Decision Tree Ensembles Utilizing Multivariate Splits Are Effective at Investigating Beta-Diversity in Medically Relevant 16S Amplicon Sequencing Data

Josip Rudar[1], G. Brian Golding[2], Stefan C. Kremer[3], Mehrdad Hajibabaei[1]

[1] Department of Integrative Biology & Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, ON, N1G 2W1, Canada

[2] Department of Biology, McMaster University, 1280 Main St. West, Hamilton, ON, L8S 4K1, Canada

[3] School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, ON, NIG 2W1, Canada

Corresponding author info:

Josip Rudar – joe.rudar@gmail.com

Mehrdad Hajibabaei – mhajibab@uoguelph.ca

Abstract

Canonical distance and dissimilarity measures can fail to capture important relationships in high-throughput sequencing datasets since these measurements are unable to represent feature interactions. By learning a dissimilarity using decision tree ensembles, we can avoid this important pitfall. We used 16S rRNA data from the lumen and mucosa of the distal and proximal human colon and the stool of patients suffering from immune-mediated inflammatory diseases and compared how well the Jaccard and Aitchison metrics preserve the pairwise relationships between samples to dissimilarities learned using Random Forests, Extremely Randomized Trees, and LANDMark. We found that dissimilarities learned by unsupervised LANDMark models were better at capturing differences between communities in each set dataset. For example, differences in the microbial communities of colon's distal lumen and mucosa were better reflected using LANDMark dissimilarity ($p \leq 0.001$, $R^2 = 0.476$) than using the Jaccard distance ($p \leq 0.001$, $R^2 = 0.313$) or Random Forest dissimilarity ($p \leq 0.001$, $R^2 = 0.237$). In addition, applying Uniform Manifold Approximation and Projection to dissimilarity matrices and transforming the result using principal components analysis created two-dimensional projections that captured the main axes of variation while also preserving the pairwise distances between samples (eg: $\rho = 0.8804$, $p \leq 0.001$ for the distal colon dissimilarities). Finally, supervised LANDMark models tend to outperform both Random Forest and Extremely Randomized Tree classifiers. Models employing multivariate splits can improve the analysis of complex high-throughput sequencing datasets. The improvements observed in this work likely result from the ability of these models to reduce noise from uninformative features. In an unsupervised setting, LANDMark models can preserve pairwise relationships between samples. When used in a supervised manner, these methods tend to learn a decision boundary that is more reflective of the biological variation within the dataset.

Author Summary

Distance and dissimilarity measures are often used to investigate the structure of biological communities. However, our investigation into two commonly used distance measures, the Jaccard and Aitchison distances, demonstrates that these measures can fail to capture important relationships in microbiome communities. This is likely due to their inability to identify dependencies between features. For example, both the Jaccard and Aitchison metrics are unable to identify subsets of samples where the presence of one feature depends on another. Previous research has found that Random Forest embeddings can be used to create an alternative dissimilarity measure for dimensionality reduction in genomic datasets. We show that dissimilarities learned by decision tree ensembles, especially those using base-estimators capable of partitioning data using oblique and non-linear cuts, can be superior since these approaches naturally model these interactions.

Keywords

Metric learning, amplicon sequencing, 16S rRNA, metabarcoding, ordination, biomarker discovery, machine learning

Introduction

Biomarkers are objectively measurable characteristics of biological systems which can identify and provide evidence in favor or against a biological process or condition (1,2). For example, organisms that are present or absent in patients suffering from a disease, such as Crohn’s Disease, could be considered a biomarker if they can be used to predict the condition

(3). Machine learning (ML) algorithms are being increasingly applied to a wide array of genomic, metagenomic, and transcriptomic data sets to identify relevant biomarkers and create predictive models of these datasets. When analyzing amplicon sequencing data one typical goal is to discover amplicon sequence variants (ASVs) associated with each of the biological communities being studied. For example, a recent study identified how impaired dopamine signaling in mice with a defective dopamine transporter gene alters the activity of metabolic pathways and the composition of the gut microbiome (4). Unlike approaches such as DESeq2 and MetagenomeSeq, ML models tend not to assume anything about the underlying distribution of each co-variate (5,6). Furthermore, many ML models, such as neural networks and Random Forests, can naturally model interactions between covariates (7,8). For these reasons, ML represents a potentially powerful way to identify biomarkers in high-throughput sequencing (HTS) data. Out of the myriad of available machine learning methods, Random Forests (RFs) and other decision tree ensembles have become very popular due to their good overall performance when working with high-throughput sequencing data. Furthermore, extensive tools and approaches have been designed which are starting to peel back the “black-box” veneer of these and other machine learning models (9). For example, RFs have been recently applied to study and identify operational taxonomic units (OTU), which can be considered a class of biomarkers, from the microbiomes of patients suffering from cardiovascular disease, chronic obstructive pulmonary disease, and various immune-mediated inflammatory diseases (3,10,11). These models, which are not linearly constrained, have been shown to generalize well to unseen data in more recent amplicon sequencing studies (12).

While machine learning has become incredibly popular and has led to important discoveries, biomarker selection using RFs and other commonly used approaches can be problematic due to the various algorithmic assumptions. For example, each decision tree in a RF uses a recursive series of axis-orthogonal splits to approximate the underlying data generating function (13,14). However, more complex oblique or non-linear splits often result in more appropriate representations of the data generating function (12,14). Another classification algorithm, k-nearest neighbors, is sensitive to the number of neighbors and the distance metric (15). Logistic regression, ridge regression, and linear support vector classifiers can only learn linear decision boundaries (12). while neural networks can require a large amount of data and time to learn appropriate weights for each parameter.

One aspect of RFs which have not been extensively explored is their ability to learn a dissimilarity measure when working in an unsupervised setting. Unsupervised RFs have previously been used to discover similar cell populations in single-cell RNAseq data, identify different classes of renal cell carcinomas tumors, study the underlying structure of a population using shared genetic variations (16–18). This body of work has demonstrated that unsupervised RFs can identify important sources of variation between samples while still being robust to noise and problems stemming from the high dimensionality of high-throughput sequencing datasets. While these results lay the groundwork and demonstrate the utility of unsupervised RFs, they do not investigate the potential of multivariate decision trees in learning a similarity function. In this study, we investigate multivariate decision trees. Specifically, we will investigate their ability to learn a similarity measure and how this similarity measure compares with distance measures. Finally, we will examine how successful multivariate trees are at classifying and identifying biomarkers in two medically important human microbiome datasets.

Methods

Dataset Selection

Two human microbiome datasets were selected for inclusion in this study. The first was derived from the colons of healthy individuals (19) using 16S rRNA amplicon sequencing. This dataset collected samples from the unprepared colons of healthy individuals and was chosen since we could divide the dataset into four sets of comparisons (19). These comparisons examined differences in the abundance of OTUs between the microbial communities of the proximal lumen (RS) and mucosa (RB), the distal lumen (LS) and mucosa (LB), between the RS and the LS, and finally between the RB and the LB. The second dataset was chosen since it contains samples from patients who suffer from immune-mediated inflammatory diseases (IMID) (3). Differences between the microbiomes of patients suffering from Chron's disease (CD), ulcerative colitis (UC), multiple sclerosis (MS), and rheumatoid arthritis (RA) were compared to healthy controls. Specifically, the work by Forbes et al. (2018) investigated if disease-specific taxonomic biomarkers, OTUs, could be identified in each patient's stool. In both studies, the authors used differential abundance testing and Random Forests to identify potential OTU biomarkers (3,13).

Bioinformatic Processing of Raw Reads

Raw sequences from two previously published datasets were obtained from the Sequence Read Archive (PRJNA450340 and PRJNA418115) (3,19). All bioinformatic processing of the raw reads was prepared using the MetaWorks v1.8.0 pipeline (available online at: https://githib.com/terrimporter/MetaWorks) (20). The default settings for merging reads were used except for the parameter controlling the minimum fraction of matching bases, which was increased from 0.90 to 0.95. This was done to remove a larger fraction of potentially erroneous reads. Merged reads were then trimmed using the default settings MetaWorks passes to CutAdapt. Since reads from PRJNA418115 were pre-processed and the primers removed

(Personal Communication with Kaitlin Flynn, Ph.D. (kjflynn06@gmail.com) in January 2019), no reads were discarded during trimming. The remaining quality-controlled sequences were then de-replicated and denoised using VSEARCH 2.15.2 to remove putative chimeric sequences (21). Finally, VSEARCH was used to construct a matrix where each row is a sample and each column an Amplicon Sequence Variant (ASV). Taxonomic assignment was conducted using the RDP Classifier (version 2.13) and the built-in reference set (22).

ASVs which are likely to be contaminants, specifically those likely belonging to chloroplasts and mitochondria, were removed. From the remaining sequences, only those belonging to the domain *Bacteria* and *Archaea* were retained for further analysis. In the IMID dataset, only sequences assigned to *Firmicutes*, *Actinobacteria*, and *Tenericutes* were retained. This was done since the original study found that operational taxonomic units assigned to other bacterial groups were underrepresented (3). Following the initial processing steps, ASVs with a bootstrap support of 0.8 or higher were chosen for further analysis. The cutoff of 0.8 for the V4 rRNA region sequenced in the 16S dataset was chosen because fragments of ~ 200 bp in length are likely to be assigned to the correct genus 95.7% of the time (23). A site by ASV count matrix, where each row is a sample and each column an ASV, was created using this data. The matrix was filtered to retain only ASVs found in three or more samples. This filtration step was taken since reducing the size of the feature space can often lead to a more generalizable model (24–26).

The filtered matrix must be transformed in such a way to minimize the impact of various technical factors, such as differences in library size (27). Our unsupervised and supervised analyses examined two transformations of the filtered matrix. The first transformation we investigated was the presence-absence transformation. This transformation is useful since it reflects if ASVs are present or absent in the sample and the impact of technical errors, such as differences in library size and the uneven amplification of DNA can be minimized. The second transformation, the centered-log-ratio (CLR) transformation, was used since it effectively addresses the fact that amplicon-sequencing data is compositional (24,28). independent. When searching for biomarkers, the transformation which resulted in the best generalization performance was used.

Training of Unsupervised Models

Tree-based models are an effective means of capturing the similarity between samples. The similarity matrix, S, can be constructed by calculating how often samples co-occur in the terminal leaves of each decision tree. This co-occurrence, $S(x_i, x_j)$, is a similarity and can be found using the following equation:

$$Equation\ One: S(x_i, x_j) = \frac{x_i x_j^T}{N}$$

Where x_i and x_j is the vector representation of all terminal node positions of samples x_i and x_j in the forest, and N is the total number of trees in the forest. The similarity matrix, S, is then converted into a dissimilarity matrix, D (Equation Two) (17). This dissimilarity measure, while not a metric such as the Jaccard distance (29), can be used to investigate beta-diversity and can be constructed using either a supervised or an unsupervised approach (17).

$$Equation\ Two: D = \sqrt{1 - S}$$

To use decision tree ensembles in an unsupervised manner a second dataset is created such that the columns (ASVs) are randomly permuted. In the case of the CLR-transformed data, the original counts were permuted before the CLR transformation. The samples in the permuted dataset are assigned a label of “0” while samples in the original dataset are assigned a label of “1”. The classifier s is then tasked to find the difference between the permuted and original data. RF and ET classifiers were used at their default settings, except for the number of trees which was set to 128 (30). LANDMark (Oracle) models were trained using 128 trees and with the number of features set to $\sqrt{n}$, where n is the number of features in the filtered dataset. This was done to generate a more diverse set of trees. To avoid generating proximity matrices that are biased due to a lucky permutation, we created 100 different unsupervised proximity matrices using equation one and combined them using equation two to create a dissimilarity matrix.

Analysis of Beta-Diversity

Dissimilarity and distance matrices were used as input for PerMANOVA and a principal coordinates analysis (PCoA). A Uniform Manifold Approximation and Projection (UMAP) using the dissimilarity and distance matrices was also conducted (31). The UMAP algorithm was chosen since it projects a high-dimensional graph of the input data into a lower-dimensional Euclidean space. This algorithm can create potentially better representations of the sampling space since high-throughput sequencing data can lie on a complex high-dimensional manifold (31). Finally, the pairwise distances between samples in the UMAP embedding were calculated and used by PCoA to embed the UMAP projection into a two-dimensional space (32). Spearman’s rho was used to measure the distortion between the embeddings and the original distances/dissimilarities.

Assessment of Supervised Model Generalization Performance

Following our investigation of beta-diversity using similarity measures derived from unsupervised models, we assessed the generalization and feature selection performance of the

LANDMark (Oracle), ET, and RF classifiers (33,34). Thirty different train-test splits, with the classes in each set being proportional to those in the original, were created for each metabarcoding data set. 50% of the original data was used to construct each training set and the random state used to create each train-test split was set to the iteration number for the split for reproducibility. RF and ET classifiers were used at their default settings, apart from the number of trees which was set to 128 (30). LANDMark (Oracle) models were also trained using 128 trees and, as in the unsupervised learning, the number of features considered at each node was set to $\sqrt{n}$. The remaining 50% of the data were used to calculate the balanced accuracy score using Scikit-Learn (33). This process was repeated for presence-absence and CLR-transformed data. Unless otherwise stated, the analysis of the IMID data was conducted using the first time point. This was done to avoid inflating the balanced accuracy scores since the microbiomes across time were found to be highly similar.

The transform (presence-absence or CLR) resulting in the best generalization performance and was used during feature selection. ASV features were selected using a combination of recursive feature elimination (RFE) followed by RFE with 5-fold stratified cross-validation. RFE was used to find a set of 200 predictive features. This step aimed to remove ASVs with little predictive value. Following this, RFE with 5-fold stratified cross-validation was used to create a more distilled subset of at least 20 predictive ASVs. The step size for each round of feature elimination was set to 5%. Each iteration's test set was used to evaluate the predictive balanced accuracy of the final model. The subsets of ASVs from the best performing iteration were chosen for further analysis and display. Shapley scores, calculated using the 'Explainer' function of the Python 'shap' package was used to identify the ASVs which strongly impacted the prediction of each sample (35). The 'shap' package was also used to generate decision heatmaps which display the impact on prediction for each ASV. When this process was used to analyze IMID data only samples from the first time point were used as input into RFE. However, Shapley scores were calculated twice. The first set of scores was calculated using each iteration's test data. The second set of scores was calculated using the first time point as the background dataset and the second time point as the testing data. A Bayesian analysis, using Nadeau and Bengio's corrected t-test implemented in the Python 'baycomp' package, was used to compare the generalization performance of models before feature selection and after feature selection (36). The region of practical equivalence (ROPE), the probability of two models having equivalent performance, was defined as a difference in score within ±0.025. Although the choice for the size of this region is arbitrary, this size was chosen since it represents the impact of two classification errors. Finally, the structure of the decision space will be investigated to ascertain how well each model learns an appropriate decision boundary (14,37).

Results

The Choice of Transformation and Dissimilarity Measure Can Result in Different Interpretations of Amplicon Sequencing Data

When LANDMark (Oracle), ET, and RF classifiers were trained to differentiate between real and randomized samples, statistically significant differences between sampling locations were detected when using each model's dissimilarity matrix (Table 1). These tests demonstrated that the most suitable choice of transformation depends on the dataset. For example, the main effect (sampling location) clearly explained a greater fraction of the variance when using the presence-absence transformation in each subset of the healthy gut data. For the IMID data, the CLR transformation was the better choice. These tests also demonstrate that unsupervised models, such as LANDMark (Oracle), can capture information that distinguishes samples, especially when trained using appropriately transformed data. To test if the number of features used has an impact on the explanatory ability of the main effect, we created multiple dissimilarity matrices where the number of features considered at each node was N, 2N, 4N, 8N, and 16N. Here N is equal to the square root of the number of ASVs. This investigation revealed that the explanatory ability of the main effect in each dataset appears to be sensitive to the number of features explored at each node (Figure 1). Interestingly, there appears to be an inverse relationship between LANDMark and the RF and ET models. Finally, the amount of explained variance along the first principal coordinate tended to be greater when using LANDMark (Oracle) dissimilarities. The spread of samples along this axis also tended to reflect differences in sampling location/disease phenotype (Figures 2 and 3). These results are particularly surprising since these matrices were created without using any of the metadata.

Table 1: PerMANOVA results for each transform on each subset of the healthy gut and IMID data. PerMANOVA results using the LANDMark dissimilarity measure are highlighted in bold.

			Presence – Absence			***Centered Log Ratio***		
Dataset	***Subset***	***Dissimilarity Measure***	***Pseudo-F***	***p-value***	***R^2***	***Pseudo-F***	***p-value***	***R^2***
Healthy Gut	LB-LS	Distance	4.05	0.001	0.313	2.48	0.001	0.146
		LANDMark	**5.72**	**0.001**	**0.476**	**2.50**	**0.001**	**0.147**
		Extra Trees	3.26	0.001	0.228	2.43	0.001	0.141
		Random Forest	3.35	0.001	0.237	2.41	0.001	0.139
	LB-RB	Distance	2.31	0.001	0.130	2.13	0.001	0.113
		LANDMark	**2.52**	**0.001**	**0.150**	**2.12**	**0.001**	**0.111**
		Extra Trees	2.01	0.003	0.101	1.74	0.007	0.077
		Random Forest	2.03	0.002	0.103	1.60	0.011	0.066
	RB-RS	Distance	1.68	0.005	0.073	0.855	0.785	0.020
		LANDMark	**1.93**	**0.004**	**0.094**	**0.903**	**0.708**	**0.022**
		Extra Trees	1.47	0.013	0.056	1.21	0.089	0.039
		Random Forest	1.51	0.010	0.060	1.25	0.072	0.042
	LS-RS	Distance	0.692	0.968	0.013	0.460	0.999	0.006
		LANDMark	**0.760**	**0.992**	**0.016**	**0.540**	**1.0**	**0.008**
		Extra Trees	0.819	0.946	0.018	0.714	0.994	0.014
		Random Forest	0.801	0.950	0.018	0.836	0.895	0.019
Immune Modulated Inflammatory Disease	CD-HC	Distance	3.23	0.001	0.220	5.61	0.001	0.460
		LANDMark	**4.03**	**0.001**	**0.305**	**6.46**	**0.001**	**0.530**
		Extra Trees	4.49	0.001	0.353	6.38	0.001	0.523
		Random Forest	4.55	0.001	0.359	4.66	0.001	0.370
	MS-HC	Distance	1.42	0.028	0.049	2.12	0.001	0.103
		LANDMark	**1.37**	**0.071**	**0.046**	**1.93**	**0.001**	**0.087**
		Extra Trees	1.46	0.013	0.052	1.68	0.001	0.067
		Random Forest	1.46	0.021	0.052	1.55	0.013	0.058
	RA-HC	Distance	1.69	0.005	0.065	2.89	0.001	0.169
		LANDMark	**1.50**	**0.027**	**0.052**	**2.74**	**0.001**	**0.155**
		Extra Trees	1.49	0.025	0.051	2.30	0.001	0.114
		Random Forest	1.52	0.011	0.054	1.79	0.001	0.073
	UC-HC	Distance	1.47	0.019	0.053	2.15	0.001	0.106
		LANDMark	**1.50**	**0.037**	**0.054**	**1.93**	**0.001**	**0.088**
		Extra Trees	1.46	0.015	0.052	1.91	0.001	0.0086
		Random Forest	1.45	0.02	0.051	1.67	0.003	0.067

Figure 1: Distribution of the PerMANOVA effect sizes (R^2) for each type of dissimilarity matrix. Each learned dissimilarity matrix is constructed using an ensemble of decision trees. The internal nodes of each decision tree examine (or use in the case of LANDMark) a subset of all ASVs while Aitchison Distances were constructed using all ASVs. The minimum number of ASVs considered, N, is the square root of the total number of ASVs. This data was generated using the Crohn's Disease subset of the IMID data.

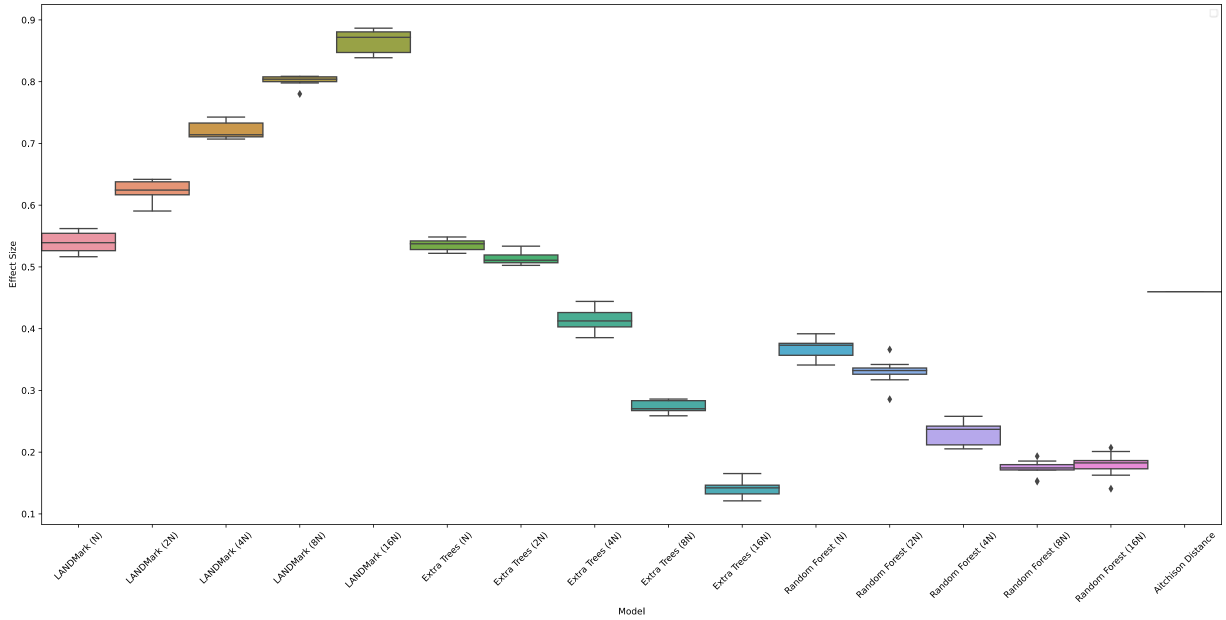


Figure 2: UMAP followed by PCoA and PCoA ordinations of the distal lumen and mucosa dataset. When only using PCoA projections of distance and dissimilarity matrices, each axis explains only a fraction of the total variation in the dataset. However, projections of the UMAP space using PCoA are more informative. In these projections, the first PCoA axis explains the vast majority of the variation in the distance and dissimilarity matrices. Furthermore, in these projections, the variation along the first axis appears to be strongly related to differences in community structure. The coloring of points serves as a visual aid and it does not affect the result. LB are samples taken from the distal mucosa while LS are samples taken from the distal lumen.

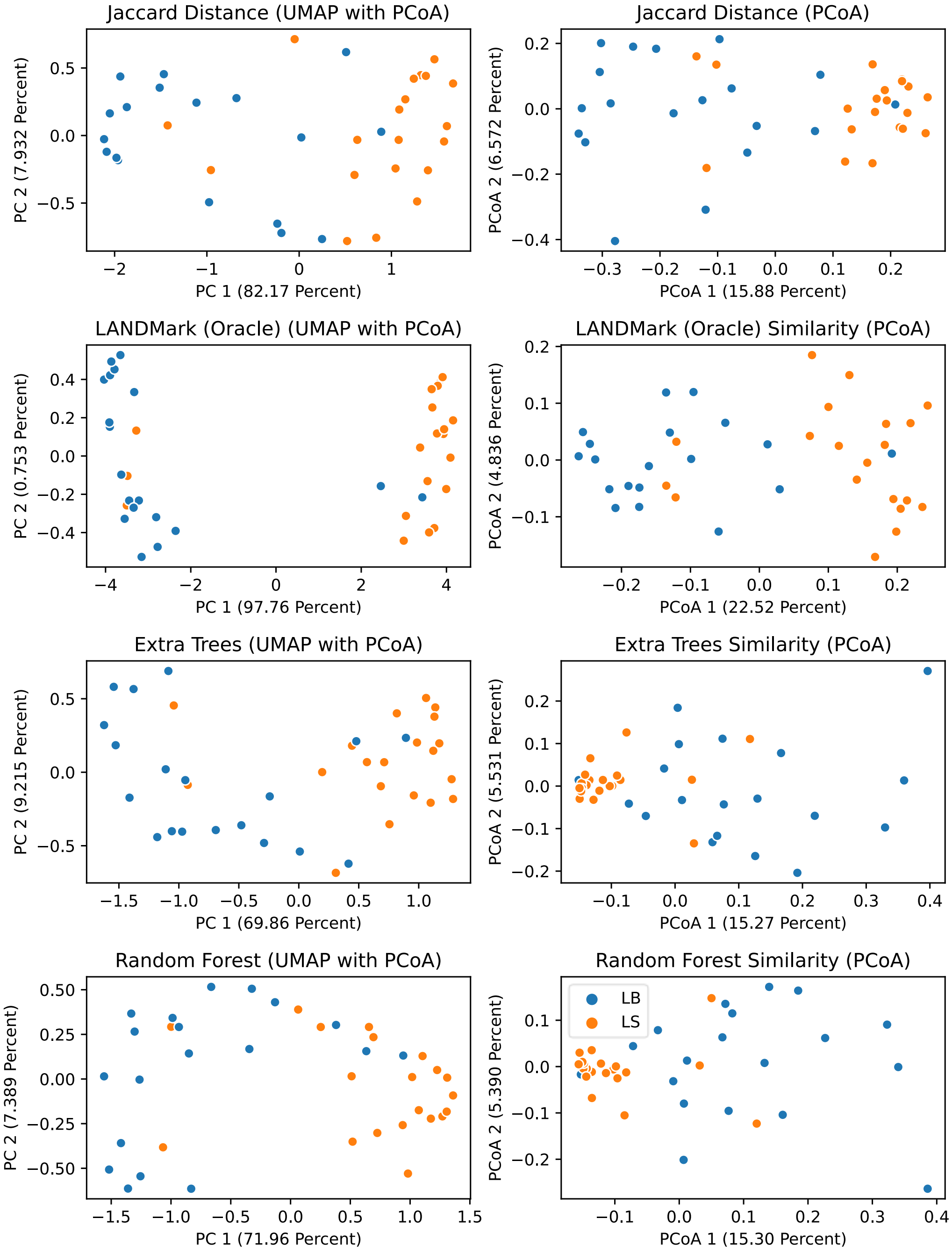
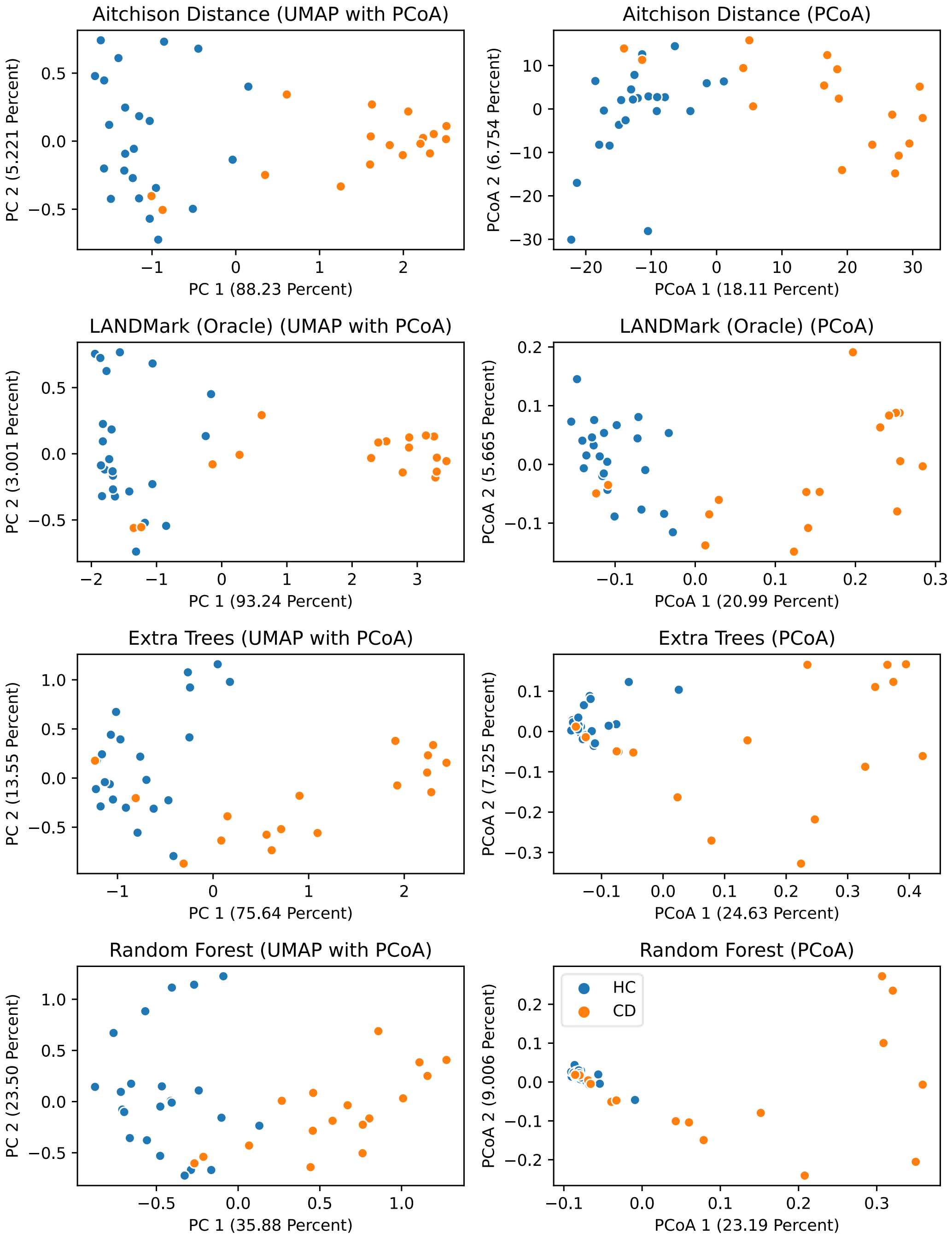


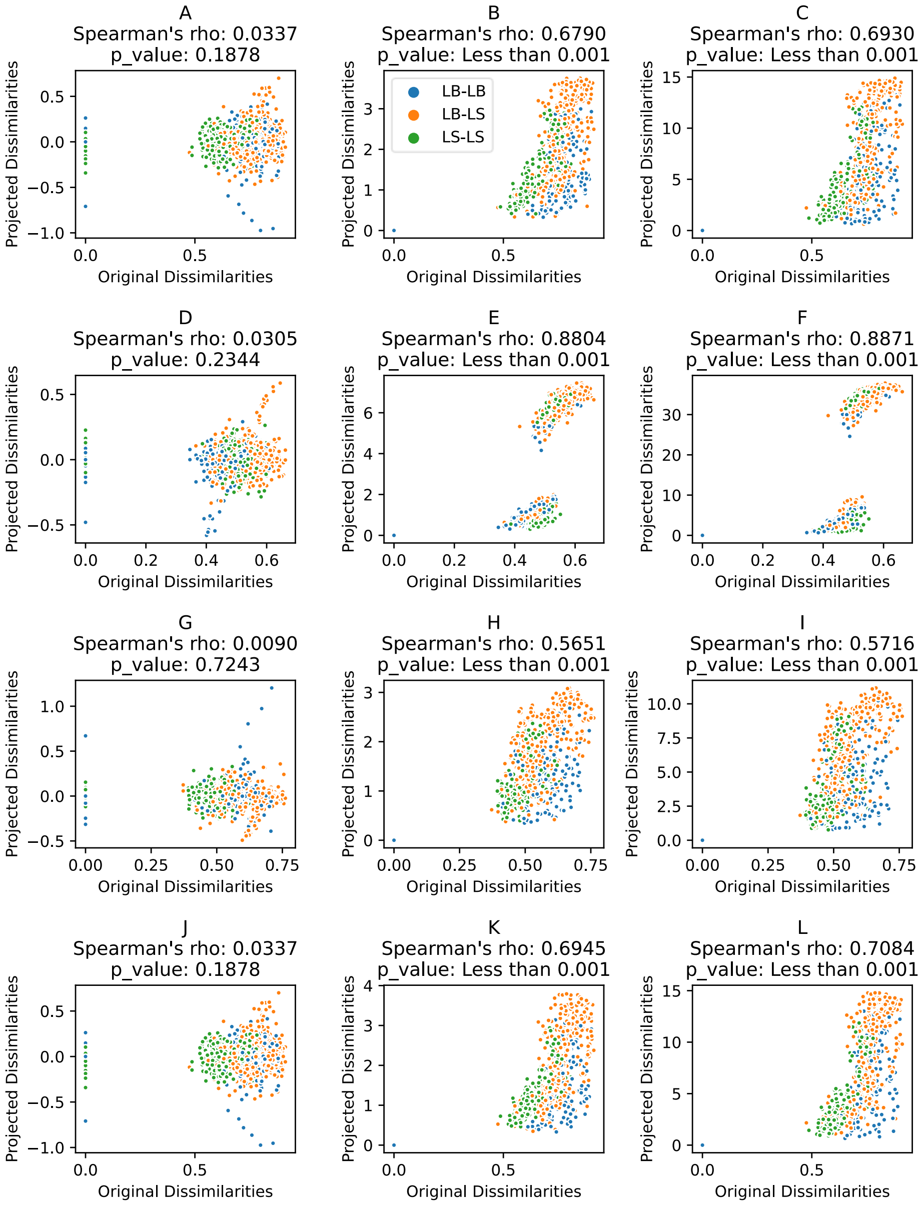
Figure 3: UMAP followed by PCoA and PCoA ordinations of the Crohn's disease subset. When only using PCoA projections of distance and dissimilarity matrices, each axis explains only a fraction of the total variation in the dataset. However, projections of the UMAP space using PCoA are more informative. In these projections, the first PCoA axis explains the vast majority of the variation in the distance and dissimilarity matrices. Furthermore, in these projections, the variation along the first axis appears to be strongly related to differences in community structure. The coloring of points serves as a visual aid and it does not affect the result. HC indicates healthy controls while CD indicates patients suffering from Crohn's Disease.



UMAP followed by PCoA is Effective at Creating Ordinations of the Investigated 16S rRNA Datasets

In PCoA projections of the original dissimilarity matrices, little to no correlation between distances in the original and projected spaces was observed (Figures 4 and 5 A, D, G, J). However, there is a trend where the most dissimilar pairs of samples could be found on the right side of each PCoA plot. Projections of each original dissimilarity matrix by UMAP, however, appear to better reflect the topology of the input space since distances between samples in the original and projected space appear to be correlated (Figures 4 and 5 B, E, H, K). Simply, this means that if the distance between two samples was large in the original space it also tended to be large in the UMAP space. Furthermore, Spearman's rho tended to be highest in the UMAP projections of LANDMark (Oracle) dissimilarities, suggesting that this approach is particularly effective at preserving relationships between samples (Figure 4 and 5 E). In one dataset (LB vs LS), the projection of the samples, pairwise comparisons between samples from the original LANDMark (Oracle) dissimilarities the projected distances resulted in the formation of two distinct groups (Figure 1). This can be easily explained as inter-class variation being greater than the intra-class variation in this subset, an observation supported by the PerMANOVA results (See Table 1). This was also observed in other subsets, though not to such an extreme degree. Finally, unlike the PCoA projections of the original dissimilarities, a two-dimensional PCoA embedding of the UMAP distances does not result in a notable difference in the pairwise dissimilarities between samples (Figure 4 and 5 C, F, I, L).

Figure 4: A visualization of how each type of projection preserves the pairwise distances between the projected and original distances in the LS-LB subset of the healthy gut data. The coloring of points serves as a visual aid and it does not affect the result. The first row visualizes the pairwise relationships between projections of the Jaccard distances into the PCoA (A), UMAP (B), and UMAP followed by PCoA space (C). The meaning of columns is the same in subsequent rows. The second (D-F), third (G-I), and fourth (J-L) visualize how each projection preserves the pairwise distances when dissimilarity matrices are constructed using LANDMark (Oracle), Extremely Randomized Trees, and Random Forests, respectively.



A Spearman's rho: 0.0337 p_value: 0.1878

B Spearman's rho: 0.6790 p_value: Less than 0.001

C Spearman's rho: 0.6930 p_value: Less than 0.001

D Spearman's rho: 0.0305 p_value: 0.2344

E Spearman's rho: 0.8804 p_value: Less than 0.001

F Spearman's rho: 0.8871 p_value: Less than 0.001

G Spearman's rho: 0.0090 p_value: 0.7243

H Spearman's rho: 0.5651 p_value: Less than 0.001

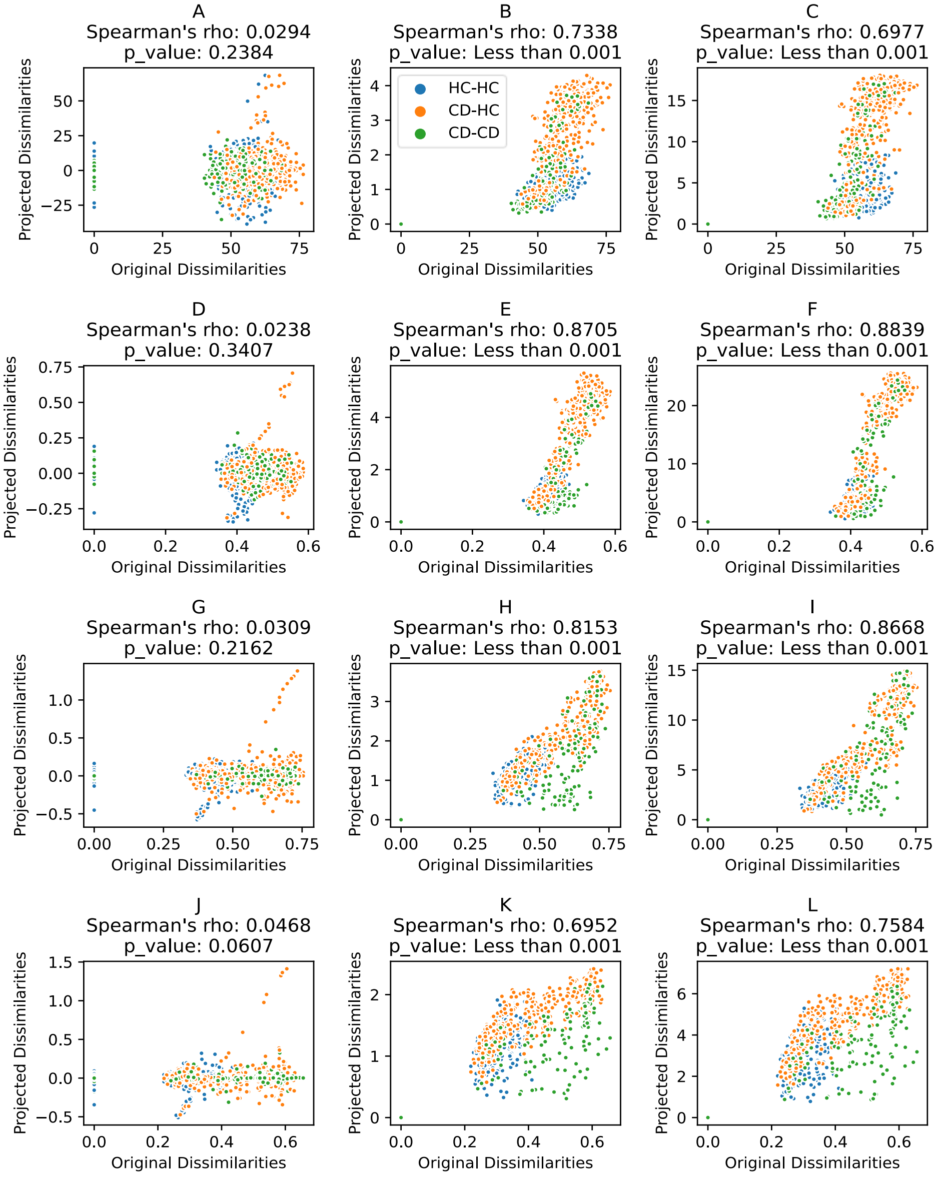
I Spearman's rho: 0.5716 p_value: Less than 0.001

J Spearman's rho: 0.0337 p_value: 0.1878

K Spearman's rho: 0.6945 p_value: Less than 0.001

L Spearman's rho: 0.7084 p_value: Less than 0.001

Figure 5: A visualization of how each type of projection preserves the pairwise distances between the projected and original distances in the Crohn's Disease subset of the IMID data. The coloring of points serves as a visual aid and it does not affect the result. The first row visualizes the pairwise relationships between projections of the Aitchison distances into the PCoA (A), UMAP (B), and UMAP followed by PCoA space (C). The meaning of columns is the same in subsequent rows. The second (D-F), third (G-I), and fourth (J-L) visualize how each projection preserves the pairwise distances when dissimilarity matrices are constructed using LANDMark (Oracle), Extremely Randomized Trees, and Random Forests, respectively.



The Choice in Data Transformation Could Impact Generalization Performance

When training using all features, generalization performance in the different subsets of the healthy gut dataset differed depending on the transformation. When training LANDMark (Oracle), ET, and RF models on the healthy-gut dataset, a Bayesian analysis showed that the presence-absence transformation is more likely to yield a model with better generalization performance in nearly all subsets of the data (Table 2). ET and RF models did perform better when trained on CLR transformed data in the RS-LS subset. However, this is unlikely to matter since no model was able to learn a way to classify RS samples from LS samples regardless of transformation. Since the PA transformed data was more likely to generate better models, we investigated if there would be any practical difference between models. In the IMID datasets, generalization performance appeared to depend on both the choice of transformation and classification model. For example, RF and ET models performed better when trained presence-absence transformed data in the MS-HC and the performance of these models are likely to be equivalent in the RA-HC and UC-HC subsets regardless of transformation (Table 2). However, the performance of LANDMark (Oracle) was best on CLR-transformed data across all subsets.

Table 2: Reporting of results investigating the effect of transformation on generalization performance. A Bayesian analysis using Nadeau and Bengio's corrected t-test was performed using each pair of transformations for each classifier. These results were obtained after training each model on all ASVs.

Dataset	*Subset*	*Model*	*Mean ± Std Dev (PA)*	*Mean ± Std Dev (CLR)*	*Probability PA > CLR*	*Probability PA = CLR*	*Probability PA < CLR*
Healthy Gut	LS-LB	**LANDMark**	**0.87 ± 0.05**	**0.73 ± 0.08**	**1.0**	**0.0**	**0.0**
		Extra Trees	0.86 ± 0.05	0.64 ± 0.13	1.0	0.0	0.0
		Random Forest	0.85 ± 0.05	0.51 ± 0.04	1.0	0.0	0.0
	RS-RB	**LANDMark**	**0.64 ± 0.10**	**0.45 ± 0.08**	**1.0**	**0.0**	**0.0**
		Extra Trees	0.66 ± 0.11	0.52 ± 0.04	1.0	0.0	0.0
		Random Forest	0.65 ± 0.10	0.49 ± 0.03	1.0	0.0	0.0
	RB-LB	**LANDMark**	**0.75 ± 0.07**	**0.72 ± 0.06**	**0.58**	**0.42**	**0.002**
		Extra Trees	0.74 ± 0.08	0.54 ± 0.09	1.0	0.0	0.0
		Random Forest	0.74 ± 0.08	0.51 ± 0.03	1.0	0.0	0.0
	RS-LS	**LANDMark**	**0.39 ± 0.09**	**0.30 ± 0.08**	**0.99**	**0.01**	**0.0**
		Extra Trees	0.37 ± 0.08	0.46 ± 0.07	0.0002	0.02	0.98
		Random Forest	0.39 ± 0.09	0.50 ± 0.02	0.0	0.001	0.99
Immune Modulated Inflammatory Disease	CD-HC	**LANDMark**	**0.83 ± 0.07**	**0.88 ± 0.06**	**0.0**	**0.08**	**0.92**
		Extra Trees	0.81 ± 0.08	0.82 ± 0.08	0.007	0.80	0.019
		Random Forest	0.82 ± 0.08	0.81 ± 0.09	0.09	0.90	0.003
	MS-HC	**LANDMark**	**0.67 ± 0.08**	**0.72 ± 0.09**	**0.0003**	**0.07**	**0.93**
		Extra Trees	0.65 ± 0.09	0.60 ± 0.12	0.76	0.23	0.02
		Random Forest	0.63 ± 0.07	0.57 ± 0.09	0.95	0.05	0.0003
	RA-HC	**LANDMark**	**0.72 ± 0.06**	**0.81 ± 0.06**	**0.0**	**0.0003**	**1.0**
		Extra Trees	0.69 ± 0.07	0.69 ± 0.10	0.12	0.62	0.26
		Random Forest	0.69 ± 0.07	0.68 ± 0.09	0.24	0.65	0.11
	UC-HC	**LANDMark**	**0.68 ± 0.08**	**0.72 ± 0.07**	**0.006**	**0.23**	**0.76**
		Extra Trees	0.68 ± 0.12	0.67 ± 0.10	0.19	0.73	0.08
		Random Forest	0.67 ± 0.09	0.65 ± 0.08	0.39	0.57	0.03

The Supervised LANDMark (Oracle) Classifier Learns Better Decision Rules than the Random Forest and Extremely Randomized Trees Classifiers

Supervised LANDMark's ability to split samples into their respective classes using multiple features resulted in clearer separations between classes (Figure 6). The decision boundaries learned by LANDMark were also less influenced by the peculiarities of the RF or ET classifiers. For example, an arcing effect was observed in the PCoA projection of the decision space of the RF classifier (Figure 6, Right Panel) while no such pattern could be observed in the decision space of the LANDMark classifier (Figure 6, Left Panel). Regardless of which classifier was used, the first principal component in each PCoA projection explained a large amount of the variance in the decision space. This suggests that each classifier can learn good decision rules which separate different classes of samples (14,37). However, due to the small number of samples, the PCoA results for the higher components should be interpreted with some caution. Finally, LANDMark (Oracle) models tend to be as good or better than RF or ET models since they appear to generalize better (Tables 3 - 5).

Table 3: Results of a Bayesian analysis that investigated the effect of feature selection on generalization performance.

Dataset	*Subset*	*Model*	*Mean ± Std Dev (Before)*	*Mean ± Std Dev (After)*	*Probability Before > After*	*Probability Before = After*	*Probability Before < After*
Healthy Gut	LS-LB	**LANDMark**	**0.87 ± 0.05**	**0.88 ± 0.04**	**0.01**	**0.93**	**0.06**
		Extra Trees	0.86 ± 0.05	0.85 ± 0.06	0.11	0.87	0.02
		Random Forest	0.85 ± 0.05	0.85 ± 0.05	0.03	0.97	0.01
	RS-RB	**LANDMark**	**0.64 ± 0.10**	**0.64 ± 0.12**	**0.11**	**0.80**	**0.08**
		Extra Trees	0.66 ± 0.11	0.68 ± 0.09	0.01	0.65	0.34
		Random Forest	0.65 ± 0.10	0.68 ± 0.10	0.01	0.34	0.65
	RB-LB	**LANDMark**	**0.75 ± 0.07**	**0.74 ± 0.09**	**0.25**	**0.73**	**0.02**
		Extra Trees	0.74 ± 0.08	0.74 ± 0.08	0.06	0.92	0.02
		Random Forest	0.74 ± 0.08	0.72 ± 0.08	0.38	0.59	0.03
Immune Modulated Inflammatory Disease	CD-HC	**LANDMark**	**0.88 ± 0.06**	**0.86 ± 0.06**	**0.29**	**0.71**	**0.0**
		Extra Trees	0.82 ± 0.08	0.85 ± 0.07	0.0	0.45	0.55
		Random Forest	0.81 ± 0.09	0.83 ± 0.09	0.003	0.59	0.40
	MS-HC	**LANDMark**	**0.72 ± 0.09**	**0.72 ± 0.10**	**0.14**	**0.79**	**0.07**
		Extra Trees	0.60 ± 0.12	0.63 ± 0.11	0.01	0.44	0.55
		Random Forest	0.57 ± 0.09	0.61 ± 0.11	0.0	0.18	0.82
	RA-HC	**LANDMark**	**0.81 ± 0.06**	**0.80 ± 0.08**	**0.17**	**0.78**	**0.05**
		Extra Trees	0.69 ± 0.10	0.75 ± 0.08	0.0	0.11	0.89
		Random Forest	0.68 ± 0.09	0.75 ± 0.08	0.0	0.05	0.95
	UC-HC	**LANDMark**	**0.72 ± 0.07**	**0.73 ± 0.07**	**0.04**	**0.72**	**0.24**
		Extra Trees	0.67 ± 0.10	0.69 ± 0.10	0.0	0.52	0.48
		Random Forest	0.65 ± 0.08	0.68 ± 0.08	0.0	0.68	0.62

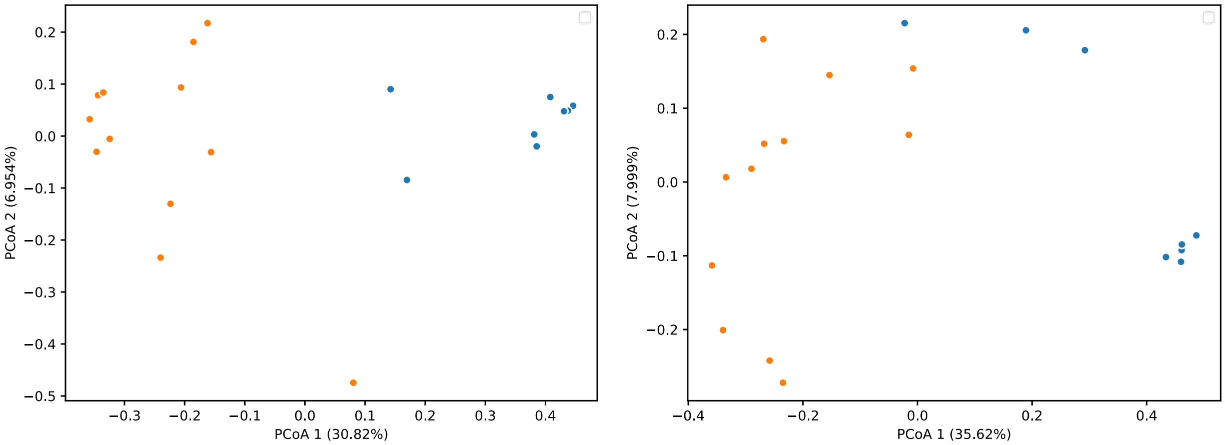
Table 4: Results of a Bayesian analysis comparing the generalization performance of different models before feature selection. These results were obtained using the best-performing transformation.

Dataset	***Subset***	***Model A***	***Model B***	***Probability Model A > Model B***	***Probability Model A = Model B***	***Probability Model A < Model B***
Healthy Gut (Presence – Absence)	LS-LB	LANDMark	Extra Trees	0.17	0.83	0.0003
		LANDMark	Random Forest	0.31	0.68	0.0
		Extra Trees	Random Forest	0.03	0.96	0.004
	RS-RB	LANDMark	Extra Trees	0.01	0.60	0.39
		LANDMark	Random Forest	0.09	0.63	0.28
		Extra Trees	Random Forest	0.15	0.84	0.01
	RB-LB	LANDMark	Extra Trees	0.18	0.80	0.02
		LANDMark	Random Forest	0.27	0.71	0.02
		Extra Trees	Random Forest	0.13	0.80	0.07
Immune Modulated Inflammatory Disease (CLR)	CD-HC	LANDMark	Extra Trees	0.93	0.07	0.0004
		LANDMark	Random Forest	0.96	0.04	0.0003
		Extra Trees	Random Forest	0.10	0.90	0.0001
	MS-HC	LANDMark	Extra Trees	0.99	0.007	0.0002
		LANDMark	Random Forest	1.00	0.0	0.0
		Extra Trees	Random Forest	0.63	0.36	0.008
	RA-HC	LANDMark	Extra Trees	1.00	0.002	0.0
		LANDMark	Random Forest	1.00	0.0004	0.0
		Extra Trees	Random Forest	0.18	0.79	0.03
	UC-HC	LANDMark	Extra Trees	0.78	0.20	0.02
		LANDMark	Random Forest	0.97	0.03	0.0004
		Extra Trees	Random Forest	0.46	0.53	0.006

Table 5: Results of a Bayesian analysis comparing the generalization performance of different models after feature selection. These results were obtained using the best-performing transformation.

Dataset	***Subset***	***Model A***	***Model B***	***Probability Model A > Model B***	***Probability Model A = Model B***	***Probability Model A < Model B***
Healthy Gut (Presence – Absence)	LS-LB	LANDMark	Extra Trees	0.58	0.42	0.002
		LANDMark	Random Forest	0.68	0.32	0.0
		Extra Trees	Random Forest	0.07	0.89	0.04
	RS-RB	LANDMark	Extra Trees	0.01	0.28	0.71
		LANDMark	Random Forest	0.0008	0.16	0.84
		Extra Trees	Random Forest	0.07	0.77	0.16
	RB-LB	LANDMark	Extra Trees	0.09	0.81	0.09
		LANDMark	Random Forest	0.36	0.63	0.01
		Extra Trees	Random Forest	0.37	0.62	0.02
Immune Modulated Inflammatory Disease (CLR)	CD-HC	LANDMark	Extra Trees	0.26	0.72	0.02
		LANDMark	Random Forest	0.59	0.40	0.01
		Extra Trees	Random Forest	0.30	0.70	0.0
	MS-HC	LANDMark	Extra Trees	0.97	0.03	0.0
		LANDMark	Random Forest	0.99	0.01	0.0
		Extra Trees	Random Forest	0.40	0.58	0.04
	RA-HC	LANDMark	Extra Trees	0.86	0.14	0.0
		LANDMark	Random Forest	0.87	0.13	0.0
		Extra Trees	Random Forest	0.16	0.75	0.09
	UC-HC	LANDMark	Extra Trees	0.68	0.30	0.03
		LANDMark	Random Forest	0.87	0.12	0.003
		Extra Trees	Random Forest	0.35	0.63	0.01

Figure 6: Principal Coordinate Analysis projections of test data can be used to assess model fit. Proximity matrices extracted from supervised LANDMark (Oracle) (Left) and Random Forest (Right) models trained on centered-log ratio transformed counts from the Crohn's Disease subset of the Immune-Mediated Inflammatory Disease dataset were projected into two dimensions using PCoA. Higher explained variation along the first principal component reflects the ability of each model to learn a simple set of decision rules. Healthy controls are colored orange while samples from patients suffering from Crohn's Disease are colored blue. Coloring of points serves as a visual aid and it does not affect the result.

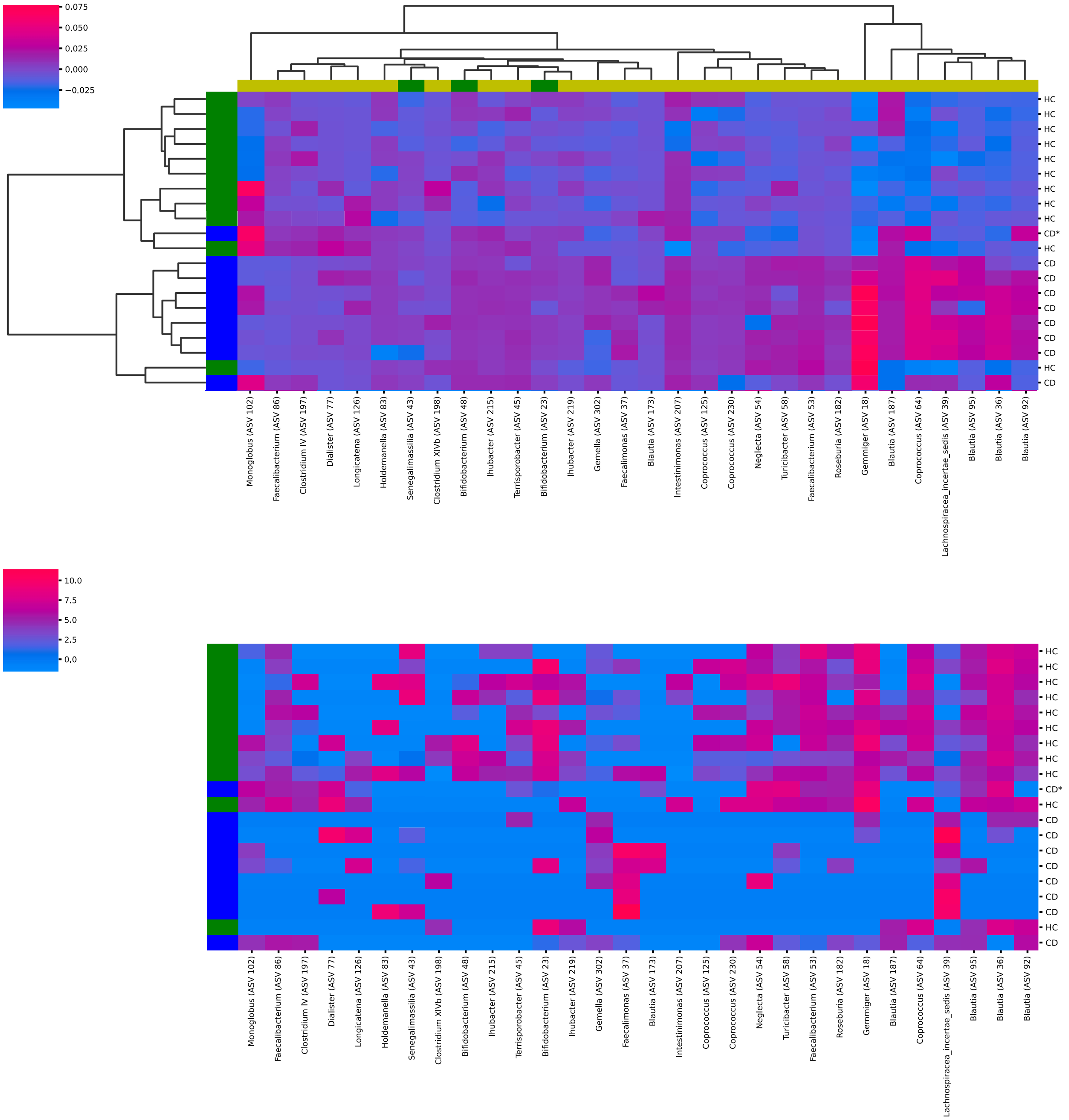


ASVs Predicted to Have a High Impact on Model Performance is Consistent with Previously Reported Results

The ASVs identified using LANDMark (Oracle) and RFE in the LB-LS subset of the healthy gut dataset are generally consistent with what was reported by Flynn et al. (19). We confirmed that *Turicibacter spp.*, *Peptoniphilus spp*., and *Finegoldia spp.* play a role in differentiating these two sites (19) (Suppl Figures 1 and 2). However, the results suggest that the individual impact that these ASVs have on classification is somewhat muted. Also, the differences in overall importance may be due to the experimental design since we built our models using 50% of the dataset. The ASV which had the strongest influence on generalization performance in test samples, ASV 317, belonged to *Schaalia spp*. and was not originally identified as important. Interestingly, ASV 576 (assigned to *Anaeromassilibacillus spp.*) was only present in one test sample but its absence strongly shifted the predictions of the model towards both types of samples, suggesting a possible interaction between one or more ASVs. Currently, it is difficult to determine interactions between ASVs using LANDMark. To investigate potential interactions involving ASV 576, an Extremely Randomized Trees model with 2048 trees was trained. This approach was chosen since it has been shown to approximate a non-linear function as the number of trees increases (33,34). While classification was not perfect (balanced accuracy score of 0.9) this follow-up analysis did confirm that ASVs 317 (*Schaalia*), 457 (*Enterocloster*), 429 (*Faecalicatena*), 120 (*Veillonella*), 610 (*Eisenbergiella*), and 249 (*Lawsonibacter*) primarily impact classification and that the effect of ASV 576 is likely an artifact (Suppl Figure 3).

We identified a group of ASVs which are important for distinguishing between CD and HC samples. ASVs belonging to *Gemmiger*, *Coprococcus*, and *Lachnospiracea incertae sedis* were included in this group. Furthermore, the genera identified by our model are consistent with those reported in the original work (3). Lower abundance in ASVs 18, 64, 36, 95, 187, and 92 - shift model predictions away from HCs. These ASVs were assigned to the genera *Gemmiger*, *Coprococcus*, and *Blautia* (for the remainder) respectively. Interestingly, a higher abundance of these ASVs did not result in a strong shift towards the prediction of a HC. An increase in the abundance of ASV 39 (*Lachnospiracea incertae sedis*) shifts predictions towards CD. A sixth ASV which was assigned to the genus *Monoglobus*, a taxon that was not previously identified as important, was identified in our analysis (Figure 7). While a detailed discussion of *Monoglobus* is outside the scope of this work, this species has been shown to be involved in pectin degradation and the metabolites produced from these pathways are important mediators of the inflammatory response (38,39). Within test samples from the first time point higher abundance of this ASV tended to shift some predictions towards healthy controls while a lower abundance of this ASV tends to shift predictions away from healthy controls. In a follow-up analysis using the second time point, however, the impact this ASV had on model predictions was considerably more muted (Suppl Figure 4). Finally, our analysis identified a group of additional ASVs (which included taxa such as *Terrisporobacter*, *Neglecta*, *Roseburia)* where a decrease in abundance tends to shift predictions towards CD. The overall influence that these ASVs exert on prediction is smaller, however.

Figure 7: Analysis of LANDMark (Oracle) models using model agnostic approaches can identify sets of predictive ASVs. These ASVs were identified using recursive feature elimination. Changes in the abundance (bottom, with pink indicating higher abundance) of specific ASVs appears to be related to how strongly (top) ASVs shift model predictions towards CD or a healthy control (HC). In the top graph, positive values (pink) indicate model predictions are shifted towards CD while negative values (blue) indicate shifts towards HCs. An asterisk denotes a sample that was not correctly predicted.



Discussion

The datasets investigated here were chosen since the human gut microbiome is an important area of medical research and is becoming increasingly linked to important disease phenotypes. Since machine learning models are becoming increasingly used to identify predictive features, it is important to understand how the quality and interpretation of results change depending on the machine learning model. This will hopefully allow greater insights into the composition and function of the human microbiome. The choice of transformation and dissimilarity measure is an important consideration when investigating microbiome data. It has long been known that the choice of dissimilarity measure can influence our measurement and interpretation of the main gradients influencing the structure of communities and taxonomic similarity between pairs of samples (40,41). For example, recent investigations have demonstrated that this choice can result in misleading results due to the sparsity inherent to the data, and differences in library size and sampling (24,27,42). To combat these problems a multitude of dissimilarity measures and ordination approaches have been developed to summarize and visualize ASV differences between sites (41). However, it remains incomplete since distance metrics and other commonly used dissimilarity measures have difficulty capturing potential interactions between ASVs. For example, the Jaccard distance simply calculates the number of shared ASVs over the total number of unique ASVs between two communities and it fails to consider how dependencies between ASVs influence the structure of a community. An example of such a dependency occurs when the presence of one ASV depends on the exclusion of another (43). Furthermore, when using measures that use abundance information, it is simple to show how differences in abundances can result in situations where the sites that share the same species are more dissimilar than sites that have no species in common. While applying transformations, such as CLR or converting to presence-absence, can help in these situations, a review of the literature suggests that there is yet to be a consensus on which approach is best (24,41,44,45). Our results are also unclear in this matter and suggest that the best choice in transformation will depend on both the dataset and model being used. For example, our results suggest that the presence-absence transformation may be better suited when samples come from (or are suspected to come from) two or more distinct ecological niches, such as the lumen and mucosa of the colon (46). This likely occurs since differences between these communities are dominated by changes in the presence and absence of specific organisms rather than abundance. However, when analyzing changes occurring within similar niches, such as those derived from stool, the CLR transformation may be more useful since it is sensitive to changes within compositions (28,47).

Alternative approaches to measuring pairwise dissimilarity, such as learning a dissimilarity measure, have also been developed and applied to the analysis of genomic and transcriptomic datasets (13,16,17,29). Unfortunately, while the properties of various dissimilarity measures have been extensively investigated, comparatively little work has been done exploring how learned dissimilarity measures can be used to investigate the same data. They are particularly interesting since they can learn a representation of the underlying manifold upon which the input samples are embedded (29,48). Given that amplicon sequencing datasets tend to lie on such manifolds, using learned dissimilarities could represent a potentially powerful way to analyze these datasets. Furthermore, since these dissimilarity matrices are derived from decision tree ensembles, interactions between ASVs are potentially accounted for, thereby overcoming one of the weaknesses of distance metrics (7,43,48). Therefore, using learned dissimilarities could result in the construction of more informative ordinations.

Our experiments show that a PCoA, on its own, is not able to adequately project samples into an appropriate embedding. This occurs since PCoA is a type of matrix factorization algorithm and it is difficult to construct linear representation in cases where the input manifold is non-linear. In these cases, PCoA cannot adequately preserve relationships between samples and the resulting projection would not effectively capture important aspects of the data. This is evident in Figures 2 and 3, which demonstrate that the first two principal axes of each PCoA projection of the original dissimilarities explain only a small fraction of the variation in each dataset. This is further underscored by the data presented in panels A, D, G and J of Figures 4 and 5 panels. These experiments clearly show that PCoA only rotates the input space and does not preserve the pairwise dissimilarities between samples in the resulting projection. Graph algorithms, such as UMAP, are an attractive alternative since these approaches are designed to learn an appropriate representation of the input manifold. Our experiments, evidenced in Figures 4 and 5, show that UMAP (and UMAP followed by PCoA) preserves the relationships between samples in the projected space since the pairwise dissimilarities in the original and projected space are correlated (31,49). Simply put, if the distance or dissimilarity between a pair of samples is large in the original space it tends to be large in the projected space. Applying these algorithms to our datasets allowed us to effectively visualize the relationships between samples, specifically differences in sampling location, with minimal distortion. Our results also support the growing body of work that shows that UMAP preserves the overall structure of HTS datasets and that it is more capable of representing sources of biological variation than PCoA (32). Finally, since the number of components used to construct the UMAP projection is arbitrary, we strongly suggest that a grid search over two UMAP parameters, the number of components and neighbors, is run so that a projection that best preserves the pairwise dissimilarity between samples can be constructed.

The dissimilarity matrices learned by unsupervised LANDMark (Oracle) resulted in projections that more clearly distinguished between the known main effects (sampling location and disease phenotype) (Table 1). Also, as the number of features used for splitting in LANDMark (Oracle) increased, the explanatory power of the main effects grew. This result demonstrates that distance metrics, such as the Jaccard or Aitchison metrics, might not capture the important differences between samples as readily as learned dissimilarities. One possible explanation for this result could be due to the inclusion of an increasing number of irrelevant dimensions as the dimensionality of the dataset increases (50,51). In amplicon sequencing datasets, irrelevant dimensions likely occur due to the inclusion of uninformative ASVs, potentially informative but highly variable ASVs, splitting a single genome, and missing data (24,27,52,53). Learned dissimilarity measures, such as those explored here, may be capable of identifying and reducing the impact uninformative ASVs exert when measuring dissimilarity. For example, in a RF classifier only ASVs which result in the best split are chosen at each node (13). Therefore, the impact of uninformative ASVs tends to be minimized since they are not selected as often. LANDMark (Oracle) extends this idea by identifying which linear or non-linear model is best at discriminating between classes using a randomly selected coalition of ASVs (37).

We show that using oblique decision tree ensemble classifiers, such as LANDMark (Oracle), can result in a highly predictive model. In this work, we show that a LANDMark (Oracle) classifier was likely to be at least as good as the ET or RF classifiers. Furthermore, when compared to RF and ET classifiers, we demonstrate that using feature selection is less likely to impact the generalization performance of a LANDMark (Oracle) classifier (Table 3). This result is important since it suggests that LANDMark (Oracle) is more robust to noise, especially when trained on CLR-transformed data. Furthermore, it is important to consider the shape of the decision boundaries learned by these classifiers. Both the RF and ET classifiers will produce a blocky boundary since each is only capable of learning axis-aligned splitting rules, although the boundary learned by ET tends to be smoother due to the random selection of cut-points (14,34). Smoother boundaries are preferred since they are likely to be a more faithful approximation of the rules which generate the data being studied (14,54). While the performance of all three models was similar in some instances, issues in the decision boundaries in these instances were noted. Specifically, we observed structures in the higher components of a PCoA using proximity matrices derived from supervised RF and ET models. In contrast, these structures did not exist in LANDMark (Oracle) models, implying the learning of a smoother boundary. This is consistent with other work involving this class of classifiers (14,37).

The generalization performance of our models tended to differ from that reported in the original work (3,19). We believe that these differences arose from differences in methodology, the use of ASVs, our choice of transformation, and our use of split-half cross-validation. Since we chose to analyze ASVs instead of OTUs, the dimensionality of our dataset substantially increased. For example, in the original IMID study the authors used 383 OTUs while our study found 702 ASVs (3). While using ASVs can provide a richer amount of information, generalization performance may degrade if ASVs artificially split bacterial genomes into different clusters (52). This occurs since the signal from one unique strain will now be spread over multiple ASVs. While this can lead to lower classification performance, this choice is justifiable since the results of our analysis are reproducible and these ASVs we identified as important can be used to generate new hypotheses for future experiments (55). The number of trees used to train our models and how generalization performance was calculated were also different. The original IMID work used 500 trees and calculated generalization performance using the out-of-bag error while the work by Flynn et al. (2018) used non-rarefied data as input and measured generalization performance using AUC scores (3,19). In contrast, we used 128 trees and split our data into training and testing sets using repeated split-half cross-validation. Previous work has demonstrated that after 128 trees the performance of a RF tends to plateau (30,37). Some additional testing using the various subsets of the IMID dataset demonstrated that adding additional trees to our analysis is unlikely to result in substantially better performance (Suppl Table 1). Finally, and likely the most significant contributor to differences in generalization performance, is our choice to use repeated split-half cross-validation. This approach is expected to result in decreased generalization performance since fewer samples are used for training. However, the advantage of this approach is that the overlap between training datasets is minimized (56). This reduces the dependence between different estimates of generalization performance thereby improving the ability to detect a true difference between the generalization performance of two classifiers (56). An additional advantage of using split-half cross-validation is that we can use more testing samples to calculate feature importance scores.

The ASVs identified as important by LANDMark (Oracle) are consistent with those identified in the original studies. This not only confirms the viability of LANDMark (Oracle) in this area of research, but it also strengthens the original work as their findings were replicated using a very different approach. Our work also demonstrates that classifiers such as LANDMark can not only validate the results of the original studies, but they can also add additional insights. For example, in the LS-LB investigation LANDMark (Oracle) identified *Schaalia spp*. as an important marker capable of distinguishing between the proximal lumen and mucosa of the colon. Finally, while detecting single ASV biomarkers is important, we should always be cognizant of the fact that these organisms interact with each other and the host. Therefore, when building and analyzing predictive models it is important to use approaches that can explore, quantify, and validate these interactions. In addition to detecting strongly predictive ASVs, our approach was also capable of detecting ASVs which have a more subtle effect on predicting CD and HC patients and whether samples originated in the distal or the proximal colon.

When looking at the ASVs identified by each model, both the RF and ET identified fewer ASVs than LANDMark (Oracle). The larger number of ASVs identified by LANDMark (Oracle) is likely due to differences in the way in which nodes are constructed. In RF and ET classifiers, only single features are used to construct each node (13,34). Therefore, only a very small fraction of features (at most n-1, where n is the number of samples) will be used to construct each tree. In practice, however, it is more likely that fewer features will be used if particularly good splits are found. It is also possible that features are reused at deeper nodes within each tree. This form of tree construction has also been shown to have a strong regularizing effect, which could limit the amount of available information upon which decisions are made (57). While it is likely that a regularization effect similar to that observed in RF and ET occurs in LANDMark, the strength of this effect may be more muted because LANDMark considers more features at each node (37). This allows a richer amount of information to be used to construct each tree but comes at the cost of including features that may have a limited impact on classification. For this reason, we believe it is particularly important to pair LANDMark models with model agnostic introspection algorithms, such as Permutation Explainer, which are capable of quantifying feature importance and interactions between features (58). It is also important to note that genome splitting could also contribute to this effect (52). For example, multiple ASVs assigned to *Peptoniphilus* in the LS-LB data and *Blautia* and *Coprococcus* in the CD-HC data. Therefore, additional work is needed to determine the extent of this issue in 16S datasets. Work is also needed to determine how best to handle this problem.

Conclusions and Future Work

Our work has shown that unsupervised LANDMark (Oracle) models can learn effective dissimilarity matrices. When paired with modern dimensionality reduction approaches, such as UMAP, the global structure of the original dissimilarity matrix is preserved. UMAP representations can then be combined with existing matrix factorization approaches to create informative ordinations. However, this comes at a cost of clarity since it is difficult to determine how variance along each axis is related to the presence/absence or abundance of each ASV. Therefore, it is important to conduct work investigating approaches capable of identifying which ASVs impact the location of samples in the transformed space. Finally, we show that LANDMark (Oracle) can learn highly predictive models after feature selection. Importantly, the ASVs identified by feature selection is consistent with contemporary work. Due to the way LANDMark constructs each tree, further investigations into the integration of feature selection and a statistical analysis of the resulting feature impact scores are necessary. This could potentially identify a small subset of highly predictive ASVs and this analysis would sidestep the need to use generalized linear models since the degree of confidence in the impact that each ASV has on classification is evaluated rather than differences in abundance/presence.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Availability of data and materials

Authors can confirm that all relevant data are included in the article and/or its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

JR is supported by funds from the Food from Thought project as part of Canada First Research Excellence Fund. MH received funding from the Government of Canada through Genome Canada and Ontario Genomics. BG is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant (RGPIN-2020-05733).

Authors' contributions

JR and MH conceived the project. JR analyzed/interpreted the results. JR wrote the draft. JR, MH, BG, and SK read, discussed, and contributed to the draft. MH provided computational resources. All authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Dr. Katie McGee and Dr. Terri M. Porter for their thoughtful discussions during the development of LANDMark.

References

1. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010 Nov;5(6):463–6.

2. Zhang Z, Liu Z-P. Robust biomarker discovery for hepatocellular carcinoma from high-throughput data by multiple feature selection methods. BMC Med Genomics. 2021 Aug 25;14(Suppl 1):112–112.

3. Forbes JD, Chen C-Y, Knox NC, Marrie R-A, El-Gabalawy H, de Kievit T, et al. A comparative study of the gut microbiota in immune-mediated inflammatory diseases-does a common dysbiosis exist? Microbiome. 2018 Dec 13;6(1):221–221.

4. DiCarlo GE, Mabry SJ, Cao X, McMillan C, Woynaroski TG, Harrison FE, et al. Autism-Associated Variant in the SLC6A3 Gene Alters the Oral Microbiome and Metabolism in a Murine Model. Front Psychiatry. 2021 Apr 15;12:655451–655451.

5. Paulson JN, Stine OC, Bravo HC, M P. Robust methods for differential abundance analysis in marker gene surveys. Nature Methods. 2013;10(12):1200–2.

6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 Dec 5;15(12):550.

7. Cutler RD, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random Forests for Classification in Ecology. Ecology. 2007;88(11):2783–92.

8. Ryo M, Rillig MC. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere. 2017;8(11):01976.

9. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.

10. Wang J, Chai J, Sun L, Zhao J, Chang C. The sputum microbiome associated with different sub-types of AECOPD in a Chinese cohort. BMC Infectious Diseases. 2020 Aug 18;20(1):610.

11. Aryal S, Alimadadi A, Manandhar I, Joe B, Cheng X. Machine Learning Strategy for Gut Microbiome-Based Diagnostic Screening of Cardiovascular Disease. Hypertension. 2020;76(5):1555–62.

12. Kubinski R, Djamen-Kepaou J-Y, Zhanabaev T, Hernandez-Garcia A, Bauer S, Hildebrand F, et al. Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease. bioRxiv [Internet]. 2021; Available from: https://www.biorxiv.org/content/early/2021/05/04/2021.05.03.442488

13. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

14. Menze BH, M K, Splitthoff DN, K K, Hamprecht FA. On oblique random forests. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, editors. Machine Learning and Knowledge Discovery in Databases. 2011. p. 453–69.

15. Ehsani R, Drabløs F. Robust Distance Measures for kNN Classification of Cancer Data. Cancer Inform. 2020 Oct 13;19:1176935120965542.

16. Pouyan MB, Kostka D. Random forest based similarity learning for single cell RNA sequencing data. Bioinformatics. 2018 Jul 1;34(13):i79–88.

17. Alhusain L, Hafez AM. Cluster ensemble based on Random Forests for genetic data. BioData Mining. 2017 Dec 15;10(1):37.

18. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012 Jun 1;99(6):323–9.

19. Flynn K, Ruffin MT IV, Turgeon K, Schloss PD. Spatial Variation of the Native Colon Microbiota in Healthy Adults. Cancer Prevention Research. 2018;11(7):393–402.

20. Porter TM, Hajibabaei M. METAWORKS: A flexible, scalable bioinformatic pipeline for multi-marker biodiversity assessments. bioRxiv. 2020;

21. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:2584.

22. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology. 2007 Aug;73(16):5261–7.

23. Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, Smidt H, et al. Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. PLOS ONE. 2009 Aug;4(8):1–15.

24. Gloor GB, Macklaim JM, Pawlovsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017;8:2224.

25. Ranasinghe JA, Stein ED, Miller PE, Weisberg SB. Performance of two Southern California benthic community indices using species abundance and presence-only data: relevance to DNA barcoding. PLoS One. 2012;7(8):40875.

26. Wallen ZD. Comparison study of differential abundance testing methods using two large Parkinson disease gut microbiome datasets derived from 16S amplicon sequencing. BMC Bioinformatics. 2021 May 25;22(1):265.

27. Hugerth LW, Andersson AF. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Frontiers in Microbiology. 2017;8:1561.

28. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, et al. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. mSystems. 2019 Feb;4(1).

29. Xiong C, Johnson D, Xu R, Corso JJ. Random Forests for Metric Learning with Implicit Pairwise Position Dependence. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2012. p. 958–66. (KDD '12). Available from: https://doi.org/10.1145/2339530.2339680

30. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 154–68.

31. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Journal of Open Source Software. 2020;3(29):861.

32. Armstrong G, Martino C, Rahman G, Gonzalez A, Vázquez-Baeza Y, Mishne G, et al. Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. mSystems. 0(0):e00691-21.

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.

34. Geurts P, Ernst D, Wehenkel L. Extremely Randomized Trees. Machine Learning. 2006;63(1):3–42.

35. Lundberg SM, Lee S. A Unified Approach to Interpreting Model Predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017 [Internet]. Long Beach; 2017. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

36. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis. Journal of Machine Learning Research. 2017;18(77):1–36.

37. Rudar J, Porter TM, Wright M, Golding GB, Hajibabaei M. LANDMark: An ensemble approach to the supervised selection of biomarkers in high-throughput sequencing data. BMC Bioinformatics. 2022;Forthcoming.

38. Elshahed MS, Miron A, Aprotosoaie AC, Farag MA. Pectin in diet: Interactions with the human microbiome, role in gut homeostasis, and nutrient-drug interactions. Carbohydrate Polymers. 2021;255:117388.

39. Kim CC, Healey GR, Kelly WJ, Patchett ML, Jordens Z, Tannock GW, et al. Genomic insights from Monoglobus pectinilyticus: a pectin-degrading specialist bacterium in the human colon. ISME J. 2019 Jun;13(6):1437–56.

40. Bacaro G, Ricotta C, Mazzoleni S. Measuring beta-diversity from taxonomic similarity. Journal of Vegetation Science. 2007;18(6):793–8.

41. Wildi O. Evaluating the Predictive Power of Ordination Methods in Ecological Context. Mathematics [Internet]. 2018;6(12). Available from: https://www.mdpi.com/2227-7390/6/12/295

42. Leite MFA, Kuramae EE. You must choose, but choose wisely: Model-based approaches for microbial community analysis. Soil Biology and Biochemistry. 2020;151:108042.

43. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2012;14(3):315–26.

44. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: An ecological perspective. Methods in Ecology and Evolution. 2019;10(3):389–400.

45. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(27).

46. Duncan K, Carey-Ewend K, Vaishnava S. Spatial analysis of gut microbiome reveals a distinct ecological niche associated with the mucus layer. null. 2021 Jan 1;13(1):1874815.

47. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. Ann Epidemiol. 2016 May;26(5):322–9.

48. Tsagkrasoulis D, Montana G. Random forest regression for manifold-valued responses. Pattern Recognition Letters. 2018;101:6–13.

49. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nature Biotechnology. 2021 Feb 1;39(2):156–7.

50. Ayesha S, Hanif MK, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion. 2020;59:44–58.

51. Koul N, Manvi SS. Machine-Learning Algorithms for Feature Selection from Gene Expression Data. In: Srinivasa KG, Siddesh GM, Manisekhar SR, editors. Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications [Internet]. Singapore: Springer Singapore; 2020. p. 151–61. Available from: https://doi.org/10.1007/978-981-15-2445-5_10

52. Schloss PD. Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. mSphere. 2021 Aug 25;6(4):e0019121.

53. Edgar RC. UCHIME2: Improved chimera detection for amplicon sequences. bioRxiv [Internet]. 2016; Available from: https://www.drive5.com/uchime/

54. Kuncheva LI, Rodriguez JJ. Classifier ensembles with a random linear oracle. IEEE Transactions on Knowledge and Data Engineering. 2007;19(4):500–8.

55. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 2017;11:2639–43.

56. Valente G, Castellanos AL, Hausfeld L, Martino FD, Formisano E. Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. NeuroImage. 2021;238:118145.

57. Mentch L, Zhou S. Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success. Journal of Machine Learning Research. 2020;21:171:1-171:36.

58. Wang C, Feng L, Qi Y. Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China. Environmental Research. 2021;202:111740.

Additional Files

Additional File 1 – Supplementary Figures 1 to 4

Additional File 2 – Supplementary Table 1

Additional File 3 – Raw ESV Table, Taxonomic Assignments, and Code