, in many cases, determining how and which samples are labeled can carry some subjectivity(51). In the case of determining host subject, the samples were collected from either host, making it easy to label samples. However, in the use-case of the work presented in *Chapter 3* regarding object provenance using geospatially collected samples, it is up to the researcher to define the boundaries of the sample distribution. More specifically, in a model trained to predict if a vessel originated in Baltimore, the researchers must decide which samples should be labeled and learned as Baltimore, and which samples are anything but Baltimore, or Norfolk signatures.

Since we now have a sound strategy for interpreting machine learning models, we plan to take the work presented in this section and apply it to the target problem of object provenance discussed in *Chapter 3* and throughout this dissertation. This will allow us to probe deeply into machine learning models built on real-world microbial communities in the Chesapeake Bay and extract enough actionable insight to construct a robust molecular detection assay.

Preliminary work suggests that it is possible. In this test case, we made the following assumptions: (1) when a vessel travels from Baltimore to Norfolk, the microbial signature 'signal' from home port is lost around half-way (~150km), and (2) samples for each class membership were determined with considerations of maintaining class balance and to account for parameterization (not having to

over/under sample) so that machine learning model fidelity, generality and inherent interpretability is not reduced (**Fig 4.6**). Although only these assumptions are presented here, there are plenty of assumptions one can make with this dataset, and we plan to try many more. The work to establish the best approach for modeling object provenance is ongoing and as we move this project forward, we plan to further validate this model for reproducibility.

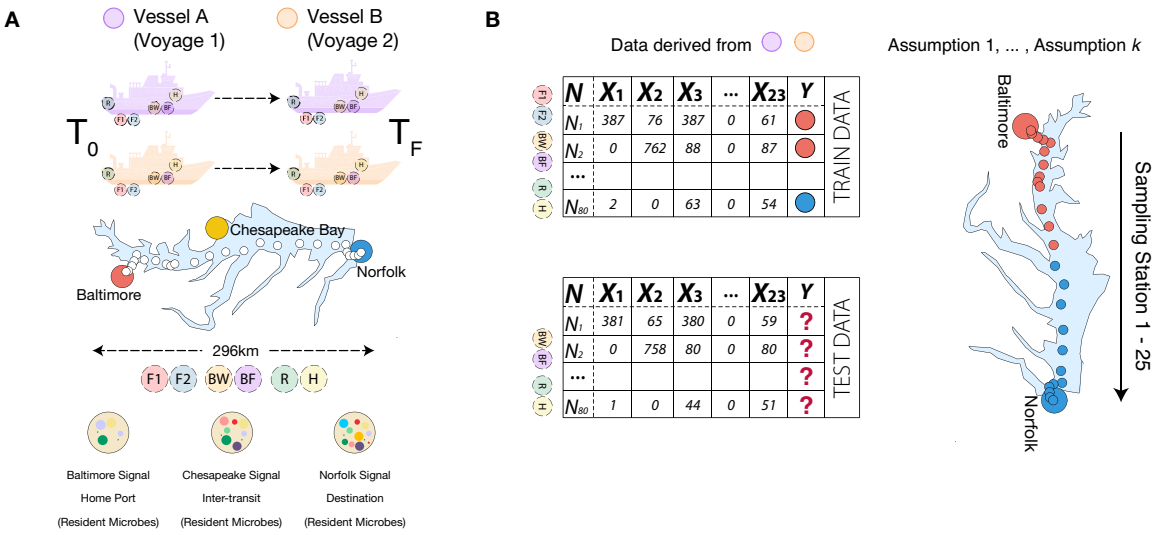


Figure 4. 6

Schematic of framework to investigate target problem (vessel provenance). To put this future modeling procedure into perspective: (**A**) Shows the system wide sampling regime along the Chesapeake Bay (as thoroughly outlined in *Chapter 3*). (**B**) Illustrates the proposed modeling procedure based on assumption (i.e., which resident signatures of each locations do we want the model to learn) and shows the data that will be used to train and validate the model. Note that assumptions are necessary for better model generality, such as using domain-knowledge to assess how to determine which samples are labeled as each class membership (Baltimore (red)/Norfolk(blue)).

4.7 Methods

4.7.1 Dataset

The workflow developed in this study is summarized using the 'moving pictures of the human microbiome' dataset(33) from two healthy host subjects, one male and one female sampled daily for 15 months and 6 months, respectively – at three body sites (gut (stool), sebum (left and right palms) and tongue (saliva)). Only stool samples were used to demonstrate this framework. This dataset is suited for this analysis as it is a dense microbiome study and has sufficient sample sizes to

model microbe-host interactions from multiple body sites and has been benchmarked in multiple other studies.

4.7.2 Interpretation metrics

Section 4.4.2 provides comprehensive methodology of the algorithms and re-implementations of the original work of Houtao Deng (29) and Sejong Oh(30), to bring interpretation to microbiome-based machine learning. The code and data used that support the findings of this study are available from the corresponding author upon reasonable request.

4.7.3 Computational analysis and visualization

All statistical analysis, machine learning models, and visualization were conducted on a local server (Red Hat Enterprise Linux server 7.3 [Maipo]; 256 Gb of random-access memory [RAM]) and on R environment version 3.6.3 using custom functions and the following packages and associated dependencies as follows (in no particular order): phyloseq, tidyverse, microbiome, eulerr, microbiomeutilities, pheatmap, Biostrings, MicrobeDS, dplyr, vegan, recipes, ggpubr, missForest, Hmisc, mi, grid, gridExtra, lattice, DMwR, purr, pROC, PRROC, reshape2, randomForest, inTrees, caret, caretEnsemble, ggplot2.

4.8 References

1. Zhou Y-H, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in genetics. 2019;10:579.
2. Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J. Embracing environmental genomics and machine learning for routine biomonitoring. Trends in Microbiology. 2019;27(5):387-97.
3. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. Mbio. 2020;11(3).
4. Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal. 2021.
5. Aasmets O, Lüll K, Lang JM, Pan C, Kuusisto J, Fischer K, et al. Machine learning reveals time-varying microbial predictors with complex effects on glucose regulation. bioRxiv. 2020.
6. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, et al. Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes. 2018;9(2):104.
7. Ghannam RB, Schaerer LG, Butler TM, Techtmann SM. Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities. Msphere. 2020;5(1).
8. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;8(8):832.
9. Bathaee Y. The artificial intelligence black box and the failure of intent and causation. Harv JL & Tech. 2017;31:889.

10. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, et al. Loop-mediated isothermal amplification of DNA. Nucleic acids research. 2000;28(12):e63-e.
11. Kellner MJ, Koob JG, Gootenberg JS, Abudayyeh OO, Zhang F. SHERLOCK: nucleic acid detection with CRISPR nucleases. Nature protocols. 2019;14(10):2986-3012.
12. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16); 2016.
13. Fenchel T. Microbial behavior in a heterogeneous world. Science. 2002;296(5570):1068-71.
14. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. SIAMCAT: user-friendly and versatile machine learning workflows for statistically rigorous microbiome analyses. bioRxiv. 2020.
15. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Scientific reports. 2020;10(1):1-9.
16. Shamsaddini A, Dadkhah K, Gillevet PM. BiomMiner: An advanced exploratory microbiome analysis and visualization pipeline. PloS one. 2020;15(6):e0234860.
17. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608. 2017.
18. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
19. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. 2019;116(44):22071-80.
20. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-77.
21. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nature Reviews Nephrology. 2020;16(7):408-21.
22. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion. 2019;50:71-91.
23. Netzer M, Hackl WO, Schaller M, Alber L, Marksteiner J, Ammenwerth E, editors. Evaluating Performance and Interpretability of Machine Learning Methods for Predicting Delirium in Gerontopsychiatric Patients. dHealth 2020–Biomedical Informatics for Health and Care: Proceedings of the 14th Health Informatics Meets Digital Health Conference; 2020: IOS Press.
24. Fellous J-M, Sapiro G, Rossi A, Mayberg H, Ferrante M. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. Frontiers in neuroscience. 2019;13:1346.
25. Aasmets O, Lüll K, Lang JM, Pan C, Kuusisto J, Fischer K, et al. Machine learning reveals time-varying microbial predictors with complex effects on glucose regulation. Msystems. 2021;6(1).
26. Bogart E, Creswell R, Gerber GK. MITRE: inferring features from microbiota time-series data linked to host status. Genome biology. 2019;20(1):1-15.
27. Richardson M, Gottel N, Gilbert JA, Lax S. Microbial similarity between students in a common dormitory environment reveals the forensic potential of individual microbial signatures. MBio. 2019;10(4).
28. Mascarenhas R, Ruziska FM, Moreira EF, Campos AB, Loiola M, Reis K, et al. Integrating computational methods to investigate the macroecology of microbiomes. Frontiers in genetics. 2020;10:1344.
29. Deng H. Interpreting tree ensembles with intrees. International Journal of Data Science and Analytics. 2019;7(4):277-87.

30. Oh S. Feature Interaction in Terms of Prediction Performance. Applied Sciences. 2019;9(23):5191.
31. Burns MB, Blekhman R. Integrating tumor genomics into studies of the microbiome in colorectal cancer. Gut microbes. 2019;10(4):547-52.
32. Clarke TH, Gomez A, Singh H, Nelson KE, Brinkac LM. Integrating the microbiome as a resource in the forensics toolkit. Forensic Science International: Genetics. 2017;30:141-7.
33. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome biology. 2011;12(5):1-8.
34. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:160605386. 2016.
35. Molnar C, Casalicchio G, Bischl B. iml: An R package for interpretable machine learning. Journal of Open Source Software. 2018;3(26):786.
36. Hosoda S, Nishijima S, Fukunaga T, Hattori M, Hamada M. Revealing the microbial assemblage structure in the human gut microbiome using latent Dirichlet allocation. Microbiome. 2020;8(1):1-12.
37. Hill M, Bunce R, Shaw M. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. The Journal of Ecology. 1975:597-613.
38. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH human microbiome project. Genome research. 2009;19(12):2317-23.
39. Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, et al. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. The FASEB Journal. 2013;27(3):1012-22.
40. Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. PloS one. 2012;7(6):e34242.
41. Dietterich T. Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR). 1995;27(3):326-7.
42. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics. 2007;8(1):1-21.
43. Debeer D, Strobl C. Conditional permutation importance revisited. BMC bioinformatics. 2020;21(1):1-30.
44. Hooker G, editor Discovering additive structure in black box functions. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining; 2004.
45. Borthagaray AI, Soutullo A, Carranza A, Arim M. A modularity-based approach for identifying biodiversity management units. Revista chilena de historia natural. 2018;91(1):1-10.
46. Yang L, Cao X, He D, Wang C, Wang X, Zhang W, editors. Modularity Based Community Detection with Deep Learning. IJCAI; 2016.
47. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Annals of Applied Statistics. 2008;2(3):916-54.
48. Agrawal R, Imieliński T, Swami A, editors. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data; 1993.
49. Kuhn M. Caret: classification and regression training. Astrophysics Source Code Library. 2015:ascl: 1505.003.
50. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC bioinformatics. 2010;11(1):1-13.
51. Culotta A, McCallum A, editors. Reducing labeling effort for structured prediction tasks. AAAI; 2005.

5 Conclusions

This dissertation summarizes several approaches to understand microbial communities in complex systems, such as marine and freshwater settings. The objectives in this body of work were aimed overall at investigating if natural microbial communities that are everywhere in nature could be used to classify a location, and the provenance of an object in the environment.

To summarize the overall contributions discussed throughout this dissertation:

5.1 Chapter 1

This section aimed to review machine learning in the context of microbes as they relate to our health and built environments. Provided was an in-depth overview of microbiome studies employing a variety of machine learning algorithms – from microbial ecology to the human microbiome and environmental monitoring. We then proceeded to compare the machine learning algorithms (supervised and unsupervised) used in these studies, with brief mention of advantages over traditional multivariate statistics. Additionally, provided was a thorough comparison of open-source toolkits that can be used for predictive and exploratory machine learning modeling of microbial datasets. Our review followed with mentions of shortcomings of common machine learning practice in the experimental microbiome literature, and how machine learning interpretation could be improved. As reproducibility is a concern in studies that employ machine learning, more interpretive open-source software should be acknowledged as an integral part of the modern workflows of investigating microbiota.

5.2 Chapter 2

This study was designed to probe microbial biogeography. We demonstrated an initial experimental field design and machine learning approach to test if natural microbial communities could discriminate geospatial location on a global scale. Here, we chartered research vessels to collect samples from 604 different locations around twenty busy ports across eight countries and three continents. We found that microbes can be used as accurate differentiators of geospatial location. Even at the lowest taxonomic resolution of phylum, our models were accurate in differentiating all twenty geospatial locations (log_{loss}, 0.58; accuracy, 0.84) and when these locations were binned into regions (log_{loss}, 0.33; accuracy, 0.90). These accuracies are well above what would be expected for random classifications taking place in our models (based on model kappa, local, 0.83; region, 0.88). This work demonstrated that microbes can be used as tools for indicating geospatial location in marine and fresh-water systems. This suggested that objects that occupy and pass through these systems may be able to pick up microbes and carry them as they move around – and was the basis for moving forward.

5.3 Chapter 3

Extending the work in *Chapter 2*, here we tested the stability and persistence of microbes that colonize vessels as they travel through the water. To characterize this, we designed a field experiment in the Chesapeake Bay, chartering two research vessels for two independent voyages from the home port of Baltimore to the destination port of Norfolk (296km). Samples were collected along the way from the surrounding open water (the seed) and from the surface of the vessel. These results show similar microbial community composition from *vessel* and *open water* despite sampling on independent *voyages* (e.g., different dates) (**Fig. 3.2**). This led us to probe for taxa that are most robust to system dynamics (abiotic parameters, etc.) and temporality. By viewing the community from the perspective of a shared 'core' set of taxa, we could still thoroughly investigate research questions and better understand the system in the aggregate (**Fig. 3.3**). Additionally, we demonstrated that our nine derived taxa are robust to the complexity of this real-world system. For example, these signatures persist at least across a distance of 296km, and are present on both vessels sampled despite the overall high variability in the microbial community between the two vessels. This data suggested that we could potentially interface machine learning modeling on these robust 'core' biomarkers as features to detect the previous known origin of a vessel.

5.4 Chapter 4

In order to properly validate our machine learning model built to detect object provenance, and whether predictions being made hold true to the biological system – a framework to extract biologically meaningful information from models that are otherwise uninterpretable was developed and employed. Otherwise, our approach would have been purely a black box form of investigation. Existing tools for machine learning interpretation is mostly focused on providing researchers with more accurate models, rather than proving interpretation metrics to help inference the data points (microbial taxa) that are driving accurate models. Here, we demonstrated that while methods for inferencing how single microbial community members influence single predictions are beneficial (local interpretations), appreciating the inner workings of multiple microbial community members and how they generally discern a system state is more robust and generalizable (global interpretation) – and can help extend work. Often a condition in a system (disease state, contamination, etc.) is not attributable to, or cannot be remediated by a single taxon, but multiple taxa. One of the strengths of machine learning is the ability to appreciate these groups of features in making a prediction. However, using traditional interpreting metrics that focus on the importance of a single feature may limit the applicability to the real-world system that is being modeled (i.e., appreciating the full microbial community rather than a single taxon, or subsets of the community). This framework provides novel ways

to extract a reasonable biological understanding of the data being used in the model. For example, likelihood estimates that a sample belongs to a particular class membership which can be observed in the context of the microbial community profile (**Fig 4.4**). This framework is robust to other microbiome-based machine learning tasks (independent of domain) – and can be employed by other researchers to better apprehend and interpret their results. Ultimately, this framework is what is going to allow us better to understand our machine learning models built to detect object provenance – along with helping extend these models to a real-world application.

6 Summary statements

This work provides a case study for the use of natural microbial community data combined with machine learning to address forensic applications. We advanced the field by showing the feasibility of using machine learning to identify previous location from microbial community data. Additionally, we developed new approaches for interrogation of machine learning models constructed from next-generation sequencing (NGS) data, allowing us, and others, to glean biologically meaningful information. This work lays the foundation for more in-depth study of the use of microbial communities in forensic applications.