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

We investigated how well decision-tree ensemble classifiers can identify and quantify differences in the compositions of microbial communities using previously sequenced 16S rRNA gene from the lumen and mucosa of the proximal human colon and the stool of patients suffering from Crohn's Disease. Tree-based approaches were then compared to ordinations produced using the Jaccard and Aitchison metrics. Since these investigations favored tree-based approaches, they were used as a basis for a workflow capable of learning dissimilarities, projecting them into a lower dimensional space, and identifying the features impacting the location of samples in the projections. When used with the centered log-ratio transformation, our new workflow (TreeOrdination) was capable of identifying differences in the microbial communities of the proximal colon and between Crohn's Disease patients and healthy controls ($9.08 \leq$ Pseudo-F ≤ 30.31, $0.001 \leq p \leq 0.004$; $21.19 \leq$ Pseudo-F ≤ 101.83, $0.001 \leq p \leq 0.002$). These projections were often better than those produced using the Aitchison distance ($8.22 \leq$ Pseudo-F ≤ 13.34, $p \leq 0.004$; $15.20 \leq$ Pseudo-F ≤ 30.70, $p \leq 0.001$). TreeOrdination models also generalized well to unseen data. In the proximal colon dataset, for example, default TreeOrdination models resulted in balanced accuracy scores that were more likely to be higher than classifiers only trained on centered-log ratio transformed data. Finally, TreeOrdination was able to identify ASVs known to be associated with each type of sample while also identifying potential new biomarkers of immune-mediated inflammatory diseases, such as *Parolsenella*, in the Crohn's Disease dataset.

Author Summary

Distance and dissimilarity measures are often used to investigate the structure of biological communities. However, this investigation demonstrates that these measures can fail to capture important relationships in microbial communities since they are unable to properly

capture the geometry of the input space. 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 the geometry of the input space.

Introduction

The statistical analysis of the composition of ecological communities, beta-diversity, is essential to determine their function within hosts and the broader environment. Central to the analysis of beta diversity is the calculation of pairwise distances or dissimilarities between samples. This calculation is carried out by carefully choosing and then applying a particular dissimilarity that is known to measure an important characteristic of the data. For example, some methods (such as correlation-based distance methods) perform well at measuring environmental gradients while others, such as the Jaccard distance, can perform well when one is interested in clustering data, assuming this is even possible (1–4). One particularly important drawback inherent to many dissimilarity measures is their inability to discover important geometric properties of the input space. A considerable amount of work has demonstrated that many genomic, transcriptomic, and metagenomic datasets are better understood if the underlying geometry of the input space is considered. This better understanding can be reached since these datasets do not often meet the “Assumption of Linearity” used by methods such as Non-Negative Matrix Factorization, Principal Components Analysis and its generalization, Principal Coordinates Analysis (5–8). Finally, these methods rarely explore correlations and other

dependencies between taxa. These dependencies and correlations commonly occur within communities and, when investigated, the investigations are often limited in scope and a result of specific actions on the part of the researcher (9).

Investigating community composition, such as detecting differentially abundant taxa, often involves the use of different tools. For example, one may use PerMANOVA to identify if there are differences between groups while then using another tool, such as ANCOM or MetagenomeSeq, to identify the operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) which differentiate groups (10–16). Each of these tools has different sets of assumptions and the violation of these assumptions could render the results meaningless. For example, if distributional assumptions are not met an analysis could suffer from higher false-positive and false-negative rates (3). This can then lead to potentially erroneous or incomplete interpretations. Furthermore, it is important to note that discovering which features can distinguish between communities is inextricably linked to the measure we use to identify if a statistically significant difference in community structure exists. Therefore, a unified approach using a single model to quantify differences between communities and samples would be very useful when investigating ecologically and medically important data.

Machine learning (ML) algorithms are being increasingly applied to a wide array of genomic, metagenomic, and transcriptomic data sets and we believe that they can be used to create a unified approach to study high-throughput sequencing data. The literature is rich in examples where ML has been used to identify biomarkers and create predictive models of these datasets. Unlike statistical models, ML models tend to be non-parametric (16,17). Furthermore, due to the way ML models such as neural networks and Random Forests learn, they are capable of identifying dependencies between covariates without the need for the user to explicitly include

these terms in the model (9,18–20). Also, many ML models, such as tree-based ensembles, are not linearly constrained. For this reason they can potentially learn a better understanding of the geometry of the input space and therefore generalize well to unseen data (5,20).

In this work we explore the ability of tree-based ensembles to investigate high-throughput sequencing data. Specifically, we focus on one, and arguably underused, ability inherent to this class of models: they can be used in an 'unsupervised' manner to learn the pairwise dissimilarities between samples (18,21). By using these classifiers in this way, it may be possible to simultaneously visualize the differences between communities while also identifying the features (ASVs, OTUs, genera, etc) which contribute most to the placement of each community within the ordination space (20,22). While this can be done using principal components analysis and similar algorithms these approaches suffer from two important limitations: transformation of new data often requires the original data (in the case of robust principal components analysis) and the algorithms are linearly constrained (5,23). Decision trees also have limitations that need to be considered. For example, the most popular methods (Random Forests, Extra Trees) make axis-orthogonal cuts. While improvements, such as the additional randomization, can help these models find better decision boundaries these improvements are not a solution to this drawback (24–26). Recently, we have introduced a decision-tree ensemble called LANDMark (Large Scale Non-Parametric Discovery of Markers – available at https://github.com/jrudar/LANDMark) that has the ability to learn oblique or non-linear splits (25). Recent and historical work demonstrates that these types of splits often result in more appropriate representations of the data generating function (5,26). Like the Random Forests, LANDMark recursively partitions samples until a stopping criterion is met. Unlike Random Forests, LANDMark can use learned linear and non-linear models to partition samples.

In our previous work we show that this results in a better representation of the decision boundary between different types of samples (25).

In this study, we investigate the quality of the dissimilarity measures learned by LANDMark and a modification of the Random Forest algorithm known as Extra Trees. These measures of dissimilarity will be compared with those calculated using commonly used dissimilarity measures such as the Jaccard and Aitchison metrics. The ability of each method to identify differences between communities and the concordance between projections will be investigated using commonly used statistics, such as PerMANOVA and Procrustes analysis. Furthermore, we will show how machine learning can be used to identify ASVs associated with patterns in the distribution of samples by applying state-of-the-art introspection methods to each unsupervised decision-tree ensemble. The code used to perform the analysis in this work can be found at: https://github.com/jrudar/Unsupervised-Decision-Trees.

Methods

Creation of Positive and Negative Controls

Positive and negative controls were created using a copy of the code found at https://github.com/cameronmartino/deicode-benchmarking/blob/master/simulations (23). This code was used in the simulations which tested robust centered log-ratio transformation introduced by Martino et al. (2019) (23). Briefly, a synthetic dataset containing 200 samples and 1000 features was according to the procedure found in Martino et al. (2019). This was a positive control. A copy of this dataset was made, and each feature column was independently permuted to create the negative control. Both datasets were then analyzed using the methods described below.

Dataset Selection

Two human microbiome datasets were selected for use as case studies in this work. The first dataset was derived from the colons of healthy individuals (27) using 16S rRNA amplicon sequencing. This dataset collected samples from the unprepared colons of healthy individuals (27). In the original work, the authors examined differences in the abundance of OTUs between the microbial communities of the proximal lumen (PL) and mucosa (PM), the distal lumen (DL) and mucosa (DM), between the PL and the DL, and finally between the PM and the DM. The second dataset was chosen since it contains samples from patients who suffer from immune-mediated inflammatory diseases (IMID) (28). Differences between the microbiomes of patients suffering from Crohn'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 could be identified in each patient's stool. In both studies, the authors used Random Forests to identify potential biomarkers and the results of this work can be compared to the potential biomarkers identified by the workflow presented here (21,28). Since the proximal region of the human colon and Crohn's Disease represents two well-investigated research areas, these subsets from the datasets described above were included as case studies in our work. Should our workflow be successful, it should (at the very least) be able to replicate the findings described in Flynn et al. (2018) and Forbes et al. (2018) while also supporting the wider body of work upon which these studies draw.

Bioinformatic Processing of Reads

Raw sequences from two previously published datasets were obtained from the Sequence Read Archive (PRJNA450340 and PRJNA418115) (27,28). All bioinformatic processing of the

raw reads was prepared using the MetaWorks v1.8.0 pipeline (available online at: https://githib.com/terrimporter/MetaWorks) (29). 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 (30). 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 (31).

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 OTUs assigned to other bacterial groups were underrepresented (28). A site by ASV count matrix, where each row is a sample and each column an ASV, was created for each subset of data under investigation. The count matrix was filtered to remove ASVs found in two or fewer samples. This filtration step was taken since reducing the size of the feature space can often lead to a more generalizable model (14,32,33).

Data Transformations

The filtered matrix must be transformed in such a way as to minimize the impact of various technical factors, such as differences in library size (10). Our analyses examined three transformations of the filtered matrix: the presence-absence transformation of rarefied samples, the centered-log ratio (CLR) transformation, and the robust centered-log ratio (rCLR) transformation (14,23,34). The presence-absence transformation was chosen since it reflects if ASVs are simply present or absent. By using this approach, the differences in counts are ignored while the impact of some effects, such as large differences in library size, can be somewhat minimized by first creating a rarefied dataset before applying the transformation (3,23). To prepare the rarefied libraries, we first calculated the sum of all counts in each sample and then calculated the 15th percentile of these sums. The probability of each ASV occurring in each sample was also calculated. ASVs were then sampled with replacement using the calculated probabilities until the specified library size was reached (3). Libraries smaller than the 15th percentile than this were removed from the analysis (35).

Normalizing library sizes by rarefying and the presence-absence transformation can be problematic since they throw away useful data due to the normalization of library sizes and, in the case of presence-absence, ignoring differences in the abundance of reads (14). Since the absolute size of each library is often irrelevant, a compositional approach is needed to analyze the data (14,36). The CLR and rCLR transformations were used to account for the compositional nature of these datasets. Filtered datasets were CLR transformed using the Scikit-Bio package (version 0.5.7) (37). The closure operation was applied to each sample and zero-counts were replaced using the multiplicative replacement procedure (13,38). Finally, the CLR transformation was applied according to equation 1 where X_i is the ith sample in X, $x_{D,i}$ is the Dth feature in X_i, and $g(X_i)$ is the geometric mean of all the taxa in sample X_i. The rCLR

transformation, which is implemented in the DEICODE package, is similar to the CLR transform and was developed to account for the sparse nature of microbiome data (23). Unlike the CLR transformation, however, the rCLR is calculated using only the non-zero data from each sample. Furthermore, for DEICODE to calculate the reconstructed matrix it must use all of the data points. This procedure is not capable of identifying a rule which can be used to transform a new set of data into the reduced space (23). In these experiments, DEICODE (version 0.2.4) was used to generate two RPCA components using 1000 iterations of the matrix completion algorithm. The *U*, *V*, and *S* matrices were extracted and used to reconstruct the rCLR transformed samples (equation 2) (23). The *U* and *V* matrices contain sample and feature loading information for each component while *S* is analogous to the matrix of eigenvalues used by singular-value decomposition. This work was carried out in a Python 3.10 environment.

$$CLR(X_i) = \left[\log\frac{x_{1,i}}{g(X_i)}, \dots, \log\frac{x_{D,i}}{g(X_i)}\right] \quad (1)$$

$$rCLR(X) = USV^T \quad (2)$$

Calculation of Dissimilarities

The Jaccard distance and Aitchison distances are used since these metrics can be used with compositional data (Equations 3 and 4). These metrics were used to calculate pairwise dissimilarities for presence-absence and (r)CLR data, respectively. In these equations X_i and X_j represent the ith and j^{th} samples in X. $x_{k,i}$ and $x_{k,j}$ represent the kth feature in samples x_i and x_j. Distances were calculated using the 'pairwise_distances' function from Scikit-Learn (version 1.0.2) (39).

$$Jaccard\ Distance\ (X_i, X_j) = \frac{X_i \cap X_j}{X_i \cup X_j} \quad (3)$$

$$Aitchison\ Distance\ (X_i, X_j) = \sqrt{\sum_{i=1}^{D} \left(CLR(X_i) - CLR(X_j)\right)^2} \quad (4)$$

Tree-based models, using algorithms such as Extra Trees, have also been found to be an effective means of capturing the dissimilarity between samples. To calculate the dissimilarities using Extra Trees, one must first construct a similarity matrix, S. This matrix is constructed by simply counting how often each pair of samples co-occur in the terminal leaves of each decision tree across the forest (Equation 5).

$$Tree\ Similarity\ \left(X_i, X_j\right) = \frac{\overline{X}_i \overline{X}_j^{\,T}}{N} \qquad (5)$$

In this equation X_i and X_j are the ith and j^{th} samples of X while $\overline{X}_i$ and $\overline{X}_j$ are one-hot encoded representations of the terminal leaves into which X_i and X_j fall. N is the number of trees in the forest. The pairwise similarity matrix is then converted into a dissimilarity matrix, D, using equation 6 (22). This dissimilarity measure can then be used to investigate community differences (22,40,41).

$$Tree\ Dissimilarity = \sqrt{1 - S} \qquad (6)$$

Unfortunately, for Extra Trees to generate a consistent set of one-hot encodings (for use in equation 5), this method of calculating dissimilarities requires that the original dataset is loaded whenever using the model for new data. While this can be addressed, developing a workaround is outside of the scope of this work and will, likely, not be done by most users who adopt this procedure. When developing LANDMark, this problem was addressed by returning the entire set of nodes for each tree in the forest (25). Therefore, the one-hot encodings produced by LANDMark can be used directly. This can be particularly useful since it allows LANDMark

to be used as a secondary transformation. Specifically, the dimensionality reduction is done using the one-hot encodings instead of the dissimilarity matrix. In these experiments, we used LANDMark version 1.1.4 and the implementation of the Extra Trees classifier found in Scikit-Learn version. 128 trees were used to train each Extra Trees classifier. LANDMark was run at its default settings except for the number of trees, which was set to 80, and the 'use_nnet' parameter, which was set to 'False' to reduce training time. All tests were run on a computer using an Intel Core i5-10400F CPU with 32 GB of RAM. This work was carried out in a Python 3.10 environment.

Regardless of which procedure is used, a second randomized dataset, X_R, is needed if one is to use decision trees to calculate unsupervised dissimilarities. To create randomized datasets, counts within each ASV are randomly sampled without replacement, thereby breaking the correlation structure between ASVs without affecting count distributions (42,43). The original and randomized samples are then combined. To prevent the supervised-learning model from using information about the original class labels, the permuted dataset is assigned a label of "Random" while original samples are assigned a label of "Original". The combined dataset is then transformed using the rCLR, CLR, and presence-absence transformations. Twenty different randomized datasets were created. The dissimilarities or one-hot encodings between pairs of real samples from each dataset were then used for dimensionality reduction.

Dimensionality Reduction Using PCoA, RPCA, and UMAP

All samples were used to generate the initial ordinations. Principal Coordinates Analysis (PCoA) was used to project presence-absence transformed data into a two-dimensional space. This was done using the Scikit-Bio and Scikit-Learn packages (37,39). This procedure was also used to transform the dissimilarities learned by the Extra Trees classifier and the one-hot

encodings learned using LANDMark. The Jaccard distance was used to calculate the distances between pairs of LANDMark encodings. PCoA, using Euclidean distances, and Robust PCA (RPCA), using DEICODE, were similarly used to project the CLR and rCLR transformed data, respectively (13,23,37). Uniform Manifold Approximation and Projection (UMAP) was also used to create projections. UMAP was chosen since it, like PCoA, can be used with any dissimilarity matrix. Also, UMAP is not linearly constrained and can be used to transform new data into the projected space. The latter is important since this is not possible when using PCoA with non-Euclidean distances or RPCA. UMAP works by creating a non-linear projection. This is accomplished by first constructing a non-linear graph of the input data and then finding an embedding of this graph by minimizing the difference in the probability distribution between neighboring points in the projection and original space via gradient descent (44). Projections created using UMAP have been shown to generate better visualizations of complex data, such as microbiome data (8,44). In these experiments, the Euclidean distance was used as the 'metric' parameter for UMAP (version 0.5.3) when projecting with CLR and rCLR transformed data. When projecting dissimilarities learned by Extra Trees and LANDMark, this parameter was set to 'precomputed' and 'jaccard', respectively. Since the datasets are small, the 'n_neighbors' parameter was set to 8. Finally, the 'min_dist' parameter was set to 0.001 to encourage the formation of small clusters.

Statistical Assessment of Each Ordination

Each dissimilarity matrix was used as input for PerMANOVA (11). The p-value was calculated using 999 replications. This analysis was completed using the 'permanova' function from the Scikit-Bio package (37). The 'seaborn' and 'matplotlib' packages were used to plot the PCoA, RPCA, and UMAP projections (45,46). A Python implementation of the PROTEST

procedure for a Procrustes analysis was used to assess the concordance between pairs of projections (PCoA-PCoA, PCoA-UMAP, etc) (47). This procedure permutes the rows of one projection and computes the probability of rejecting the null hypothesis (concordance between datasets) using equation 7. In this equation m_{perm} represents the permuted test statistic and k is the number of permutations (999 in this experiment). A 1 is added in the numerator and denominator and this represents that fact that the observed statistic, m_{obs}, is a valid permutation and therefore must be considered. For a Procrustes test, the test statistic is the sum of the squared residual deviations between projections.

$$p - value = \frac{\sum_{i=1}^{k} 1\left(m_{perm,i} \leq m_{obs)}\right) + 1}{k + 1} \quad (7)$$

TreeOrdination: Automating the Identification of Informative ASVs When Using Tree Dissimilarities

It can be challenging to identify which ASVs are responsible for differentiating communities based on projections of the sample by ASV matrix. Machine learning represents a unique direction through which this problem can be addressed since decision tree ensembles can be used to efficiently capture important information about the geometric structure of the data (20,25,26). To demonstrate how these approaches can be used we have developed a workflow, TreeOrdination, which uses a relatively simple set of steps to create and analyze projections of high-dimensional amplicon sequencing data. The TreeOrdination source code and documentation are available at https://github.com/jrudar/TreeOrdination.

TreeOrdinatoin begins by creating N bootstrapped datasets. This parameter can be configured by the user and is set to 20 by default. A randomized dataset, X_R, is then constructed as described above. The original and randomized datasets are combined and used to train a

LANDMark classifier consisting of 80 decision trees (25). The neural network functionality of the classifier is disabled in the default implementation to reduce training time. This classifier is only used to distinguish the real data from its randomized counterpart. Upon completion of training, the original training data is passed through the classifier to create an embedding that describes the terminal leaves into which each sample falls. Twenty embeddings are created and concatenated together. These are then used to train an Extra Trees classifier to predict the original class labels. In addition to this, the embeddings are transformed into a smaller 15-dimensional UMAP projection (44). To create this projection the 'min_dist' parameter is set to 0.001, while the 'neighbors' and 'metric' parameters are set to 8 and "jaccard", respectively. Scikit-Learn's 'PCA' functionality is then used to construct a two-dimensional PCA projection of this space (39). The original data is then used to train an Extra Trees regression model to predict the location of each sample in the projected space. Features important for predicting the location of samples in this space are then identified using SAGE and Shapley values (48–50). These are calculated using the 'sage' (version 0.0.4) and 'shap' (version 0.41.0) packages.

Assessment of Generalization Performance

Presence-absence, center-log ratio, and robust center-log ratio transformed data were projected using TreeOrdination. To compare if the generalization performance differed between projections, each dataset was also projected into a two-dimensional space using PCoA (from Scikit-Bio) or DEICODE's RPCA function in the case of rCLR transformed data (23). The projected and original samples were then used to train Extra Trees, K-nearest neighbor, and LANDMark classifiers (24,25,39). Before training, samples were split into training and testing sets using five-fold cross-validation with two repeats. The generalization performance of each model was then calculated using the test set (39). To test if generalization performance differed

between models and transforms, a Bayesian analysis, using Nadeau and Bengio's corrected t-test implemented in the Python 'baycomp' package (version 1.0.2), was used (51). The region of practical equivalence (ROPE), which represents the probability of two models having equivalent performance, was defined as a difference in score within ±0.05. Although the choice for the size of this region is arbitrary, this size was chosen since it represents the impact of roughly two classification errors. Visualizations of the projected spaces were created using 'matplotlib' (version 3.5.2) and 'seaborn' (version 0.11.2) in Python 3.10. Finally, we performed a literature search of the best five ASVs identified using TreeOrdination to check if these ASVs are associated with the sampling locations and diseases of interest.

Results

TreeOrdination Projections Result in Well Separated Groups in Simulated Data

In the simulated tests, the analysis using the negative controls confirmed that randomized data will usually result in PerMANOVA F-statistics that are not significant regardless of transformation. However, significant differences between groups was detected in several PCoA projections of CLR transformed data (Figure S1). Upon further examination using cross-validated balanced accuracy scores, the results demonstrated that the generalization performance of supervised models trained on various splits of the negative control data was consistent with the performance of random models (Table 1, Supplementary File 2). Substantial differences between groups could be detected when positive controls were used to create TreeOrdination, PCoA, or RPCA projections (Figure S2, Table 1, Supplementary File 2). For example, the range in the F-statistics associated with the CLR transformed data, TreeOrdination projections of the CLR transformed data, and PCoA projections of the CLR transformed data was 38.56 ± 0.94, 20512.55 ± 7364.14, and 749.32 ± 36.66, respectively. Pairwise Bayesian t-tests suggested that

these differences were not due to chance and the probability of TreeOrdination producing F-statistics greater than those produced using PCoA projected transformed data is approximately 100%. This was also the case when using the presence-absence and rCLR transformed data (See Supplementary File 2). On simulated data, at least, these results support our hypothesis that learned dissimilarities and non-linear projections should be used to investigate beta diversity.

Table 1: Cross-validated generalization performance of each classification model when trained using positive and negative control data. The average score and standard deviation of the average score is reported.

Control	**Classification Model**	**Centered Log-Ratio**	**Presence-Absence**	**Robust Centered Log-Ratio**
Positive	3-KNN (PCoA/RPCA)	1	1	1
	3-KNN (Transformed Only)	0.98 ± 0.01	0.96 ± 0.04	1
	5-KNN (PCoA/RPCA)	1	1	1
	5-KNN (Transformed Only)	0.98 ± 0.01	0.98 ± 0.02	1
	Extra Trees (PCoA/RPCA)	1	1	1
	Extra Trees (Transformed Only)	0.98 ± 0.02	0.98 ± 0.02	1
	LANDMark (PCoA/RPCA)	1	1	1
	LANDMark (Transformed Only)	0.99 ± 0.02	0.99 ± 0.01	1
	Random (PCoA/RPCA)	0.49 ± 0.09	0.49 ± 0.09	0.505 ± 0.08
	Random (Transformed Only)	0.54 ± 0.08	0.50 ± 0.08	0.535 ± 0.10
	3-KNN (TreeOrdination)	1	1	1
	5-KNN (TreeOrdination)	1	1	1
	Extra Trees (TreeOrdination)	1	1	1
	LANDMark (TreeOrdination)	1	1	1
	Random (TreeOrdination)	0.53 ± 0.07	0.48 ± 0.09	0.53 ± 0.08
Negative	3-KNN (PCoA/RPCA)	0.55 ± 0.07	0.55 ± 0.07	0.49 ± 0.06
	3-KNN (Transformed Only)	0.50 ± 0.07	0.50 ± 0.09	0.49 ± 0.06
	5-KNN (PCoA/RPCA)	0.55 ± 0.08	0.55 ± 0.08	0.45 ± 0.06
	5-KNN (Transformed Only)	0.48 ± 0.07	0.48 ± 0.09	0.45 ± 0.06
	Extra Trees (PCoA/RPCA)	0.52 ± 0.08	0.47 ± 0.09	0.48 ± 0.08
	Extra Trees (Transformed Only)	0.51 ± 0.08	0.53 ± 0.06	0.47 ± 0.05
	LANDMark (PCoA/RPCA)	0.48 ± 0.06	0.50 ± 0.09	0.51 ± 0.05
	LANDMark (Transformed Only)	0.53 ± 0.08	0.55 ± 0.07	0.50 ± 0.04
	Random (PCoA/RPCA)	0.48 ± 0.07	0.54 ± 0.08	0.47 ± 0.08
	Random (Transformed Only)	0.53 ± 0.08	0.50 ± 0.07	0.51 ± 0.11
	3-KNN (TreeOrdination)	0.49 ± 0.09	0.53 ± 0.08	0.51 ± 0.07
	5-KNN (TreeOrdination)	0.47 ± 0.08	0.51 ± 0.05	0.49 ± 0.08

	Extra Trees (TreeOrdination)	0.53 ± 0.07	0.52 ± 0.09	0.49 ± 0.06
	LANDMark (TreeOrdination)	0.49 ± 0.05	0.49 ±0. 12	0.46 ± 0.09
	Random (TreeOrdination)	0.53 ± 0.06	0.49 ± 0.11	0.53 ± 0.08

All Methods Produce Projections That Will Result in Similar Interpretations

When using CLR and rCLR transformed data, a PROTEST analysis from each case study implied that the configuration of samples within each of the projected spaces is similar (Figures 1A and 2A). A small amount of disparity was detected between the projections of rCLR transformed data which likely arises due to the concatenation of randomized data to the original samples. However, this did not appear to strongly impact the overall configuration of the projections since we were able to reject the null hypothesis (permuted projections are more concordant than the observed Procrustes statistic). The p-values calculated after permutation PROTEST were below the 0.05 threshold. Due to the randomization in the selection of samples and/or features during training, there is a greater disparity between the projections of tree-based dissimilarities. Projections of LANDMark dissimilarities were usually consistent, except for a small fraction of projections created using UMAP. The extra variability impacted UMAP projections of Extra Trees dissimilarities the most and we were unable to reject the null hypothesis for a large fraction of these projections. However, this is not completely surprising given the tendency of this algorithm to produce different projections when given the same set of data (52). Also, concordance between the different pairs of projections does not imply that these projections are informative.

Fig 1: PROTEST and PerMANOVA results for the proximal colon case study. The PROTEST statistic uses the Procrustes test to measure the concordance between two multivariate datasets. One dataset is then permuted many times and the permuted statistic is calculated. The null hypothesis (no concordance between projections) can be rejected if the p-value after

permutation is less than 0.05. The PerMANOVA F-statistic is the ratio of between groups separation to within group separation. A higher F-statistic suggests greater separation between groups. Boxen plots are used to visualize the distribution of test statistics between each pair of projection methods (20 different projections per method). The results for the proximal colon are reported. In panel (A) we show the distribution of the PROTEST statistic on the top and the distribution of the p-values associated with this statistic on the bottom while panel (B) reports the distribution of the F-statistic on the top and the p-values associated with this statistic on the bottom.

Fig 2: PROTEST and PerMANOVA results for the proximal Crohn's Disease case study. The PROTEST statistic uses the Procrustes test to measure the concordance between two multivariate datasets. One dataset is then permuted many times and the permuted statistic is calculated. The null hypothesis (no concordance between projections) can be rejected if the p-value after permutation is less than 0.05. The PerMANOVA F-statistic is the ratio of between groups separation to within group separation. A higher F-statistic suggests greater separation between groups. Boxen plots are used to visualize the distribution of test statistics between each pair of projection methods (20 different projections per method). The results for the proximal colon are reported. In panel (A) we show the distribution of the PROTEST statistic on the top and the distribution of the p-values associated with this statistic on the bottom while panel (B) reports the distribution of the F-statistic on the top and the p-values associated with this statistic on the bottom.

To answer this question, it was important to consider the separation between different groups within each projection using PerMANOVA (Figures 1B and 2B). Generally, when the CLR and rCLR transformation were applied to each dataset, statistically significant differences between groups could be detected regardless of the projection method. In a small number of instances, we were not able to reject PerMANOVA's null hypothesis and this only occurred when using the UMAP projection on rCLR transformed data (Figure 2B). In the proximal colon case study, when using tree-based dissimilarities alongside the presence-absence transformation we were unable to detect differences between groups. Although this also occurred in the Crohn's disease case study, this problem was usually restricted to projections created using UMAP. Using UMAP to project the CLR, rCLR resulted in an increase in the magnitude of the F-statistic in both case studies (Figures 1B and 2B). For example, the F-Statistic for the PCoA projection of

CLR transformed CD-HC samples was observed to be 23.68 ($p \leq 0.001$) while a UMAP projection of the same data resulted in an F-Statistic of 36.58 ± 3.36 with p-values ranging between 0.001 and 0.007. However, increases in the magnitude of the F-statistic after using UMAP to transform learned dissimilarities of presence-absence transformed data were inconsistent. Finally, the F-statistics calculated using projections of tree-based dissimilarities were competitive with those produced using projections of the transformed data. Bayesian t-tests comparing the differences in the distribution of the F-statistic suggested that these observations are not due to chance (Tables S1-S6). While our results are encouraging, they should be interpreted conservatively since we only used two case-studies

The Choice of Classification Model Can Result in Significant Differences in Generalization Performance

While concordance between projections and statistics which measure the separation between groups of samples is important, one must also consider how well each transformation and projection method can model the data. This is an important consideration since it may be possible to extract useful information about which ASVs are responsible for correctly predicting the class label of unseen data and the location of this data in the projected space. Unfortunately, applying the rCLR and/or RPCA or using PCoA with non-Euclidean distances cannot transform new data into the projected space. The only way that this could occur is if the samples used to create the initial projection are combined with the testing samples. Therefore, it is likely the results for rCLR transformed data could be slightly optimistic since some data leakage occurs. TreeOrdination is more flexible since it can work with any type of transformed data.

In this work, we trained supervised models on each set of transformations and measured the generalization performance of the trained models using the balanced accuracy score. We

found that this score differed considerably depending on the classifier and transformation (Figure 3). To start, we can confidently claim that all classifiers (regardless of transformation and projection methods applied) performed better than random. When compared to training supervised models on projected or full data, we observed that applying a secondary transformation using TreeOrdination resulted in highly competitive balanced accuracy scores (Figure 3) and that this is unlikely to be due to chance. For example, in the PL-PM dataset, the generalization performance of the CLR-TreeOrdination transformation was always more likely to generate a better model than the CLR transformation alone (Table 2). In the Crohn's Disease dataset, the CLR-TreeOrdination approach was also favored although the likelihood of approaches being equivalent was greater (Table 3). In addition, applying the presence-absence transformation to the Crohn's Disease data resulted in the best generalization performance (Figure 3. A full report of the results can be found in Tables S7-S12. Finally, the first two axes of the projections created using TreeOrdination accounted for nearly all the multivariate variability in each dataset and these projections and TreeOrdination projections tended to result in projections with the largest PerMANOVA F-statistics (Figures 4 and 5). This figure also shows that representations learned by TreeOrdination are capable of transforming unseen samples and placing them near or into regions corresponding to the condition of interest. TreeOrdiantion is one of the few approaches capable of taking in any type of transformed dataset and learning a mapping that can then be applied to unseen data.

Table 2: Bayesian t-test results using balanced accuracy scores of centered log-ratio transformed proximal colon data. The best result is highlighted in bold.

Comparison A	Comparison B	Mean A ± Std Deviation	Mean B ± Std Deviation	P(A > B)	P(A = B)	P(A < B)
TreeOrdination (Default)	LANDMark (Raw)	0.864 ± 0.092	0.811 ± 0.082	**0.520**	0.442	0.038
	LANDMark (Projection)		0.817 ± 0.163	**0.491**	0.304	0.206
	3-KNN (Raw)		0.756 ± 0.067	**0.849**	0.145	0.006

	Extra Trees (Raw)		0.808 ± 0.109	**0.532**	0.398	0.070
	TreeOrdination (5-KNN)		0.878 ± 0.099	0.054	**0.771**	0.174
	Extra Trees (Projection)		0.758 ± 0.108	**0.815**	0.175	0.011
	5-KNN (Raw)		0.758 ± 0.093	**0.761**	0.209	0.030
	TreeOrdination (3-KNN)		0.856 ± 0.093	0.207	**0.664**	0.129
	5-KNN (Projection)		0.739 ± 0.081	**0.902**	0.095	0.003
	3-KNN (Projection)		0.758 ± 0.081	**0.840**	0.154	0.006
	TreeOrdination (LANDMark)		0.831 ± 0.094	0.348	**0.619**	0.033

Table 3: Bayesian t-test results using balanced accuracy scores of centered log-ratio transformed Crohn's Disease data. The best result is highlighted in bold.

Comparison A	Comparison B	Mean A ± Std Deviation	Mean B ± Std Deviation	P(A > B)	P(A = B)	P(A < B)
TreeOrdination (Default)	LANDMark (Raw)	0.856 ± 0.120	0.863 ± 0.130	0.101	**0.733**	0.166
	LANDMark (Projection)		0.847 ± 0.109	0.047	**0.941**	0.011
	3-KNN (Raw)		0.822 ± 0.137	0.376	**0.574**	0.051
	Extra Trees (Raw)		0.838 ± 0.099	0.339	**0.467**	0.194
	TreeOrdination (5-KNN)		0.878 ± 0.130	0.023	**0.770**	0.207
	Extra Trees (Projection)		0.817 ± 0.114	0.400	**0.579**	0.022
	5-KNN (Raw)		0.823 ± 0.140	0.381	**0.538**	0.081
	TreeOrdination (3-KNN)		0.856 ± 0.120	0	**1**	0
	5-KNN (Projection)		0.856 ± 0.120	0	**1**	0
	3-KNN (Projection)		0.81 ± 0.131	0.462	**0.522**	0.016
	TreeOrdination (LANDMark)		0.856 ± 0.120	0	**1**	0

Fig 3: Distribution of Balanced Accuracy Scores for each type of transformation and model
The distribution of balanced accuracy scores changes depending on the type of transformation and model in the (A) proximal colon and (B) Crohn's Disease datasets.

Fig 4: Comparison of each projection method. Data from the proximal colon is found along the top row while Crohn's disease data is found along the bottom row. (A) Presence-Absence transformation. (B) Robust Centered Log-Ratio transformation. (C) TreeOrdination transformation, (D) Centered Log-Ratio transformation. (E) Presence-Absence transformation. (F) Robust Centered Log-Ratio transformation. (G) TreeOrdination transformation, (H) Centered Log-Ratio transformation.

Fig 5: Distribution of the PerMANOVA F-Statistic when using TreeOrdination, PCoA projection, and no projection. The distribution of the PerMANOVA F-Statistic under each condition when using data from the (A) proximal colon and (B) Crohn's Disease datasets.

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

Similar to the RPCA (or PCA in the case of CLR transformed data), TreeOrdination can identify which features were used by the model to generate the ordination. Unfortunately, the calculation of SAGE values using the full implementation of TreeOrdination is very expensive. To address this issue, we employed an approximation that uses the original transformed data to train an Extra Trees regression model to predict the location of each sample in the projected space created using the UMAP/PCA transformations. By taking this approach, it becomes feasible to calculate SAGE scores. These scores quantify how important each ASV is to the model (Figures 6A and 7A). Test samples projected using the approximation roughly appear near other samples of the same class (Figure 6B and Figure 7B). However, the configuration of test points in the approximate projection does not fully mirror that which is produced using the full TreeOrdination pipeline. To test if there is a significant difference in configurations between the full and approximate projections, we conducted a PROTEST analysis. PROTEST results suggested that the evidence for the alternative hypothesis (no difference between configurations) exists for both case studies (Crohn's Disease: $m_{obs} = 0.02$, $p \leq 0.001$; Proximal Colon: $m_{obs} =$

0.10, $p \leq 0.001$). Since both projections can be interpreted similarly, it is logical to ask which factors impact the location of each sample in the projected space. Shapley scores represent a powerful way to do this and were used in this work to identify these ASVs (Figure 6CD and 7CD).

Fig 6: Global and local feature importance scores for the proximal lumen dataset. (A) Global feature importance scores calculated using SAGE. More important features will have a higher SAGE value. (B) Configuration of the actual (left) and approximate (right) projections the projection produced using TreeOrdination. Using the final TreeOrdination projection, an Extra Trees Regressor was trained to predict the location of samples in the projected space given the original centered log-ratio transformed training data. (C) A waterfall plot illustrating which ASVs impact the projected location of an individual mucosal sample along the x-axis of approximate projection in panel (B). ASVs with positive Shapley scores push the model towards the positive part of the axis while those with negative Shapley scores push the model towards the negative part of the axis. (D) Same as panel (C), only this time the sample is from the lumen.

Fig 7: Global and local feature importance scores for the Crohn's Disease dataset. (A) Global feature importance scores calculated using SAGE. More important features will have a higher SAGE value. (B) Configuration of the actual (left) and approximate (right) projections the projection produced using TreeOrdination. Using the final TreeOrdination projection, an Extra Trees Regressor was trained to predict the location of samples in the projected space given the original centered log-ratio transformed training data. (C) A waterfall plot illustrating which ASVs impact the projected location of an individual Crohn's Disease sample along the x-axis of approximate projection in panel (B). ASVs with positive Shapley scores push the model towards the positive part of the axis while those with negative Shapley scores push the model towards the negative part of the axis. (D) Same as panel (C), only this time the sample is from a healthy control.

For the microbial community of the distal human colon, TreeOrdination suggested that several bacteria could be used to predict the location of proximal mucosal and luminal samples in the projected space (Figure 6A). While not an exhaustive list, our approach suggested that *Ihubacter*, *Lachnospiracea*, *Firmicutes*, and *Lactobacillus* are necessary to place these samples. To understand the samples in each case study in better detail, it was important to quantify how ASVs impact the location of individual samples (Figures 6CD and 7CD). We found that the

presence of an unclassified ASV (ASV 1141), strongly determines the location of samples across the first principal axis. Furthermore, other ASVs belonging to the *Eubacteriaceae*, *Lachnospiracea*, and *Clostridiales* all strongly impact the ability of the model to place samples. Finally, we observed that the combined contribution of the other features helped impact the overall prediction (Figures 6 C and D). For example, in the approximate projection learned by the Extra Trees model, the model predicted that a mucosal sample should have a value of 0.902 along the PC1. Without any consideration of the contribution of the other 1421 features, the 10 features with the highest magnitude Shapley values would place this sample closer to luminal samples (Figure 6C). These results demonstrate that the importance of an individual ASV should not be investigated in isolation, but its role needs to be considered in a broader context. We also identified several ASVs which are important for distinguishing between CD and HC samples. *Parolsenella*, *Ruminococcaceae*, *Corynebacterium*, *Clostridiales*, *Coprococcus*, *Catenibacterium*, and *Blautia* were among the top ASVs which can be used to predict the location of each sample in the projected space (Figure 7A, C, and D). Interestingly, *Parolsenella* was not identified by Forbes et al. (2018) as an important indicator of Crohn's Disease. Since then, evidence has been found to suggest that it may have a role in the development of MS, which is itself a class of IMID (53). Of the other bacteria identified by TreeOrdination, *Ruminococcaceae*, *Corynebacterium*, *Clostridiales*, *Coprococcus*, *Catenibacterium*, and *Blautia* have known associations with Crohn's Disease and/or other human (immune-mediated) inflammatory diseases (53–57).

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. The choice of transformation and dissimilarity measure is an important consideration when investigating microbiome data. It has long been known that these choices can influence our measurement and interpretation of the main gradients influencing the structure of communities and taxonomic similarity between pairs of samples (3,10,14,58–60). To combat these problems a multitude of transformations, dissimilarity measures, and ordination approaches have been developed or proposed. Each combination of these approaches has its strengths and can be used to summarize and investigate different aspects of community composition (3,59,61). For example, some dissimilarity measures, such as Jaccard and Canberra, perform well for detecting clusters and others are more applicable to detecting ecological gradients (4). While the application of these transformations and/or using specific dissimilarities can help bring clarity to the analysis, a review of the literature suggests there is no consensus on which approach is best (3,14,59,61). Likely, the best combination of approaches will be dataset-dependent.

Recent innovations and developments in the field of machine learning have resulted in a rapid increase in the number of studies employing these approaches. In the context of this work, approaches which use tree-based dissimilarities to identify the pairwise relationships between samples are very interesting since these dissimilarities are learned (20–22,25,26). Since decision-tree models are not linearly constrained, these models could arrive at a better representation of the underlying geometry of the samples (5,20,25,42). Our work supports this hypothesis due to the improvements we observed in the F-statistics and overall configuration of samples in the projected spaces when using non-linear transformations. Using the workflow developed here, TreeOrdination, we show that complex machine learning models can be used to generate good projections and identify important features which were used to create these projections. In addition, the predictive models learned using TreeOrdination transformed data can learn an

internal representation that generalizes well to unseen data. This is an important capability as RPCA and PCoA (with non-Euclidean distances) cannot be used to project unseen data into a lower dimensional space (23). Our results support the growing body of work that demonstrates that tree-based dissimilarities are useful for capturing and visualizing relationships between samples and that biological datasets tend to lie on a non-linear manifold (7,8). For example, the first two principal axes in each case study's TreeOrdination projection accounted for nearly all the variance in each dataset and the placement of samples along this axis was usually associated with sampling location or disease state. At first glance, it appears as though this is true for the other projections (Figure 4). However, in each case study the first two principal axes of each PCoA projection accounted for a small fraction of the total variance, suggesting that these projections do not adequately explain the data. Also, while the first two principal axes of each RPCA explained 100% of the variation within each dataset, the configuration of samples within RPCA projections was more ambiguous than those produced using TreeOrdiantion projections. In particular, there appeared to be more overlap between samples from each class.

We conjecture that the linearly constrained methods, such as PCoA, cannot account for the complex geometry of the input space since these methods usually result in projections with low amounts of explained variation along each principal axis and are associated with smaller lower PerMANOVA F-Statistics. Indeed, it is trivial to show that PCoA is incapable of generating good projections of highly non-linear toy data. When moving beyond toy data, there exists a considerable amount of work advocating for the adoption of newer graph-based methods for the analysis of biological data. For example, Elhaik (2022) demonstrates that the validity of the results produced from one commonly used linear dimensionality reduction algorithm, PCA, can be easily called into question due to artifacts within the data and the ease by which it can be

manipulated to produce favorable results (6). In addition, there is an ongoing debate around how useful a projection is when the amount of explained variance for the first few axes is very low (arbitrarily defined as less than 60%). Investigations using non-linear transformations of real microbiome data using UMAP also supports this hypothesis. Armstrong et al. (2021) demonstrated the ability of UMAP to resolve biologically relevant clusters and ecological gradients using data from the human microbiome project and the ’88 soils’ dataset while Dorrity et al. (2020) showed that clustering samples in UMAP space identified protein-protein interactions more effectively than clustering the same data in PCA space (7,8).

In addition to working well with non-linear data, learned dissimilarities can be advantageous since more “traditional” measures of dissimilarity have difficulty capturing potential dependencies between ASVs. LANDMark and Extra Trees reduce this problem due to their hierarchical nature. For example, in an Extra Trees model if the abundance and/or presence of one ASV depends on another we should expect to see sub-trees using these ASVs to occur more frequently. We also conjecture that LANDMark and other tree-dissimilarity measures work well due to their ability to handle noisy data. Noise is often introduced due to the inclusion of uninformative ASVs, potentially informative but highly variable ASVs, splitting a single genome, and missing data due to incomplete sampling or the limitations of library amplification and sequencing (10,14,62,63). Learned dissimilarity measures, such as those explored here, may be capable of identifying and reducing the impact that uninformative ASVs exert when measuring dissimilarity. For example, in an Extra Trees classifier, only the best randomly selected cut point from a set of randomly selected ASVs is chosen to split each node (21). When choosing splits in this manner the impact of uninformative ASVs tends to be minimized since they are not selected as often. In addition, after the decision for the node is made, each set of

samples is partitioned into one of two child nodes irrespective of the magnitude of the unselected features. This likely results in a strong regularizing effect across the entire forest (64). LANDMark extends this idea by identifying which (regularized) linear or non-linear model is best at discriminating between classes using a randomly selected subset of ASVs (25).

While a deeper and more detailed analysis of the ASVs identified here is outside the scope of this work, TreeOrdination identified ASVs belonging to taxa known to be associated with the proximal colon and Crohn's Disease. The primary focus of this work is to begin an investigation into the role, advantages, and disadvantages that learned dissimilarities and non-linear embeddings have in the interpretation of amplicon sequencing data. Our main contributions are our results showing that tree-based dissimilarities can effectively capture biologically meaningful information about samples and a proof-of-concept workflow that uses LANDMark to learn a tree-based dissimilarity and low-dimensional representation of the original samples. These representations can then be used to create predictive models which can subsequently be analyzed using modern introspection methods to identify predictive sets of ASVs. We would also like to note that, although simple dissimilarity and distance measures are unlikely to be expressive enough to account for the nuances and complexities of modern amplicon sequencing datasets, their use should not be outright dismissed. For one, the calculation of pairwise dissimilarities using formulae is orders of magnitude faster than TreeOrdination or other learned dissimilarities. These measures and metrics can also be used for exploratory purposes, an analysis that may be time-consuming to do with learned dissimilarities. Finally, if the projections produced using these dissimilarities account for a significant amount of explained variation, there is likely little to be gained by applying more complex approaches.

Conclusions and Future Work

Our work has shown that unsupervised LANDMark models can learn effective dissimilarity matrices. When paired with modern dimensionality reduction approaches, such as UMAP, an informative lower-dimensional projection can be created. However, our work does have some important limitations. First, the datasets used here are small. While we believe that our results should hold with larger datasets, further testing is needed to determine if this is true. Secondly, we only tested transformations and metrics which represent a more statistically sound way to analyze compositional datasets. Other transformations, such as the geometric mean of pairwise ratios, the isometric log-ratio, and cumulative sum scaling transformations were not tested since we wanted to focus on three commonly used approaches, two of which account for the compositionality inherent to this type of data (1,16,65). However, we hope to conduct a more comprehensive analysis using a wider variety of transformation procedures in a more extensive future comparative analysis. Finally, while we believe that tools such as LANDMark and TreeOrdination are capable, the work presented here speaks to a broader conversation that is occurring within the scientific community analyzing ecological and microbiome data. Specifically, more evidence is coming to light which demonstrates the inadequacy of using PCoA in the study of beta-diversity and that better insights into the data are likely to be generated using non-linear tools. In addition, our work demonstrates the important role that ML will play in the near future since ML can be used as an alternative to validate the results from statistical approaches, such as PerMANOVA. For example, statistically significant PerMANOVA results were recorded in some of the negative control tests. If taken in isolation, these results could be misinterpreted. By using additional lines of evidence, such as the generalization performance of cross-validated predictive models, we can come to a conclusion that suggests that the data may not have predictive value. Therefore, we advocate for taking a

more holistic investigative approach. In doing so better conclusions can be drawn since multiple measures of data quality are combined and weighed together. Although this is often more difficult when working with ML models, since they are notoriously difficult to interpret, an increasingly growing number of tools have been created to allow us to peek into these black boxes.

Acknowledgments

We would like to thank Dr. Katie McGee and Dr. Terri M. Porter for their thoughtful discussions during the development of TreeOrdination and LANDMark. We would also like to thank Dr. Katie McGee for her suggestions on where to improve the message of our manuscript. Finally, we extend our thanks to our reviewers for their helpful and insightful comments.

References

1. Silverman JD, Washburne AD, S M, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife. 2017;6:21887.

2. Chen B, He X, Pan B, Zou X, You N. Comparison of beta diversity measures in clustering the high-dimensional microbial data. PLoS One. 2021;16(2):e0246893.

3. 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).

4. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Methods. 2010 Oct;7(10):813–9.

5. Kubinski R, Djamen-Kepaou JY, 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

6. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Scientific Reports. 2022 Aug 29;12(1):14683.

7. Dorrity MW, Saunders LM, Queitsch C, Fields S, Trapnell C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nature Communications. 2020 Mar 24;11(1):1537.

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

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

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

701 11. Anderson MJ, Walsh DCI. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous
702 dispersions: What null hypothesis are you testing? Ecological Monographs. 2013;83:557–74.

703 12. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nature
704 Communications. 2020 Jul 14;11(1):3514.

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

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

709 15. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker
710 discovery and explanation. Genome Biology. 2011;12:60.

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

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

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

717 19. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life
718 Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform.
719 2012;14(3):315–26.

720 20. Rhodes JS, Cutler A, Moon KR. Geometry- and Accuracy-Preserving Random Forest Proximities
721 [Internet]. arXiv; 2022. Available from: https://arxiv.org/abs/2201.12682

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

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

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

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

728 25. Rudar J, Porter TM, Wright M, Golding GB, Hajibabaei M. LANDMark: An ensemble approach to the
729 supervised selection of biomarkers in high-throughput sequencing data. BMC Bioinformatics.
730 2022;23(1):110.

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

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

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

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

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

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

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

749 33. Wallen ZD. Comparison study of differential abundance testing methods using two large Parkinson
750 disease gut microbiome datasets derived from 16S amplicon sequencing. BMC Bioinformatics.
751 2021 May 25;22(1):265.

752 34. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.
753 PLoS Computational Biology. 2014;10(4):1003531.

754 35. Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential
755 abundance analysis. npj Biofilms and Microbiomes. 2020 Dec 2;6(1):60.

756 36. Greenacre M, Martínez-Álvaro M, Blasco A. Compositional Data Analysis of Microbiome and Any-
757 Omics Datasets: A Validation of the Additive Logratio Transformation. Frontiers in Microbiology
758 [Internet]. 2021;12. Available from:
759 https://www.frontiersin.org/article/10.3389/fmicb.2021.727398

760 37. team T scikit-bio development. scikit-bio: A Bioinformatics Library for Data Scientists, Students,
761 and Developers [Internet]. 2022. Available from: http://scikit-bio.org

38. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology. 2003 Apr 1;35(3):253–78.

39. 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.

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

41. 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

42. Dalleau K, Couceiro M, Smail-Tabbone M. Unsupervised Extremely Randomized Trees. In: Phung D, Tseng VS, Webb GI, Ho B, Ganji M, Rashidi L, editors. Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing; 2018. p. 478–89.

43. Camargo A. PCAtest: testing the statistical significance of Principal Component Analysis in R. PeerJ. 2022;10:e12967.

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

45. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007;9(3):90–5.

46. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021.

47. Jackson DA. PROTEST: A PROcrustean Randomization TEST of community environment concordance. Écoscience. 1995;2(3):297–303.

48. 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.

49. 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.

50. Covert IC, Lundberg S, Lee SI. Understanding Global Feature Contributions with Additive Importance Measures. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. (NIPS'20).

51. 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.

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

798 53. Perez-Muñoz ME, Sugden S, Harmsen HJM, Hart BA 't, Laman JD, Walter J. Nutritional and
799 ecological perspectives of the interrelationships between diet and the gut microbiome in multiple
800 sclerosis: Insights from marmosets. iScience. 2021;24(7):102709.

801 54. Assa A, Butcher J, Li J, Elkadri A, Sherman PM, Muise AM, et al. Mucosa-Associated Ileal Microbiota
802 in New-Onset Pediatric Crohn's Disease. Inflammatory Bowel Diseases. 2016 Jun;22(7):1533–9.

803 55. The HC, Florez de Sessions P, Jie S, Pham Thanh D, Thompson CN, Nguyen Ngoc Minh C, et al.
804 Assessing gut microbiota perturbations during the early phase of infectious diarrhea in Vietnamese
805 children. Gut Microbes. 2018 Jan 2;9(1):38–54.

806 56. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, Antibiotics,
807 and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host
808 Microbe. 2015 Oct 14;18(4):489–500.

809 57. Diling C, Xin Y, Chaoqun Z, Jian Y, Xiaocui T, Jun C, et al. Extracts from Hericium erinaceus relieve
810 inflammatory bowel disease by regulating immunity and gut microbiota. Oncotarget. 2017 Oct
811 17;8(49):85838–57.

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

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

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

818 61. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for
819 normalizing microbiome data: An ecological perspective. Methods in Ecology and Evolution.
820 2019;10(3):389–400.

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

823 63. Edgar RC. UCHIME2: Improved chimera detection for amplicon sequences. bioRxiv [Internet].
824 2016; Available from: https://www.drive5.com/uchime/

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

827 65. Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. GMPR: A robust normalization method for
828 zero-inflated count data with application to microbiome sequencing data. PeerJ. 2018;6:e4600.

829
830

Supporting Information

Additional File 1 – Supplementary Tables 1 to 12

Additional File 2 – Positive and Negative Control Results

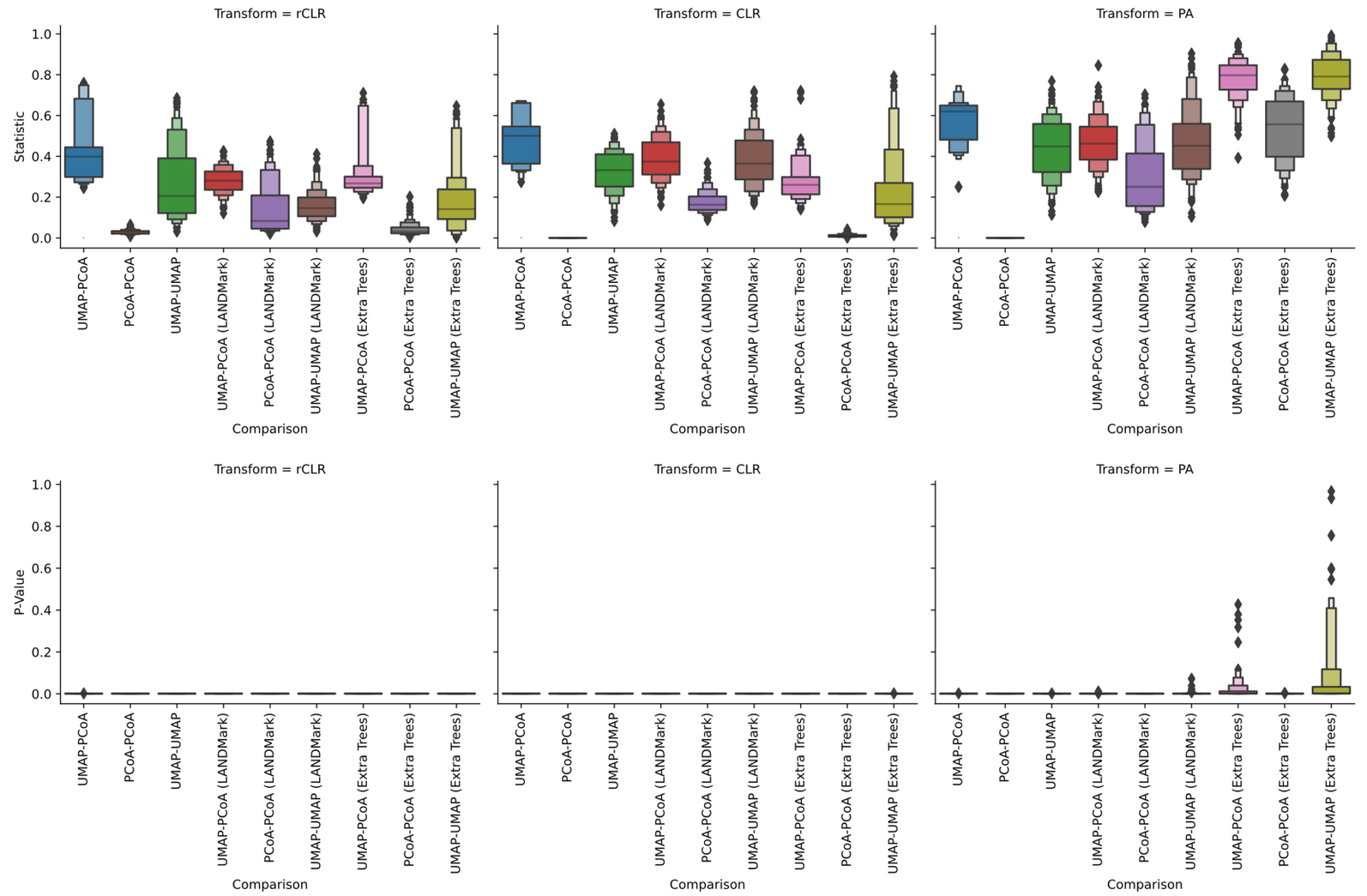
Additional File 3 – Supplementary Figures 1 and 2

GitHub Link 1 – Raw Data and Code Used to Run the Analysis

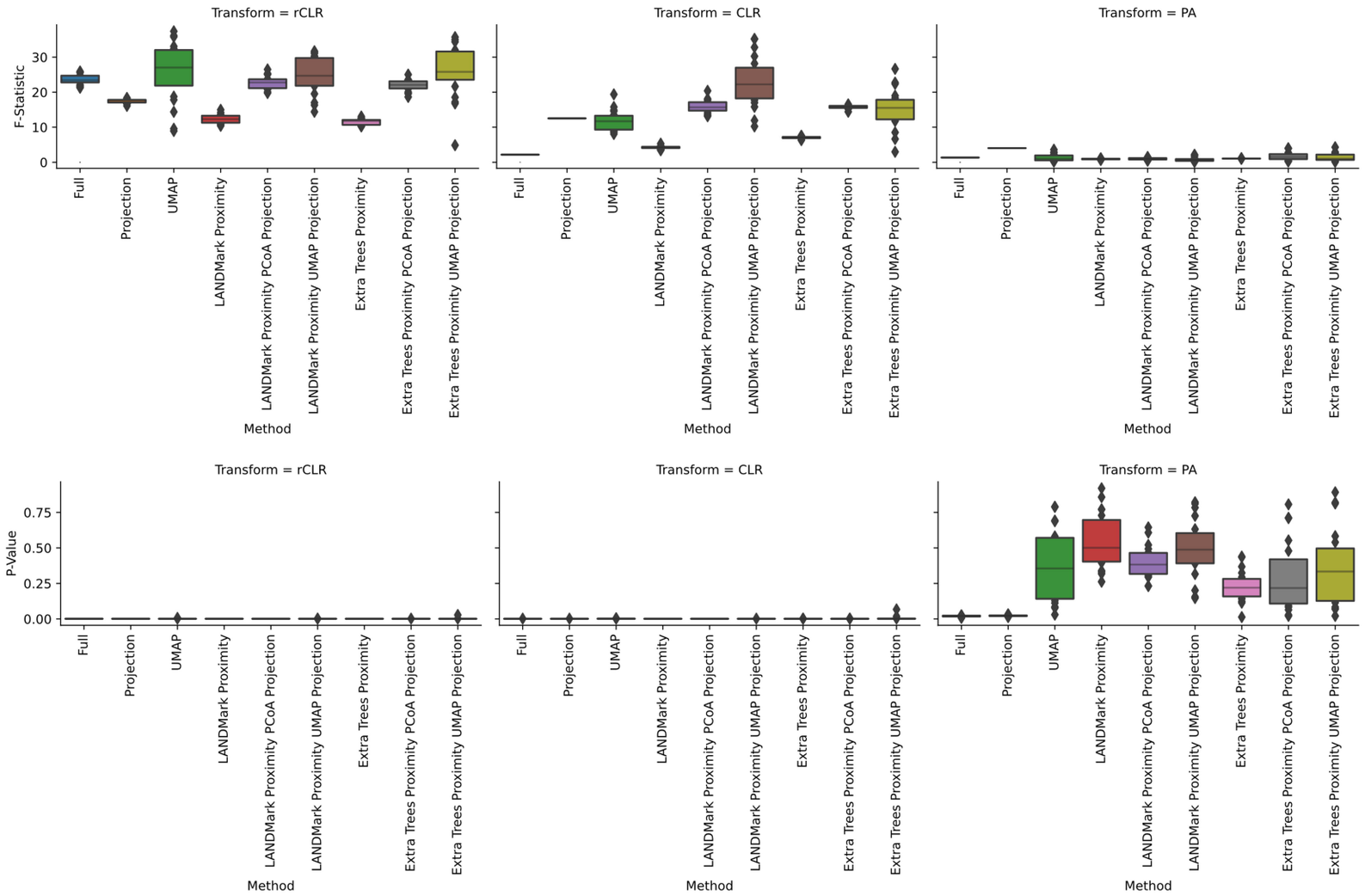
(https://github.com/jrudar/Unsupervised-Decision-Trees)

GitHub Link 2 – TreeOrdination Package (https://github.com/jrudar/TreeOrdination)

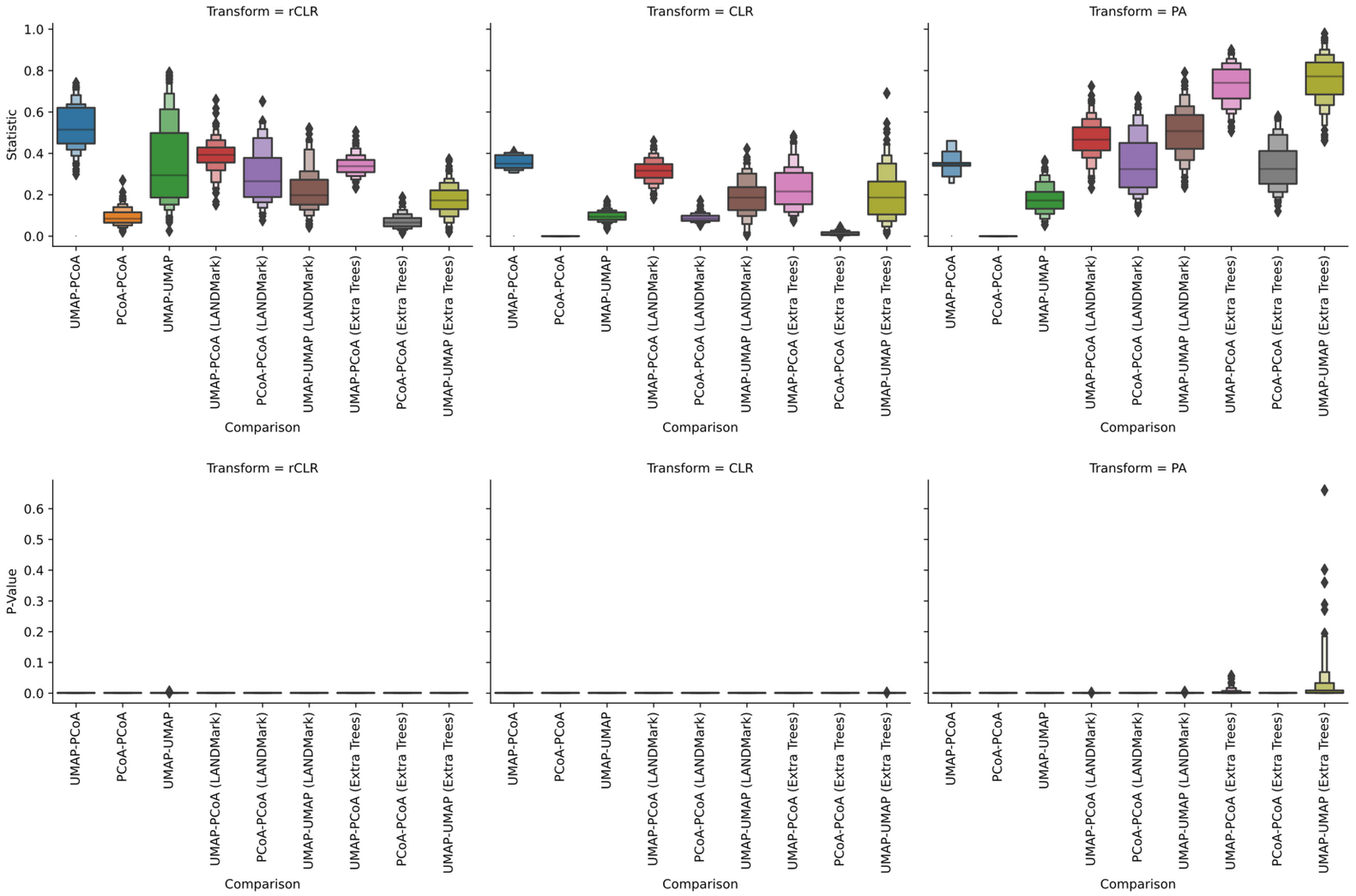
A


Transform = rCLR
Transform = CLR
Transform = PA
Statistic
P-Value
Comparison
UMAP-PCoA
PCoA-PCoA
UMAP-UMAP
UMAP-PCoA (LANDMark)
PCoA-PCoA (LANDMark)
UMAP-UMAP (LANDMark)
UMAP-PCoA (Extra Trees)
PCoA-PCoA (Extra Trees)
UMAP-UMAP (Extra Trees)

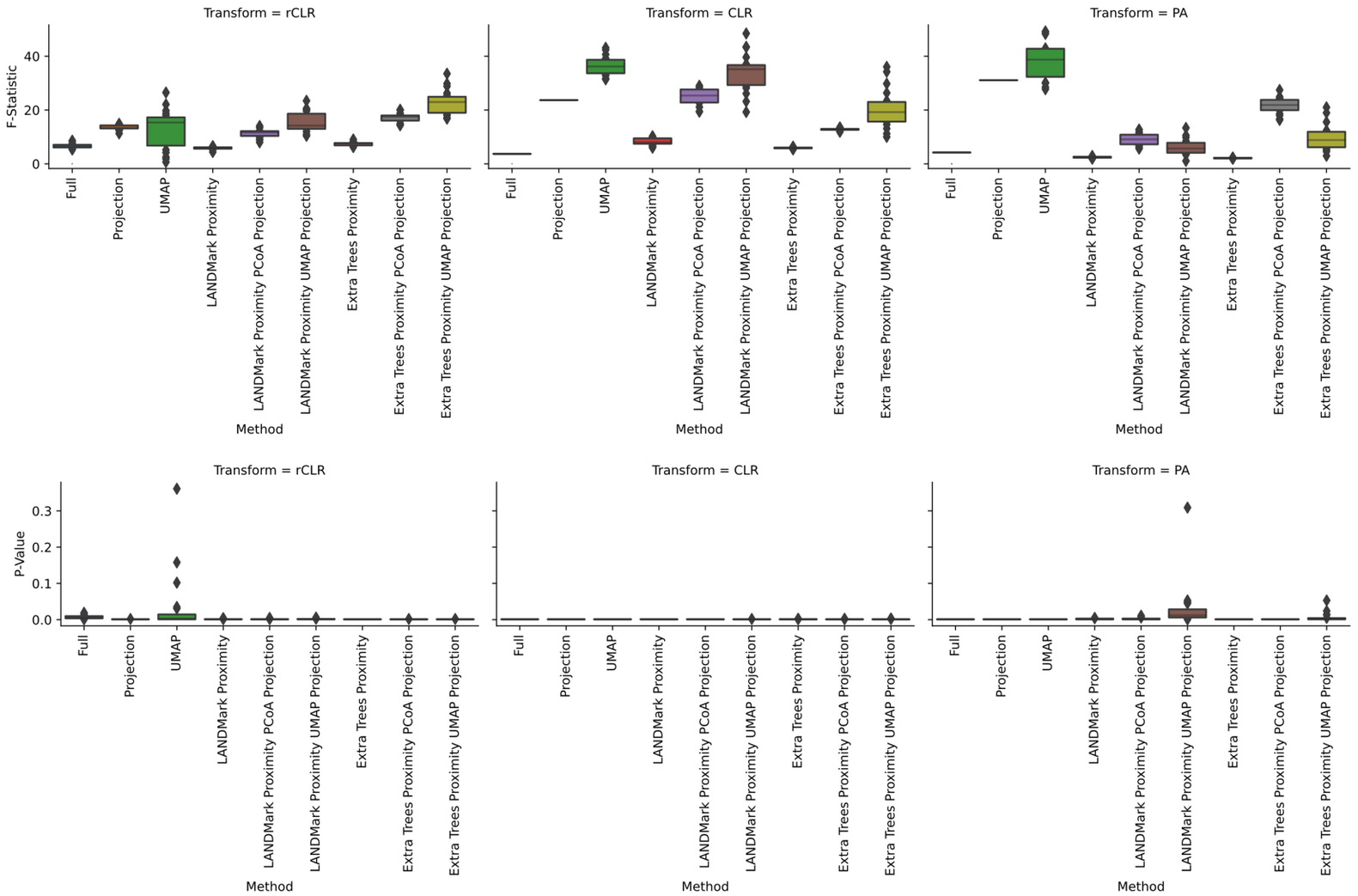
B


Transform = rCLR
Transform = CLR
Transform = PA
F-Statistic
P-Value
Method
Full
Projection
UMAP
LANDMark Proximity
LANDMark Proximity PCoA Projection
LANDMark Proximity UMAP Projection
Extra Trees Proximity
Extra Trees Proximity PCoA Projection
Extra Trees Proximity UMAP Projection

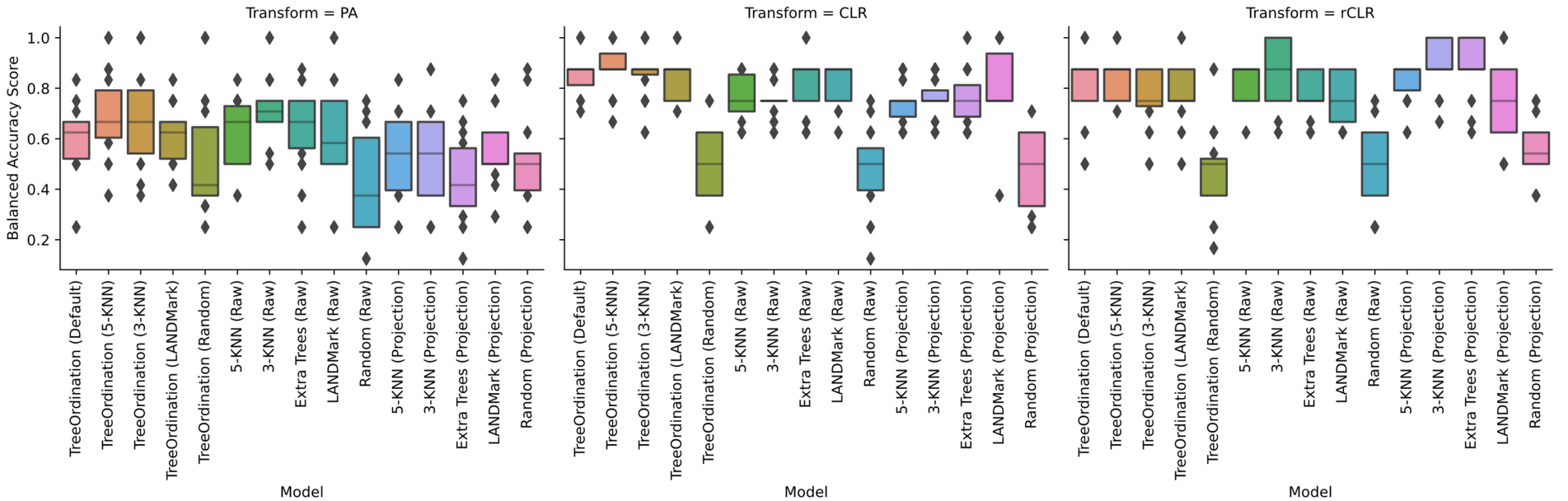
A


Transform = rCLR
Transform = CLR
Transform = PA
Statistic
P-Value
Comparison
UMAP-PCoA
PCoA-PCoA
UMAP-UMAP
UMAP-PCoA (LANDMark)
PCoA-PCoA (LANDMark)
UMAP-UMAP (LANDMark)
UMAP-PCoA (Extra Trees)
PCoA-PCoA (Extra Trees)
UMAP-UMAP (Extra Trees)

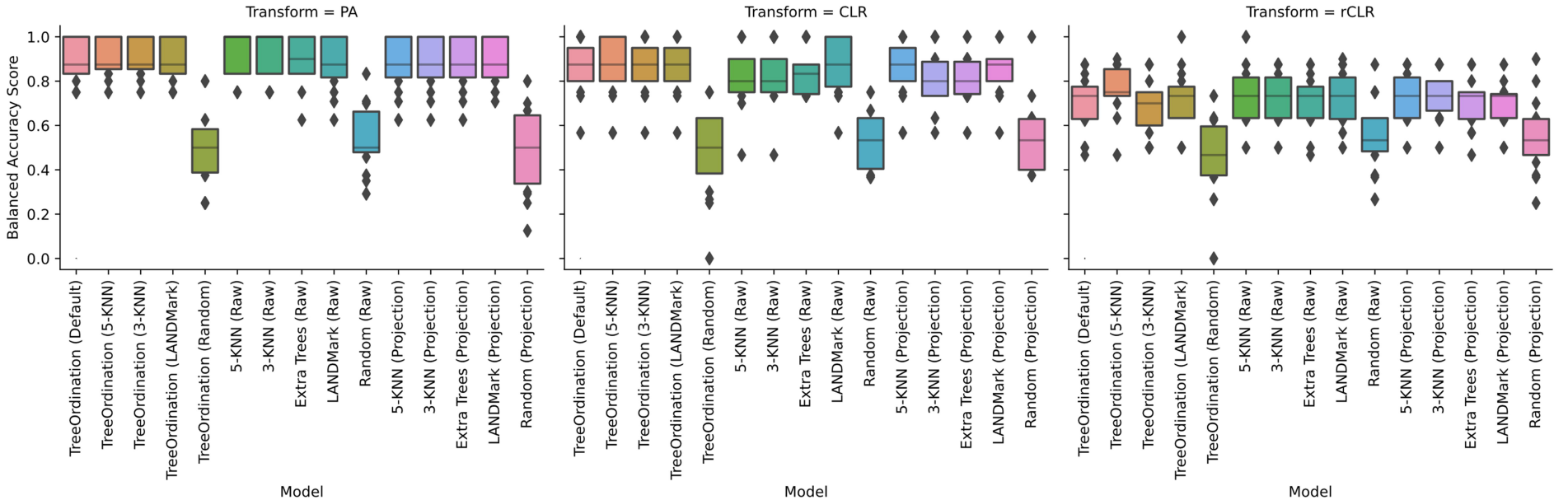
B

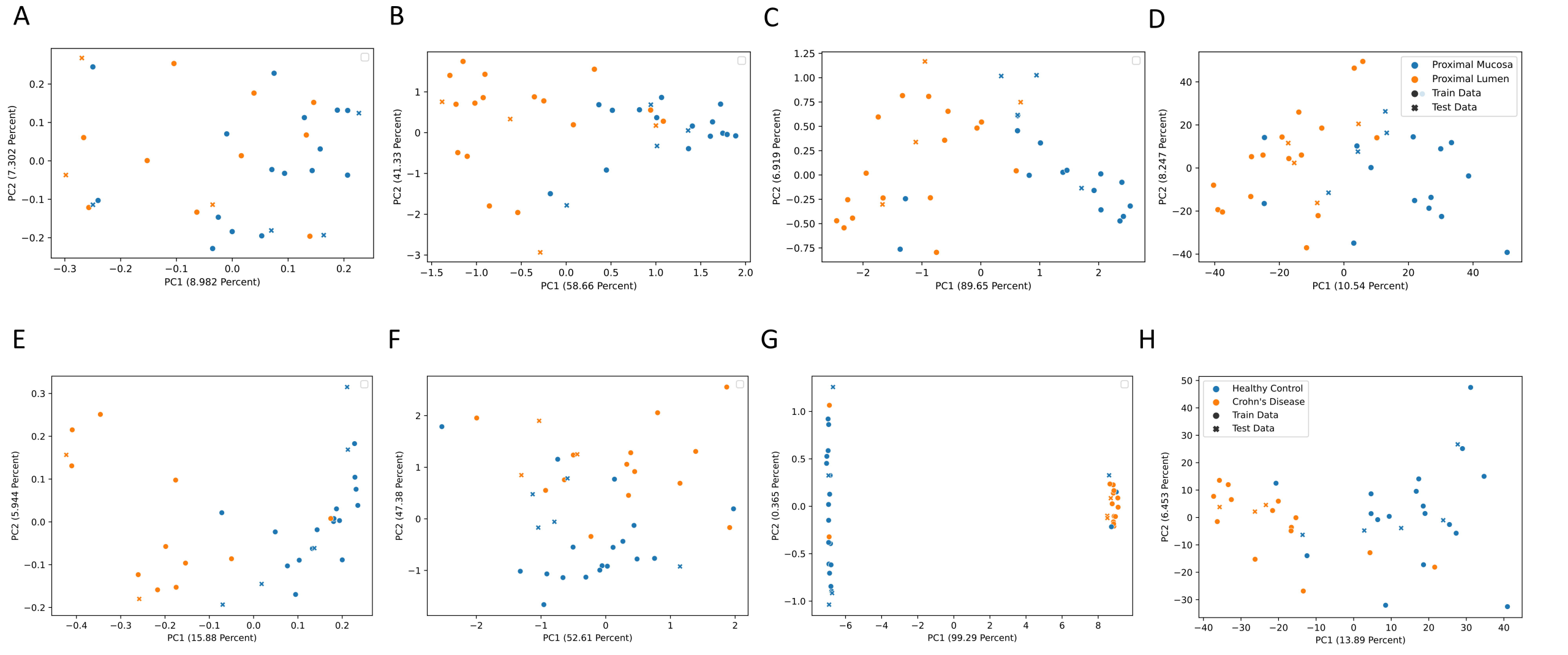

Transform = rCLR
Transform = CLR
Transform = PA
F-Statistic
P-Value
Method
Full
Projection
UMAP
LANDMark Proximity
LANDMark Proximity PCoA Projection
LANDMark Proximity UMAP Projection
Extra Trees Proximity
Extra Trees Proximity PCoA Projection
Extra Trees Proximity UMAP Projection

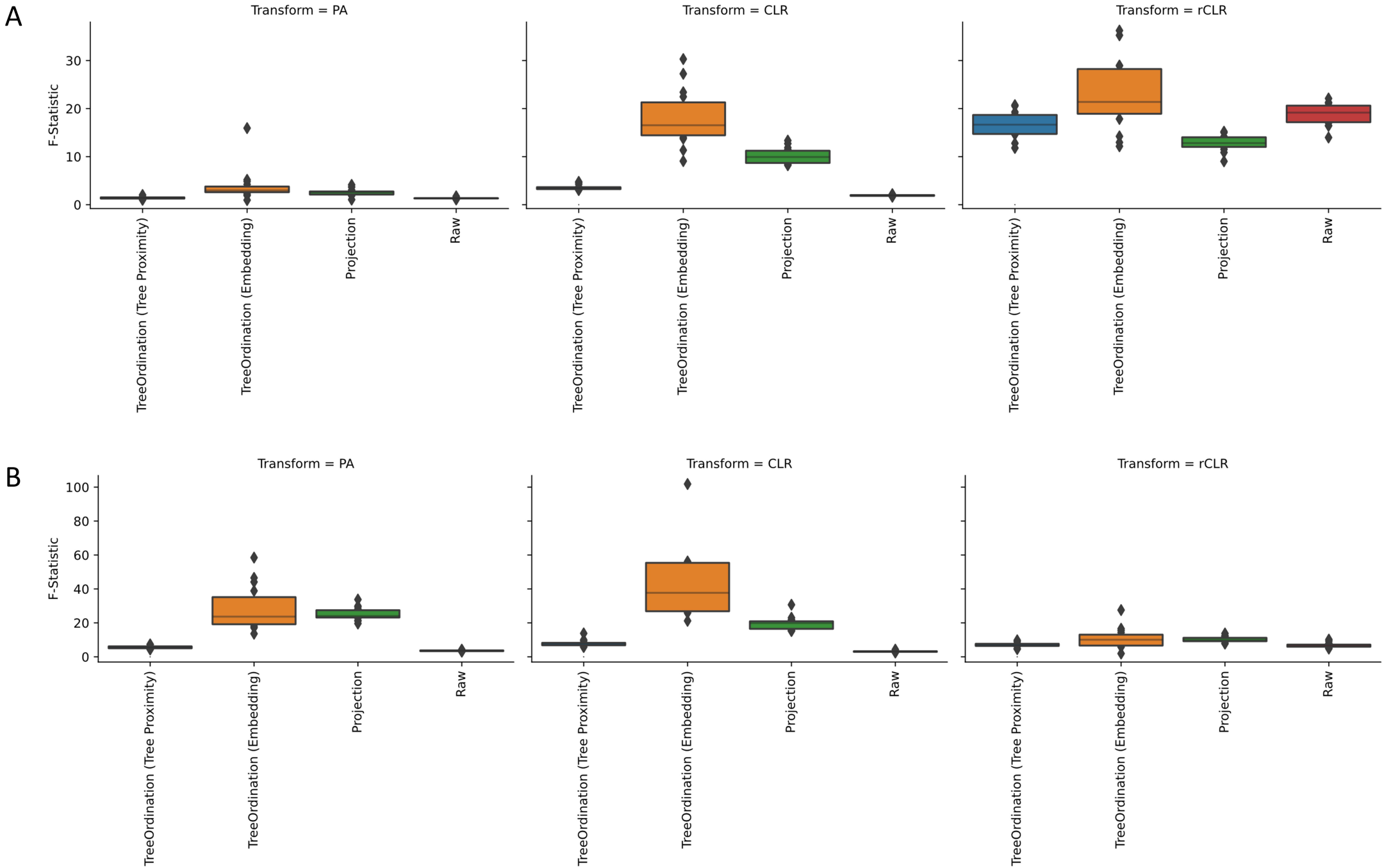
A

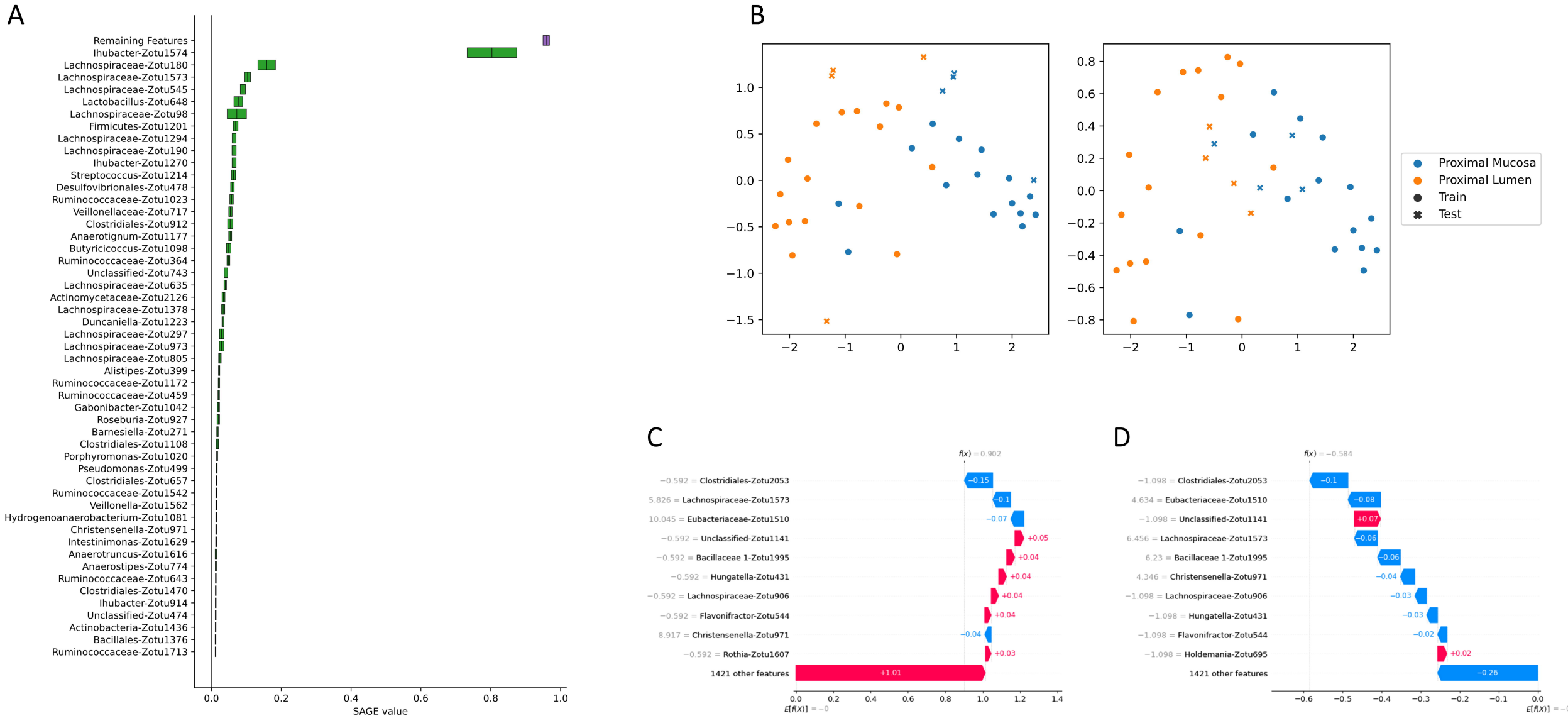

Transform = PA
Transform = CLR
Transform = rCLR
Balanced Accuracy Score
1.0
0.8
0.6
0.4
0.2
TreeOrdination (Default)
TreeOrdination (5-KNN)
TreeOrdination (3-KNN)
TreeOrdination (LANDMark)
TreeOrdination (Random)
5-KNN (Raw)
3-KNN (Raw)
Extra Trees (Raw)
LANDMark (Raw)
Random (Raw)
5-KNN (Projection)
3-KNN (Projection)
Extra Trees (Projection)
LANDMark (Projection)
Random (Projection)
Model

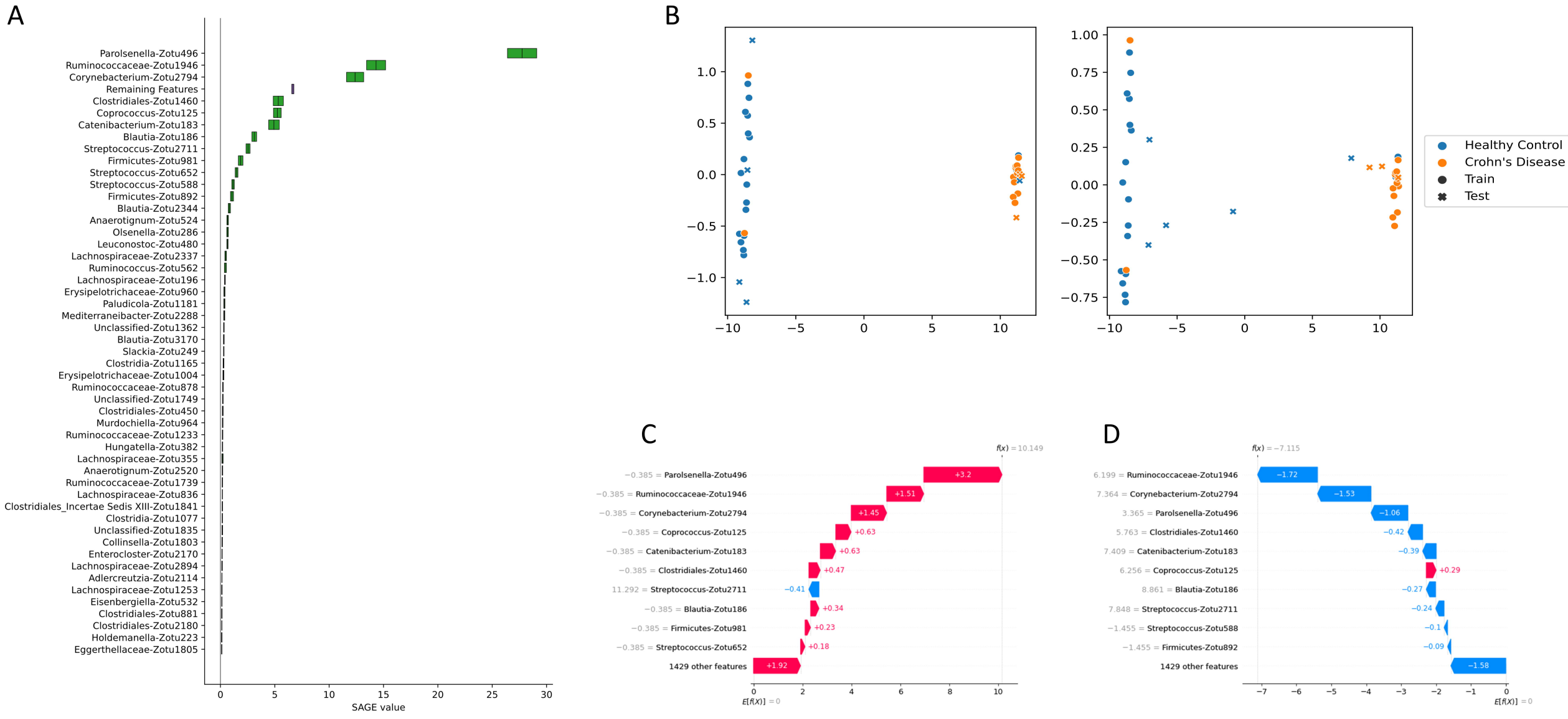
B


Transform = PA
Transform = CLR
Transform = rCLR
Balanced Accuracy Score
1.0
0.8
0.6
0.4
0.2
0.0
TreeOrdination (Default)
TreeOrdination (5-KNN)
TreeOrdination (3-KNN)
TreeOrdination (LANDMark)
TreeOrdination (Random)
5-KNN (Raw)
3-KNN (Raw)
Extra Trees (Raw)
LANDMark (Raw)
Random (Raw)
5-KNN (Projection)
3-KNN (Projection)
Extra Trees (Projection)
LANDMark (Projection)
Random (Projection)
Model


A
PC2 (7.302 Percent)
PC1 (8.982 Percent)
B
PC2 (41.33 Percent)
PC1 (58.66 Percent)
C
PC2 (6.919 Percent)
PC1 (89.65 Percent)
D
PC2 (8.247 Percent)
PC1 (10.54 Percent)
Proximal Mucosa
Proximal Lumen
Train Data
Test Data
E
PC2 (5.944 Percent)
PC1 (15.88 Percent)
F
PC2 (47.38 Percent)
PC1 (52.61 Percent)
G
PC2 (0.365 Percent)
PC1 (99.29 Percent)
H
PC2 (6.453 Percent)
PC1 (13.89 Percent)
Healthy Control
Crohn's Disease
Train Data
Test Data


A
Transform = PA
Transform = CLR
Transform = rCLR
F-Statistic
0
10
20
30
TreeOrdination (Tree Proximity)
TreeOrdination (Embedding)
Projection
Raw
B
Transform = PA
Transform = CLR
Transform = rCLR
F-Statistic
0
20
40
60
80
100
TreeOrdination (Tree Proximity)
TreeOrdination (Embedding)
Projection
Raw


A
Remaining Features
Ihubacter-Zotu1574
Lachnospiraceae-Zotu180
Lachnospiraceae-Zotu1573
Lachnospiraceae-Zotu545
Lactobacillus-Zotu648
Lachnospiraceae-Zotu98
Firmicutes-Zotu1201
Lachnospiraceae-Zotu1294
Lachnospiraceae-Zotu190
Ihubacter-Zotu1270
Streptococcus-Zotu1214
Desulfovibrionales-Zotu478
Ruminococcaceae-Zotu1023
Veillonellaceae-Zotu717
Clostridiales-Zotu912
Anaerotignum-Zotu1177
Butyricicoccus-Zotu1098
Ruminococcaceae-Zotu364
Unclassified-Zotu743
Lachnospiraceae-Zotu635
Actinomycetaceae-Zotu2126
Lachnospiraceae-Zotu1378
Duncaniella-Zotu1223
Lachnospiraceae-Zotu297
Lachnospiraceae-Zotu973
Lachnospiraceae-Zotu805
Alistipes-Zotu399
Ruminococcaceae-Zotu1172
Ruminococcaceae-Zotu459
Gabonibacter-Zotu1042
Roseburia-Zotu927
Barnesiella-Zotu271
Clostridiales-Zotu1108
Porphyromonas-Zotu1020
Pseudomonas-Zotu499
Clostridiales-Zotu657
Ruminococcaceae-Zotu1542
Veillonella-Zotu1562
Hydrogenoanaerobacterium-Zotu1081
Christensenella-Zotu971
Intestinimonas-Zotu1629
Anaerotruncus-Zotu1616
Anaerostipes-Zotu774
Ruminococcaceae-Zotu643
Clostridiales-Zotu1470
Ihubacter-Zotu914
Unclassified-Zotu474
Actinobacteria-Zotu1436
Bacillales-Zotu1376
Ruminococcaceae-Zotu1713
0.0 0.2 0.4 0.6 0.8 1.0
SAGE value
B
Proximal Mucosa
Proximal Lumen
Train
Test
C
f(x) = 0.902
−0.592 = Clostridiales-Zotu2053 −0.15
5.826 = Lachnospiraceae-Zotu1573 −0.1
10.045 = Eubacteriaceae-Zotu1510 −0.07
−0.592 = Unclassified-Zotu1141 +0.05
−0.592 = Bacillaceae 1-Zotu1995 +0.04
−0.592 = Hungatella-Zotu431 +0.04
−0.592 = Lachnospiraceae-Zotu906 +0.04
−0.592 = Flavonifractor-Zotu544 +0.04
8.917 = Christensenella-Zotu971 −0.04
−0.592 = Rothia-Zotu1607 +0.03
1421 other features +1.01
E[f(X)] = −0
D
f(x) = −0.584
−1.098 = Clostridiales-Zotu2053 −0.1
4.634 = Eubacteriaceae-Zotu1510 −0.08
−1.098 = Unclassified-Zotu1141 +0.07
6.456 = Lachnospiraceae-Zotu1573 −0.06
6.23 = Bacillaceae 1-Zotu1995 −0.06
4.346 = Christensenella-Zotu971 −0.04
−1.098 = Lachnospiraceae-Zotu906 −0.03
−1.098 = Hungatella-Zotu431 −0.03
−1.098 = Flavonifractor-Zotu544 −0.02
−1.098 = Holdemania-Zotu695 +0.02
1421 other features −0.26
E[f(X)] = −0


A
Parolsenella-Zotu496
Ruminococcaceae-Zotu1946
Corynebacterium-Zotu2794
Remaining Features
Clostridiales-Zotu1460
Coprococcus-Zotu125
Catenibacterium-Zotu183
Blautia-Zotu186
Streptococcus-Zotu2711
Firmicutes-Zotu981
Streptococcus-Zotu652
Streptococcus-Zotu588
Firmicutes-Zotu892
Blautia-Zotu2344
Anaerotignum-Zotu524
Olsenella-Zotu286
Leuconostoc-Zotu480
Lachnospiraceae-Zotu2337
Ruminococcus-Zotu562
Lachnospiraceae-Zotu196
Erysipelotrichaceae-Zotu960
Paludicola-Zotu1181
Mediterraneibacter-Zotu2288
Unclassified-Zotu1362
Blautia-Zotu3170
Slackia-Zotu249
Clostridia-Zotu1165
Erysipelotrichaceae-Zotu1004
Ruminococcaceae-Zotu878
Unclassified-Zotu1749
Clostridiales-Zotu450
Murdochiella-Zotu964
Ruminococcaceae-Zotu1233
Hungatella-Zotu382
Lachnospiraceae-Zotu355
Anaerotignum-Zotu2520
Ruminococcaceae-Zotu1739
Lachnospiraceae-Zotu836
Clostridiales_Incertae Sedis XIII-Zotu1841
Clostridia-Zotu1077
Unclassified-Zotu1835
Collinsella-Zotu1803
Enterocloster-Zotu2170
Lachnospiraceae-Zotu2894
Adlercreutzia-Zotu2114
Lachnospiraceae-Zotu1253
Eisenbergiella-Zotu532
Clostridiales-Zotu881
Clostridiales-Zotu2180
Holdemanella-Zotu223
Eggerthellaceae-Zotu1805
0 5 10 15 20 25 30
SAGE value
B
Healthy Control
Crohn's Disease
Train
Test
C
f(x) = 10.149
−0.385 = Parolsenella-Zotu496 +3.2
−0.385 = Ruminococcaceae-Zotu1946 +1.51
−0.385 = Corynebacterium-Zotu2794 +1.45
−0.385 = Coprococcus-Zotu125 +0.63
−0.385 = Catenibacterium-Zotu183 +0.63
−0.385 = Clostridiales-Zotu1460 +0.47
11.292 = Streptococcus-Zotu2711 −0.41
−0.385 = Blautia-Zotu186 +0.34
−0.385 = Firmicutes-Zotu981 +0.23
−0.385 = Streptococcus-Zotu652 +0.18
1429 other features +1.92
E[f(X)] = 0
D
f(x) = −7.115
6.199 = Ruminococcaceae-Zotu1946 −1.72
7.364 = Corynebacterium-Zotu2794 −1.53
3.365 = Parolsenella-Zotu496 −1.06
5.763 = Clostridiales-Zotu1460 −0.42
7.409 = Catenibacterium-Zotu183 −0.39
6.256 = Coprococcus-Zotu125 +0.29
8.861 = Blautia-Zotu186 −0.27
7.848 = Streptococcus-Zotu2711 −0.24
−1.455 = Streptococcus-Zotu588 −0.1
−1.455 = Firmicutes-Zotu892 −0.09
1429 other features −1.58
E[f(X)] = 0