Effects of Data Transformation and Model Selection on Feature Importance in Microbiome Classification Data

Zuzanna Karwowska[1,2*], Oliver Aasmets[3*], Tomasz Kosciolek[1,4&], Elin Org[3&]

* - co-first

& - co-corresponding

1 - Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

2 - Doctoral School of Exact and Natural Sciences, Jagiellonian University, Krakow, Poland

3 - Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia

4 - Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland

tomasz.kosciolek@uj.edu.pl

elin.org@ut.ee

Abstract

Accurate classification of host phenotypes from microbiome data is essential for future therapies in microbiome-based medicine and machine learning approaches have proved to be an effective solution for the task. The complex nature of the gut microbiome, data sparsity, compositionality and population-specificity however remain challenging, which highlights the critical need for standardized methodologies to improve the accuracy and reproducibility of the results. Microbiome data transformations can alleviate some of the aforementioned challenges, but their usage in machine learning tasks has largely been unexplored. Our aim was to assess the impact of various data transformations on the accuracy, generalizability and feature selection by analysis using more than 8,500 samples from 24 shotgun metagenomic datasets. Our findings demonstrate the feasibility of distinguishing between healthy and diseased individuals using microbiome data with minimal dependence on the algorithm and transformation selection. Remarkably, presence-absence transformation performed comparably well to abundance-based transformations, and only a small subset of predictors is crucial for accurate classification. However, while different transformations resulted in comparable classification performance, the most important features varied significantly, which highlight the need to reevaluate machine-learning based biomarker detection. Our research provides valuable guidance for applying machine learning on microbiome data, offering novel insights and highlighting important areas for future research.

Introduction

Human microbiome carries a vast amount of information that can be used to improve the understanding of our functioning and be potentially used to improve clinical practice and public health. Harnessing this information on the other hand is not trivial due to the complexity of the microbial ecosystem, which comprises hundreds of species and involves intricate interactions between the ecosystem members. Similarly to other fields of biology, machine learning (ML) approaches have become pivotal in microbiome research as they can inherently account for the high dimensionality and versatile data types. Predicting an outcome based on the taxonomic or functional profile is perhaps the most widespread use of ML in the microbiome field, however thanks to its versatility, ML can be used for taxonomic assignment, functional profiling and others (Marcos-Zambrano et al. 2021). ML has been successfully used to build classification models for diseases such as colorectal cancer (Wirbel et al. 2019) and pancreatic cancer (Kartal et al. 2022), but also for predicting the disease outcome in the future such as for liver diseases (Liu et al. 2022), type 2 diabetes (Ruuskanen et al. 2022) or all cause mortality (Salosensaari et al. 2021).

Currently, analysis of the microbiome data lacks standards and best approaches are yet to be identified (Kubinski et al. 2022; Hernández Medina et al. 2022). For example, differential abundance analysis, a common analysis step to identify members of the microbiome whose abundance is different between the study groups, has been shown to produce remarkably varying results depending on the analysis methodology used (Nearing et al. 2022). Such conflicting results can be explained by the unique properties of microbiome data, such as compositionality, high dimensionality, and excessive zeros, which pose challenges for standard statistical methods (Gloor et al. 2017). To address these limitations, various data transformations like total-sum-scaling (TSS), arcsine-square-root (aSIN), and log-ratio transformations such as centered-log-ratio (CLR) or additive log-ratio (ALR) are commonly employed in microbiome research. However, the impact of data transformations on prediction and classification tasks employing machine learning algorithms remains poorly understood.

Recently, Giliberti *et al* carried out an extensive analysis to compare the performance of models based on TSS with the models using only the information about the presence or absence as predictors (Giliberti et al. 2022). Intriguingly, they found that presence-absence of the microbes as features in a predictive model leads to equivalent performance when compared to relative abundance transformation. However, there are indications that other data transformations, especially log-ratio based transformations can outperform the TSS in predictive tasks. For example, CLR has been shown to improve the prediction accuracy over TSS (Kubinski et al. 2022; "(17) (PDF) On Machine Learning Algorithms and

Compositional Data" n.d.). Nevertheless, in light of the recent results, it remains unclear whether the common data transformations can help to improve the prediction accuracy.

Here, we systematically evaluate the impact of various data transformations on the binary classification performance to determine the optimal modeling strategies for shotgun metagenomics data. We employ 7 data transformations in combination with two ML algorithms (random forest and elastic net) and assess their performance on 24 metagenomic datasets across various disease outcomes to ensure unbiased and robust assessment. In addition, we investigate how the selection of the data transformation impacts the external generalizability and feature selection, which is essential for biomarker discovery.

Results

Study design

To investigate the impact of data transformation on the classification performance, we employed publicly available shotgun metagenomic sequencing datasets present in the *curatedMetagenomicData* R package, which encompass more than 6,000 samples across various disease cohorts (**Supplementary Table 1**) (Pasolli et al. 2017). Additionally, we took advantage of the available in-house deep metagnomic data along with various health data from the Estonian microbiome cohort (N = 2,509) (Aasmets et al. 2022). **Figure 1a** shows the design and objectives of the study. Firstly, each dataset was transformed using seven distinct data transformations. The transformations are: presence-absence transformation (PA), relative abundance (total sum scaling, TSS), logarithm of TSS, arcsine square root transformation (aSIN), and four compositional transformations (centered log-ratio (CLR); robust centered log-ratio (rCLR) and additive log-ratio (ALR). The transformed datasets were then used in a binary classification setting using two learning algorithms, random forest (RF) and elastic net (ENET).

Our primary aim was to assess classifier performance across various data transformations within the analyzed datasets (within-study (WS) setting). Secondly, we aimed to assess the impact of data transformations in different analytical scenarios. In addition to analyzing the performance in the within-study setting, we evaluated the external generalizability of the models by carrying out a leave-one-study-out cross-validation for colorectal cancer (CRC; 11 datasets, **Supplementary Table 1**) and obesity that is subjects with BMI > 30 (BMI30; 5 datasets, **Supplementary Table 1**). Furthermore, we analyzed whether the data transformations benefit from a larger sample size or whether they are dependent on the number of predictors by changing the data dimensionality and evaluating the effects on

the classification performance. Lastly, we analyzed the features selected by the models to gain additional insights about how data transformations impact the conclusions of the analysis.

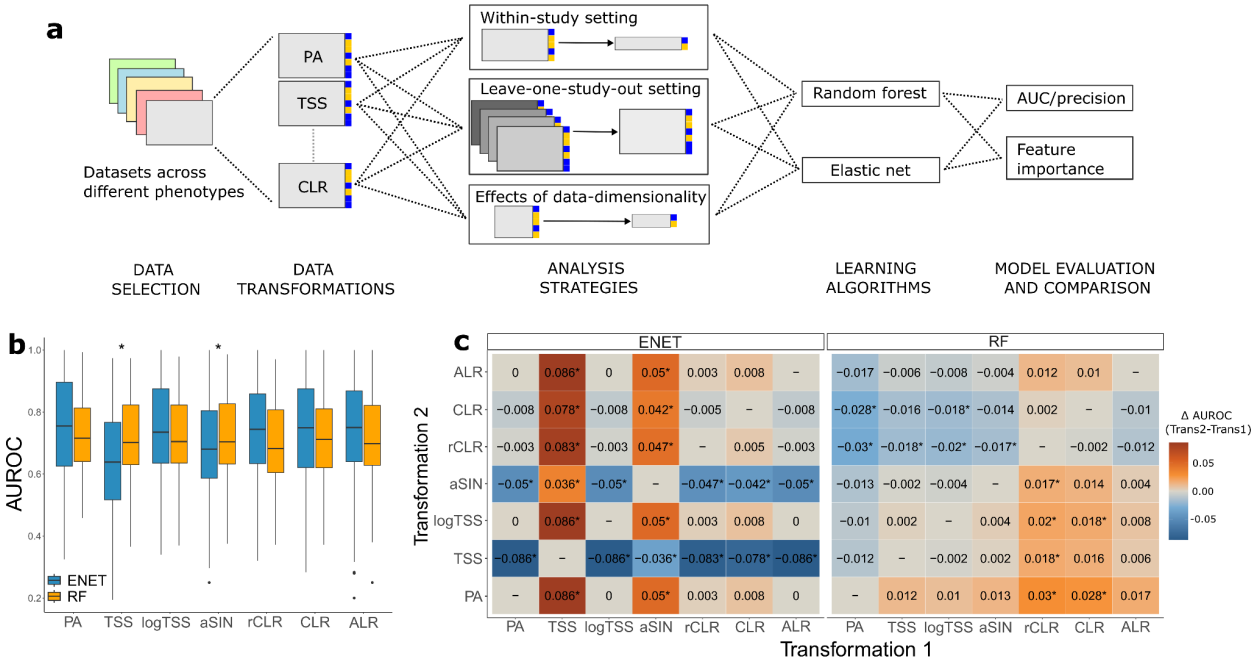


Figure 1. **STUDY DESIGN AND CLASSIFICATION RESULTS.** Panel **a** - Overview of the study design. **b** - Classification performance in the within-study setting for every data transformation and algorithm. **c** - Statistical analysis results between the used data transformations (paired t-test) for elastic net (ENET) and random forest (RF); Values and colors correspond to change in the AUROC between Transformation 2 and Transformation 1; * indicates a statistically significant difference in the AUROC (Wilcoxon signed-rank test, FDR <= 0.05). Abbreviations: ENET - elastic net logistic regression; RF - random forest; PA - presence-absence; TSS - total-sum scaling; logTSS - logarithm of TSS; aSIN - arcsine square root; CLR - centered log-ratio; rCLR - robust CLR; ALR - additive log-ratio.

General overview of classifier performance on data transformations

Our primary aim was to analyze whether the machine learning performance in a binary classification task is dependent on the data transformation and whether the data transformations are leveraged differently by different learning algorithms. The comparison of area under the receiver operating characteristic (AUROC) for random forest (RF) and elastic net logistic regression (ENET) by data transformations are shown in **Figure 1b-c**. Most importantly, there are no significant differences in classification performances between data transformations that are universal for the two learning algorithms considered (**Figure 1b, Supplementary Figure 1**). Thus, presence-absence (PA) as a data transformation leads to equivalent classification performance as any other abundance based

transformation, consistent with the results shown by Giliberti *et al (Giliberti et al. 2022)*. In the case of ENET, both aSIN and TSS resulted in inferior performance compared to other transformations. In contrast, RF did not exhibit such pronounced differences, showing moderate variations in performance across transformations (**Figure 1b-c**). Furthermore, our findings indicate that when TSS and aSIN are employed as the data transformation method, ENET performs significantly worse than RF on average. However, for other data transformations, ENET outperforms RF, though these differences are not statistically significant (**Figure 1b**).

Data transformation effects in different analytical scenarios

We were surprised to find no significant differences in classifier performance between different data transformations, despite their distinct properties. To understand whether the data transformations could give advantage in other analytical scenarios, we conducted several follow-up analyses. Initially, we assessed how sample size and feature filtering influenced binary classification performance. We hypothesized that some data transformations were more effective with larger sample sizes. To assess that, we applied different feature prevalence thresholds to the publicly available metagenomics datasets and to the Estonian Microbiome Cohort dataset (N = 2509), and carried out the classification task. Furthermore, for the Estonian Microbiome Cohort dataset, we additionally subsampled the cases and controls to 10%, 20%, 40%, and 60% to study the impact of sample size and feature size at once. As expected, we observed that larger sample sizes lead to improved classification performance. Surprisingly, the filtering of features did not significantly affect performance (**Supplementary Figure 2a**). Additionally, we found no substantial interaction effect between sample size and the number of features on classifier performance between different transformations. The only noteworthy exception was the TSS and aSIN transformations, which consistently resulted in lower performance with ENET (**Figure 2a, Supplementary Figure 2**).

Next, we assessed whether data transformations impacted the model's ability to generalize by measuring its classification performance on an external dataset. To do this, we employed a leave-one-study-out validation method for both colorectal cancer (CRC) and obesity defined by BMI > 30 (BMI30) datasets. This involved training a model on a combined set of samples from all studies except one and evaluating its performance on the omitted study. Our external validation results echoed those from within-sample validation, showing no significant differences when employing various data transformations (**Figure 2b**). Interestingly, in both datasets, the TSS transformation combined with ENET consistently exhibited inferior performance compared to alternative transformations. This highlights the

problem-specific nature of the analytical approach, as classification performance varied between the obesity and cancer datasets. Nevertheless, a straightforward approach combining the PA transformation with ENET consistently yielded favorable results in most target diseases.

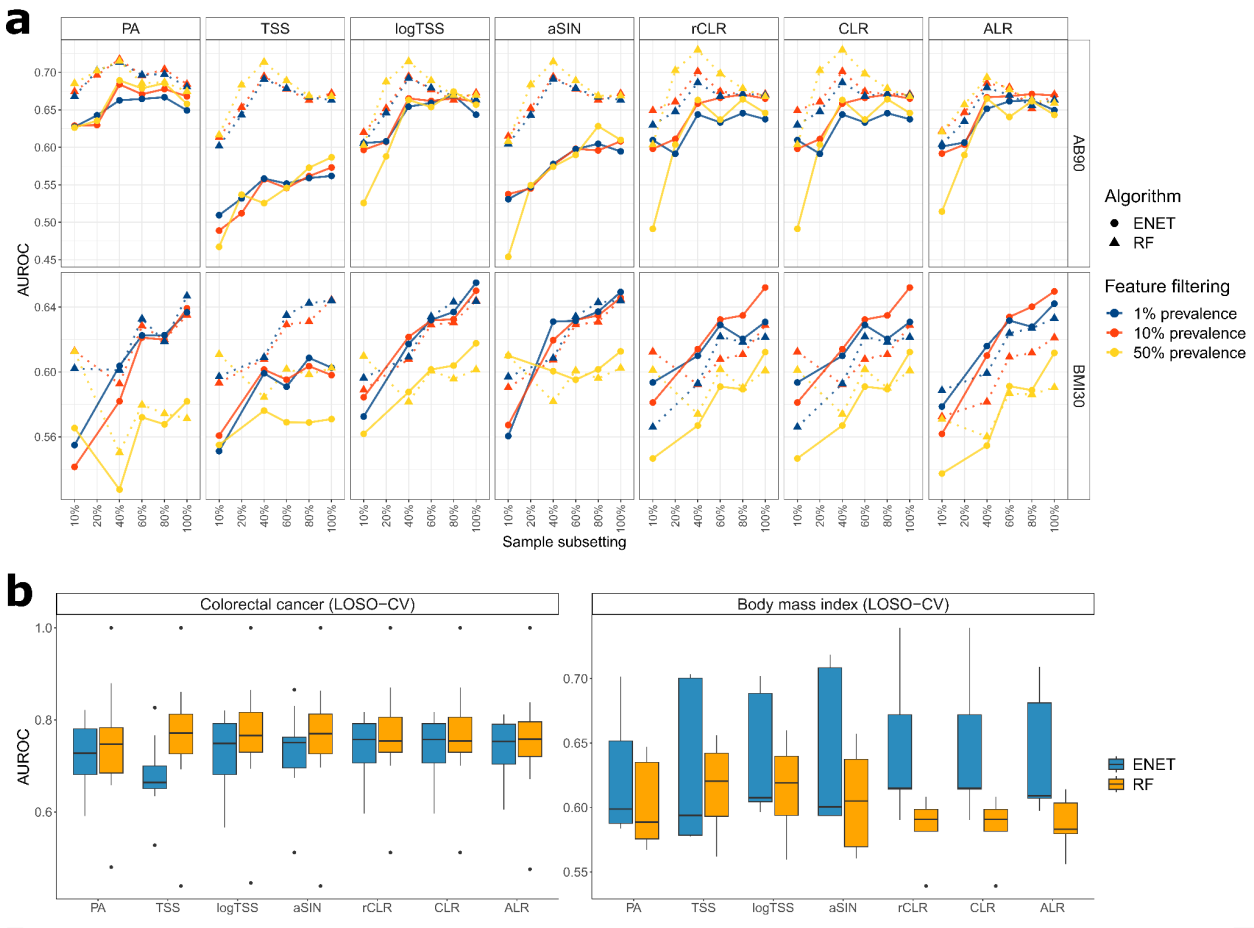


Figure 2. **CLASSIFICATION PERFORMANCE OF DIFFERENT TRANSFORMATIONS UNDER DIFFERENT SCENARIOS.** Panel **a** - Classification performance across varied data dimensions and sample sizes for random forest (RF) and elastic net logistic regression (ENET) and every transformation. Panel **b** - Transformation and Model Outcomes in Leave-One-Study-Out cross-validation (LOSO-CV) setting.

Feature importance

As data transformations showed no remarkable differences in the classification performance, we were interested in whether the predictive features are similar. To assess feature importance, we calculated mean absolute SHapley Additive exPlanations (SHAP) values for each microbe. SHAP values are a method used in machine learning to explain the contribution of each feature to the prediction of a model

(Lundberg and Lee 2017). They provide a way to understand the impact and importance of individual features in the model's output. Using mean absolute SHAP values, we found that the number of crucial features was more dataset-dependent than transformation-specific. Compositional transformations like ALR, CLR, and rCLR selected more features, particularly with RF, while PA, TSS, logTSS, and aSIN transformations selected fewer (**Figure 3a**). However, regardless of the transformation, only a small subset of features held significant importance (**Supplementary Figure 3** and **Supplementary Figure 4**). To validate this, we built classifiers on PA-transformed data from the Estonian Microbiome cohort, focusing on the most significant features. Surprisingly, using just 20 to 30 bacteria sufficed for effective classification across various phenotypes, with performance decreasing as more features were added. This underscores the challenge of dealing with the high dimensionality and complexity of microbiome data and highlights the importance of feature selection (**Figure 3b**).

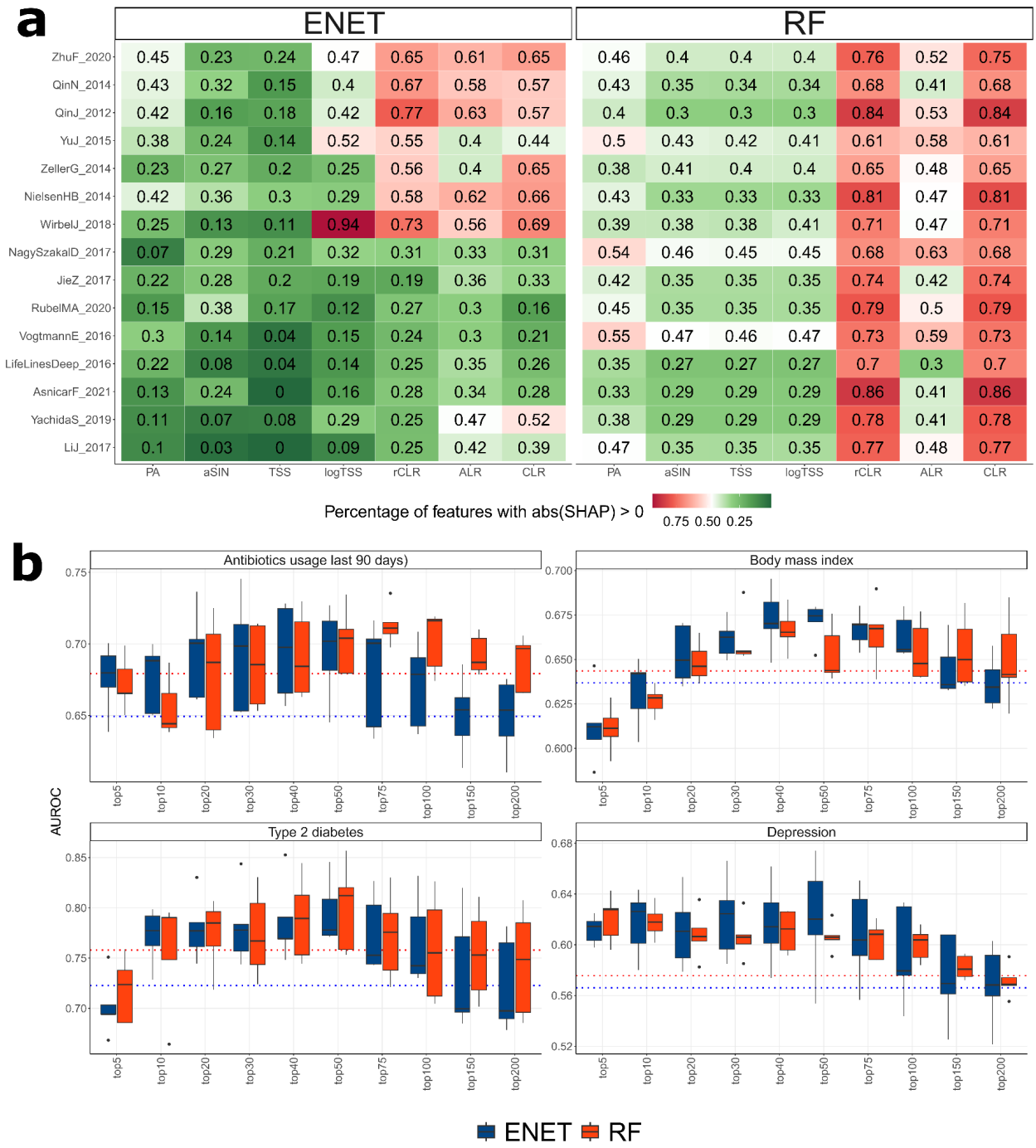


Figure 3. FEATURES SELECTION AMONG DATA TRANSFORMATIONS. Panel **a** - Number of Selected Features: Proportion for ENET and RF. Panel **b** - ENET and RF Classification Performance on presence-absence (PA) data using top-N features. The dashed line indicates classifier performance on all features.

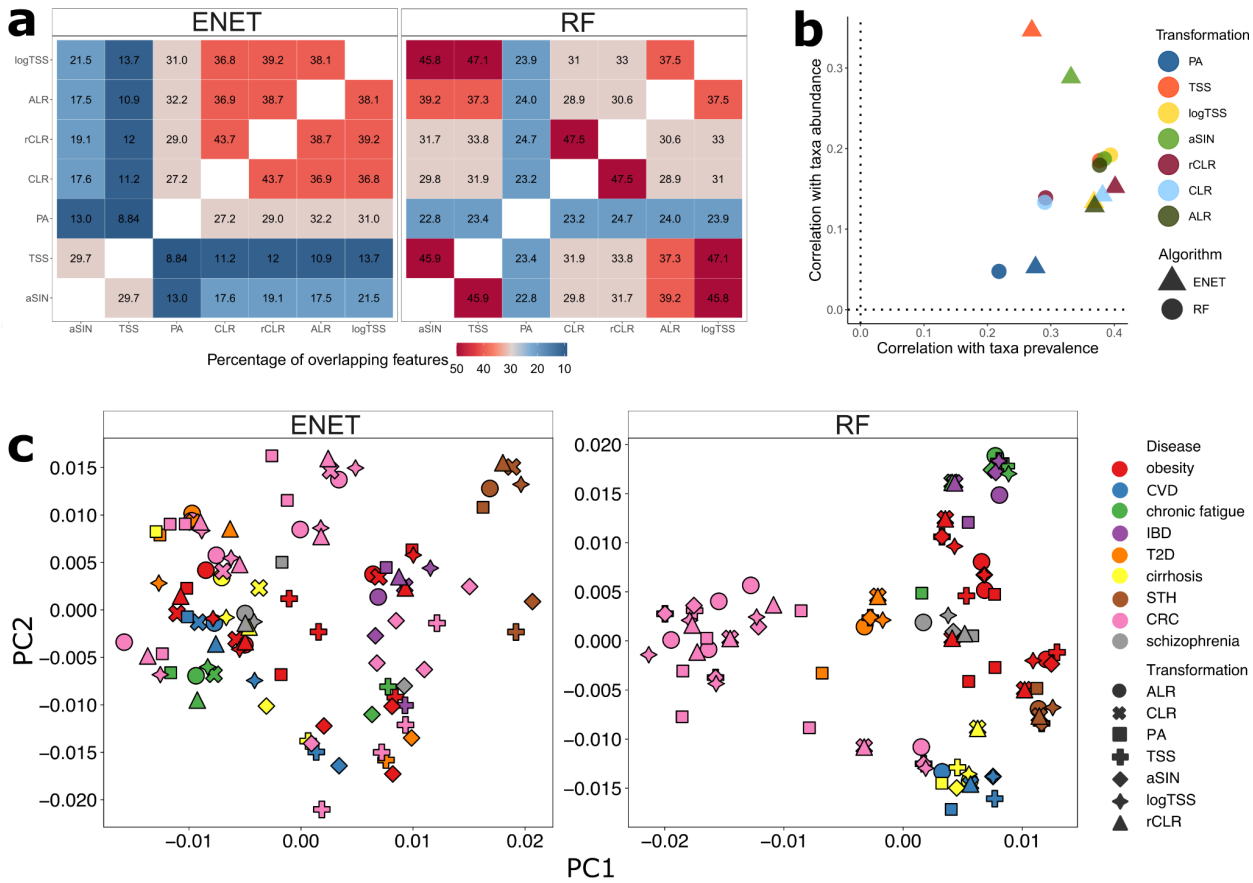


Figure 4. STABILITY OF FEATURE IMPORTANCE RESULTS. Panel **a** - Overlap of Top 25 Features Between Transformations for Each Algorithm and Dataset. Panel **b** - Spearman Correlation Between Feature's SHAP Value, Prevalence, and Abundance for Each Transformation in Both Algorithms. Panel **c** - Dimensionality Reduction of Cosine Distance Matrix for Top 25 Features Across All Datasets in ENET and RF

We examined the similarity of selected features across different data transformation-model combinations. Building on our earlier analysis that highlighted a small subset of features with high SHAP values, we focused on the overlap among the top 25 features exhibiting the highest SHAP values. Our findings revealed limited agreement, with approximately 30% overlap observed between transformations within the same algorithm. As seen in Figure 4a, TSS and aSIN transformations with ENET exhibited significantly lower overlap compared to other transformations, while the highest agreement was found between compositional transformations. In the case of RF, feature overlap was similar across all transformations, with slightly lower overlap observed between PA and other transformations. Overall, we did not observe consistent trends between transformations, except for TSS and aSIN with ENET, which showed notably lower overlap (**Figure 4a**). The most remarkable finding was the even smaller overlap of around 10% between RF and ENET **(Figure 4a, Supplementary Figure 5**).

We validated our results by analyzing feature importance with dimensionality reduction. We found that as the number of top features increased, the cosine distances decreased, except for TSS and aSIN transformations with ENET, which remained high. ENET transformations generally had greater distances compared to RF transformations (Supplementary Figure 5b). We also performed Principal Coordinates Analysis (PCoA) to reduce dimensionality, revealing distinct separation between ENET and RF along the first principal axis. Within ENET, compositional transformations clustered closely, while aSIN and TSS showed significant separation from other transformations, and PA exhibited a smaller degree of separation in RF (**Supplementary Figure 6)** .

Given the differences in feature selection (particularly with TSS and aSIN combined with ENET), we explored the relationship between feature selection and microbial characteristics. We calculated Spearman correlations between SHAP values and feature prevalence/abundance. We found a strong correlation between SHAP values and taxa relative abundance, especially for ENET + TSS/aSIN (Figure 4b, Supplementary Figure 7). We also assessed if different transformations yield similar bacteria selections and if these crucial bacteria cluster within the same dataset. We calculated cosine distances between SHAP values for 25 bacteria for each transformation within a specific model and used Principal Coordinates Analysis (PCoA) to reduce this distance matrix. Results showed that RF consistently clustered transformations within their datasets, indicating robust variable selection. In contrast, ENET had weaker agreement between transformations, with aSIN and TSS selecting different features compared to the rest (Figure 4c).

Additionally, we saw that the top 10 features chosen by models using TSS and aSIN transformations remained relatively consistent across different datasets and diseases. Among the top 10, *P. copri, B. vulgatus, B. uniformis*, and *F. prausnitzii* were often reported, all well-known and abundant gut commensals (**Supplementary Figure 9**). This indicates that TSS and aSIN in combination with ENET emphasize bacterial abundance in their feature importance assessments which might potentially complicate the identification of disease-specific markers.

In conclusion, our findings underscore that models assign utmost importance to a limited set of microbes, largely independent of the chosen data transformation. However, there can be larger differences in the most informative features chosen by the transformations. This variability could impact the development of gut microbiome health indices, assessment of microbiome dysbiosis, and biomarker discovery. Together with the equivalency in the resulting classification performance, this highlights the need to improve the feature identification, validation and stability.

Discussion

Our goal was to evaluate the impact of data transformations on machine learning performance in microbiome binary classification (Marcos-Zambrano et al. 2021). We compared two algorithms, seven data transformations, and assessed dimensionality, model generalization, and feature importance across diverse metagenomic datasets.

Our results showed that there was no significant difference in classification performance across the data transformations considered. Notably, both presence-absence (PA) and abundance-based transformations exhibited comparable performance consistent with the results reported by Giliberti *et al.* (Giliberti et al. 2022). Similarly, we observed that an elastic net algorithm (ENET) performs better with PA when compared to TSS, but there is no major difference when RF is used. Nevertheless, as Giliberti *et al.* focused on the comparison of PA and TSS, our results didn't confirm the benefit of using other abundance-based data transformations as reported previously. For example, CLR has been shown to outperform TSS with RF as the learning algorithm and CLR has been shown to outperform logTSS (Kubinski et al. 2022). The improved performance has also been reported when CLR instead of TSS is used in combination with LASSO regression (Quinn and Erb 2020). Our analysis utilized a large array of different datasets where such results couldn't be confirmed. Moreover, based on our results there was no significant effect of the transformation selection on the model generalizability nor interaction with the model performance and data dimensionality. Thus, the performance seems to be more reliant on the characteristics of the dataset than on the specific algorithm or transformation employed, which corresponds to the well known "no free lunch theorem" in machine learning.

Our feature analysis showed that the number of significant microbes was more tied to the dataset than the transformations used. Importantly, there was minimal overlap among the top features chosen by different methods, indicating diverse microbial features can yield similar classification performance. Thus, microbiome studies could take advantage of the methodologies such as statistically equivalent signatures, which aim to identify variable sets with equal predictive power (Lagani et al. 2017). In contrast, analyzing the properties of the selected features shows that ENET in combination with TSS and aSIN focuses primarily on the most abundant microbes such as *P. copri*, *F. prausnitzii, B. vulgatus* and *B. uniformis*. As the combination of ENET with TSS is one of the most popular choices for applying ML on the microbiome data (Marcos-Zambrano et al. 2021), we suggest taking caution when focusing on the feature importance as the selection might not yield a signal specific to the condition of interest and it could potentially obscure the interpretation of classification outcomes. These findings underscore the need for further research and multiple analytical strategies to validate microbial feature identification,

especially when different approaches yield contrasting feature sets with similar classification performance, raising critical questions for biomarker and dysbiosis research.

Based on our findings, we recommend adopting the presence-absence (PA) transformation for microbiome data classification tasks. PA-based classifiers demonstrate strong performance, offer a simpler interpretation, and are computationally efficient for integration into applications. However, selecting a threshold for microbe absence (e.g., setting it at zero) introduces challenges related to structural zeros and sequencing depth, which warrants further investigation (Silverman et al. 2020).

Our study's strengths lie in its systematic approach, employing two learning algorithms on diverse datasets, which enhances generalizability. However, we acknowledge limitations. Our focus is on classification tasks, potentially limiting relevance to other analyses. Dataset constraints and unexplored confounding variables are also noted. Lack of ground truth of which bacteria readly changed. In the future, we plan to explore additional questions. Firstly, we would like to assess whether using only presence-absence (PA) data is sufficient for accurate classification or if combining PA with variations in the abundance of key bacteria enhances performance. We would also investigate if the observed similarity between PA-based classifiers and abundance-based transformations is influenced by technical factors, such as shallow sequencing, which might mask certain features. These inquiries highlight the complexity of our findings and provide directions for further research.

Methods

Data acquisition

Open data. The binary classification task for comparing the different data transformations relied on datasets available in the *curatedMetagenomicData* R package (version 3.6.2) (Pasolli et al. 2017). The *curatedMetagenomicData* package contains uniformly processed shotgun metagenomic sequencing human gut microbiome data of healthy and diseased subjects. The preprocessing of the provided microbiome data, including taxonomic profiling, was carried out using the bioBakery 3 toolkit *(Beghini et al. 2021)*. For the within-study (WS) evaluation setting, we focused on 18 distinct human stool metagenomics datasets, which had primary endpoints available and included at least 50 cases and 50 controls to allow proper model evaluation. In some instances, filtering was applied to the original dataset to achieve a binary classification task. The selected datasets, defined binary classification task and the filtering procedures are shown in **Supplementary Table 1**.

For the leave-one-study-out (LOSO) evaluation setting, we included additional 6 colorectal cancer datasets with less than 50 cases or controls and 4 datasets, where primary endpoint was not available.

Estonian microbiome cohort. The Estonian Microbiome cohort (EstMB) is a volunteer based cohort currently including 2509 subjects, who have provided blood, oral and stool samples. Being part of a larger nation-wide Estonian Biobank (EstBB), linkings to various electronic health records (EHR) and questionnaires covering the lifestyle and dietary preferences are available for all of the subjects. The cohort is described in detail in Aasmets & Krigul *et al.(Aasmets et al. 2022)*. For the binary classification, three target variables were considered: antibiotics usage within the previous 90 days before the microbiome sample collection (AB90), obesity as BMI > 30 (BMI30) and classification of type 2 diabetes cases from healthy controls (T2D).
Taxonomic profiling on EstMB was carried out using *Metaphlan3 (Beghini et al. 2021)* to comply with the profiling done in the *curatedMetagenomicData* R package.

Data transformations

Numerous data transformations have been proposed to be used in the analysis of microbiome data. In the current manuscript, the following data transformations were considered:

1. Relative abundance/total-sum-scaling (TSS): The standard and most widely used technique, which scales data to relative abundance.
2. Log(TSS): A logarithmic transformation applied to TSS-normalized data.
3. Presence-absence (PA): changes abundance data into binary data.
4. Arcsin square root (aSIN ("(17) (PDF) Biometry : The Principles and Practice of Statistics in Biological Research / Robert R. Sokal and F. James Rohlf" n.d.)): Which involves applying the arcsin (inverse sine) function to each value, which can be useful for normalizing and stabilizing data that represents proportions or percentages, particularly in statistical analyses of compositional data.
5. Centered log-ratio transformation (CLR (Aitchison 1982)): procedure that enhances compositional data for standard statistical analysis by dividing each value by the geometric mean and applying a logarithmic transformation .
6. Robust CLR transformation (rCLR (Martino et al. 2019)): a robust version of CLR that handles zeros by using only observed taxa for geometric mean calculations.
7. Additive log-ratio transformation (ALR (Aitchison 1982)): in which each feature in a dataset is divided by a selected reference feature and then logarithmically transformed. To ensure analysis

robustness, we employed a data transformation approach in which we randomly selected 5 features as references for each iteration. We arrived at this choice based on our observation that adding more than 5 references did not yield significant improvements **(Supplementary Figure 10).**

Machine learning pipeline

Each of the considered transformations was applied to a dataset and a binary classification task was carried out. Random forests (RF) and elastic net (ENET) penalized regression were used as the learning algorithms in combination with the data transformations. ENET logistic regression is a machine learning algorithm that combines L1 (Lasso) and L2 (Ridge) regularization techniques to perform logistic regression with variable selection, making it suitable for high-dimensional data by minimizing overfitting and selecting the most relevant features. Random Forest, on the other hand, is an ensemble learning method that builds multiple decision trees and combines their predictions to improve classification accuracy and handle complex relationships in data while reducing the risk of overfitting.

These algorithms were chosen due to their popularity and competitive performance in the microbiome field (Zhou and Gallins 2019) and as RF being a nonlinear and ENET a linear method, they can provide a contrast for studying transformation-algorithm interactions. Followingly, the model fitting and evaluation is described.

Within the study (WS) setting. For each 21 datasets used for the within-study evaluation, the following repeated holdout validation procedure for parameter tuning and model evaluation was carried out:

1. *Data is split to training/test set (75%-25%) stratified by the target variable*
2. *Hyperparameter tuning on the training set (75%) using 5-fold cross-validation with grid search (10 parameter combinations)*
3. *Model with optimal hyperparameters is fit on the whole training data (75%)*
4. *Model is evaluated on the test set (25%)*

The initial data test/train split and model evaluation were carried out on 10 random data splits to assess the variation arising from sampling resulting in 10 performance estimates per evaluation.

Leave-One-Study-Out (LOSO) setting. The LOSO setting was carried out for the 11 available colorectal cancer and 5 obesity (BMI >= 30) datasets. The aim was to understand whether the

dataset-to-dataset generalization performance might be dependent on the chosen data transformation. For the model fitting and evaluation, the following procedure was carried out:

1. *Data is split to training/test set so that one dataset works as the test set and other datasets as a combined training set*
2. *Hyperparameter tuning on the training set using 5-fold cross-validation with grid search (10 parameter combinations)*
3. *Model with optimal hyperparameters is fit on the whole training data (75%)*
4. *Model is evaluated on the test set - left out dataset*

The model evaluation was carried out using each dataset per target variable as a test set. This resulted in 11 performance estimates for colorectal cancer and 5 for obesity.

Model comparison

To test the differences in the binary classification results, we used the Wilcoxon signed-rank test for each pair of data transformations. For that, we first averaged the results over different folds and target variables to account for the overrepresentation of certain phenotypes like colorectal cancer and obesity. After that, a paired test within each target was carried out. Thus, we tested the hypothesis that the difference in the performance of *transformation1* and *transformation2* is not 0. To account for the multiple testing, Benjamini-Hochbergi procedure was applied to the nominal p-values.

Filtering subjects and features

Due to the large sample size, EstMB dataset was further used to study the effects of sample size and number of bacteria used by the model on the performance of the classifiers, focusing on differences between the data transformations. For that reason, the cases and controls of antibiotics usage (AB90) and obesity (BMI30) were subsampled to 10/20/40/60/80/100% of the initial number of cases and controls. Additionally, in combination with the sample subsetting, prevalence filtering (1/10/50%) for the microbial taxa was applied to study the impact of the number of predictors. Thus, in total of 6*3 = 18 scenarios per algorithm and data transformation were analyzed.

Feature importance analysis

Feature importance evaluation was based on the SHapley Additive exPlanations (SHAP) values. SHAP values are a method used in machine learning to explain the contribution of each feature to the

prediction of a model (Lundberg and Lee 2017). We first determined the percentage of features within each dataset that exhibited SHAP values exceeding zero. Subsequently, we focused on the bacteria displaying the highest SHAP values across all models. For feature overlap assessment, we quantified the degree of overlap among the top 25 features between various transformations, as well as between identical transformations across different algorithms. The average feature overlap was calculated by taking the average across different folds and datasets.

To evaluate model performance when only the most important features are used as predictors, we utilized ENET and RF models on presence-absence transformed data, considering the top 5 to top 200 most important features. The resulting Area Under the Receiver Operating Characteristic Curve (AUC) and precision scores were computed for all publicly available datasets as described in the modelling pipeline section for the WS setting.

Additionally, we calculated the Spearman correlation between SHAP values and bacterial relative abundance and prevalence for each transformation and algorithm. We further explored the similarity of the selected features across the algorithm-transformation pairs by computing the cosine and Hamming distance for the feature importance matrix and analyzing the distances in a reduced space. To reduce dimensionality and visualize relationships, Principal Coordinate Analysis (PCoA) was applied separately to cosine and Hamming distances for algorithm.

Acknowledgments

The authors would like to thank Mari-Liis Tammeorg, Marili Palover, Anu Reigo, Neeme Tõnisson, Liis Leitsalu, and Esta Pintsaar for participating in the sample collection process of the Estonian Microbiome cohort. We thank Steven Smit, Rita Kreevan, Martin Tootsi, and Triinu Temberg for the DNA extraction process. We thank Reidar Andreson for bioinformatic support and Witold Wydmański for fruitful discussions during project planning. We also thank all the study participants.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

E.O. and T.K., designed and supervised the study. Z.K. and O.A. performed the data analysis, interpreted the data and prepared the figures. Z.K. and O.A. wrote the manuscript. All authors read and approved the final paper.

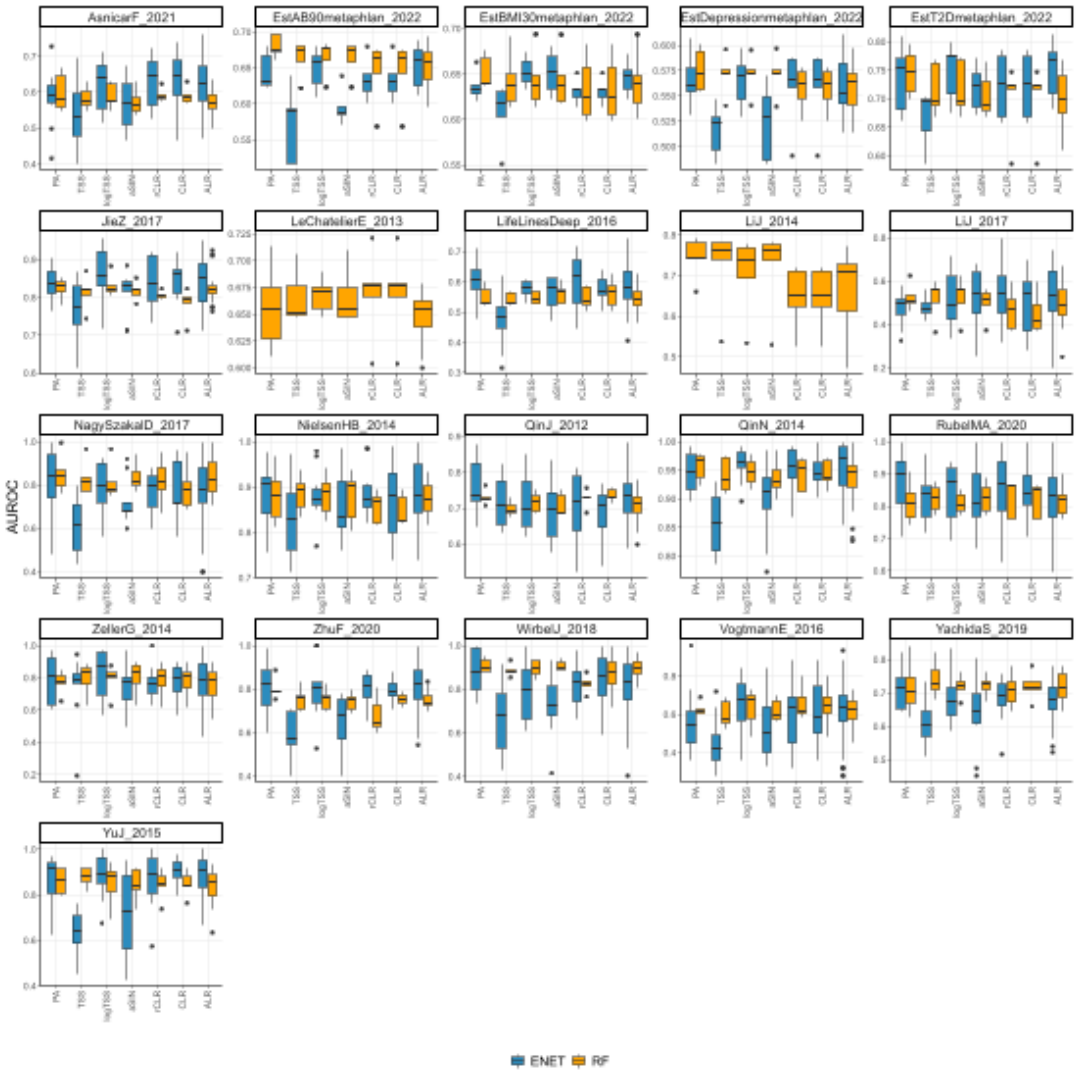
Funding

This work was funded by an Estonian Research Council grant (PRG1414 to EO) and an EMBO Installation grant (No. 3573 to EO). TK and ZK are funded by the National Science Centre, Poland grant 2019/35/D/NZ2/04353.

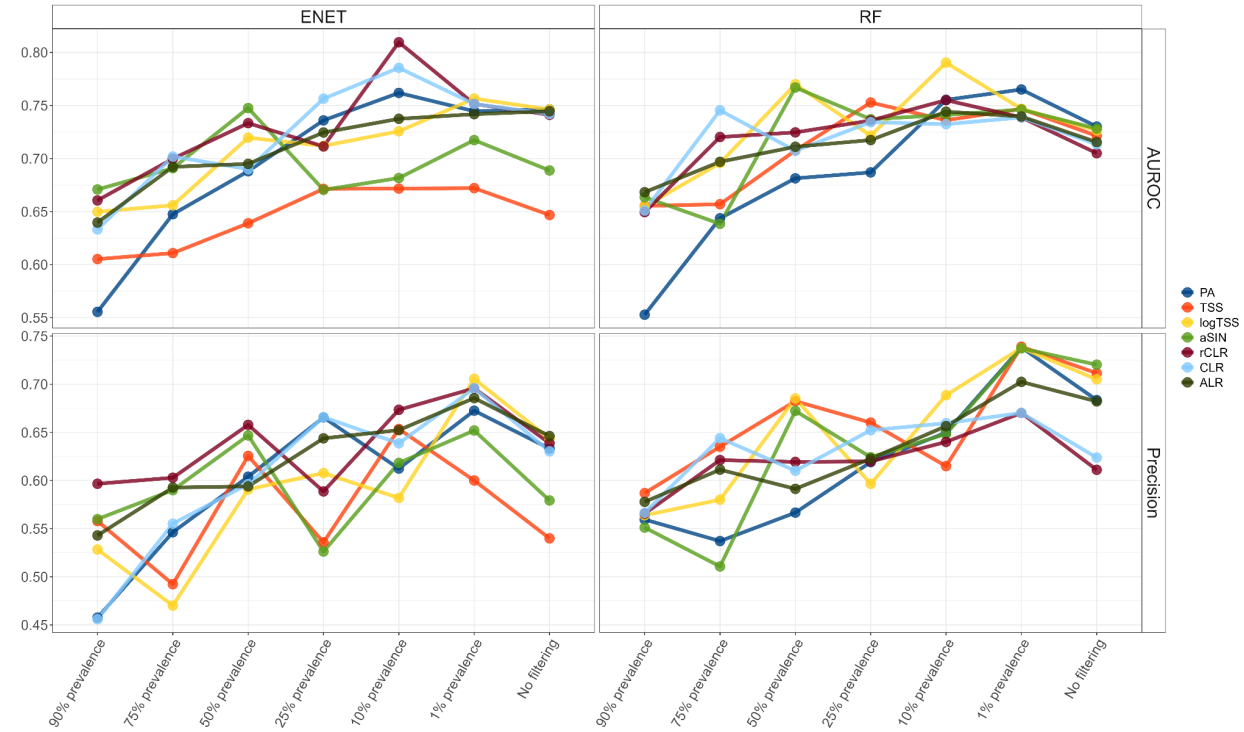
Supplementary Figures

SOURCE	NR	DATASET	TARGET VARIABLE	N_cases	N_controls	FILTERING	COMMENT
Publicly available datasets	1	QinJ_2012	0 - control 1 - T2D	170	174		
	2	YuJ_2015	0 - control 1 - CRC	74	50		
	3	YachidaS_2019	0 - control 1 - CRC	258	251		
	4	VogtmannE_2016	0 - control 1 - CRC	52	52		
	5	WirbelJ_2018	0 - control 1 - CRC	60	65		
	6	ZellerG_2014	0 - control 1 - CRC	53	61	study_condition %in% c("control", "CRC")	
	7	ThomasAM_2019c	0 - control 1 - CRC	40	40		Only used in the Leave-One-Study-Out setting
	8	ThomasAM_2018b	0 - control 1 - CRC	32	28		Only used in the Leave-One-Study-Out setting
	9	ThomasAM_2018a	0 - control 1 - CRC	29	24	study_condition %in% c("control", "CRC")	Only used in the Leave-One-Study-Out setting
	10	HanniganGD_2017	0 - control 1 - CRC	27	28	study_condition %in% c("control", "CRC")	Only used in the Leave-One-Study-Out setting
	11	GuptaA_2019	0 - control 1 - CRC	30	30		Only used in the Leave-One-Study-Out setting
	12	FengQ_2015	0 - control 1 - CRC	46	61	study_condition %in% c("control", "CRC")	Only used in the Leave-One-Study-Out setting
	13	RubelMA_2020	0 - control 1 - STH (parasites)	89	86		
	14	QinN_2014	0 - control 1 - cirrhosis	123	114		
	15	NielsenHB_2014	0 - control 1 - IBD	81	236	days_from_first_collection == 0	
	16	NagySzakalD_2017	0 - control 1 - myalgic encephalomyelitis/chronic fatigue syndrome	50	50		
	17	LiJ_2017	0 - pre-hypertension 1 - hypertension	99	56	study_condition %in% c("pre-hypertension", "hypertension")	
	18	JieZ_2017	0 - control 1 - atherosclerotic cardiovascular disease	214	171		
	19	ZhuF_2020	0 - control 1 - schizophrenia	91	81		
	20	LifeLinesDeep_2016	0 - control 1 - BMI > 30	136	999		
	21	AsnicarF_2021	0 - control 1 - BMI > 30	160	937		
	22	LiJ_2014	0 - control 1 - BMI > 30	59	189		
	23	LeChatelierE_2013	0 - control 1 - BMI > 30	169	123		
Estonian Microbiome Cohort	24	BMI30_2022	0 - control 1 - BMI > 30	543	1966		
	25	AB90_2022	0 - control 1 - AB used within 90 days	264	2245		
	26	T2D_2022	0 - control 1 - T2D	118	2391		
	27	Depression_2022	0 - control 1 - Depression	620	1889		

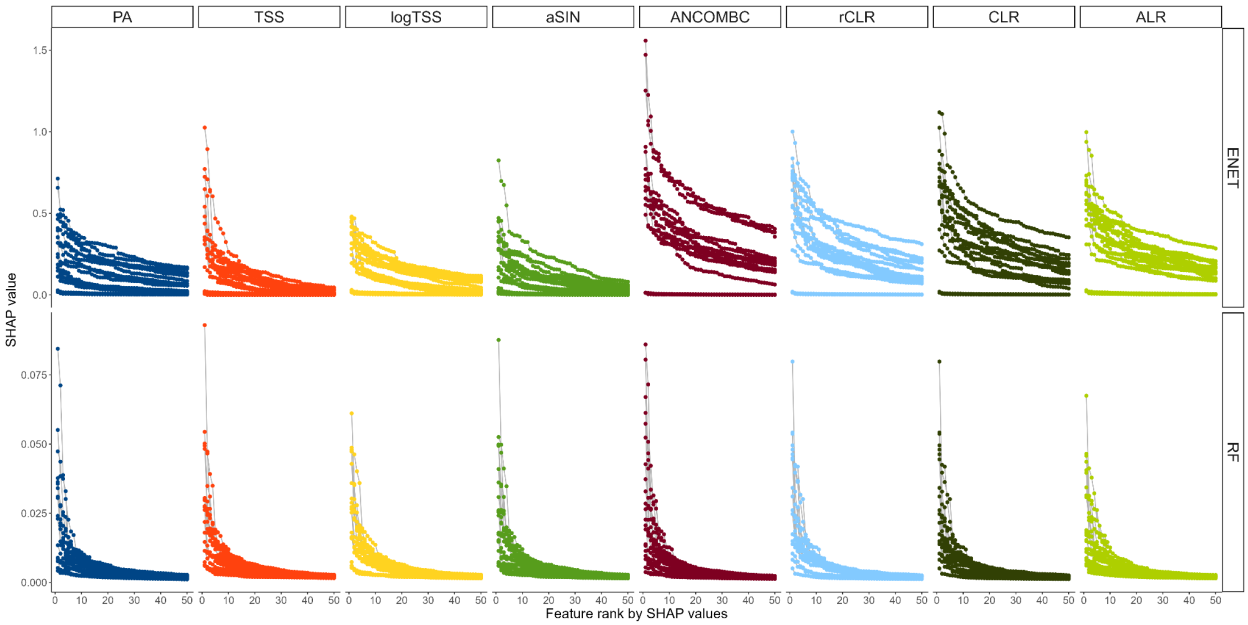
Supplementary Table 1. Overview of the datasets used



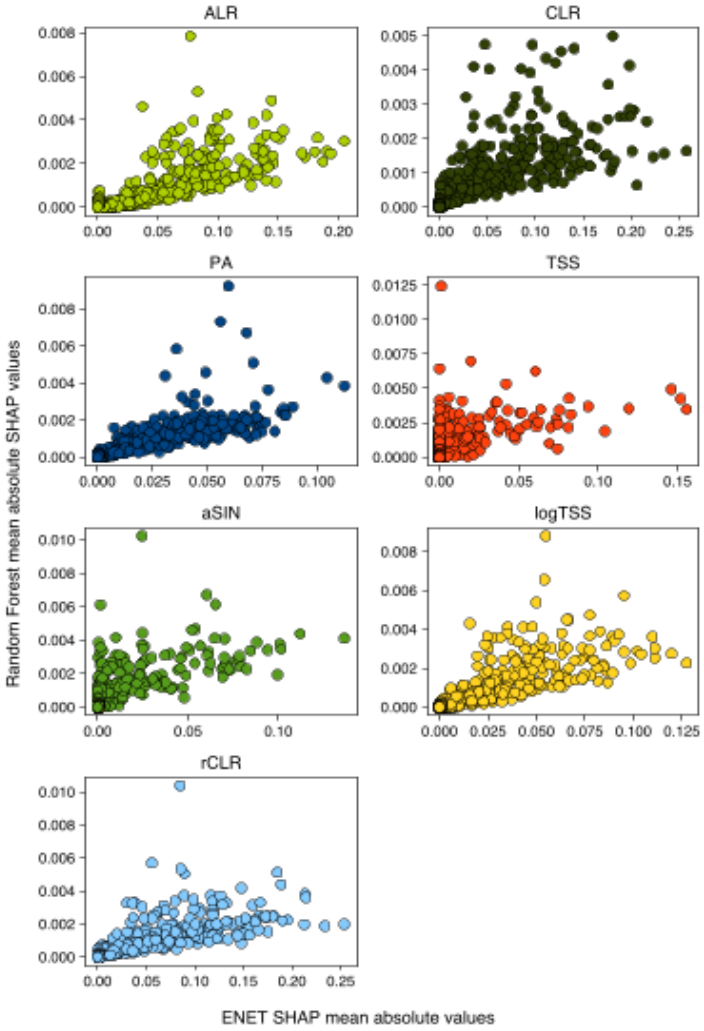
Supplementary Figure 1. Classification performance (AUROC) by studied targets. Abbreviations: *ACD - atherosclerotic cardiovascular disease;* BMI30 - *body mass index > 30; CRC - colorectal cancer; IBD - inflammatory bowel disease; STH - soil-transmitted helminths*; *T2D - type 2 diabetes*



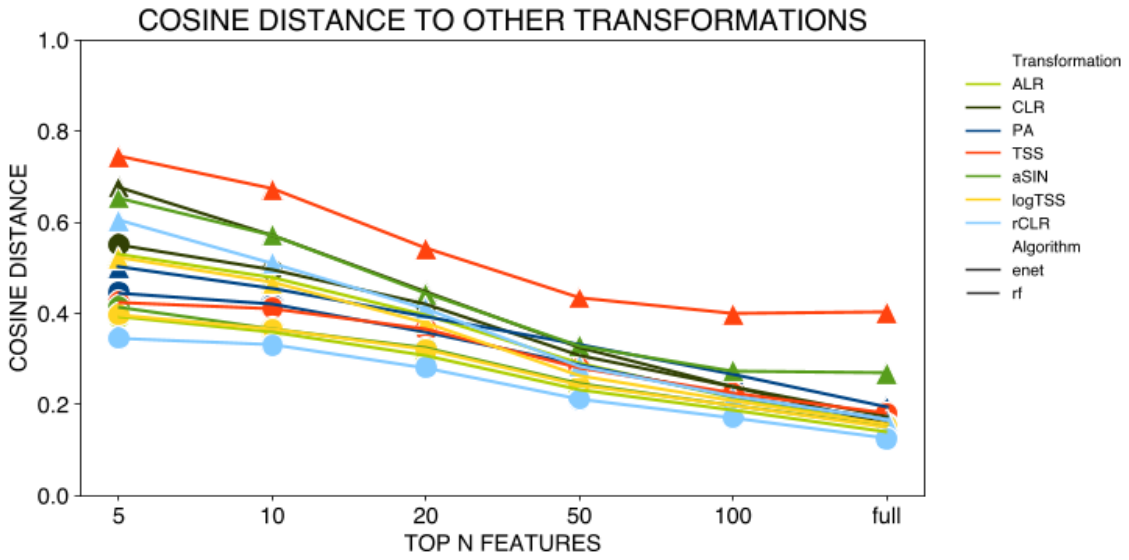
Supplementary Figure 2. Classifier performance dependence on the data dimensionality. Colors represent three different feature filtering cutoffs, solid line refers to elastic net (ENET) performance, dashed line shows random forest (RF) performance.



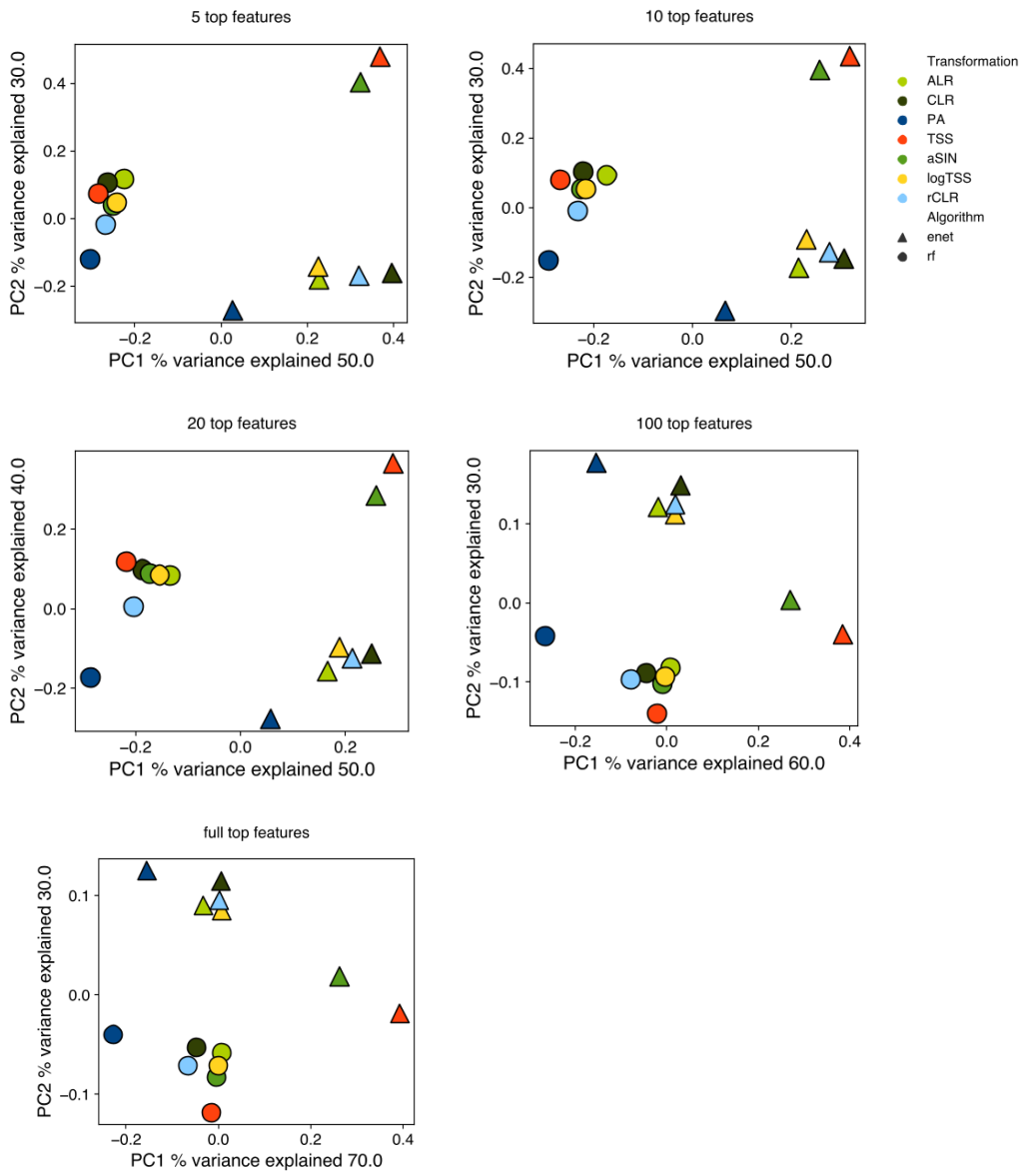
Supplementary Figure 3. Mean absolute SHAP values distribution for across all datasets plotted separately for each transformation for top 50 features;



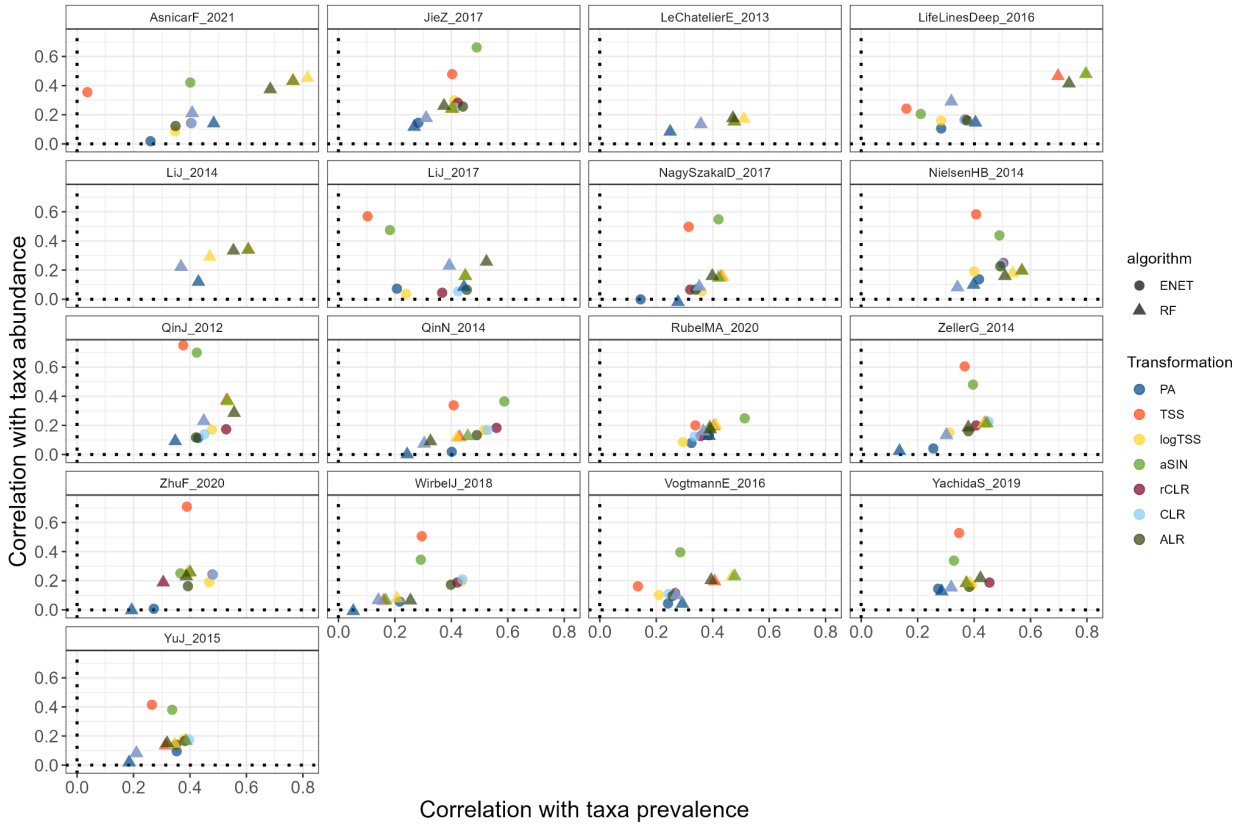
Supplementary Figure 4. Mean absolute SHAP values for the same feature from ENET and RF model across all datasets plotted separately for each transformation.



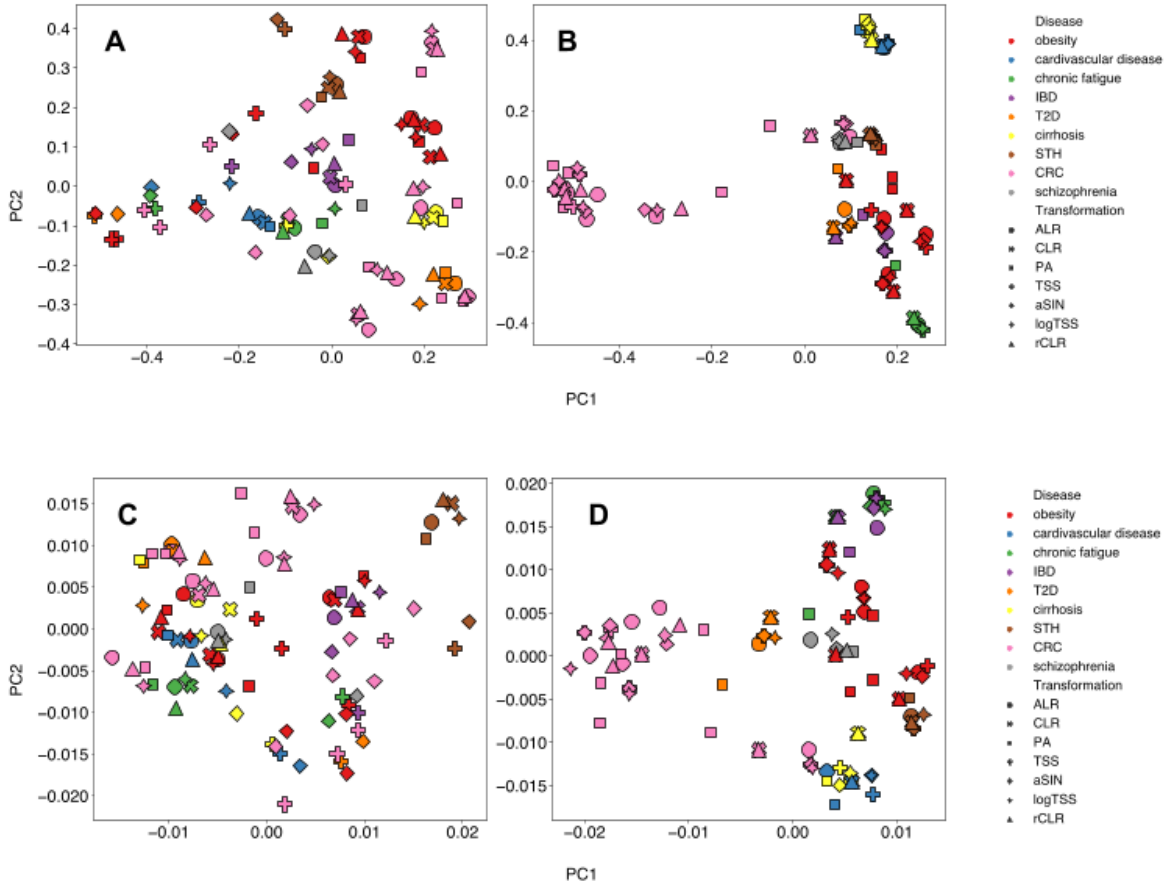
Supplementary Figure 5. Cosine distance between SHAP values between transformations for the same dataset calculated for top 5, 10, 02, 50, 100 bacteria with the highest SHAP values as well as all bacteria in the dataset.



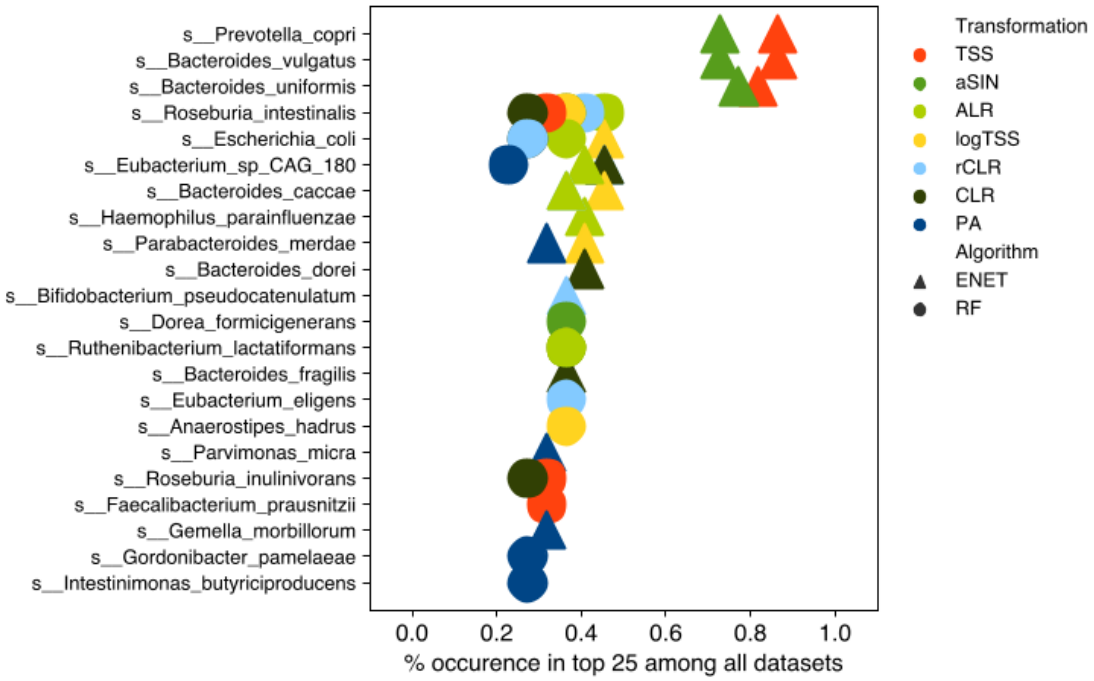
Supplementary Figure 6. PCoA on cosine distance between 10, 20, 50 and 100 features with the highest SHAP values.



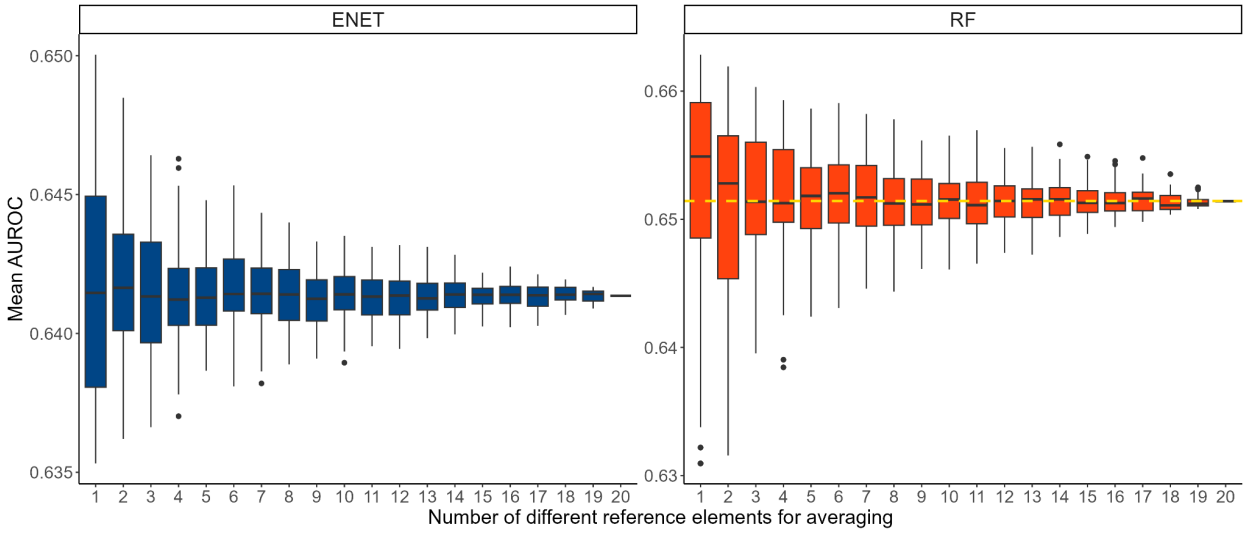
Supplementary Figure 7. Correlation between the mean absolute SHAP values and taxa prevalence/mean relative abundance. Abbreviations: *ACD - atherosclerotic cardiovascular disease; BMI - body mass index; CRC - colorectal cancer; IBD - inflammatory bowel disease; STH - soil-transmitted helminths*; *T2D - type 2 diabetes*



Supplementary Figure 8. A. PCoA on cosine distance between SHAP values from ENET model across all datasets and transformations coloured by dataset name; **B.** PCoA on cosine distance between SHAP values from RF model across all datasets and transformations coloured by dataset name; **C.** PCoA on cosine distance between SHAP values from ENET model across all datasets and transformations coloured by disease name; **D.** PCoA on cosine distance between SHAP values from RF model across all datasets and transformations coloured by dataset name;



Supplementary Figure 9. Overview of bacteria that were the most often in top 25 features with highest mean absolute SHAP values.



Supplementary Figure 10. Classifier performance in respect to the number of different reference elements used for averaging for the additive log-ratio (ALR) transformation.

References

"(17) (PDF) Biometry : The Principles and Practice of Statistics in Biological Research / Robert R. Sokal and F. James Rohlf." n.d. ResearchGate. Accessed September 14, 2023. https://www.researchgate.net/profile/F-Rohlf/publication/44554870_Biometry_the_principles_and_practice_of_statistics_in_biological_research_Robert_R_Sokal_and_F_James_Rohlf/links/0a85e5314d1bf07070000000/Biometry-the-principles-and-practice-of-statistics-in-biological-research-Robert-R-Sokal-and-F-James-Rohlf.pdf.

"(17) (PDF) On Machine Learning Algorithms and Compositional Data." n.d. ResearchGate. Accessed September 7, 2023. https://www.researchgate.net/profile/Hassan-Talebi-2/publication/334225471_On_machine_learning_algorithms_and_compositional_data/links/5dd8a4c5a6fdccdb445c4a8b/On-machine-learning-algorithms-and-compositional-data.pdf.

Aasmets, Oliver, Kertu Liis Krigul, Kreete Lüll, Andres Metspalu, and Elin Org. 2022. "Gut Metagenome Associations with Extensive Digital Health Data in a Volunteer-Based Estonian Microbiome Cohort." *Nature Communications* 13 (1): 869.

Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 44 (2): 139–77.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Giliberti, Renato, Sara Cavaliere, Italia Elisa Mauriello, Danilo Ercolini, and Edoardo Pasolli. 2022. "Host Phenotype Classification from Human Microbiome Data Is Mainly Driven by the Presence of Microbial Taxa." *PLoS Computational Biology* 18 (4): e1010066.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And This Is Not Optional." *Frontiers in Microbiology* 8 (November): 2224.

Hernández Medina, Ricardo, Svetlana Kutuzova, Knud Nor Nielsen, Joachim Johansen, Lars Hestbjerg Hansen, Mads Nielsen, and Simon Rasmussen. 2022. "Machine Learning and Deep Learning Applications in Microbiome Research." *ISME Communications* 2 (1): 1–7.

Kartal, Ece, Thomas S. B. Schmidt, Esther Molina-Montes, Sandra Rodríguez-Perales, Jakob Wirbel, Oleksandr M. Maistrenko, Wasiu A. Akanni, et al. 2022. "A Faecal Microbiota Signature with High Specificity for Pancreatic Cancer." *Gut* 71 (7): 1359–72.

Kubinski, Ryszard, Jean-Yves Djamen-Kepaou, Timur Zhanabaev, Alex Hernandez-Garcia, Stefan Bauer, Falk Hildebrand, Tamas Korcsmaros, et al. 2022. "Benchmark of Data Processing Methods and Machine Learning Models for Gut Microbiome-Based Diagnosis of Inflammatory Bowel Disease." *Frontiers in Genetics* 13 (February): 784397.

Lagani, Vincenzo, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, and Ioannis Tsamardinos. 2017. "Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets." *Journal of Statistical Software* 80 (September): 1–25.

Liu, Yang, Guillaume Méric, Aki S. Havulinna, Shu Mei Teo, Fredrik Åberg, Matti Ruuskanen, Jon Sanders, et al. 2022. "Early Prediction of Incident Liver Disease Using Conventional Risk Factors and Gut-Microbiome-Augmented Gradient Boosting." *Cell Metabolism* 34 (5): 719–30.e4.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874.

Marcos-Zambrano, Laura Judith, Kanita Karaduzovic-Hadziabdic, Tatjana Loncar Turukalo, Piotr Przymus, Vladimir Trajkovik, Oliver Aasmets, Magali Berland, et al. 2021. "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment." *Frontiers in Microbiology* 12 (February): 634511.

Martino, Cameron, James T. Morton, Clarisse A. Marotz, Luke R. Thompson, Anupriya Tripathi, Rob Knight, and Karsten Zengler. 2019. "A Novel Sparse Compositional Technique Reveals Microbial Perturbations." *mSystems* 4 (1). https://doi.org/10.1128/mSystems.00016-19.

Nearing, Jacob T., Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, et al. 2022. "Microbiome Differential Abundance Methods Produce Different Results across 38 Datasets." *Nature Communications* 13 (1): 342.

Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, et al. 2017. "Accessible, Curated Metagenomic Data through ExperimentHub." *Nature Methods* 14 (11): 1023–24.

Quinn, Thomas P., and Ionas Erb. 2020. "Interpretable Log Contrasts for the Classification of Health Biomarkers: A New Approach to Balance Selection." *mSystems* 5 (2). https://doi.org/10.1128/mSystems.00230-19.

Ruuskanen, Matti O., Pande P. Erawijantari, Aki S. Havulinna, Yang Liu, Guillaume Méric, Jaakko Tuomilehto, Michael Inouye, et al. 2022. "Gut Microbiome Composition Is Predictive of Incident Type 2 Diabetes in a Population Cohort of 5,572 Finnish Adults." *Diabetes Care* 45 (4): 811–18.

Salosensaari, Aaro, Ville Laitinen, Aki S. Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, et al. 2021. "Taxonomic Signatures of Cause-Specific Mortality Risk in Human Gut Microbiome." *Nature Communications* 12 (1): 2671.

Silverman, Justin D., Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. 2020. "Naught All Zeros in Sequence Count Data Are the Same." *Computational and Structural Biotechnology Journal* 18 (September): 2789–98.

Wirbel, Jakob, Paul Theodor Pyl, Ece Kartal, Konrad Zych, Alireza Kashani, Alessio Milanese, Jonas S. Fleck, et al. 2019. "Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer." *Nature Medicine* 25 (4): 679–89.

Zhou, Yi-Hui, and Paul Gallins. 2019. "A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction." *Frontiers in Genetics* 10 (June): 579.