Supervised Machine Learning Enables Geospatial Microbial Provenance

Chandrima Bhattacharya[1,2,10], Braden T. Tierney[2,3], Krista A. Ryon[2,3], Malay Bhattacharyya[4,5], Jaden J. A. Hastings[2,3], Srijani Basu[6], Bodhisatwa Bhattacharya[7], Debneel Bagchi[8], Somsubhro Mukherjee[9], Lu Wang[9], Elizabeth M. Henaff[10,], Christopher E. Mason[2,3,11,*]*

[1] Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine, New York, NY, USA
[2] The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[3] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[4] Center for Artificial Intelligence and Machine Learning, Indian Statistical Institute, Kolkata, India
[5] Machine Intelligence Unit, Indian Statistical Institute, Kolkata,India
[6] Department of Medicine, Weill Cornell Medicine, New York, NY, USA
[7] Department of Electrical and Electronics Engineering, Birla Institute of Technology, Mesra, Ranchi, India
[8] Department of Metallurgy & Materials Engineering, Indian Institute of Engineering Science & Technology, Shibpur, Howrah, India
[9] Department of Biological Sciences, National University of Singapore, Republic of Singapore
[10] Integrated Design and Media, Center for Urban Science and Progress, NYU Tandon School of Engineering, Brooklyn, NY, USA
[11] WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA

***Corresponding author**
chm2042@med.cornell.edu and ehenaff@nyu.edu

Keywords: *Microbial fingerprint, Microbial forensics, Supervised Machine Learning, Metagenomics, Bioindicator species*

Abstract:

The recent increase in publicly available metagenomic datasets with geospatial metadata has made it possible to determine location-specific, microbial fingerprints from around the world. Such fingerprints can be useful for comparing microbial niches for environmental research, as well as for applications within forensic science and public health. To determine the regional specificity for environmental metagenomes, we examined 4305 shotgun-sequenced samples from the MetaSUB Consortium dataset – the most extensive public collection of urban microbiomes, spanning 60 different cities, 30 countries, and 6 continents. We were able to identify city-specific microbial fingerprints using supervised machine learning (SML) on the taxonomic classifications, and we also compared the performance of ten SML classifiers. We then further evaluated the five algorithms with the highest accuracy, with the city and continental accuracy ranging from 85-89% to 90-94%, respectively. We then used these results to develop Cassandra, a random-forest-based classifier that identifies indicator species to aid in fingerprinting and can infer higher-order microbial interactions at each site. We further tested the Cassandra algorithm on the Tara Oceans dataset, the largest collection of marine-based microbial genomes, where it classified the oceanic sample locations with 83% accuracy. These results and code show the utility of SML methods and Cassandra to identify bioindicator species across both oceanic and urban environments, which can help guide ongoing efforts in biotracing, environmental monitoring, and microbial forensics.

1. Introduction

The decreasing cost of DNA sequencing has made the study of microbial and other species increasingly possible, ranging from the human-built environments of cities to isolated and extreme natural environments. Large-scale genomics projects such as the Earth Microbiome Project (EMP) (Gilbert et al., 2014) and the Human Microbiome Project (HMP) (Turnbaugh et al., 2007) have pioneered the comparative study of microbial species across targeted ecosystems. In addition, the Tara Oceans Study (Pesant et al., 2015), the Metagenomic and Metadesign of Subway and Urban Biomes (MetaSUB) Consortium (Danko et al., 2021), and the Extreme Microbiome Project (XMP) (Tighe et al., 2017, Sierra et al, 2022, Tighe et al., 2015) have all broadly expanded the collection of diverse environmental microbial samples, revealing biodiversity and genetic differences between environments at a global scale.

The advancement of DNA sequencing technologies, along with global efforts generating large-scale molecular and annotated data sets, played an integral role in expanding traditional forensic methods to those that can utilize DNA-based material evidence in criminal investigations (Elhaik et al., 2021; Robinson et al., 2021). Historically, there is evidence of microbes providing ancillary evidence in criminal cases, including bioterrorism, pathogenic outbreak, and transmission, post-mortem analysis, microbial tracing as well as identification of humans through microbial "fingerprints", which are signatures characteristic of a particular environmental location or niche. (Mason-Buck et al., 2020). Previous work has also revealed that using large-scale genomics datasets can provide the statistical confidence needed for forensics (Jäger et al., 2017). Outside of human-associated ecosystems, microbial forensics could feasibly be used -- in the oceans, for example -- for tracking deleterious algal blooms or organisms with unknown migration patterns (Günther et al., 2022). This would be complementary to existing approaches in which oceanographers use discarded host DNA to ascertain the presence of a given species (Székely et al., 2021).

The goal of microbial forensics is to be able to identify the point of origin, provenance, or likely history of any given sample using microbial fingerprints: distinct microbiomes that can identify their place of origin (Cordier et al., 2018). Prior work in MF has revealed that there is a distance decay relationship between soil microbiome and geolocation, which could be exploited for biotracing (Habtom et al., 2019). Work done by Jesmok et al., 2016 and Chase et al., 2016 demonstrated the successful assignment of places of origin using microbial profiles with the help of supervised classification methods, such as k-nearest neighbor (k-NN) and support vector machines (SVM). Microbial fingerprinting methods have also been used to implement microbial surveillance, as demonstrated by Sanachai et al. (2016), who identified the site of origin from soil microbes in the sole of the shoe using 16S rDNA profiles. Additionally, Segata et al., 2011,

developed Linear Discriminant Analysis (LDA) effect size (LEfSe) method and used LDA-based methods for finding biomarkers, including phenotypic indicators from 16S samples. Using ensemble machine learning approaches in application to forensics, Kim et al., 2015 exhibited that gene expression profiles from the bacterium *Escherichia coli* can be used for determining environmental conditions, including abiotic and biotic components, and when additional genome-scale transcriptional information was provided, the classifier outperformed the previous results. Ideally, metagenomic datasets could be used to develop discriminating microbial signatures or fingerprints, but there is a lack of universally accepted tools for exploring microbial fingerprints and identifying bioindicator species from metagenomics data (Robinson et al., 2021, Cordier et al., 2019).

Though it is clear that microbial interactions play a significant role in many environments, methods associated with applications within microbial forensics (MF) remain limited as the field is still in a nascent stage (Mason-Buck et al., 2020, Goodwin et al.,2020). While promising, microbial forensics has historically used reductive or targeted methods for data processing, including Operational Taxonomic Units (OTUs), which focus solely on sequence variation in targeted regions such as *16S* and 18S ribosomal RNA (rRNA) or the internal transcribed spacer (ITS). By leveraging whole-genome sequencing data generated by short-read, long-read, and linked-read methods, a more comprehensive view of samples can be leveraged for biotracing purposes. Even though we can generalize all microbial sequence (16S, long-reads, WGS) data to be a sparse vector quantifying frequencies (relative/absolute) within Euclidean space (Quinn, et al., 2018) or large kmers (Danko et al., 2021), the variation of sequencing technologies makes it computationally challenging to quantify and interpret cross-platform and cross-study data (Li, 2015). Moreover, current forensics tools do not account for differences due to variations in the sampling method (Robinson et al., 2021), extraction methods (Hall, et al., 2020), and possible noise due to amplification, sequencing, or contamination. Also, most microbial forensics studies focus on large-scale environmental profiling of natural sources and only recently have data been related that focuses on the urban microbiome (MetaSUB, 2016, Danko et al., 2021).

Here we present ***Cassandra***, an SML tool to address these unmet needs of metagenomics-based microbial forensics, which can predict bioindicator species from shotgun sequence datasets. To build this tool we first integrated six pre-processing methods to address the common challenges of sparsity and composition affecting data representation of sequence data processed by commonly used metagenomics toolkits in the pipeline. Our open-source tool, ***Cassandra***, is a random-forest-based algorithm that deploys machine learning to identify bioindicator species using quantitative microbiome profiles as features. Using the data available through the MetaSUB Consortium – which includes extensive metadata including geolocation, surface of origin, and temperature – we first explored the presence of microbial fingerprints by testing multiple approaches, including SML methods, traditional clustering, and statistical approaches, using a protocol developed for MetaSUB (Danko et al., 2021, Bhattacharya C., 2020). We were further able to identify multiple

species that could be used as bioindicator species with high confidence using Cassandra (81.3% for cities, 90.3% accuracy for continents using WGS-based metagenome profiles). Comparison of the signatures of different genomes provides a measure of similarity, but is extremely challenging (Campbell et al., 1999), thus to confirm the tool's utilization for predicting bioindicator species, we tested on the orthogonal dataset (Tara Oceans, 86.5% using OTU-based microbial profiles).

2. Methods

2.1. Dataset Details

The MetaSUB Consortium aims "to create geospatial metagenomic and forensic genetic maps" (MetaSUB, 2016). The organization has been collecting annual metagenomic data since 2015, providing the largest to-date, built-environment microbial study. The dataset provides a rich resource of urban microbiomes covering more than 60 cities located across 6 continents, and it also provides longitudinal metagenomics data for 16 cities. Prior MetaSUB work has explored the diversity of microorganisms in a built environment consisting of mass transit systems, including subways, buses, and waterways, across various geographical locations (Danko et al., 2021). The dataset used for this analysis contains 4,305 samples processed by KrakenUniq from 60 unique cities (including negative controls and positive controls), 9 continental divisions, and surface materials (including some controls).

The Tara Oceans Foundation has been exploring the water for scientific knowledge since 2003. We validated our tool with the publicly available Tara Oceans dataset consisting of microbes processed using the 16S framework (Pesant et al., 2015), with the supplementary dataset was downloaded from the publication. We used the metagenomic data [referred to as MetaG], consisting of taxon classified from gene-coding sequences isolated. We trained our model separately on Phylum, Class, Order, Family, Genus, and OTU from the reported taxa classification. We trained on the feature "OS_region," which divided the dataset based on locations, including, the Red Sea, Mediterranean Sea, Indian Ocean, South and North Atlantic Ocean, North and South Pacific Ocean, Arctic Ocean, and the Southern Ocean.

2.2. Data Preprocessing and Adopted Classification Methodologies

Since different metagenomic tools including QIIME (Bolyen et al., 2019), MetaPhlan (Beghini et al., 2021), Kraken (Wood et al., 2019), and others report taxonomic assignments to use different methods, the species data produced have different quantifiable attributes with respect to compositionality, sparsity. Given that microbiome datasets can be of multiple formats, we tested the 6 methods for preprocessing to reduce the biases and quantitative challenges.

- *binary*: Binarizing to 0,1 based on a threshold value (default=0.0001)

- *clr*: Transformation function for compositional data based on Aitchison geometry to the real space
- *multiplicative replacement*: Transformation function for compositional data uses the multiplicative replacement strategy for replacing zeros such that compositions still add up to 1.
- *raw*: No preprocessing
- *standard scalar*: Standardizing the data by removing the mean and scaling to unit variance
- *total-sum*: Converts all the samples to relative abundance

We selected 10 models from 7 different families based on their previous successful use in genomic studies (Goallec et al.,2020, Pasolli et al., 2016) and further developed a Voting Classifier using 3 of the top performing models. Specifically, this included:

(1) Logistic Regression, which classifies estimating the probability of an event occurring based on a priori knowledge,
(2) Linear Discriminant Analysis (LDA), a generalization of Fisher's linear discriminant for separating multiple classes based on linear combination of features,
(3) Ensemble Methods, which are an amalgamation of multiple models to produce an optimal classification (Random Forest Classifier, Extra Tree, AdaBoost),
(4) Tree-based Classifiers including Decision Tree that use branching methods to evaluate the outcomes for classification,
(5) K-nearest neighbor (kNN), a Supervised Learning method which classifies it by a vote of K nearest training objects as determined by some distance metric,
(6) Bayesian Classifiers, including Gaussian NBC which update their posterior probability after ingesting new kinds of data,
(7) Support Vector Machines (SVM), which separates classes by using a maximal margin hyperplane based on nonlinear decision boundary (Support Vector Classifier and Linear Support Vector Classifier (lSCV)).
The Voting Classifier, another ensemble classifier was created using multiple models and predicted on the basis of aggregating the findings of each base estimator, which was created using RF, lSVC, and Logistic Regression. Table 1 shows the parameters tuned for machine learning prediction; the user-defined parameters can be set manually. If it is blank, it refers to the default setting for the ML model.

Machine Learning Methods	**Defined Parameters**
AdaBoost	n_estimators=1000 (User Defined), learning_rate=4
Decision Tree	-
Extra Tree	n_estimators=1000 (User Defined), criterion='entropy'

Gaussian Naive Bayes	-
K-nearest neighbors	n_neighbors=21 (User Defined)
Linear Discriminant Analysis	-
Linear Support Vector Classifier	kernel='linear', probability=True
Logistic Regression	solver='lbfgs', C=1e5, max_iter= 1000000
Voting Classifier	voting="soft"
Random Forest Classifier	n_estimators=1000 (User Defined), criterion="entropy", bootstrap=True, warm_start=True
Support Vector Machine Classifier	gamma='scale', decision_function_shape='ovo', kernel="rbf", probability=True

Table 1: ***Machine Learning parameters used for training.*** *The method name and the code parameters are listed based on the Python3 SciPy package. If left blank, it means default parameters were used.*

A common problem in ML predictions is overfitting the test data. To eliminate overfitting, we added random Gaussian noise with a mean 0 and standard deviation ranging between 10E-10 and 10E+3 which was done using the "--noise" value set to be TRUE. The result includes precision, and recall value along with top 1, 2, 3, 5, and 10 accuracies (the number of times where the correct label is among the top n labels predicted). The accuracy is the accuracy for a subset computer for multilabel classification. The tools also report the confusion matrix for multiclass multilabel classification.

2.3. Addressing overfitting by cross-validation

Beyond the standard methodologies listed in Table 1, it is also necessary to accommodate and correct small human and machine errors, such as sampling variability, DNA heterogeneity, and missing metadata. To address these issues, we test for overfitting of the data by methods such as adding Gaussian Noise, cross-validations (k-fold cross-validation), and cross-parameter analysis (leave-one-group-out). The k-fold cross-validation makes the model more generalizable, by estimating the model performance in new sets of data created by different model splits, thus providing a more accurate estimation of accuracy. Leave-one-group-out uses a third-party provided group to create a split, where the group information encodes arbitrary domain-specific stratifications of the samples. Both the cross-validations report the best accuracy, average accuracy scores, and standard deviation in accuracy. Machine learning approaches be used for spatial interpolation. In Section 3.3, we first use k-fold cross-validation to classify the accuracy of parameters, and then use leave-one-group-out cross-validation, choosing city and continents as the

group, while looking into the prediction accuracy for the parameters. Here we use the cross-validation methods to validate critical parameters that could be affected by interpolation accuracy.

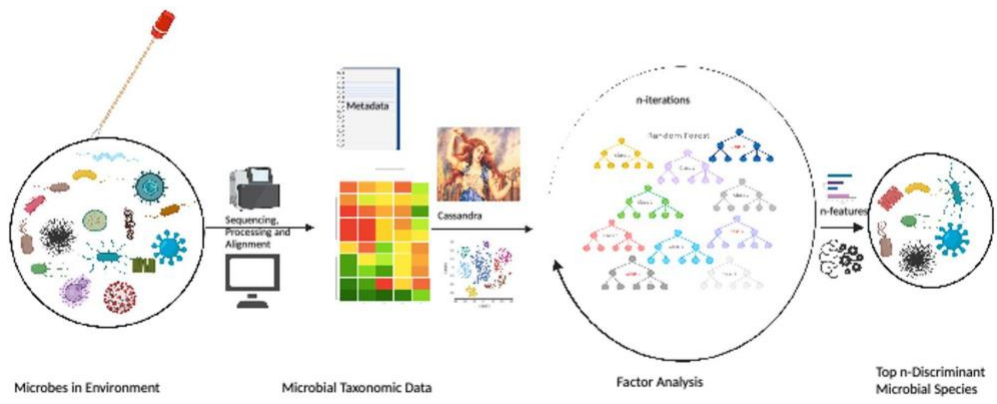
2.4. Cassandra predicts bioindicator species providing explainability and interpretability of datasets

Cassandra was developed as a factor-analysis method to find the bioindicator microbial species. This can be defined as the species chosen by random forest to differentiate between the various sub-environments. Of note, the top five methods compared here (Table 2) displayed high accuracy and low standard deviation, which matches some prior work. Bioinformatic-based benchmark studies by Couronné et al.,2018 have shown Random Forest classifiers generally perform better than Logistic Regression in 69% of the datasets. Other studies designed for metagenomic datasets by Statnikov et al.,2013 supported RF, SVM, kernel ridge regression, and logistic regression are the best methods for microbiome datasets. We chose to use a Random Forest because it yields output that is simple to interpret, stable against Gaussian noise, does not overfit in unbalanced datasets, and the standard deviation between the various preprocessor methods is extremely low.

The input to Cassandra takes the metadata feature of interest, like geolocation, temperature, and the minimum accuracy for prediction expected by the user for the dataset. Using standard machine learning protocols, Cassandra first does an 80-20 train-test split, creates a new tree based on the training set, and trains the model. It evaluates the accuracy of the test set. If the desired accuracy is achieved, Cassandra selects the run and reports the attributed features (microbes) for the run. The algorithm keeps running till we have 1000 instances chosen based on performance with OTU/WGS datasets to achieve consensus (which can be modified by users) where Cassandra achieves the desired accuracy. The weight of all the species is reported for each of those instances, which helps us understand higher-level interaction between species and the geolocation. Also, another output file is generated which selects the top n-species of interest (which can be user modifiers) of interest which are bioindicator species.

If the accuracy cutoff provided is unachievable in 15,000 runs, Cassandra archives arbitrary accuracy ranging from 0 to less than the desired value, based on the dataset. Hence, exploratory analysis for classification was run, based on which we selected the accuracy parameter. We used the entire MetaSUB dataset to predict bioindicator species with >80% accuracy for cities and >90% accuracy for continents and >85% accuracy for predicting OTUs for the Tara Oceans dataset.

A.



B.

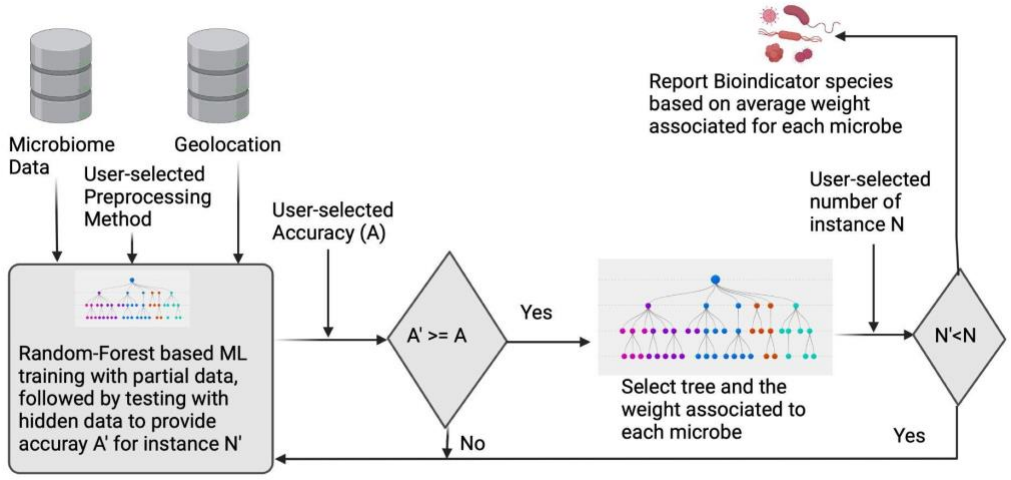


Figure 1: ***Working principle of Cassandra.*** *The Random Forest-based method is designed to select bioindicator species for applications to microbial forensics. (A) Diagrammatic schematic showing a conceptual interpretation of how Cassandra selects top bioindicator species for discriminating location from microbial data and geolocation and (B) Algorithmic Schema that Cassandra uses for reporting species of interest*

3. Results

3.1 Microbial Fingerprints can be observed from the MetaSUB dataset both at the City and Continent levels

For evaluating the potential of geographical profiling to identify the city and continent of origin using the microbiome, we compared the performance of 10 models on the MetaSUB dataset (Figure 2). The best classifiers for city prediction were Logistic Regression, Linear Support Vector Classifier (LSVC), Support Vector Machine (SVM), Extra Tree, and Random Forest (RF), with most showing >80% accuracy (Table 2 and Figure 2A). Using the top 3 models, namely LSVC, Logistic Regression, and RF, we developed an ensemble voting classifier, which was able to reach an accuracy of 89.51% and had a standard deviation of 0.0175. Overall, these data showed that microbial fingerprints are a good indicator for determining geolocation.

However, metagenomics data can have technical noise and other biases introduced by experimental protocols, sequencing methods, or data-cleaning approaches, and this can create challenges for classifier algorithms (McIntyre et al., 2017). We simulated noise in the natural environment by adding Gaussian noise to gauge the impact on the models' performance (Figure 2C). Most models maintained their performance metrics, even with the addition of random Gaussian noise, with a mean change of 0 and standard deviation ranging between 10E-10 to 10E+1), indicating our model does not likely suffer from overfitting, but still demonstrates the importance of preprocessing and model selection (Figure 2D). In general, we observed that all the classifiers performed better than random chance (1/no_of_distinct_classes = 1/60 = 0.0167 or 1.67% accuracy) which was achieved with a training time of fewer than 10 minutes for each instance (Figure 2C). Of note, classifications for continents using shotgun metagenomics data outperformed those for cities with (Figure 2B) Linear Regression, Logistic Regression, and Voting Classifiers, reaching over 90% accuracy.

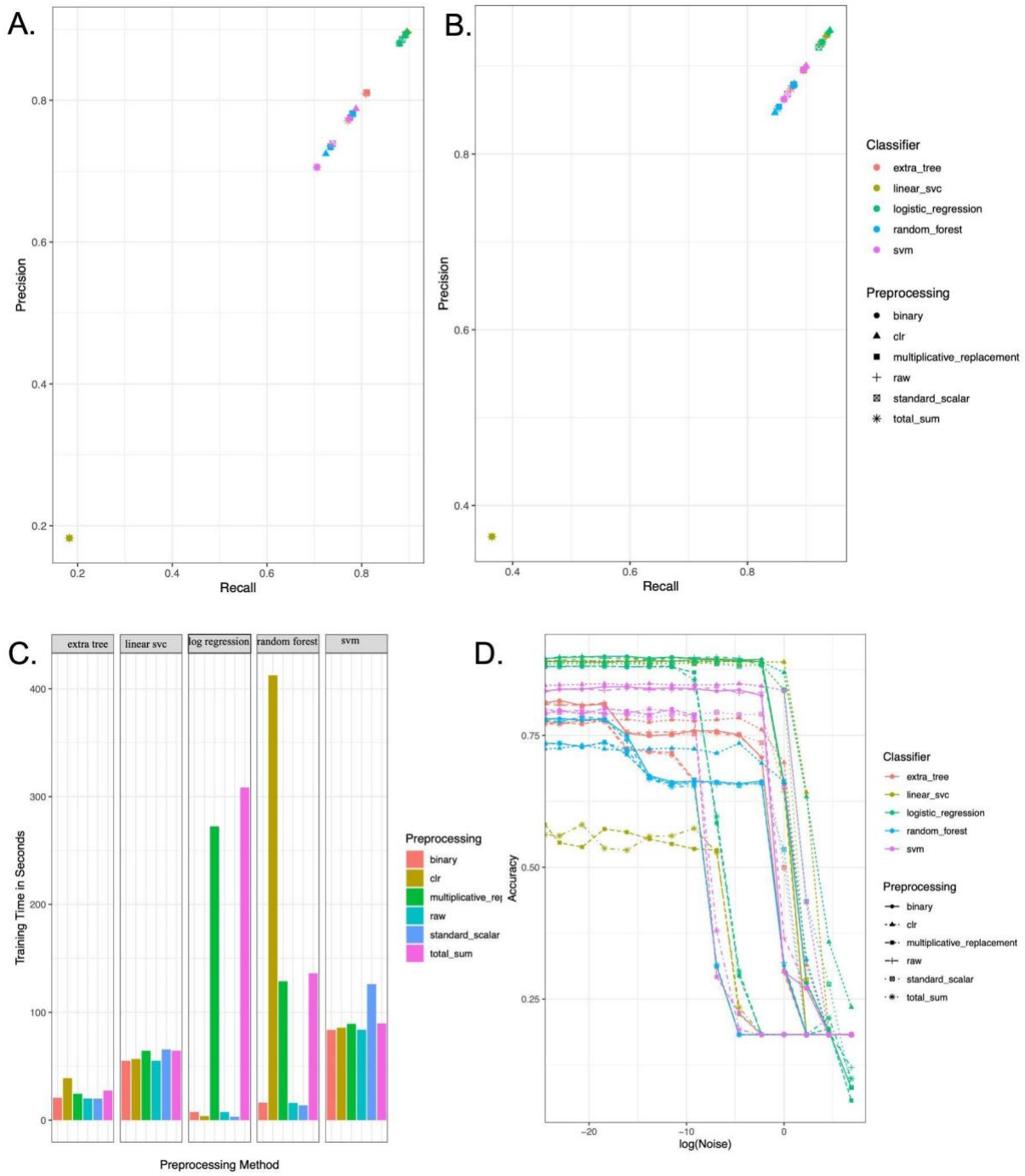


Figure 2: Performance of the ML tools for predicting geolocation from MetaSUB dataset. *Top 5 methods to detect microbial fingerprints of cities with high precision and recall for (A) city. (B) continents. Micro averaging has been used for calculating the precision and recall values to account for class imbalances. (C) Training time required for city classification (D) Gaussian noise (used to mimic metagenomic noises) for the best preprocessing for each model to predict city*

	Preprocessing method						
Classifier	***Binary***	***CLR***	***Multiplicative Inverse***	***Raw***	***Standard Scalar***	***Total-Sum***	***Standard Deviation***
Extra tree	**0.812**	0.775	0.776	0.809	**0.812**	0.771	0.018328
Linear SVC	0.889	**0.893**	0.581	0.890	0.883	0.562	0.14990
Logistic Regression	**0.8952**	0.890	0.879	**0.895**	0.8856	0.881	0.00617
Random Forest	**0.781**	0.734	0.778	0.778	0.779	0.735	0.02126
SVM	0.834	**0.845**	0.790	0.833	0.792	0.798	0.02226

Table 2: Performance of the top 5 models for city classification: *The accuracy and standard deviation between the model accuracy for the 5 five models.*

3.2. Adding interpretability and explainability to machine-learned microbial fingerprints by characterizing bioindicator microbes with Cassandra

Prior work by Danko et al., 2021 showed that RF models can be used for city classification from shotgun sequence data, and as such, we decided to use an RF method to develop Cassandra, a new tool to identify bioindicator species associated with geolocations. Microbial species selected by Cassandra over multiple iterations were considered "bioindicator microbial species" for microbial prioritization, meaning that they were the top taxa unique or enriched to one location. Cassandra requires user input of "accuracy", a parameter that defines the minimum accuracy expected for uniquely identifying cities. As we have observed from previous experiments, RF could achieve an accuracy score of around 80% for classifying cities. Thus, we set Cassandra to achieve 80% accuracy for classifying our desired biomarker species. We also selected *clr* preprocessing for our dataset, as it was most stable against noise, as indicated in Figure 2D. Cassandra took 1,108 runs for a feature selection to achieve 80% accuracy across 1,000 iterations.

We next examined the output of Cassandra after this processing, which details the weight associated with each microbe for classification with minimum accuracy. We found that microbial abundance (the amount of a given microbe) and microbial prevalence (the number of samples where the microbes were found) had a linear relationship (Figure 3A). Also, we found the bioindicator species displayed a strong correlation with microbial abundance and prevalence (Figure 3B, C). The abundance of the top 50 discriminant species, however, showed a wide variance (Figure 3D). Out of the 31 core microbial taxa in the MetaSUB dataset, we observed 4 of

them in the top 50 discriminant species, namely: *Streptococcus mitis*, *Brevundimonas sp. GW460*, *Brevundimonas naejangsanensis*, *Cutibacterium acnes*.

These results thus revealed an interesting relationship between discriminant microbes: the algorithm selected abundant microbes (which have a differential presence across cities), unique microbes, as well as a few core microbes with differential abundances across cities, all as discriminant features. Hence, both the abundance and presence of microbes are utilized by machine learning tools for estimating unique city-based fingerprints. Interestingly, when we compared the top 50 species of microbes selected for predicting continents, we found that around 50% (22/50) of species were shared with the top 50 species selected for predicting cities (Figure 3E), and the feature ranks differed across selection for species at both continent and city levels.

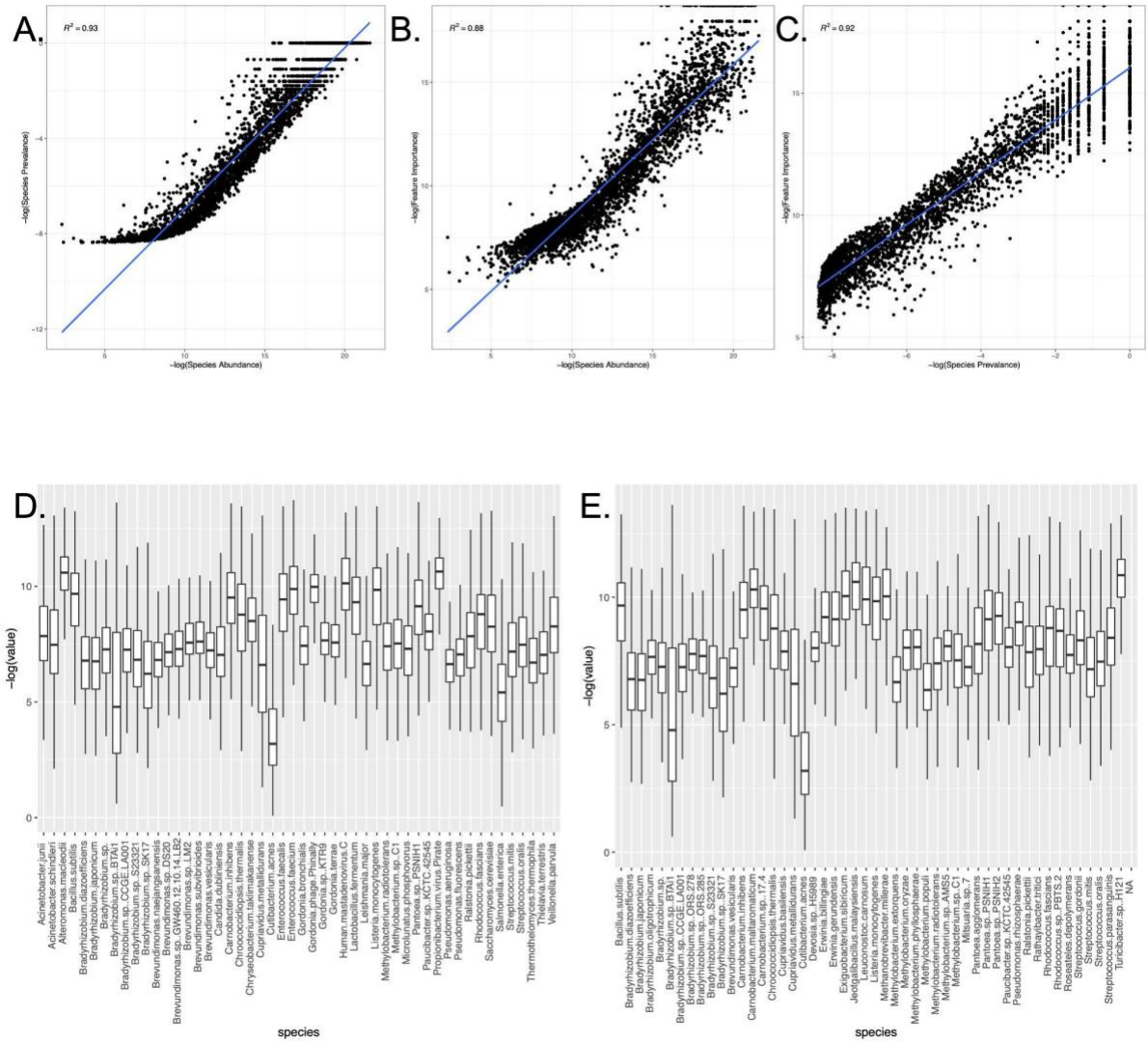


Figure 3: Bioindicator Species and their association with other metrics. *(A-C): The feature importance (weight assigned to bioindicator species) of the microbes as a bioindicator for cities, species prevalence (number of samples the species is present in), and species abundance (relative abundance of species across all samples) shows a linear relationship when plotted against one another (D) Boxplot depicting the abundance of the top 50 discriminant microbial species for cities in the original MetaSUB data (E) Boxplot depicting the abundance of the top 50 discriminant microbial species for the continent in the original MetaSUB data*

3.3 Feature interpolation for Microbial Forensics can be achieved using microbial data

Using the MetaSUB dataset, we next examined the ability to predict other forensic features, including surface material (e.g., metal, glass, plastic), temperature, climate, and other clues of sample provenance. To develop these methods, we trained the MetaSUB species data on eight unique features, including surface type, sampling type, coastal city, climate, and others (Table 3). Each feature was analyzed independently using 3 cross-validation methods (k-fold, leave one group out for the city, and one for the continent), which repeatedly uses part of the samples for learning the model, and the remainder for validating the predictions (Methods 3.3).

Three models were separately trained for each feature and cross-validation methods using the following (classifier + preprocessing) methods: random forest with standard scalar, the voting classifier (mixed model) with clr, and logistic regression with multiplicative replacement. After compiling all outcomes, we observed that surface material always outperformed other metadata types (Table 3). In general, microbial data can also classify other associated features with geolocation, like elevation, coastal association, and climate. To confirm this, we also tested classification with LOGO, where we also saw some features like surface material (coarse) outperforming k-fold validation.

Feature	**#Variables**	**Desc. of feature**	**10-fold av. accuracy (%)**	**Leave one group out (city) av. accuracy (%)**	**Leave one group out (continent) av. accuracy (%)**	**Random Chance (%)**
City climate	13	Categorized climate Types	93.0 ± 0.0168	34.5 ± 0.3537	12.7 ± 0.1338	7.69
Coastal	3	Coastal city and altitude for non-coastal city	90.9 ± 0.0123	59.1 ± 0.3197	40.8 ± 0.1323	33.33

Coastal City	2	Binary values	90.9 ± 0.0108	52.4 ± 0.2409	40.5 ± 0.1009	50
Location Type	68	Location for collection	85.7 ± 0.0216	44.2 ± 0.3358	40.6 ± 0.3221	1.47
Surface	460	Type of surface	47.2 ± 0.0177	8.7 ± 0.1107	3.8 ± 0.0253	.22
Surface Material	115	Standardizedin g Surface parameter	54.3 ± 0.0178	24.4 ± 0.1920	32.5 ± 0.2603	.87
Surface Ontology (Fine)	6	Type of surface: stone, biological, etc	66.2 ± 0.0248	50.3 ± 0.2495	61.8 ± 0.1532	16.67
Surface Ontology (Coarse)	3	Surface permeability/ Control	85.4 ± 0.0127	87.1 ± 0.1722	83.0 ± 0.0980	33.33

Table 3: Cross-validation methods for data interpolation: *For 8 features, we perform cross-validation based on 10-fold and leave-one-group-out split based on cities/continent as groups. The table explains the features and reports average accuracy and standard deviation for the different splits for the best method amongst the above 3 mentioned. In general, we find the accuracy for predicting a feature from an unknown city higher than that from an unknown continent*

3.4. Modeling the microbial fingerprint of the Tara Oceans Dataset

We next tested our ability to detect microbial signatures in non-human-associated ecosystems, specifically, the Tara Oceans dataset. We trained separately for all the separate taxon levels to classify the oceanic location from which the sample is collected. In general, we found that we were able to predict all taxon levels, with a precision and recall of greater than 65% (Figure 4), including Operational Taxonomic Units (OTUs). For class, genus, and OTU, the accuracy was greater than 80%. We then used OTUs to train on Cassandra, which showed >85% accuracy (the top 15 bioindicators' OTUs are depicted in Figure 4B).

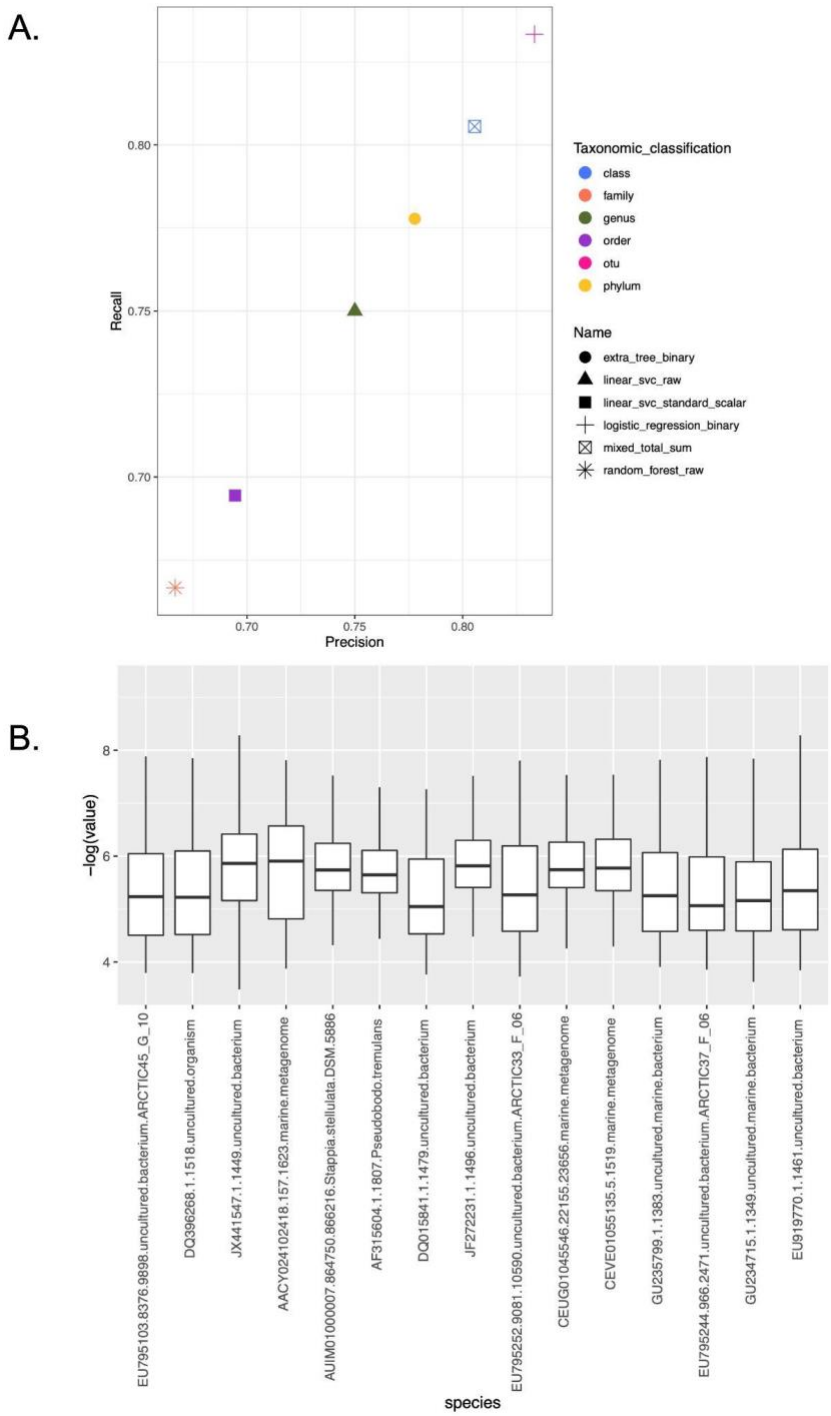


Figure 4: Validation using the Tara Oceans dataset shows the presence of microbial fingerprints in the water sample*: (A) Precision vs Recall for the best model for prediction at each category to predict geolocation based on ocean region. We observe that irrespective of being the same dataset when we try classifying based on different domains, the best model/preprocessing differ, even if they are able to achieve similar accuracy. (B) Top 15 OTUs selected by Cassandra from Tara datasets along with their weight assigned by Cassandra.*

4. Discussion

In this study, we describe the utility of SML tools and further developed Cassandra in identifying and attributing bio-indicator species in oceanic and urban environments that can help guide future microbial forensics efforts. This investigation demonstrates that, even from varied mixtures of microbial communities, we can uniquely predict the provenance of a given metagenomic sample by microbial fingerprint. Deploying the MetaSUB protocol (Danko et al., 2021), we were able to classify the continent and city of origin from a WGS metagenomic sample from MetaSUB, with 94% and 89% accuracy, respectively, and the sea or ocean of origin with 83% accuracy when using OTUs from the Tara Oceans dataset. In general, scientific investigators need to carefully and appropriately address the probative classic evidence while studying the biological agent as evidence. Cassandra can be utilized for metagenomic samples of forensic value, for example, in the comparison of samples from known and questioned origins.

We further cross-validated methods that utilize a microbial fingerprint for interpolating location-based metadata – such as identifying an urban environment or proximity to the coast – noting that the texture of the sampled environmental surface performed similarly at both the continent and city scale from the MetaSUB dataset (Table 3). Well-curated metadata for microbial studies at a global scale make such associations possible, and are required to provide ancillary evidence for the purpose of using these methods in microbial forensics, but could also be used to reduce boundary negotiating artifacts amongst different study cohorts (Beitz et al., 2009). Irrespective of growing evidence suggesting the usage of microbial forensics to complement traditional forensics methods, a more detailed categorization of the environmental association of microbes is required, along with more publicly available tools for analysis and more reproducible quality of research (Metcalf et al., 2017).

The use of Cassandra as a tool for the identification of species can be extended to multiple circumstances including pathogenic propagation as a biodetector (McCarthy et al.,2002), and the spread of antimicrobial resistance (Chng et al., 2020). Looking ahead, future integration using tools like MetaMeta (Piro et al., 2017) can homogenize analysis for cross-study data integration. Species classified by Cassandra can be further used to monitor pathogenic outbreaks as well as the spread of AMR using additional resources. Databases like Global Biodiversity Information Facility (GBIF) (Yesson et al., 2007), TerristrialMetagenomeDB (Corrêa et al, 2020), and Microbe Directory (MD2) (Sierra et al., 2020) can be used as ancillary information for defining phenotypes, and other attributes of the bioindicators species. Thus, the uses of this tool can help improve source tracing and phylogenetic reconstruction to determine disease transmission in geographical locations in situations of bio-crime and biothreat cases.

Using metagenomic samples, this model could provide more power to microbial forensics analyses by helping with the characterization of the origin of a sample based on its microbial signature, such as at the scene of a crime. Although a daunting endeavor, Cassandra could be trained to identify the place of the crime, and the place of the death based on the geographic distribution of the microbial communities. While Cassandra can have myriad applications, proper interpretation necessitates the requirement of proper guidelines, orthogonal validation, and statistical ranking.

A key objective of microbial forensics is to utilize microbial analyses and other evidence to resolve a sample's provenance, and geospatial attribution is often more challenging than simply identification. For example, increasing the resolution of a location estimate gets harder as the desired specificity increases (e.g., city vs. continent), and cities have varying degrees of autochthonous and unique species. Moreover, noise elimination, auto-correlation, and sample mapping are dependent on many factors, and in this paper, we have tested a range of options for data cleaning and their impact on SML tools. Microbial forensics deals with a wide range of microorganisms, including viruses, fungi, parasites, bacteria, and small eukaryotes, and future work can tease out more interactions between species as well.

5. Conclusion

Microbial forensics has a wide range of applications, and work in the field is still in its infancy. Nonetheless, forensics approaches leveraging microbes have been used in tracking crimes (Arenas et al., 2017), the provenance of samples in a city (e.g. MetaSUB), and also for sexually transmitted diseases like HIV and HCV (Network, B). Efforts led by the Human Microbiome Project were integral in understanding the microbial fingerprint across different body sites (Turnbaugh et al., 2007, Lloyd-Price et al., 2017), and this work is now being leveraged as well in microbial forensics, such as the human skin signature in the MetaSUB dataset. New studies conducted by Franzosa et al., 2015 explored the use of the human microbiome as a personal identifier, which also can aid microbial forensics. The authors reported that body site-specific microbial profiles of individuals could be used to match an individual, even several months later. Other applications include exploring the thanatomicrobiome, in determining the time elapsed since microbial death (Javan et al., 2016).

Microbial forensics may have future uses in other areas, such as identification of an individual's past exposures or countries/geographical locations previously visited (Schmidt 2017), which can bring a temporal dimension to the microbial tracking. This information could be used to highlight whether a subject is a person of interest for law enforcement services, as well as a means by which to eliminate suspects, which would make it a powerful tool that could be used as ancillary evidence in criminal cases. However, such use requires caution, as it could contribute to errors, socioeconomic discrimination, or stigmatization, and the results would need to be verified, as with all other methods in forensics. Nonetheless, the tools, data, and methods are finally emerging to make such methods of tracking and provenance a reality.

6. **Data Availability:**

Code for Cassandra is available here:
Cassandra: https://github.com/Chandrima-04/Cassandra

The machine learning methods mentioned in the paper are available as a GitHub package:
Meta_Pred: https://github.com/Chandrima-04/meta_pred

Data availability: The data is based on MetaSUB Consortium dataset and Tara Oceans dataset. All the jobs were run in the Weill Cornell Medicine cluster Curie maintained by the Scientific Computing Unit (https://scu.med.cornell.edu/site). The reported time for running machine learning models is subject to resource availability at a given instance and might require additional time if run locally.

Abbreviations:

DNA: Deoxyribonucleic acid
kNN: k-Nearest Neighbors
LDA: Linear Discriminant Analysis
LOGO: Leave one Group Out
lSVC: Linear Support Vector Classifier
MF: Microbial Forensics
NBC: Naive Bayes Classifier
OTU: Operational Taxonomic Unit
RF: Random Forest (Classifier)
SML: Supervised Machine Learning
SVM: Support Vector Machine
WGS: Whole genome sequencing

Author Contribution:

C.B., C.E.M., E.M.H. designed the ideas behind the computational tools. C.B developed the tool. M.B. helped with tool concept design. C.B. ran all analyses. C.B., B.T.T., K.A.R., J.J.H., S.B., B.B., D.B., S.M., and L.W., M.B, E.E., E.M.H, and C.E.M. edited the manuscript. All authors read the manuscript.

Acknowledgments:

The authors would like to thank Dr. Mikael Bruce Henaff, Dr. Eran Elhaik, and Dr. David C. Danko for their input. The authors would also like to acknowledge Roshni Biswas for her contribution to the GitHub repository. Richa Batra for her suggestions on manuscript editing.

Funding:
E.M.H. was funded by New York University Tandon School of Engineering startup grant for this project. We also thank Testing for America (501c3), OpenCovidScreen Foundation, the Bert L and N Kuggie Vallee Foundation, Igor Tulchinsky and the WorldQuant Foundation, Bill Ackman and Olivia Flatto and the Pershing Square Foundation, Ken Griffin and Citadel, the US National Institutes of Health (R01AI125416, R21AI129851, R01AI151059, U01DA053941), and the Alfred P. Sloan Foundation (G-2015-13964).

Conflict of Interests:
C.E.M. is a co-Founder of Biotia, Inc.

References:

Arenas, M., Pereira, F., Oliveira, M., Pinto, N., Lopes, A.M., Gomes, V., Carracedo, A. and Amorim, A., 2017. Forensic genetics and genomics: Much more than just a human affair. *PLoS genetics*, *13*(9), p.e1006960. DOI: https://doi.org/10.1371/journal.pgen.1006960

Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M. and Valles-Colomer, M., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, *10*, p.e65088. DOI:https://doi.org/10.7554/eLife.65088

Bhattacharya, Chandrima. "Decoding the cryptic metagenome: a deep dive into gene clusters and taxonomy of microbiome." PhD diss., Weill Medical College of Cornell University, 2020

Bietz, M.J. and Lee, C.P., 2009. Collaboration in metagenomics: Sequence databases and the organization of scientific work. In *ECSCW 2009* (pp. 243-262). Springer, London.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F. and Bai, Y., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, *37*(8), pp.852-857. DOI: https://doi.org/10.1038/s41587-019-0209-9

Butler, D., Mozsary, C., Meydan, C., Foox, J., Rosiene, J., Shaiber, A., Danko, D., Afshinnekoo, E., MacKay, M., Sedlazeck, F.J. and Ivanov, N.A., 2021. Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *Nature communications*, *12*(1), pp.1-17. DOI: https://doi.org/10.1038/s41467-021-21361-7

Campbell, Allan, Jan Mrázek, and Samuel Karlin. "Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA." Proceedings of the National Academy of Sciences 96, no. 16 (1999): 9184-9189. DOI: https://doi.org/10.1073/pnas.96.16.9184

Chase, John, Jennifer Fouquier, Mahnaz Zare, Derek L. Sonderegger, Rob Knight, Scott T. Kelley, Jeffrey Siegel, and J. Gregory Caporaso. "Geography and location are the primary drivers of office microbiome composition." *MSystems* 1, no. 2 (2016): e00022-16. DOI: https://doi.org/10.1128/mSystems.00022-16

Chng, K.R., Li, C., Bertrand, D., Ng, A.H.Q., Kwah, J.S., Low, H.M., Tong, C., Natrajan, M., Zhang, M.H., Xu, L. and Ko, K.K.K., 2020. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nature medicine*, *26*(6), pp.941-951. DOI: https://doi.org/10.1038/s41591-020-0894-4

Cordier, T., Forster, D., Dufresne, Y., Martins, C. I., Stoeck, T., & Pawlowski, J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. *Molecular Ecology Resources*, *18*(6), 1381-1391. (2018). DOI: https://doi.org/10.1111/1755-0998.12926

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J., Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in microbiology*, *27*(5), 387-397. (2019). DOI: https://doi.org/10.1016/j.tim.2018.10.012

Corrêa, F.B., Saraiva, J.P., Stadler, P.F. and da Rocha, U.N., 2020. TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic acids research*, *48*(D1), pp.D626-D632. DOI: https://doi.org/10.1093/nar/gkz994

Couronné, R., Probst, P. and Boulesteix, A.L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*(1), pp.1-14. DOI: https://doi.org/10.1186/s12859-018-2264-5

Danko, David, Daniela Bezdan, Evan E. Afshin, Sofia Ahsanuddin, Chandrima Bhattacharya, Daniel J. Butler, Kern Rei Chng et al. "A global metagenomic map of urban microbiomes and antimicrobial resistance." *Cell* 184, no. 13 (2021): 3376-3393. DOI: https://doi.org/10.1016/j.cell.2021.05.002

Denisko, D. and Hoffman, M.M., 2018. Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), pp.1690-1692. DOI: https://doi.org/10.1073/pnas.1800256115

Elhaik, E., Ahsanuddin, S., Robinson, J.M. *et al.* The impact of cross-kingdom molecular forensics on genetic privacy. *Microbiome* **9,** 114 (2021). DOI: https://doi.org/10.1186/s40168-021-01076-z

Franzosa, E.A., Huang, K., Meadow, J.F., Gevers, D., Lemon, K.P., Bohannan, B.J. and Huttenhower, C., 2015. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, *112*(22), pp.E2930-E2938. https://doi.org/10.1073/pnas.142385411

Gilbert, Jack A., Janet K. Jansson, and Rob Knight. "The Earth Microbiome project: successes and aspirations." *BMC biology* 12, no. 1 (2014): 1-4. DOI: https://doi.org/10.1128/mSystems.00217-17

Goodwin, K., Davis, J., Strom, M. and Werner, C., 2020. NOAA'Omics Strategy: Strategic Application of Transformational Tools. DOI: https://doi.org/10.25923/1swn-rj62

Günther, Babett, Eve Jourdain, Lindsay Rubincam, Richard Karoliussen, Sam L. Cox, and Sophie Arnaud Haond. "Feces DNA analyses track the rehabilitation of a free-ranging beluga whale." *Scientific Reports* 12, no. 1 (2022): 1-7. DOI: https://doi.org/10.1038/s41598-022-09285-8

Habtom, Habteab, Zohar Pasternak, Ofra Matan, Chen Azulay, Ron Gafny, and Edouard Jurkevitch. "Applying microbial biogeography in soil forensics." *Forensic Science International: Genetics* 38 (2019): 195-203. DOI: https://doi.org/10.1016/j.fsigen.2018.11.010

Hall, C.L., Zascavage, R.R., Sedlazeck, F.J. and Planz, J.V., 2020. Potential applications of nanopore sequencing for forensic analysis. *Forensic science review*, *32*(1), pp.23-54.

Ionas Erb, Mark F Richardson, Tamsyn M Crowley, Understanding sequencing data as compositions: an outlook and review, *Bioinformatics*, Volume 34, Issue 16, 15 August 2018, Pages 2870–2878, https://doi.org/10.1093/bioinformatics/bty175

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., et al. Developmental validation of the MiSeq FGx forensic genomics system for targeted next-generation sequencing in forensic DNA casework and database laboratories. *Forensic Science International: Genetics*, *28*, 52-70. (2017). DOI: https://doi.org/10.1016/j.fsigen.2017.01.011

Javan, G.T., Finley, S.J., Abidin, Z. and Mulle, J.G., 2016. The thanatomicrobiome: a missing piece of the microbial puzzle of death. *Frontiers in microbiology*, *7*, p.225. DOI: https://doi.org/10.3389/fmicb.2016.00225

Jesmok, Ellen M., James M. Hopkins, and David R. Foran. "Next-generation sequencing of the bacterial 16S rRNA gene for forensic soil comparison: a feasibility study." *Journal of forensic sciences* 61, no. 3 (2016): 607-617. https://doi.org/10.1111/1556-4029.13049

Keim, P.S., Budowle, B. and Ravel, J., 2011. Microbial forensic investigation of the anthrax-letter attacks. In *Microbial forensics* (pp. 15-25). Academic Press. DOI: https://doi.org/10.1016/B978-0-12-382006-8.00002-5

Kim, M., Zorraquino, V., & Tagkopoulos, I. Microbial forensics: predicting phenotypic characteristics and environmental conditions from large-scale gene expression profiles. *PLoS computational biology*, *11*(3), e1004127. (2015). DOI: https://doi.org/10.1371/journal.pcbi.1004127

Le Goallec, A., Tierney, B.T., Luber, J.M., Cofer, E.M., Kostic, A.D. and Patel, C.J., 2020. A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLoS computational biology*, *16*(5), p.e1007895. DOI: https://doi.org/10.1371/journal.pcbi.1007895.t001

Li, H., 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics and Its Application, 2, pp.73-94. DOI: https://doi.org/10.1146/annurev-statistics-010814-020351

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G. and McDonald, D., 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, *550*(7674), pp.61-66. DOI: https://doi.org/10.1038/nature23889

Mason-Buck, G.; Graf, A.; Elhaik, E.; Robinson, J.; Pospiech, E.; Oliveira, M.; Moser, J.; Lee, P.K.H.; Githae, D.; Ballard, D.; Bromberg, Y.; Casimiro-Soriguer, C.S.; Dhungel, E.; Ahn, T.; Kawulok, J.; Loucera, C.; Ryan, F.; Walker, A.R.; Zhu, C.; Mason, C.E.; Amorim, A.; Syndercombe Court, D.; Branicki, W.; Labaj, P. DNA Based Methods in Intelligence - Moving Towards Metagenomics. Preprints 2020, 2020020158

McCarty, L.S., Power, M. and Munkittrick, K.R., 2002. Bioindicators versus biomarkers in ecological risk assessment. *Human and Ecological risk assessment*, *8*(1), pp.159-164. DOI: https://doi.org/10.1080/20028091056791

McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason CE. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology. 2017 Sep 21;18(1):182.

Metcalf, J.L., Xu, Z.Z., Bouslimani, A., Dorrestein, P., Carter, D.O. and Knight, R., 2017. Microbiome tools for forensic science. *Trends in biotechnology*, *35*(9), pp.814-823. DOI: https://doi.org/10.1016/j.tibtech.2017.03.006

Network, B., Toward a National Biomonitoring System.

Pasolli, E., Truong, D.T., Malik, F., Waldron, L. and Segata, N., 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS computational biology*, *12*(7), p.e1004977. DOI: https://doi.org/10.1371/journal.pcbi.1004977

Pesant, Stéphane, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone et al. "Open science resources for the discovery and analysis of Tara Oceans data." *Scientific data* 2, no. 1 (2015): 1-16. DOI: https://doi.org/10.1038/sdata.2015.23

Piro, V.C., Matschkowski, M. and Renard, B.Y., 2017. MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, *5*(1), pp.1-11. DOI: https://doi.org/10.1186/s40168-017-0318-y

Robinson JM, Pasternak Z, Mason CE and Elhaik E (2021) Forensic Applications of Microbiomics: A Review. *Front. Microbiol.* 11:608101. DOI: https://doi.org/10.3389/fmicb.2020.608101
Sanachai, A., Katekeaw, S., and Lomthaisong, K. (2016). Forensic soil investigation from the 16S rDNA profiles of soil bacteria obtained by denaturing gradient gel electrophoresis. *Chiang Mai J. Sci.* 43, 748–755.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. "Metagenomic biomarker discovery and explanation." *Genome biology* 12, no. 6 (2011): 1-18. DOI: https://doi.org/10.1186/gb-2011-12-6-r60

Sierra, M.A., Bhattacharya, C., Ryon, K., Meierovich, S., Shaaban, H., Westfall, D., Mohammad, R., Kuchin, K., Afshinnekoo, E., Danko, D.C. and Mason, C.E., 2019. The microbe directory v2. 0: An expanded database of ecological and phenotypical features of microbes. *BioRxiv*. DOI: https://doi.org/10.1101/2019.12.20.860569

Sierra, Maria, Krista Ryon, Braden Tierney, Jonathan Foox, Chandrima Bhattacharya, Evan Afshin, Daniel Butler et al. "Cross-kingdom metagenomic profiling of Lake Hillier reveals pigment-rich polyextremophiles and wide-ranging metabolic adaptations." *bioRxiv* (2022). DOI: https://doi.org/10.1101/2022.02.17.480683

Schmidt, Charles. "Living in a microbial world." *Nature Biotechnology* 35.5 (2017): 401-404. DOI: https://doi.org/10.1038/nbt.3868

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M.J., Aliferis, C.F. and Alekseyenko, A.V., 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, *1*(1), pp.1-12. DOI: https://doi.org/10.1186/2049-2618-1-11

Székely, Dóra, Natasja Lykke Corfixen, Louise Løvschall Mørch, Steen Wilhelm Knudsen, Morgan L. McCarthy, Jonas Teilmann, Mads Peter Heide Jørgensen, and Morten Tange Olsen. "Environmental DNA captures the genetic diversity of bowhead whales (Balaena mysticetus) in West Greenland." *Environmental DNA* 3, no. 1 (2021): 248-260. DOI: https://doi.org/10.1002/edn3.176

Tighe, S., Baldwin, D., Green, S. and Reyero, N., 2015. Next-generation sequencing and the extreme microbiome project (XMP). *Next Generat Sequenc & Applic*, *2*(115), p.2. DOI: https://dx.doi.org/10.4172/jngsa.1000115

Tighe, Scott, Ebrahim Afshinnekoo, Tara M. Rock, Ken McGrath, Noah Alexander, Alexa McIntyre, Sofia Ahsanuddin et al. "Genomic methods and microbiological technologies for profiling novel and extreme environments

for the Extreme Microbiome Project (XMP)." *Journal of biomolecular techniques: JBT* 28, no. 1 (2017): 31. DOI: 10.7171/jbt.17-2801-004

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. and Gordon, J.I., 2007. The human microbiome project. *Nature*, *449*(7164), pp.804-810. DOI: https://doi.org/10.1038/nature06244

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. "The human microbiome project." *Nature* 449, no. 7164 (2007): 804-810. DOI: https://doi.org/10.1038/nature06244

Wood, D.E., Lu, J. and Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*(1), pp.1-13. DOI: https://doi.org/10.1186/s13059-019-1891-0

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. and Culham, A., 2007. How global is the global biodiversity information facility?. *PloS one*, *2*(11), p.e1124. DOI: https://doi.org/10.1371/journal.pone.000112

Zhan, Q., Babler, K.M., Sharkey, M.E., Amirali, A., Beaver, C.C., Boone, M.M., Comerford, S., Cooper, D., Cortizas, E.M., Currall, B.B. and Foox, J., 2022. Relationships between SARS-CoV-2 in wastewater and COVID-19 clinical cases and hospitalizations, with and without normalization against indicators of human waste. ACS ES&T Water. DOI: https://doi.org/10.1021/acsestwater.2c00045

"The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report." *Microbiome* 4, no. 1 (2016): 24. DOI: https://doi.org/10.1186/s40168-016-0168-z