Identifying Optimal Machine Learning Approaches for Microbiome–Metabolomics Integration with Stable Feature Selection

Suzette N. Palmer[1-3*], Animesh Mishra[1,4], Shuheng Gan[3], Dajiang Liu[5], Andrew Y. Koh[1,4,6], Xiaowei Zhan[3,6*]

Author information:

[1]Division of Hematology/Oncology, Department of Pediatrics, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]Department of Biomedical Engineering, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[3]Peter O'Donnell Jr. School of Public Health, Quantitative Biomedical Research Center, Center for the Genetics and Host Defense, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[4]Department of Microbiology, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[5]Department of Public Health Sciences, Pennsylvania State University College of Medicine, Hershey, PA 17033, USA

[6]Harold C. Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Contributions

S.N.P., A.Y.K., and X.Z. planned the research; S.N.P. performed the experiments; S.N.P. analyzed the data; A.M. and S.G. contributed technical support. S.N.P., A.M., X.Z., D.L., and A.Y.K wrote the paper.

*Corresponding authors

Correspondence to Xiaowei Zhan at xiaowei.zhan@utsouthwestern.edu and/or 214-648-5194.

Keywords: microbiome; metabolomics; multi-omics integration; feature selection; machine learning

Abstract

Microbiome research has been limited by methodological inconsistencies. Taxonomy-based profiling presents challenges such as data sparsity, variable taxonomic resolution, and the reliance on DNA-based profiling, which provides limited functional insight. Multi-omics integration has emerged as a promising approach to link microbiome composition with function. However, the lack of standardized methodologies and inconsistencies in machine learning strategies has hindered reproducibility. Additionally, while machine learning can be used to identify key microbial and metabolic features, the stability of feature selection across models and data types remains underexplored, despite its importance for downstream experimental validation and biomarker discovery. Here, we systematically compare Elastic Net, Random Forest, and XGBoost across five multi-omics integration strategies: Concatenation, Averaged Stacking, Weighted Non-negative Least Squares (NNLS), Lasso Stacking, and Partial Least Squares (PLS), as well as individual ‘omics models. We evaluate performance across 588 binary and 735 continuous models using microbiome-derived metabolomics and taxonomic data. Additionally, we assess the impact of feature reduction on model performance and feature selection stability. Among the approaches tested, Random Forest combined with NNLS yielded the highest overall performance across diverse datasets. Tree-based methods also demonstrated consistent feature selection across data types and dimensionalities. These results demonstrate how integration strategies, algorithm selection, data dimensionality, and response type impact both predictive performance and the stability of selected features in multi-omics microbiome modeling.

Introduction

The gut microbiome plays a critical role in host homeostasis and is implicated in the development of various diseases[1,2]. However, microbiome research has largely relied on correlative studies, with limited reproducibility due to methodological discrepancies and data-specific challenges[3–5]. Most microbiome studies focus on taxonomy-based profiling (e.g., 16S rRNA sequencing, metagenomic shotgun sequencing), which introduces challenges such as 1) data sparsity (i.e., a large proportion of zeros representing absent taxa), 2) varying levels of taxonomic resolution, and 3) reliance on DNA-based profiling, which does not necessarily reflect functional microbial activity, such as metabolite or protein production[6,7]. These limitations hinder efforts to move beyond association-based findings toward mechanistic insights.

To address these challenges, multi-omics integration — which combines taxonomic data with functional 'omics (e.g., metabolomics, proteomics, transcriptomics) — has emerged as a promising strategy for uncovering disease-relevant microbial interactions[8–10]. Integrative approaches increase analytical power, allowing researchers to link microbiome composition with functional microbial activity. However, microbiome research methodologies vary widely in data processing, resolution, and integration strategies[11–16] (**Supplemental File 1**). The lack of standardized methodologies can lead to inconsistencies in microbiome-derived findings[17–19]. Differences in machine learning models and integration strategies can yield highly variable results, making it difficult to establish reproducible biomarkers or validate disease associations across studies[10,20,21]. Without a systematic evaluation of these methods, it remains unclear which approaches provide the most biologically meaningful and clinically translatable insights.

To address these gaps, we conducted a multi-faceted analysis to evaluate performance trends and feature selection stability across various multi-omics data integration techniques. Specifically, we analyzed 4 multi-omics datasets with matched microbiome and metabolomics data and assessed integration performance for both binary and continuous outcomes. Our analysis applied Random Forest, XGBoost, and Elastic Net models to both full dimensional and feature reduced datasets. We compared model performance across single 'omics and multi-omics approaches, focusing on two integration strategies: concatenation and stacked generalization. For stacked generalization, we implemented Averaged Stacking, Weighted Non-Negative Least Squares (NNLS)[22], Lasso Stacking (LS)[23], and Partial Least Squares (PLS)[24]. To evaluate the impact of

81 dimensionality reduction, we filtered and reduced features within each 'omics dataset and compared model
82 performance against the full-dimensional datasets (**Figure 1A**, **Tables 1–2**). Subsequently, we assessed the
83 performance of concatenation, stacked generalization, and single-omics models (**Figure 1B**). Finally, we
84 examined feature selection consistency across machine learning models and integration strategies, comparing
85 stability between single 'omics and concatenated approaches (**Figure 1C**).

86 **Methods**

87 Data Collection and Processing

88 The data used in this study were originally obtained from Franzosa et al. (2019)[11], Yachida et al. (2019)[14],
89 Erawijantari et al. (2020)[25], and Wang et al. (2020)[26]. Processed taxonomy relative abundance and
90 metabolomics processed data were obtained from Muller et al. (2022)[37] and details can also be found at
91 https://github.com/borenstein-lab/microbiome-metabolome-curated-data/wiki5 and in **Supplemental File 2**.

92 Data Processing for the Integrated Analysis

93 Processed data files for metabolomics, metadata, species level relative abundance and genus level relative
94 abundance were further processed for downstream analysis. Taxonomic data at the genus and species levels
95 were concatenated for each dataset, and taxa with an overall abundance across samples that was less than
96 0.1% were removed. For metabolite differential intensity analysis, Limma 3.64.1[33] was applied, which employs
97 linear modeling and empirical Bayes methods to identify significant differences while adjusting for multiple testing
98 using Benjamini-Hochberg procedure. For taxonomic differential analysis, the Wilcoxon rank-sum test was used
99 to compare taxa abundance across experimental groups (More details for data processing can be found in
100 **Supplemental File 2**).

101 Overview of Multi-omics Integrative Analysis

102 The integrative analysis was developed using a custom in-house script that is readily available at
103 https://github.com/suziepalmer10/Multiomics-Integrative-Pipeline/tree/main. The packages tidyverse 2.0.0 and
104 argparse 2.2.3 were used to parse and pre-process the data for this integrative analysis[38]. For continuous
105 outcomes, performance was assessed with Root Mean Squared Error (RMSE), and binary classification was
106 evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC). For the integrative
107 analysis, AUROC is calculated using pROC 1.18.0 and Kappa is calculated using irr 0.84.1[39,40].

For this analysis, each dataset (as described above, in **Supplemental File 2,** and presented in **Tables 1–2**) was processed through the Integrated Pipeline using Elastic Net, Random Forest, and XGBoost scripts. The data was split into 80% training and 20% testing, and three repeats of five-fold cross-validation were performed. To ensure both generalizability and reproducibility, the random seed for each fold was set by multiplying the repeat index and the fold index (**Supplemental File 2**). Performance was evaluated on validation and test sets. Grid search with a predefined grid was used for Elastic Net, Random Forest, and XGBoost (**Supplemental File 2**).

Selection of Machine Learning Methods and Integrative Strategies: Elastic Net, Random Forest, and XGBoost were chosen because they have been widely used in microbiome research for building predictive models and identifying the features for experimental validation[41–43]. For multi-omics integration, we adopted feature concatenation and stacked generalization, two approaches frequently used in this field[11–16]. Collectively, these algorithms and integration techniques offer interpretable models that can be readily validated through literature review and experimental follow-up (see **Supplemental File 1** for more details).

Elastic Net Overview and Grid: The Elastic Net algorithm estimates β coefficients to minimize the error term $L(\beta) = \|Y - X\beta\|^2$, where X represents the data matrix (concatenated, taxonomy, or metabolomics), and Y represents the response variable. The regularization penalty combines L_1 (Lasso) and L_2 (Ridge) penalties. The penalty term added to the loss function is $\lambda(\alpha \sum|\beta_j| + (1-\alpha)\sum \beta_j^2)$, where $\sum |\beta_j|$ represents the L_1 penalty, $\sum \beta_j^2$ represents the L_2 penalty, λ determines the overall regularization strength, and α controls the balance between L_1 and L_2 regularization. For this analysis, the caret 6.0-93 and glmnet 4.1-6 R packages were used[23,44] (See **Supplemental File 2** for more details).

Random Forest Overview and Grid: The Random Forest algorithm constructs an ensemble of decision trees to enhance prediction accuracy and reduce overfitting. Each tree is trained on a bootstrapped sample of the data, and predictions are obtained via averaging (for regression) or majority voting (for classification). To introduce randomness and reduce feature correlation, Random Forest selects a subset of features at each split. For this analysis, the caret and randomForest 4.7-1.1 R packages were used[44,45] (See **Supplemental File 2** for more details).

XGBoost Overview and Grid: XGBoost is an optimized gradient boosting algorithm that iteratively refines weak learners to minimize the loss function. It incorporates regularization to mitigate overfitting and utilizes efficient parallel processing for computational efficiency. For this analysis, the caret and xgboost 1.7.8.1 R packages were used[44,46] (See **Supplemental File 2** for more details).

Data integration using direct concatenation: To integrate taxa and functional data using direct concatenation, we will horizontally concatenate the two matrices, where A represents taxonomy data and B represents functional data. Both matrices share the same number of rows (m), corresponding to the samples and disease phenotype that is used as the response variable. However, these matrices differ in the number of columns, n_A, which represents the number of taxa and, n_B, which represents the number of functional data (i.e., metabolites). Mathematically, let $A = \begin{pmatrix} a_{11} & \cdots & a_{1n_A} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn_A} \end{pmatrix}$ and $B = \begin{pmatrix} b_{11} & \cdots & b_{1n_B} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn_B} \end{pmatrix}$. Concatenation results in C, where $C = [A|B]$. This creates a unified matrix, where $C \in \mathbb{R}^{m\times(n_A+n_B)}$, which enables the integration of the distinct datatypes while preserving the sample structure. We then use the combined matrix, C, for machine learning model development using Elastic Net, XGBoost and Random Forest algorithms.

Data integration using averaged stacking of the predictors: Data integration using averaged stacking combines predictions from individual machine learning models (Random Forest, XGBoost or Elastic Net) for each dataset (taxonomy and metabolomics) into a single, averaged prediction for each sample. For averaged stacking, let $P_A \in \mathbb{R}^{m\times 1}$ represent predictions from a machine learning model trained on taxonomy data, A and let $P_B \in \mathbb{R}^{m\times 1}$ represent predictions from a machine learning model trained on functional data. Let m represent the number of samples and response variables that are shared across both datasets. The averaged stacking prediction, P_{avg}, for each sample is computed as $P_{Avg} = \frac{1}{2}(P_A + P_B)$, where $P_{Avg} \in \mathbb{R}^{m\times 1}$ represents the integration prediction vector across all samples.

Data integration using weighted non-negative least squares (NNLS): Data integration using weighted nnls combines predictions from each model by assigning weights to minimize the error between the combined predictions and the observed values, with the constraint that weights are non-negative. As previously described above, let $P_A \in \mathbb{R}^{m\times 1}$ and $P_B \in \mathbb{R}^{m\times 1}$, represent predictions from a machine learning model trained on taxa data, A, and functional data, B. Let $y \in \mathbb{R}^{m\times 1}$ represent response variables and let w_A and w_B represent non-negative

weights assigned to P_A and P_B. The combined prediction $P_{NNLS} = w_A P_A + w_B P_B$, where the weights w_A and w_B are determined by solving the following optimization problem, where $\min_{w_A, w_B \geq 0} \|y - (w_A P_A + w_B P_B)\|^2$. The nnls 1.5 package was used for non-negative least squares analysis[22].

<u>Data integration using least absolute shrinkage and selection operator (LASSO) stacked generalization:</u> Data integration using LASSO combines predictions by assigning weights to minimize the error between the combined prediction and the observed values while enforcing sparsity in the rights through L_1 regularization. As previously described, let $P_A \in \mathbb{R}^{m \times 1}$ and $P_B \in \mathbb{R}^{m \times 1}$, represent predictions from a machine learning model trained on taxa data, A, and functional data, B, and let $y \in \mathbb{R}^{m \times 1}$ represent response variables. For LASSO, let w_A and w_B represent weights assigned to P_A and P_B. The combined prediction, $P_{LASSO} = w_A P_A + W_B P_B$, where the weights w_A and w_B are determined by solving the following optimization problem using L_1 regularization: $\min_{w_A, w_B \geq 0} \|y - (w_A P_A + w_B P_B)\|^2 + \lambda(|w_A| + |w_B|)$, where λ is the weight of the L_1 regularization term and controls solution sparsity. The glmnet 4.1-6 package was using for LASSO[23].

<u>Data integration using partial least squares (PLS):</u> PLS is a regression method that combines the predictors from separate machine learning models by projecting the predictors into a lower dimensional space to maximize the covariance between the predictors and the response variables. This integrative method can handle multicollinearity and captures the shared variation between different datatypes. PLS operates in three stages. The first stage is latent component extraction, where PLS identifies latent components, T, by projecting the combined predictor matrix, P, onto a set of weight vectors that maximize the covariance between the latent components, T, and the response variable, y. Mathematically, $T = PW$, and as previously described, let $P_A \in \mathbb{R}^{m \times 1}$ and $P_B \in \mathbb{R}^{m \times 1}$, represent predictions from a machine learning model trained on taxa data, A, and functional data, B, and let $y \in \mathbb{R}^{m \times 1}$ represent response variables. Let $P = [P_A | P_B] \in \mathbb{R}^{m \times 2}$, represent the concatenated matrix of predictors. Let $T \in \mathbb{R}^{m \times k}$ represent the latent variables extracted by P, where k represents the number of components. Let $W \in \mathbb{R}^{2 \times k}$ represent a weight matrix that defines the linear combination between P_A and P_B. The next stage of PLS is to create a linear regression of the response variable. This is accomplished following extraction of latent components, T, and the response variable, y, are then modeled as a linear function of T. Mathematically, $y = TC + \varepsilon$, where $C \in \mathbb{R}^{k \times 1}$ represent regression coefficients for T to predict y and ε represents the residual error term. The final stage is to combine the projection and the

regression stages, which can be mathematically represented as $y_{PLS} = PWC$. The pls 2.8-4 package[24] was used for partial least squares model generation.

Performance Heatmap Generation: Following completion of each machine learning model (Elastic Net, XGBoost or Random Forest), the performance metrics mean and standard deviation are provided as an output file. The average AUROC values for binary response variables and the average RMSE values for the continuous response variables were extracted for each dataset and model using an in-house script. Following creation of the Heatmap using the Complexheatmap R package[47,48], the top 3 performing models across each dataset for Elastic Net (if completed), Random Forest and XGBoost. These are represented by the "*", "**", and "***", and do not include the single 'omics performance metrics. For some of the datasets, the Elastic Net model was unable to run due to insufficient grid and/or the likely need for further data processing which we decided not to further optimize for this study.

Performance Pie Charts: The R package ggplot2 3.4.1 was used to create the performance heatmaps for Figures 2-3 and Supplemental Figures 1-2[49]. These metrics were calculated by quantifying the number of models for each dataset that were part of the top 3 performing models (represented as '*', "**" and "***") (described above for Performance Heatmap Generation).

Feature Extraction:

Feature importance values were extracted from Elastic Net, XGBoost and Random Forest across all 15 cross-validation models. For Elastic Net, feature importance is accessed through the magnitude of its regression coefficients, where larger absolute values reflect greater predictive contributions. For random forest mean decrease in impurity (MDI) is applied to quantify feature importance and reflects the reduction in impurity achieved by splitting on a given feature, averages across all decision trees generated. For XGBoost, the Gain metric is used to access how effectively a feature improves the objective function by reducing prediction error during tree splits.

Since different metrics were used to quantify feature importance, across Elastic Net, Random Forest and XGBoost models, we decided to take the absolute value of each variable and perform min-max normalization (data scaled between 0 and 1) for each model. Mathematically, let i denote the feature ($i = \{1, 2 \dots n\}$), where n is the total number of features. Let j denote the model ($j = \{1,2 \dots C\}$), where C represents the number of folds

214 and repeats from 5-fold cross validation. Let $|S_{i,j}|$ represent the absolute value for the importance score of feature
215 i in model j. Next min-max normalization, $N_{i,j}$, for feature i for model j is calculated by $N_{i,j} =$
216 $\frac{|S_{i,j}|-min_k(|S_{k,j}|)}{max_k(|S_{k,j}|)-min_k(|S_{k,j}|)}$, where k indexes over all features in model j. I next compute the averaged normalized
217 importance for each feature across all models, $\overline{N_i} = \frac{1}{C}\sum_{j=1}^{C} N_{i,j}$. This allows for direct comparison between
218 features in each model.

219 For this analysis, the number of top features selected per model was determined by varying feature bin
220 sizes of 1, 5, 10, 20 and 50. For each bin size, only the top n features (where n equals the bin size) were retained
221 to represent the most important features. The features contained in each bin have the greatest importance values
222 and represent the top features that might be explored using peer review and/or by experimentation. Violin plots
223 were generated using ggplot2 to illustrate the overlap and uniqueness of features selected by Elastic Net,
224 Random Forest and XGBoost[49]. The proportions of overlapping and unique features for each dataset are
225 represented as individual circles within a violin plot. Models with fewer than n feature ranked above 0, or those
226 that were not able to finish, were excluded from the analysis.

227 **Results**

228 Creation of integrative analysis pipeline to analyze continuous and binary datasets

229 We collected and processed four multi-omics datasets consisting of matched metagenomic shotgun sequencing
230 and metabolomics data **(Table 1-2)**[11,14,25,26], and we evaluated 18 continuous and 13 binary matching
231 response variables. We then developed an integrative analysis pipeline to assess the performance of three
232 machine-learning algorithms (Elastic Net, Random Forest, XGBoost) in combination with five integration
233 approaches (Concatenated, Averaged Stacked, Weighted NNLS, Lasso Stacked, and PLS). For concatenation,
234 the metabolomics and metagenomic abundances for each matched sample were appended to form a single long
235 feature vector that served as input for the analysis. The remaining four approaches relied on stacked
236 generalization, where separate base models were first trained on the metabolomics block and on the
237 metagenomics block. The out-of-fold predictions from the two base models were stacked to form two meta-
238 features, which were combined by a meta-learner using all four strategies: a simple mean that assigns equal
239 weight to each prediction (Averaged Stacking), weighted non-negative least square (NNLS) regression, lasso-

240 penalized regression (Lasso stacking), and partial least squares (PLS). We evaluated each stacking approach
241 on feature reduced and full dimensional datasets to assess how well these models performed under different
242 feature scenarios (**Table 1-2**). Each analysis incorporated a grid search with 3 repeated 5-fold cross-validation
243 to optimize the base models, and performance was assessed using root-mean-squared error (RMSE) for
244 continuous outcomes and area under the receiver-operating-characteristic curve (AUROC) for binary outcomes.
245 Uniform grid parameters were applied across all analyses, and features (or, for stacked models, base model-
246 level importances) were extracted for downstream selection analysis (see **Methods** for more details).

247 Random Forest with Non-Negative Least Squares Integration Emerges as the Top Performer on Datasets with
248 Continuous Response Variables

249 We analyzed 364 feature reduced and 371 full dimensional models with continuous response variables
250 across 18 datasets, leveraging 5 integration approaches and 2 individual 'omics datasets (Metabolomics, MSS).
251 To assess model performance, we selected the top 3 models, marked as "*" for the top performer, "**" as the
252 second best, and "***" as the third best performing model for each dataset (**Figure 2**) (See **Methods** for more
253 details).

254 For feature reduced and full dimensional datasets, Random Forest emerged as the top-performing
255 algorithm. For feature reduced datasets, Random Forest achieved 44.1% of the top results, surpassing XGBoost
256 (31.5%) and Elastic Net (24.1%) (**Figure 2A-B**). For full dimensional datasets, 46.3% of the best performing
257 models were Random Forest, compared to 24.1% Elastic Net and 29.6% XGBoost (**Figure 2C-D**). Among
258 integration methods, Non-Negative Least Squares (NNLS) dominated for both feature reduced and full
259 dimensional datasets (**Figure 2**). For feature reduced datasets, NNLS achieved superior performance in 72.2%
260 of cases, exceeding Averaged Stacked (5.6%), Lasso Stacked (11.1%), PLS (7.4%), and Concatenated (3.7%)
261 (**Figure A-B**). Similarly, in full dimensional datasets, NNLS dominated with 75.9%, while Averaged Stacked
262 (7.4%), Lasso Stacked (9.3%), PLS (7.4%), and Concatenated (0%) fell behind (**Figure 2C-D**).
263 Interestingly, 9 individual metabolomics models and 1 MSS model outperformed several top integration methods
264 in the features reduced datasets, with 10 metabolomics models and 1 MSS model showing similar trends in full
265 dimensional datasets (**Figure 2A**, **Figure 2C**). Despite significant feature reduction, performance comparisons
266 between feature reduced and full dimensional datasets remained consistent, underscoring the robustness of the
267 Random Forest-NNLS combination for integrating continuous datasets (**Figure 2**).

When examining validation Root Mean Squared Error (RMSE) means, Weighted NNLS demonstrated high performance in both feature reduced (40.7%) and full dimensional (38.9%) datasets, while PLS showed severe overfitting (50.0%) with low testing performance (4.7% for both dataset types) (**Figure 2, Supplemental Figure 1**). Overall, the Random Forest-NNLS combination consistently delivered superior results for datasets with continuous response variables.

Random Forest Maintains Dominance in Binary Outcome Datasets

We analyzed 217 feature reduced and 210 full dimensional models with binary response variables across 13 datasets, applying the same integration approaches and machine learning algorithms as the continuous dataset analysis (**Figure 3**, **Supplemental Figure 2**). Random Forest dominated for both the feature reduced and full dimensional datasets (**Figure 3**). 84.2% of top performing feature reduced models and 72.7% of the top performing full dimensional models were Random Forest (**Figure 3**).

For feature reduced datasets, the integration methods performed comparably among the top models: Lasso Stacked achieved 24.2%, followed by Weighted NNLS and Averaged Stacked each at 21.2%, Concatenated at 18.2%, and PLS at 15.2% (**Figure 3A-B**). In contrast, for full dimensional datasets, Averaged Stacked performed slightly better, making up 30.3% of the top performing models, followed by Lasso Stacked (21.2%), Weighted NNLS (18.2%), PLS (15.2%), and Concatenated (15.2%) (**Figure 3C-D**).

Notably, several single-omics models outperformed the integrative methods in both the feature reduced and full dimensional datasets. Among the feature reduced datasets, 2 metabolomics models and 7 MSS models showed superior performance (**Figure 3A**). In the full dimensional datasets, 4 metabolomics and 4 MSS models outperformed the integrative approaches (**Figure 3C**). Validation AUROC means indicate comparable performance across integration methods, with the exception of PLS (36.1%), which overfit (**Supplemental Figure 2**), similar to the continuous analysis results (**Supplemental Figure 1**). Elastic Net and XGBoost also displayed overfitting, as evidenced by validation AUROC means being significantly higher than testing performance (**Figure 3**, **Supplemental Figure 2**). Overall, the Random Forest algorithm demonstrated robustness across all scenarios, irrespective of integration approach (**Figure 2-3**).

Uncovering Trends in Feature Selection Across Elastic Net, Random Forest and XGBoost

Following model development, we next sought to determine whether our base models (Elastic Net, Random Forest, and XGBoost) identified similar features across concatenated and single-omics (Metabolomics

296 and Taxa) datasets. Understanding how different models select features across various dataset conditions
297 (feature reduced vs. full dimensional, binary vs. continuous) is crucial for evaluating model consistency and the
298 impact of preprocessing steps on feature selection.

299 To assess feature selection across models, we first extracted feature importance metrics from each
300 cross-validation step (see **Methods** for more details). For each run, we computed the absolute value of each
301 feature's importance score and applied Min-Max Normalization. This normalization ensures that feature
302 importance values are comparable across models, preventing any single model from dominating due to
303 differences in scale. We then used the Mean Importance to identify the top overall features for each model. For
304 this analysis, we grouped features into bins based on their importance scores within each model, selecting those
305 with the highest feature importance. To ensure rigor, we used predefined bin sizes, which represent the top 1,
306 5, 10, 20 and 50 features. These discrete bins create a manageable list of high-priority candidates that can be
307 feasibly validated through peer review, or experimentally through *in vivo* and *in vitro* studies. These predefined
308 lists sizes allow us to compare the most influential features across models without the impractical task of vetting
309 each variable across every dataset.

310 Models were excluded from the analysis if they had fewer than the required number of features with non-
311 zero importance scores, or if Elastic Net failed to complete using the default parameter grid (see **Methods** for
312 more details). Violin plots are used for the following analysis to depict the distribution of the proportions of
313 features selected that are shared and/or unique across Elastic Net, Random Forest and XGBoost for each
314 dataset.

315 <u>Metabolite Features Selected by Tree-Based Methods Show Strong Concordance, in Contrast to the Distinct</u>
316 <u>Features Selected by Elastic Net</u>

317 We examined metabolite feature selection patterns from Elastic Net, Random Forest, and XGBoost
318 across feature reduced and full dimensional datasets with binary and continuous response variables (**Figure**
319 **4A**). Across all conditions, Elastic Net consistently identified a large proportion of unique features, regardless of
320 dataset dimensionality, response variable type or bin size (1, 5, 10, 20 or 50) (**Figure 4B-E**).

321 Overlap among the top-ranked features selected by all three models was minimal (**Figure 4B-E**). In
322 nearly all cases, the top feature selected by Elastic Net, Random Forest and XGBoost differed, irrespective of

323 whether the response variable was binary or continuous, and whether the dataset was full dimensional or feature
324 reduced. Only one dataset with a continuous response variable and reduced features (**Figure 4B**) and two
325 datasets with binary response with full dimensional features (**Figure 4E**) showed any overlap in the top ranked
326 feature across all three models.

327 Another consistent trend across all conditions were that overlap between Elastic Net and the tree-based
328 models were limited (**Figure 4**). For feature bins ranging from 5 to 50, overlapping features between Elastic Net
329 and either, Random Forest or XGBoost accounted for less than 25% of the selected features across all dataset
330 types. Two exceptions to the trend occurred in the feature-reduced continuous dataset (**Figure 4B**), where these
331 comparisons showed approximately 40% overlap in the top 5 features for both Elastic Net and Random Forest
332 and Elastic Net and XGBoost.

333 Interestingly, in feature reduced datasets with continuous response variables, we observed higher levels
334 of feature overlap among the three models compared to full dimensional datasets with continuous response
335 variables (**Figure 4B-C**). This trend, visualized through violin plots of the overlapping distributions, suggests that
336 dimensionality reduction may increase agreement in feature selection across different modeling approaches for
337 metabolites with continuous outcomes. This was observed across all conditions, regardless of response variable
338 type or dimensionality. Violin plots of overlap proportions across bins 1 to 50 also showed a clear pattern of
339 increased agreement between the two tree-based models (**Figure 4**). These results suggest that Random Forest
340 and XGBoost tend to select a stable subset of shared features, regardless of dataset structure or outcome type,
341 in contrast to the increased unique selections made by Elastic Net.

342 <u>Tree-Models Reveal Partial Concordance in Taxa Selection While Elastic Net Remains Distinct</u>

343 To assess the consistency and overlap of taxa selected by Elastic Net, Random Forest, and XGBoost,
344 we analyzed feature selection patterns across full dimensional and feature reduced datasets under binary and
345 continuous response variables (**Figure 5A**). Across all datasets, there were no instances where the top-ranked
346 feature was shared among all three models and there were no cases where the top feature was shared between
347 Elastic Net and Random Forest or between Elastic Net and XGBoost (**Figure 5B-E**). This lack of consensus for
348 the top feature highlights the distinct feature selection behavior for Elastic Net, XGBoost and Random Forest.
349 Consistent with these findings, overlap in selected features across bins 5-50 remained low for all model
350 comparisons with fewer than 25% of features overlapping across all dataset conditions (**Figure 5B-E**).

Patterns of top feature overlap between the tree-based models varied depending on response type and dimensionality. For continuous response variables, Random Forest and XGBoost exhibited greater agreement in their top features when applied to full-dimensional datasets compared to feature reduced datasets (**Figure 5B-E**). This trend was reversed for binary response variables, where tree-based models showed higher concordance in their top selected features in feature reduced datasets than in full dimensional datasets (**Figure 5D-E**).

Violin plot distributions of feature overlap between Random Forest and XGBoost for bins 5-50 maintained a consistent pattern across all conditions (**Figure 5B-E**). The distributions of shared features closely resembled those of features uniquely identified by each model. On average, approximately half of the selected features in these bins were shared between the two tree-based methods, while the other half were distinct to either Random Forest or XGBoost (**Figure 5B-E**). This indicates that, although the tree-based models exhibit some level of agreement, each still captures a large subset of unique taxa features.

In contrast, Elastic Net consistently selected a largely distinct set of features across all datasets and bin sizes. For bins 1-50, more than 75% of the selected features were unique to Elastic Net for the majority of datasets (**Figure 5B-E**). The high proportion of unique selections further underscores the importance of model selection and comparison of features selected by each model.

<u>Distinct Elastic Net Selections Contrast with Consistent Tree-Based Features for Concatenated Models</u>

To assess the consistency and overlap of features selected from concatenated models, which combine both taxa and metabolite data, we evaluated feature selection patterns from Elastic Net, Random Forest and XGBoost across full dimensional and feature reduced datasets, using binary and continuous response variables (**Figure 6A**). The overall trends for continuous response variables for the concatenated models closely mirrored those observed in the taxa-only models (**Figure 5B-C, Figure 6B-C**). Specifically, overlap in the top feature selection across all three models was rare (**Figure 6B-C**). Only one feature reduced dataset showed agreement among all three models for the top ranked feature, and no overlap was observed in full-dimensional datasets or for any of the binary response datasets (**Figure B-E**).

Only one continuous dataset showed overlap between Elastic Net and XGBoost for the top feature selected, while no overlap was observed between Elastic Net and Random Forest across any dataset, regardless of response type or dimensionality (**Figure 6B-E**). In contrast, Random Forest and XGBoost exhibited relatively

379 high agreement for their top selected features across binary and continuous datasets, with approximately half of
380 the proportion of features being shared (**Figure 6B-E**). Elastic Net consistently selected distinct top features for
381 continuous datasets, with the exception of two feature reduced datasets and one full dimensional dataset (**Figure**
382 **6B-C**).

383 When examining the top selected features (bins 5-50, which represent the top 5 through 50 features
384 selected per dataset), this trend of limited agreement between Elastic Net and tree-based models continued.
385 The proportion of overlapping features among all three models, as well as between Elastic Net and either tree-
386 based methods remained low across most datasets. Overlapping proportions exceeding 25% were rare, and the
387 majority of distributions were centered near zero, further confirming that Elastic Net largely selected features
388 distinct from those chosen by Random Forest and XGBoost (**Figure 6B-E**). Similar to the taxa-only models,
389 Random Forest and XGBoost demonstrated greater consistency in their feature selections for the concatenated
390 datasets (**Figure 5-6**). Across bins 5-50, approximately 50% of the selected features were shared between the
391 two tree-based model, while the remaining features were largely unique to each method (**Figure 6B-E**). These
392 results reinforce the pattern that Elastic Net tends to prioritize a separate set of features, while Random Forest
393 and XGBoost show more concordance in their selections, even when analyzing integrated taxonomic and
394 metabolomics data.

395 <u>Dimensionality Reduction Influences Feature Selection Stability in Machine Learning Models</u>

396 To evaluate the impact of dimensionality reduction on model behavior, we compared feature selection
397 between full dimensional and feature reduced datasets for each of the machine learning methods, Elastic Net,
398 Random Forest and XGBoost, across all metabolomics, taxa and concatenated datasets using binary and
399 continuous response variables (**Figure 7A**). Our goal was to determine whether dimensionality affected feature
400 selection stability and whether the top features selected remained consistent across the different machine
401 learning algorithms.

402 Across all data types and response variables, Elastic Net consistently demonstrated minimal agreement
403 in top feature selection for the full dimensional and feature reduced models (**Figure 7B-G**). Regardless of
404 whether the dataset involved metabolomics, taxonomy or concatenated or the response type, only 1-2 datasets
405 showed overlap in the top feature selected by Elastic Net (**Figure 7B-G**). In contrast, the tree-based models,
406 with greater emphasis on Random Forest, showed stronger consistency in the top features selected for the

407 metabolomics datasets (**Figure 7B, Figure 7E**). Approximately half of the Random Forest models identified the
408 same top feature in both the full and feature reduced datasets for both binary and continuous response variables.
409 XGBoost followed a similar trend for metabolomics data, with slightly lower proportions of agreement in top
410 features (**Figure 7B, Figure 7E**).

411 In contrast, for taxa and concatenated datasets, top feature agreement between full dimensional and
412 reduced models was rare, even for the tree-based methods (**Figure 7C-F**, **Figure 7D-G**). Across these analyses,
413 Random Forest and XGBoost typically shared the same top feature in only 1-4 datasets indicating that
414 dimensionality reduction substantially influences which features are prioritized for these data types (**Figure 7C-**
415 **F**, **Figure 7D-G**).

416 When examining bins 5-50, further insights emerged regarding the broader set of selected features. For
417 metabolomics data, regardless of response type, the majority of datasets showed less than 50% overlap in
418 features between the full dimensional and feature reduced models (**Figure 7B, Figure 7E**). Most distributions
419 were centered around zero, with emphasis on Elastic Net, which displayed major differences in selected features.
420 Notably, Random Forest models applied to metabolomics datasets demonstrated relatively even distributions of
421 overlap across bins 5-50, while XGBoost distributions differed based on response type (**Figure 7B, Figure 7E**).
422 For continuous outcomes, XGBoost showed a concentration of overlapping proportions in the 0.0-0.20 range,
423 whereas binary outcomes exhibited a more even spread across all feature bins (**Figure 7B, Figure 7E**).

424 For taxa data, feature overlap across bins 5-50 was sparse, with the highest density of violin plot
425 distributions concentrated at zero for both binary and continuous response types (**Figure 7C, Figure 7F**). Some
426 exceptions were observed, which include one continuous taxa dataset at bin 5 and two binary taxa datasets at
427 bin 5. These showed complete agreement in selected features between full and reduced models (**Figure 7C,**
428 **Figure 7F**). Random Forest models developed using taxa data showed a more evenly distributed pattern of
429 overlap, whereas XGBoost had a concentration near zero but still demonstrated higher concordance than Elastic
430 Net across most bins (**Figure 7C**, **Figure 7F**).

431 Trends observed in the concatenated models closely mirrored these seen in the taxa data (**Figure 7D**,
432 **Figure 7G**). However, distributions for bins 5-50 were generally shifted closer to zero across all three modeling
433 approaches, indicating even lower agreement in selected features between full dimensional and feature reduced
434 models (**Figure 7D**, **Figure 7G**). This suggests that when combining multiple data types, feature selection

435 becomes increasingly sensitive to dimensionality, especially for Elastic Net, which consistently demonstrated the
436 least concordance between full and feature reduced models.

437 Differences in Concatenated and Single 'Omics Feature Selection

438 To determine whether concatenated models incorporating both metabolomics and taxa data could
439 recover features identified by single 'omics models, we compared feature selection patterns across Elastic Net,
440 Random Forest and XGBoost for both full dimensional and feature reduced datasets with continuous and binary
441 response types (**Figure 8A**, **Figure 8F**). For metabolomics data, Elastic Net models trained on feature reduced
442 concatenated datasets showed no overlap in features with the corresponding single 'omics metabolomics models
443 across bins 1-50, regardless of response type (**Figure 8B-E**). In the full dimensional datasets, there was limited
444 agreement between the concatenated and metabolomics models for continuous outcomes, with most overlap
445 proportions falling below 0.4 and the highest density occurring near zero **(Figure 8C**). For full dimensional data
446 with binary outcomes, consensus remained minimal with many datasets showing no overlapping features
447 between the concatenated and metabolomics models (**Figure 8D-E**).

448 In contrast, Random Forest and XGBoost models demonstrated a greater ability to identify overlapping
449 metabolite features between concatenated and metabolomics single 'omics datasets (**Figure 8B-E**). Across top
450 feature bins 1-50, both models recovered shared features more proportionally, and in several datasets, the same
451 top features were selected. For feature reduced datasets, however, this agreement diminished, particularly for
452 continuous response variables (**Figure 8B-C**). The strongest agreement was observed in the full dimensional
453 models, where overlap between the concatenated and metabolomics datasets were more evenly distributed
454 across bin sizes (**Figure 8C**, **Figure 8E**). A notable exception was seen in the top 50 metabolite bin for XGBoost,
455 where the distribution was reduced (**Figure 8C**, **Figure 8E**).

456 When comparing taxa features between concatenated and single-omics models, we observed a higher
457 degree of consensus relative to the metabolite comparisons. Elastic Net models, surprisingly, displayed
458 substantial overlap across all bins, with the majority of feature overlap proportions falling in the 0.75 to 1.0 range
459 across both full dimensional and feature reduced datasets, regardless of response type (**Figure 8G-J**). An
460 exception occurred in the top feature comparison for the full dimensional binary dataset, where only one dataset
461 showed agreement between the concatenated and taxa models (**Figure 8J**).

For Random Forest and XGBoost, the proportion of overlapping taxa features between concatenated and single ‘omics datasets remained proportional across feature bins 5-50, regardless of feature dimensionality (**Figure 8G-J**). Slightly less agreement was observed in the top feature selection for datasets with continuous responses when compared to binary, where the latter demonstrated higher and more proportionate overlap (**Figure 8G-J**).

To further investigate how different data types influence model behavior for concatenated data, we evaluated the proportion of taxa versus metabolite features selected across bins sizes 1-50 (**Supplemental Figure 3A**). For Elastic Net, taxa consistently dominated feature selection across all datasets, with top features in every bin originating from the taxa data, regardless of dimensionality or response type (**Supplemental Figure 3B-E**). This suggests Elastic Net may be more sensitive to high dimensional features within concatenated models. In contrast, Random Forest and XGBoost selected a more balanced mix of metabolites and taxa features across bins, indicating a more even consideration of both data modalities in feature importance rankings (**Supplemental Figure 3B-E**).

Discussion

This study presents a comprehensive evaluation of machine learning performance and feature selection stability across multi-omics microbiome and metabolomics datasets using Elastic Net, Random Forest and XGBoost. By systematically comparing these models across both continuous and binary response variables, full dimensional and feature reduced datasets, and integration strategies including concatenation and ensemble-based methods, we aimed to uncover generalizable trends in model behavior and feature selection stability (**Figure 1**). One of the most consistent findings was the robust performance of Random Forest, particularly when combined with Non-Negative Least Squares (NNLS) integration, for datasets with continuous outcomes (**Figure 2**). This is consistent with prior studies highlighting the capacity of Random Forest to handle noisy, sparse, nonlinear and high dimensional data, such as microbiome data, effectively[5,27,28]. NNLS further enhanced model performance, reinforcing recent evidence that integration methods leveraging weighted combinations of predictions can outperform simpler or more complex stacking strategies for microbiome multi-omics data[16].

Across binary outcomes, Random Forest also outperformed other models, though the advantage was less pronounced with more comparable performances across the stacked integration methods (**Figure 3**). Notably, several individual ‘omics models outperformed integrative models for both binary and continuous

490 responses, particularly metabolomics models. This suggests that in certain cases, single ‘omics models may
491 identify strong predictive features and/or relationships that are lost during integration. We also observed that
492 model performance results were largely consistent between full dimensional and feature reduced datasets, with
493 Random Forest maintaining top performance across both response types (**Figure 2-3**). This suggests that,
494 despite substantial reductions in input features, overall predictive performance can be preserved when robust
495 models and integration strategies are applied[29].

496 In terms of feature selection, a key trend that emerged across all datasets is that Elastic Net consistently
497 identified a distinct and largely non-overlapping set of features compared to Random Forest and XGBoost
498 (**Figure 4-6**). This divergence is likely attributable to the linear and sparsity-inducing nature of Elastic Net, which
499 tends to select a small subset of features that are linearly associated with the outcome and includes groups of
500 correlated predictors. Alternatively, tree-based models are more adept at capturing nonlinear, higher-order
501 interactions and distributed signals[30].

502 This trend was particularly evident in the metabolomics datasets, where Random Forest and XGBoost
503 frequently shared more than 50% of their top-ranked features across bins, while Elastic Net consistently selected
504 a largely distinct set of features (**Figure 4**). The divergence in feature selection became even more pronounced
505 in the concatenated models that combined metabolomics and taxonomic data (**Figures 5–6**). Feature selection
506 stability was also more sensitive to dimensionality, with Elastic Net showing minimal overlap in top-ranked
507 features between full and feature-reduced datasets, especially in metabolomics, whereas Random Forest and
508 XGBoost demonstrated moderate consistency (**Figure 7**). These findings suggest that dimensionality reduction
509 alters the feature landscape more substantially for linear models, and that feature interpretability and
510 reproducibility in Elastic Net may be more influenced by preprocessing choices compared to tree-based
511 approaches.

512 When comparing feature selection between single ‘omics and concatenated models, tree-based methods
513 retained a relatively high degree of consistency, while Elastic Net rarely recovered the same features—especially
514 in metabolomics analyses (**Figure 8**). Interestingly, for taxonomic features, Elastic Net showed much higher
515 agreement between concatenated and single ‘omics models. This pattern likely reflects higher dimensionality in
516 the concatenated datasets, where taxa features outnumber metabolite features. As a result, Elastic Net appears

517 biased towards selecting taxa features, while Random Forest and XGBoost maintained a more balanced
518 selection across both data types. These results are consistent with previous findings suggesting that tree-based
519 models are more resilient to greater numbers of features and sparsity in high-dimensional data[30].

520 While this work provides several new insights into model behavior and feature selection for microbiome
521 multi-omics data, there are several limitations. First, the study was conducted using four multi-omics microbiome
522 datasets[11,14,25,26], limiting generalizability. Broader applications across a wider range of microbiome-related
523 diseases would help validate these findings. Additionally, taxonomic profiling was performed using Kracken
524 v2.1.1 and Bracken v2.8, which may produce different results when compared to other classification tools[31,32].
525 We also only selected one method for feature reduction analyses. We used Limma[33] for metabolomics data
526 and Wilcoxon signed-rank test for the taxonomic relative abundance data. These might not be the best method
527 for feature reduction and/or may not accurately reflect feature signal structure across all datasets. Exploration
528 into other methods of microbiome data transformation and its effect on multi-omics data integration should be
529 explored[34]. Furthermore, the concatenated datasets suffered from greater dimensionality, with taxa feature
530 outnumbering metabolomics for all datasets. This imbalance likely contributed to Elastic Net's consistent
531 selection of taxa-only features. Finally, a uniform hyperparameter grid was used across all models, which
532 provided standardization but may not have captured optimal tuning for individual datasets. Future works should
533 incorporate additional datasets exploring diverse diseases, consider additional integration methods such as
534 DIABLO[35] or MOFA+[36], explore alternative dimensionality reduction strategies and taxa classification
535 strategies and include refined hyperparameter optimization. Additionally, biological validation of selected
536 features, through literature mining or experimental follow-up, is crucial to ensure biological relevance. Overall,
537 this study offers a robust framework for evaluating machine learning performance and feature selection in multi-
538 omics microbiome data, highlighting the strengths and limitations of current modeling strategies and laying the
539 groundwork for more interpretable and scalable integrative analysis.

540 **Key Points**

541 - A total of 1,323 models were developed to comprehensively evaluate prediction performance and the
542 robustness of feature selection using metabolomics and taxonomy datasets. These models included
543 three widely used machine learning algorithms – Elastic Net, Random Forest and XGBoost – applied

544 across five integration strategies and single 'omics approaches on datasets with binary and continuous
545 outcomes.

546 - For continuous outcomes, Random Forest combined with NNLS integration achieved the highest
547 performance and maintained strong predictive performance across full dimensional and feature reduced
548 datasets.
549 - For binary outcomes, Random Forest consistently performed well regardless of the integration strategy.
550 Notably, single 'omics models, especially those using metabolomics data, outperformed integrative
551 approaches.
552 - Tree-based models demonstrated greater consistency in feature selection across different
553 dimensionalities and integration strategies.

554 **Data Availability**

555 Data is free and available through: https://github.com/borenstein-lab/microbiome-metabolome-curated-
556 data/wiki/Data-overview#datasets-included[33]. Code for this integrative analysis and associated analysis are free
557 and available through: https://github.com/suziepalmer10/Multiomics-Integrative-Pipeline.

558 **Funding**

559 This work is supported by National Institute of Allergy and Infectious Disease [grant numbers AI179406 to A.Y.K.,
560 AI169298 to X.Z., 5T32AI005284-43 to S.N.P.]; the University of Texas Southwestern Medical Center and
561 Children's Health Cellular and ImmunoTherapeutics Program (to A.Y.K.); National Human Genome Research
562 Institute [grant number HG011035 to D.L.]; and National Institute of General Medical Sciences [grant number
563 GM126479 to X.Z.].

564 **Competing Interests**

565 AYK is a consultant for Prolacta Bioscience. AYK receives research funding from Novartis. AYK is co-founder of
566 Aumenta Biosciences.

567

Table 1: Details for Data With Continuous Response Variables

			Featured Reduced Details					Full Dimensional Details				
					Run Time (HH:MM)					Run Time (HH:MM)		
Study Name (Subset)	Response Variable	Sample Size	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost
Erawijantari	Age	96	71	775	00:09	00:01	01:55	524	3103	00:25	00:05	03:16
Erawijantari	Total Cholesterol	82	71	775	00:09	00:01	01:50	524	3103	00:25	00:06	02:48
Erawijantari	Glucose	84	71	775	00:09	00:01	01:47	524	3103	00:25	00:06	02:40
Franzosa (Control vs. CD)	Age	144	2175	3603	00:30	00:13	04:50	8848	5137	01:34	00:47	09:56
Franzosa (Control vs. CD)	Fecal Calprotectin (Fp)	107	2175	3603	00:30	00:10	04:10	8848	5137	01:06	00:32	06:50
Franzosa (Control vs. IBD)	Age	220	2519	3849	00:39	00:24	06:15	8848	5137	01:18	01:00	10:11
Franzosa (Control vs IBD	Fecal Calprotectin (Fp)	153	2519	3849	00:33	00:16	04:54	8848	5137	01:08	00:42	07:48
Franzosa (Control vs. UC)	Age	132	1268	1816	00:18	00:06	03:43	8848	5137	01:31	00:45	09:25
Franzosa (Control vs. UC)	Fecal Calprotectin (Fp)	88	1268	1816	00:16	00:04	03:01	8848	5137	01:25	00:36	06:33
Wang	Age	287	71	2261	00:13	00:09	03:22	276	5293	00:26	00:26	04:50
Wang	BMI	280	71	2261	-	00:09	03:06	276	5293	00:26	00:27	04:26
Wang	Creatinine	281	71	2261	00:12	00:10	02:49	276	5293	00:25	00:30	04:10
Wang	eGFR	281	71	2261	00:12	00:11	03:37	276	5293	00:25	00:32	05:11
Wang	Urea	281	71	2261	00:13	00:09	03:33	276	5293	00:25	00:27	04:59
Yachida (All Samples)	Age	347	173	1823	00:23	00:11	03:22	450	5770	00:32	00:34	05:27
Yachida (All Samples)	Alcohol	347	173	1823	00:24	00:11	03:43	450	5770	00:31	00:33	05:47
Yachida (All Samples)	BMI	347	173	1823	-	00:11	03:05	450	5770	-	00:35	05:05
Yachida (All Samples)	Brinkman Index	347	173	1823	00:25	00:12	03:41	450	5770	00:32	00:36	05:46

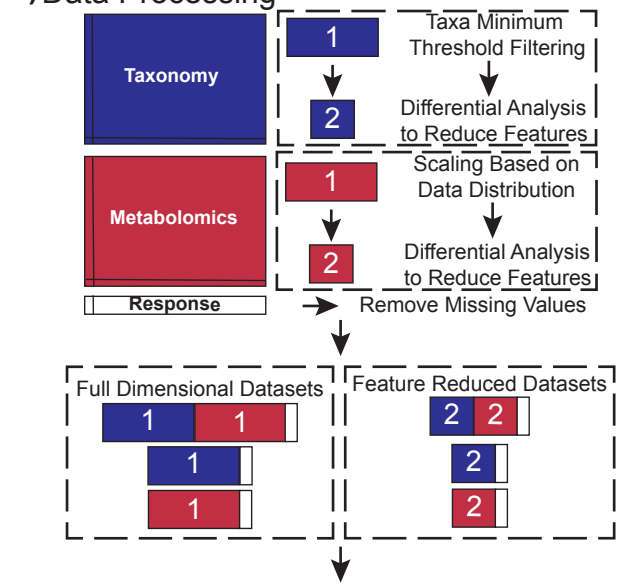
'-' represents models that failed to complete.

Table 2: Details for Data With Binary Response Variables

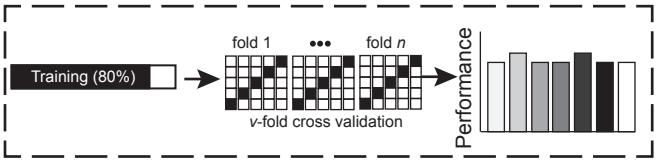
			Featured Reduced Details					Full Dimensional Details				
					Run Time (HH:MM)					Run Time (HH:MM)		
Study Name (Subset)	Response Variable	Sample Size	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost
Erawijantari	No Alcohol vs. Alcohol	96 (47/49)	71	775	00:09	00:01	01:55	524	3103	00:25	00:05	03:16
Erawijantari	Female vs. Male	96 (32/64)	71	775	-	00:04	01:16	524	3103	00:31	00:28	02:14
Erawijantari	Control vs. Gastrectomy	96 (54/42)	71	775	00:11	00:04	01:08	524	3103	00:23	00:27	02:09
Franzosa (Control vs. CD)	Control vs. CD	144 (56/88)	2175	3603	00:38	01:13	02:43	8848	5137	01:52	04:52	05:29
Franzosa (Control vs. IBD)	Control vs. IBD	220 (56/164)	2519	3849	00:54	02:19	03:16	8848	5137	01:48	07:34	05:34
Franzosa (Control vs. UC)	Control vs. UC	132 (56/76)	1268	1816	00:22	00:30	02:07	8848	5137	01:53	04:45	05:21
Wang	Female vs. Male	287 (107/180	71	2261	-	01:03	02:09	276	5293	-	03:49	03:06
Wang	Control vs. Kidney Failure	287 (67/220)	71	2261	00:17	00:50	01:36	276	5293	00:34	03:11	02:29
Yachida (All Samples)	Female vs. Male	347 (143/204	173	1823	00:18	01:08	02:12	450	5770	00:42	05:16	03:39
Yachida (Control vs MP & CRC)	Control vs. MP and CRC	277 (127/150)	173	1823	00:16	00:51	01:58	450	5770	00:35	04:02	03:15
Yachida (Control vs MP & Stage 0)	Control vs. MP & Stage 0 CRC	194 (127/67)	173	1823	00:07	00:09	01:24	450	5770	-	02:28	02:46
Yachida (Control vs Stage 1 & 2)	Control vs. Stage 1 & 2 CRC	196 (127/67)	173	1823	00:05	00:05	01:13	450	5770	-	02:30	02:47
Yachida (Control vs Stage 3 & 4)	Control vs. Stage 3 & 4 CRC	181 (127/54)	173	1823	00:07	00:08	01:15	450	5770	-	02:08	02:34

'-' represents models that failed to complete.

A) Data Processing

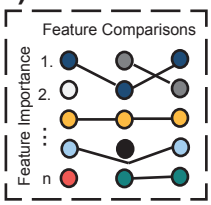


B) Model Training and Testing



Input Models: Elastic Net, Random Forest, XGBoost
Integration Techniques: Single Omics (Taxonomy and Metabolomics), Concatenatation, Averaged Stacking, Weighted Non-Negative Least Squares (NNLS), Lasso Stacking, Partial Least Squares.

C) Feature Evaluation

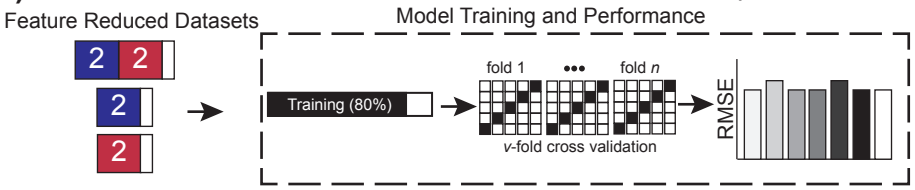


Comparisons:
Elastic Net vs. Random Forest vs. XGBoost
Full Dimensionall vs. Feature Reduced Datasets
Concatenated vs. Taxonomy/Metabolomics

Figure Legends

Figure 1. Pipeline for the Multi-omics Integrative Analysis. A) Matched taxonomy and metabolomics data are processed separately. Taxa with an abundance below 0.001 across all samples are removed. Depending on the study design and the method used for metabolomics, data are appropriately scaled. Response variables with missing values are excluded from downstream analyses. Differential expression analyses are conducted separately for metabolomics (indicated in red) and taxa (indicated in blue). Both the original datasets (taxa – 1, metabolomics – 1) and their corresponding response variables are concatenated. Similarly, the feature reduced datasets (taxa – 2, metabolomics – 2) and their response variables are concatenated, enabling comparisons between original and feature-reduced data. **B)** The processed datasets are then analyzed using the multi-omics integration pipeline. The data are first split into 80% training and 20% testing sets. The training data undergo 3 × 5-fold cross-validation to optimize model performance. Performance metrics—area under the receiver operating characteristic (ROC) curve (AUROC) for binary outcomes and root mean squared error (RMSE) for continuous outcomes—are used to evaluate both validation and testing performance. The base machine learning models are Elastic Net, XGBoost, and Random Forest. Integration strategies include simple concatenation, averaged stacking, weighted non-negative least squares (NNLS), Lasso stacking, and Partial Least Squares (PLS). Additionally, single ‘omics models are generated separately for taxonomy and metabolomics datasets. For each response variable, 21 models are generated (or attempted), representing various combinations of base learners, integration methods, and data types. **C)** After model generation, feature extraction is performed. Top features selected by each machine learning model (Elastic Net, Random Forest, XGBoost) are compared to identify consistently important variables. Further comparisons are conducted to assess feature selection differences between full dimensional and feature reduced datasets. Lastly, the features selected from concatenated taxonomy and metabolomics base models are compared to those identified in the integrated concatenated models to evaluate the impact of data integration on feature importance.

A) Performance of Feature Reduced Data with Continuous Response Variables



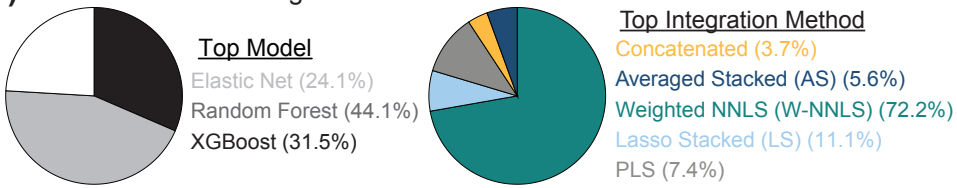
Testing RMSE

$\log_{10}$(RMSE): 3, 2, 1, 0

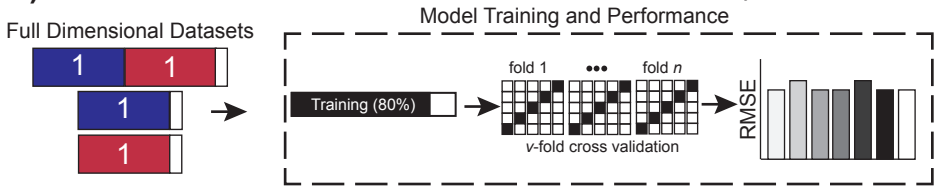
Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
		11.26	21.60	16.89	15.04	10.85**	11.36	11.62
		11.63	11.74	11.63	11.59	11.09***	11.76	12.49
		11.55	13.20	13.12	11.90	10.39*	12.78	13.27
		14.83>*	157.15	138.14	79.39	18.01**	29.80	46.06
		16.98>*	19.61	19.38	17.67*	18.50***	18.98	20.98
		20.43	21.66	21.94	18.68	19.26	20.22	22.14
		12.11	94.64	107.47	49.80	10.99**	13.01	18.21
		11.91	11.39	11.34	11.56	10.95*	11.83	12.97
		13.94	12.46	11.92	12.43	10.99***	12.69	12.94
		13.67	182.15	64.57	90.76	12.56*	14.21	24.54
		13.60	13.54	13.46	13.52	12.96**	13.46	14.14
		14.14	14.57	13.83	13.83	13.16***	13.78	14.08
		1.75	2.46	1.81	1.99	1.70**	1.78	1.88
		1.83	2.04	1.91	1.92	1.73***	1.84	1.90
		1.76	2.23	1.83	1.90	1.68*	1.82	1.87
		15.38>***	106.68	45.96	55.77	15.08*	15.61***	15.95
		16.27	17.10	16.63	16.56	16.08	16.40	17.55
		15.93	17.76	16.84	16.14	15.55**	15.91	16.32
		1.26>***	2.03	1.62	1.52	1.24*	1.28**	1.28***
		1.45	1.78	1.53	1.58	1.37	1.48	1.48
		1.49	1.83	1.53	1.50	1.42	1.49	1.55
		16.84	46.81	51.34	27.74	15.83**	17.40	18.96
		17.82	17.27	17.59	17.32	16.05***	17.40	20.08
		16.49	19.04	18.04	16.98	15.42*	17.41	17.32
		1.49>****	2.56	1.92	1.96	1.28*	1.54	1.45
		1.74	2.19	1.88	1.95	1.36**	1.75	1.63
		1.75	2.11	1.86	1.85	1.52***	1.79	1.84
		11.43	17.46	17.24	12.53	11.15	11.65	12.06
		11.11	10.95	10.90***	10.95	10.86**	10.98	11.30
		11.94	11.22	11.05	11.06	10.85*	11.05	11.04
		3.47	3.54	3.52	3.47	3.38**	3.46***	3.55
		3.81	3.80	3.73	3.61	3.38*	3.67	3.63
		1.39	1.64	1.41	1.39	1.37	1.37	1.44
		0.81	1.19	1.03	0.97	0.63***	0.81	0.66
		0.62>****	1.13	0.71	0.77	0.61*	0.62**	0.64
		39.12	263.71	187.58	133.68	31.88	45.28	46.89
		16.94	28.08	23.80	21.82	10.68*	14.88	12.30
		13.09>****	25.49	16.95	16.47	12.56**	13.34***	13.71
		9.51	37.27	16.47	20.24	9.44	11.28	10.93
		7.55	9.00	8.50	8.14	6.74*	7.42	7.02**
		7.37	8.80	7.61	7.54	7.02***	7.24	7.26
		2.84>****	2.90	2.90	2.85	2.82*	2.85***	2.86
		3.25	3.08	3.09	3.05	2.82**	3.06	2.94
		4.63	4.61	4.61	4.62	4.53	4.63	4.62
		4.46	4.52	4.52	4.45***	4.39*	4.50	4.45**
		4.84	4.61	4.65	4.57	4.47	4.61	4.65
		11.11	17.79	19.95	13.11	10.83	12.36	12.64
		11.35	9.56>****	9.60***	10.31	9.12*	9.74	9.53**
		12.18	10.04	9.88	10.52	10.15	10.31	10.33
		4.20	4.24	4.24	4.21	4.18	4.23	4.43
		4.09>****	4.21	4.21	4.13	4.07**	4.15	4.18
		4.18	4.39	4.27	4.12***	4.03*	4.17	4.17

Row labels: Age; Cholesterol; Glucose; Control vs. CD Age; Control vs. CD Fp; Control vs. IBD Age; Control vs. IBD Fp; Control vs. UC Age; Control vs. UC Fp; Age; BMI; Creatinine; eGFR; Urea; BMI; Brinkman Index; Age; Alcohol

B) Feature Reduced Categorical Distributions with the Lowest RMSE



C) Performance of Full Dimensional Data with Continuous Response Variables



Testing RMSE

$\log_{10}$(RMSE): 4, 3, 2, 1, 0

Dataset: Erawijantari, Franzosa, Wang, Yachida

Model: Elastic Net, Random Forest, XGBoost

Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
		10.58>***	582.21	463.18	291.49	10.96***	129.24	104.22
		11.23	11.85	11.75	11.47	10.87**	11.77	12.88
		11.00	12.85	12.48	11.44	10.31*	12.20	12.95
		15.78>**	1536.37	1353.00	768.17	19.32***	229.59	371.82
		18.33>**	20.83	20.32	19.27**	18.96*	19.44	19.85
		21.36	23.55	22.16	20.34	20.29	21.93	24.67
		13.60	841.64	457.05	422.08	10.75**	29.37	119.57
		11.93	12.08	11.94	11.91	10.71*	12.19	13.27
		14.56	13.99	14.61	13.31	10.78***	13.99	12.99
		14.10	287.77	142.76	143.60	13.08***	14.26	16.93
		13.75	13.47	13.58	13.55	13.03**	13.65	14.50
		13.80	13.99	14.02	13.39	12.96*	13.57	13.72
		1.84	2.35	1.99	1.97	1.78*	1.87	1.96
		1.86	2.01	1.88	1.92	1.82***	1.88	2.02
		1.91	2.25	2.01	1.97	1.80**	1.96	2.00
		14.91>***	127.82	39.25	66.16	14.56*	15.15**	15.49
		15.97	17.03	16.22	16.36	15.87	16.23	17.05
		15.88	17.42	16.28	15.92	15.54***	15.86	16.05
		1.36>**	2.03	1.57	1.57	1.34*	1.37***	1.38
		1.49	1.80	1.53	1.62	1.37**	1.51	1.48
		1.51	1.86	1.53	1.53	1.43	1.50	1.63
		15.64>****	142.99	32.77	74.16	15.18*	18.95	27.73
		16.21	17.63	16.60	16.77	15.75***	16.92	20.05
		16.55	18.64	16.61	16.90	15.54**	16.92	17.90
		1.65>****	2.56	1.99	2.00	1.49*	1.69***	1.78
		1.95	2.23	1.99	2.08	1.62**	1.96	1.79
		1.92	2.19	1.90	1.94	1.78	1.92	1.93
		11.70	40.69	39.72	23.70	11.12	11.52	11.58
		10.92	10.82	10.79	10.79	10.65*	10.84	10.93
		11.33	11.09	10.76	10.66***	10.57*	10.85	10.98
		3.53	3.93	3.86	3.67	3.42***	3.57	3.70
		3.49	3.54	3.53	3.49	3.39*	3.48	3.54
		3.69	3.73	3.63	3.54	3.39**	3.59	3.59
		1.28	1.64	1.41	1.36	1.31	1.32	1.36
		0.83	1.24	1.10	1.00	0.65*	0.83	0.68
		0.66>***	1.19	0.71	0.83	0.65**	0.67***	0.70
		28.49	343.64	220.18	171.48	28.70	36.18	41.98
		17.26	29.38	25.61	22.69	10.73*	15.58	12.08**
		13.36	26.22	15.70	17.25	13.26***	14.21	14.85
		9.00	21.85	13.74	13.65	8.96	10.80	10.22
		7.64	9.20	8.76	8.27	6.78*	7.51	7.04***
		7.15	9.19	7.64	7.64	6.85**	7.18	7.22
		2.84>****	2.90	2.90	2.86	2.82**	2.85***	2.86
		3.09	3.03	3.02	2.95	2.82*	2.98	2.90
		4.63	4.61	4.61	4.62	4.53	4.63	4.64
		4.40	4.46	4.45	4.40***	4.26*	4.44	4.33**
		4.61	4.64	4.56	4.47	4.42	4.53	4.54
		11.15	21.82	26.56	15.04	10.88	12.43	12.05
		11.16	9.80>****	9.83	10.37	9.39*	9.83	9.65**
		11.51	9.91	9.99	10.24	9.82***	10.03	10.07
		4.21	4.23	4.23	4.21	4.17	4.22	4.34
		4.09>***	4.21	4.20	4.13***	4.07**	4.15	4.20
		4.18	4.37	4.30	4.14	4.06*	4.14	4.17

Row labels: Age; Cholesterol; Glucose; Control vs. CD Age; Control vs. CD Fp; Control vs. IBD Age; Control vs. IBD Fp; Control vs. UC Age; Control vs. UC Fp; Age; BMI; Creatinine; eGFR; Urea; BMI; Brinkman Index; Age; Alcohol

D) Full Dimensional Categorical Distributions with the Lowest RMSE

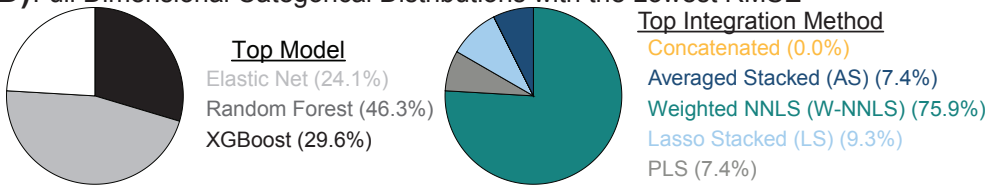
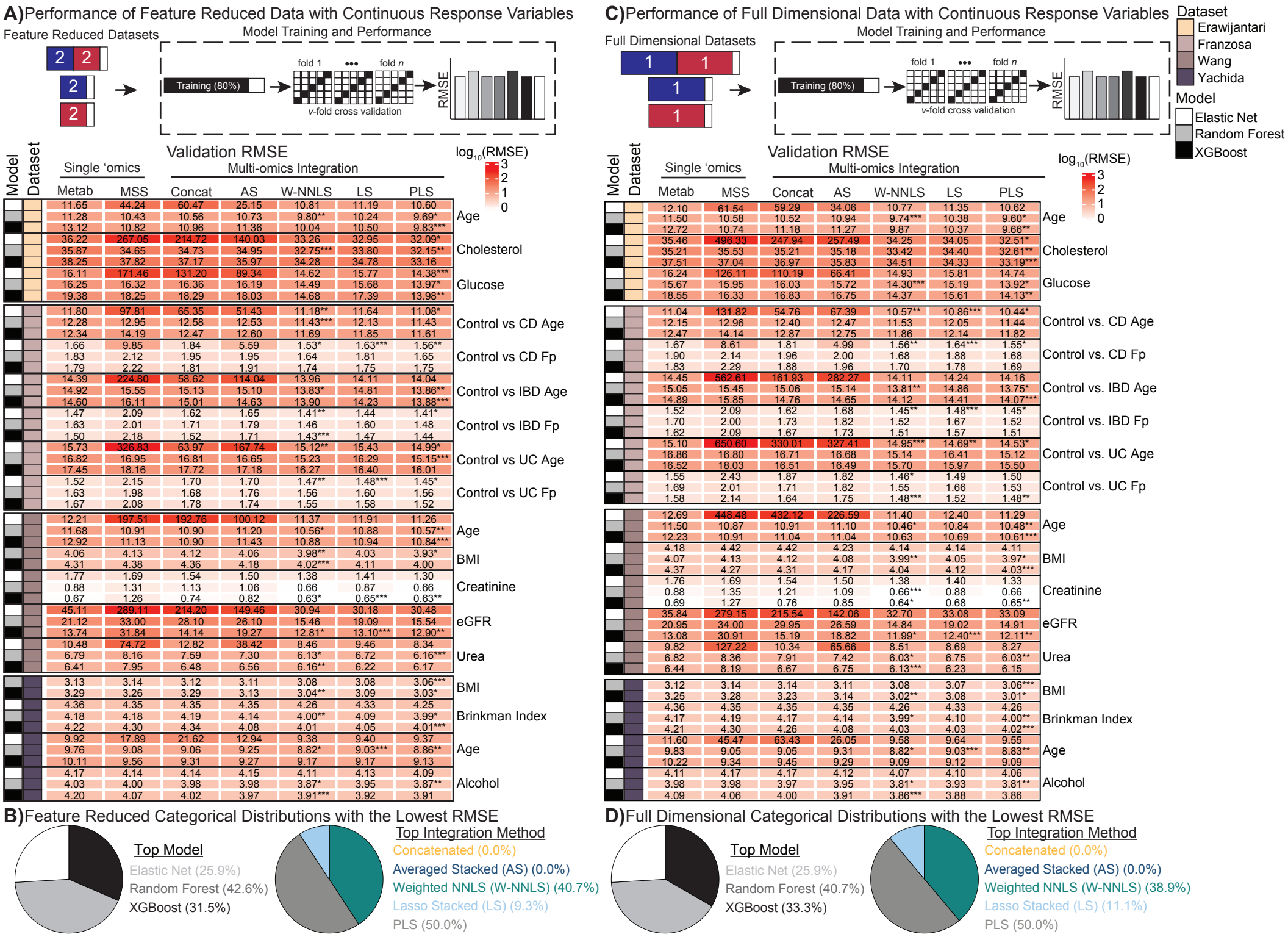


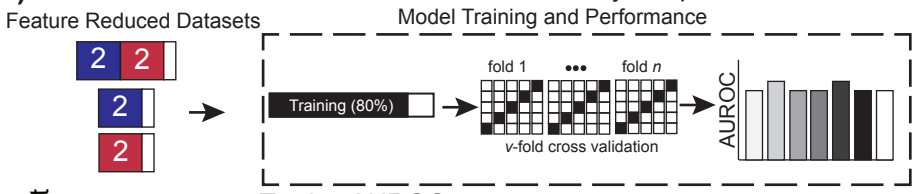
Figure 2. Testing Performance Metrics Across Continuous Feature Reduced and Full Dimensional Datasets. Testing performance for feature reduced matched microbiome and metabolomics datasets with continuous response variables are represented by **A)**. The top performing machine learning models and top integration methods for each of the feature reduced datasets are represented as pie charts in **B)**. Testing performance for full dimensional matched microbiome and metabolomics datasets with continuous response variables are represented by **C)**. The top performing machine learning models and top integration methods for each of the full dimensional datasets are represented as pie charts in **D)**. Performance is assessed using RMSE Mean. Actual RMSE values are shown in each cell of the heatmap with $\log_{10}$ scaling for the coloring. Response variables used for each model are shown on the right of each heatmap. The top performing models per response variable are represented by '*', where '*' is the best performing model, '**' is the second-best performing model and '***' is the third-best performing model. The proportions are calculated using the top three best performing models (*, **, ***) per response variable. 'T' represents ties in performance; response variables with majority 'T' were not used for pie chart calculations. Individual 'omics (Metabolomics and MSS) were not represented in the pie charts. '>' or 'T' represents increased or tied performance of an individual 'omics to one of the top performing integration models ('*', '**', and '***').


A) Performance of Feature Reduced Data with Continuous Response Variables
Feature Reduced Datasets
Model Training and Performance
Training (80%)
fold 1
fold n
v-fold cross validation
RMSE
Validation RMSE
log10(RMSE)
Model
Dataset
Single 'omics
Multi-omics Integration
Metab
MSS
Concat
AS
W-NNLS
LS
PLS
Age
Cholesterol
Glucose
Control vs CD Age
Control vs CD Fp
Control vs IBD Age
Control vs IBD Fp
Control vs UC Age
Control vs UC Fp
Age
BMI
Creatinine
eGFR
Urea
BMI
Brinkman Index
Age
Alcohol
B) Feature Reduced Categorical Distributions with the Lowest RMSE
Top Model
Elastic Net (25.9%)
Random Forest (42.6%)
XGBoost (31.5%)
Top Integration Method
Concatenated (0.0%)
Averaged Stacked (AS) (0.0%)
Weighted NNLS (W-NNLS) (40.7%)
Lasso Stacked (LS) (9.3%)
PLS (50.0%)
C) Performance of Full Dimensional Data with Continuous Response Variables
Full Dimensional Datasets
Model Training and Performance
Training (80%)
fold 1
fold n
v-fold cross validation
RMSE
Validation RMSE
log10(RMSE)
Model
Dataset
Single 'omics
Multi-omics Integration
Metab
MSS
Concat
AS
W-NNLS
LS
PLS
Age
Cholesterol
Glucose
Control vs. CD Age
Control vs. CD Fp
Control vs. IBD Age
Control vs. IBD Fp
Control vs. UC Age
Control vs. UC Fp
Age
BMI
Creatinine
eGFR
Urea
BMI
Brinkman Index
Age
Alcohol
D) Full Dimensional Categorical Distributions with the Lowest RMSE
Top Model
Elastic Net (25.9%)
Random Forest (40.7%)
XGBoost (33.3%)
Top Integration Method
Concatenated (0.0%)
Averaged Stacked (AS) (0.0%)
Weighted NNLS (W-NNLS) (38.9%)
Lasso Stacked (LS) (11.1%)
PLS (50.0%)
Dataset
Erawijantari
Franzosa
Wang
Yachida
Model
Elastic Net
Random Forest
XGBoost

608 **Supplemental Figure 1**. **Validation Performance Metrics Across Continuous Feature Reduced and Full**
609 **Dimensional Datasets.** Validation performance for feature reduced matched microbiome and metabolomics
610 datasets with continuous response variables are represented by **A)**. The top performing machine learning models
611 and top integration methods for each of the feature reduced datasets are represented as pie charts in **B)**.
612 Validation performance for full dimensional matched microbiome and metabolomics datasets with continuous
613 response variables are represented by **C)**. The top performing machine learning models and top integration
614 methods for each of the full dimensional datasets are represented as pie charts in **D)**. Performance calculated
615 using RMSE Mean. Actual RMSE values are shown in each cell of the heatmap with $\log_{10}$ scaling for the coloring.
616 Response variables used for each model are shown on the right of each heatmap. The top performing models
617 per response variable are represented by '*', where '*' is the best performing model, '**' is the second-best
618 performing model and '***' is the third-best performing model. The proportions are calculated using the top three
619 best performing models (*, **, ***) per response variable. 'T' represents ties in performance; response variables
620 with majority 'T' were not used for pie chart calculations. Individual 'omics (Metabolomics and MSS) were not
621 represented in the pie charts. '>' or 'T' represents increased or tied performance of an individual 'omics to one
622 of the top performing integration models ('*', '**', and '***').

623

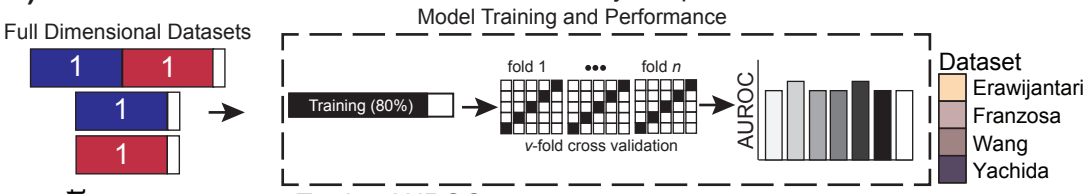
A) Performance of Feature Reduced Data with Binary Response Variables



Testing AUROC

Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
	No Alcohol vs. Alcohol	0.54	0.60	0.60	0.58	0.55	0.59	0.58
		0.56	0.74>*	0.71*	0.66	0.62	0.68**	0.64
		0.59	0.67>***	0.64	0.67***	0.52	0.66	0.65
	Female vs. Male	0.58	0.75>*	0.73*	0.52	0.72**	0.63***	0.60
		0.51	0.59	0.61	0.54	0.56	0.54	0.57
	Control vs. Gastrectomy	1.00T	0.83	0.99	1.00T	1.00T	1.00T	1.00T
		1.00T	0.91	0.96	1.00T	1.00T	1.00T	1.00T
		1.00T	0.89	1.00T	0.99	1.00T	0.99	1.00T
	Control vs. Chrohn's Disease	0.97>***	0.92	0.96	0.97	0.97	0.97	0.97
		0.97	0.97	0.97	0.97	0.97*	0.97**	0.97***
		0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Control vs. IBD	0.94	0.84	0.91	0.94	0.95	0.94	0.92
		0.94	0.91	0.95	0.95***	0.95	0.94	0.94
		0.93	0.93	0.95	0.96*	0.96**	0.95	0.95
	Control vs. Ulcerative Colitis	0.93	0.83	0.95	0.93	0.94	0.94	0.91
		0.97	0.96	0.97	0.97	0.98*	0.98***	0.98**
		0.95	0.95	0.96	0.97	0.97	0.96	0.95
	Female vs. Male	0.60>****	0.55	0.54	0.63*	0.57	0.57	0.57
		0.55	0.56	0.56	0.60**	0.57	0.58***	0.57
	Control vs. Kidney Failure	1.00T	0.95	0.98	1.00T	1.00T	1.00T	1.00T
		1.00T	0.98	1.00T	1.00T	1.00T	1.00T	1.00T
		1.00T	0.98	1.00T	1.00T	1.00T	1.00T	1.00T
	Female vs. Male	0.69	0.64	0.68	0.70	0.72	0.69	0.69
		0.75	0.64	0.73	0.75***	0.76*	0.75	0.76**
		0.71	0.67	0.75	0.74	0.74	0.73	0.74
	Control vs. MP & CRC	0.74	0.65	0.66	0.71	0.73	0.73	0.73
		0.73	0.80>***	0.80**	0.80	0.77	0.80***	0.81*
		0.70	0.74	0.76	0.76	0.77	0.76	0.74
	Control vs. MP & Stage 0 CRC	0.69	0.59	0.61	0.71	0.64	0.69	0.69
		0.74	0.83>***	0.83***	0.86*	0.77	0.84**	0.82
		0.65	0.75	0.72	0.75	0.65	0.74	0.72
	Control vs. Stage 1 & 2 CRC	0.76	0.72	0.76	0.78	0.78	0.77	0.77
		0.65	0.86>***	0.87*	0.81	0.86**	0.85***	0.83
		0.64	0.83	0.81	0.75	0.81	0.78	0.77
	Control vs. Stage 3 & 4 CRC	0.82	0.81	0.83	0.87	0.86	0.85	0.83
		0.86	0.94>***	0.94**	0.93	0.94*	0.93	0.93***
		0.79	0.87	0.86	0.86	0.84	0.87	0.87

C) Performance of Full Dimensional Data with Binary Response Variables



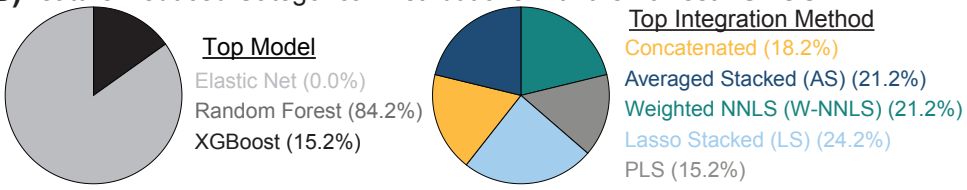
Testing AUROC

Model: Elastic Net, Random Forest, XGBoost

AUROC: 1, 0.5, 0

Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
	No Alcohol vs. Alcohol	0.53	0.59	0.55	0.56	0.54	0.54	0.53
		0.61	0.73>*	0.71**	0.71*	0.62	0.63	0.68***
		0.55	0.66	0.67	0.63	0.62	0.61	0.60
	Female vs. Male	0.55	0.55	0.58	0.56	0.58	0.55	0.56
		0.70>*	0.58	0.59	0.63***	0.59	0.67**	0.68*
		0.57	0.58	0.56	0.56	0.53	0.57	0.55
	Control vs. Gastrectomy	1.00T	0.75	0.99	0.98	1.00T	1.00T	1.00T
		1.00T	0.87	0.94	1.00T	1.00T	1.00T	1.00T
		1.00T	0.87	0.99	1.00T	1.00T	1.00T	1.00T
	Control vs. Chrohn's Disease	0.98>*	0.94	0.96	0.97	0.98**	0.97	0.98*
		0.97	0.97>****	0.97	0.97	0.97	0.97***	0.97
		0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Control vs. IBD	0.94	0.85	0.92	0.94	0.95	0.94	0.93
		0.93	0.91	0.94	0.94	0.94	0.93	0.93
		0.93	0.94	0.95	0.96**	0.96*	0.95***	0.95
	Control vs. Ulcerative Colitis	0.93	0.75	0.95	0.92	0.92	0.93	0.92
		0.95	0.96	0.97	0.97	0.98*T	0.96	0.96
		0.94	0.96	0.95	0.97***	0.98*T	0.96	0.95
	Female vs. Male	0.59>*	0.53	0.52	0.58*	0.56	0.55	0.56
		0.56	0.56	0.54	0.57**	0.57	0.57***	0.56
	Control vs. Kidney Failure	1.00T	0.93	0.98	1.00T	1.00T	1.00T	1.00T
		1.00T	0.98	1.00T	1.00T	1.00T	1.00T	1.00T
		1.00T	0.97	1.00T	1.00T	1.00T	1.00T	1.00T
	Female vs. Male	0.72	0.66	0.70	0.73	0.75	0.72	0.72
		0.81>**	0.68	0.72	0.78	0.81*	0.80**	0.80***
		0.78	0.68	0.76	0.77	0.79	0.78	0.78
	Control vs. MP & CRC	0.71	0.56	0.55	0.63	0.69	0.69	0.67
		0.71	0.72	0.74***	0.75*	0.72	0.73	0.75**
		0.63	0.69	0.69	0.69	0.70	0.69	0.68
	Control vs. MP and Stage 0 CRC	0.69	0.75>***	0.76**	0.76*	0.57	0.74***	0.69
		0.63	0.60	0.62	0.64	0.56	0.62	0.61
	Control vs. Stage 1 & 2 CRC	0.58	0.72>**	0.72*	0.68**	0.66***	0.63	0.64
		0.61	0.65	0.63	0.64	0.59	0.65	0.62
	Control vs. Stage 3 & 4 CRC	0.72	0.66	0.69	0.75	0.71	0.72	0.73
		0.89	0.89	0.91**	0.94*	0.85	0.90***	0.89
		0.85	0.80	0.82	0.88	0.86	0.83	0.82

B) Feature Reduced Categorical Distributions with the Lowest AUROC



D) Full Dimensional Categorical Distributions with the Lowest AUROC

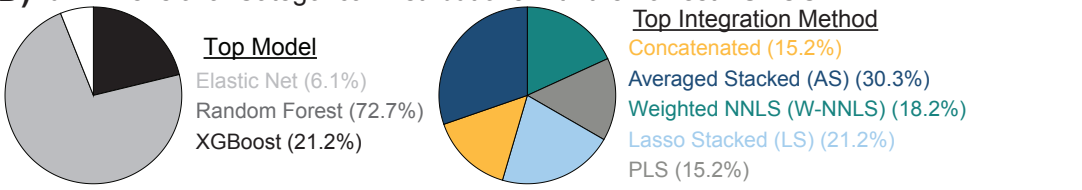
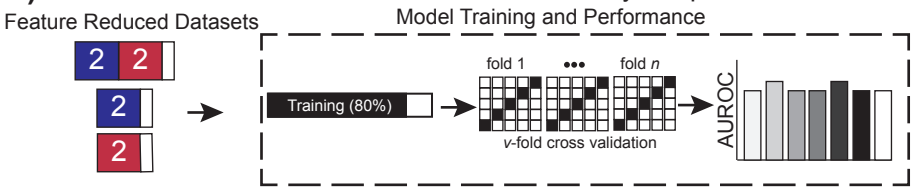


Figure 3. Test Performance Metrics Across Binary Feature Reduced and Full Dimensional Datasets. Testing performance for feature reduced matched microbiome and metabolomics datasets with binary response variables are represented by **A)**. The top performing machine learning models and top integration methods for each of the feature reduced datasets are represented as pie charts in **B)**. Testing performance for full dimensional matched microbiome and metabolomics datasets with binary response variables are represented by **C)**. The top performing machine learning models and top integration methods for each of the full dimensional datasets are represented as pie charts in **D)**. Performance is calculated using AUROC mean, which are shown in the heatmap. Response variables used for each model are shown on the right of each heatmap. The top performing models per response variable are represented by '*', where '*' is the best performing model, '**' is the second-best performing model and '***' is the third-best performing model. The proportions are calculated using the top three best performing models (*, **, ***) per response variable. 'T' represents ties in performance; response variables with majority 'T' were not used for pie chart calculations. Individual 'omics (Metabolomics and MSS) were not represented in the pie charts. '>' or 'T' represents increased or tied performance of an individual 'omics to one of the top performing integration models ('*', '**', and '***').

A) Performance of Feature Reduced Data with Binary Response Variables

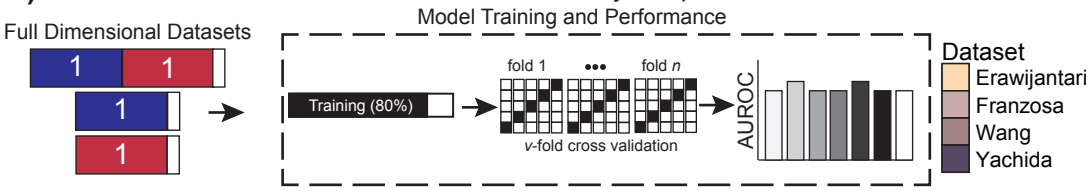


Validation AUROC

Model	Dataset	Metab	MSS	Concat	AS	W-NNLS	LS	PLS
		Single 'omics		Multi-omics Integration				
		0.59	0.60	0.62	0.60	0.59	0.59	0.65
		0.63	0.64	0.63	0.67***	0.60	0.70*	0.69**
		0.54	0.57	0.58	0.60	0.53	0.60	0.63
		0.72	0.71	0.69	0.76***	0.75	0.79**	0.80*
		0.64	0.62	0.68	0.70	0.65	0.71	0.72
		0.97	0.76	0.87	0.93	0.97	0.96	0.97
		0.95	0.90	0.93	0.98**	0.98*	0.98	0.98***
		0.92	0.87	0.94	0.96	0.98	0.96	0.97
		0.94	0.87	0.93	0.93	0.94	0.94	0.94
		0.93	0.94	0.94	0.94	0.94	0.94	0.94
		0.94	0.93	0.93	0.94	0.95***	0.95**	0.95**
		0.95	0.83	0.90	0.94	0.95***	0.95**	0.95*
		0.91	0.88	0.91	0.91	0.91	0.92	0.91
		0.92	0.88	0.92	0.92	0.93	0.93	0.92
		0.96	0.75	0.93	0.93	0.96***	0.96	0.96
		0.94	0.85	0.94	0.94	0.94	0.95	0.94
		0.96	0.85	0.95	0.95	0.96*	0.96	0.96**
		0.57	0.63	0.62	0.60	0.61	0.64***	0.65*
		0.60	0.62	0.60	0.64	0.63	0.64**	0.63
		1.00[T]	0.89	0.97	1.00[T]	1.00[T]	1.00[T]	1.00[T]
		1.00[T]	0.90	0.98	1.00[T]	1.00[T]	1.00[T]	1.00[T]
		1.00[T]	0.92	1.00[T]	1.00[T]	1.00[T]	1.00[T]	1.00[T]
		0.70	0.71	0.73	0.75	0.76	0.76	0.76
		0.77	0.64	0.74	0.77	0.79**	0.78	0.79*
		0.77	0.64	0.76	0.77	0.79	0.79	0.79***
		0.68	0.65	0.69	0.68	0.72	0.73	0.72
		0.70	0.77	0.77	0.78***	0.78	0.79**	0.79*
		0.64	0.75	0.74	0.73	0.76	0.75	0.76
		0.68	0.51	0.54	0.66	0.66	0.67	0.68***
		0.56	0.67	0.65	0.67	0.67	0.68**	0.67
		0.59	0.67	0.65	0.65	0.65	0.68	0.69*
		0.80	0.74	0.78	0.80	0.83***	0.83**	0.83
		0.73	0.80	0.80	0.81	0.83*	0.82	0.82
		0.70	0.77	0.77	0.78	0.78	0.79	0.79
		0.77	0.69	0.68	0.76	0.78	0.79	0.79
		0.69	0.82	0.83	0.80	0.83**	0.84*	0.83
		0.64	0.83	0.83	0.80	0.81	0.82	0.83***

No Alcohol vs. Alcohol
Female vs. Male
Control vs. Gastrectomy
Control vs. Chrohn's Disease
Control vs. IBD
Control vs. Ulcerative Colitis
Female vs. Male
Control vs. Kidney Failure
Female vs. Male
Control vs. MP & CRC
Control vs. MP & Stage 0 CRC
Control vs. Stage 1 & 2 CRC
Control vs. Stage 3 & 4 CRC

C) Performance of Full Dimensional Data with Binary Response Variables



Validation AUROC

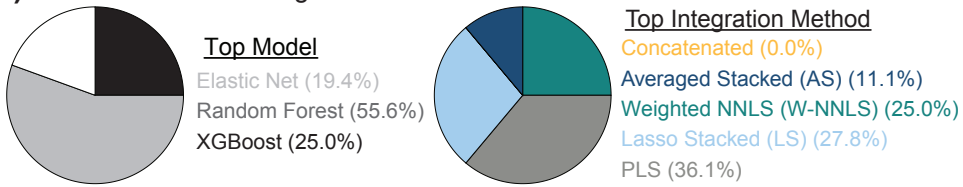
Model	Dataset	Metab	MSS	Concat	AS	W-NNLS	LS	PLS
		Single 'omics		Multi-omics Integration				
		0.66	0.60	0.63	0.61	0.63	0.67	0.69
		0.64	0.60	0.60	0.61	0.67	0.67	0.70*[T]
		0.68	0.60	0.59	0.66	0.66	0.69***	0.70*[T]
		0.73[T]***	0.54	0.69	0.74**	0.73	0.73***	0.75*
		0.68	0.61	0.64	0.66	0.65	0.68	0.70
		0.65	0.62	0.60	0.64	0.59	0.66	0.68
		0.97	0.68	0.88	0.90	0.98*	0.97**	0.97***
		0.93	0.87	0.92	0.96	0.97	0.95	0.97
		0.94	0.84	0.91	0.94	0.96	0.96	0.96
		0.94	0.89	0.93	0.93	0.94**	0.94***	0.94
		0.93	0.93	0.94	0.94	0.94	0.94	0.94
		0.94	0.93	0.94	0.94	0.95*	0.94	0.94
		0.93[T]***	0.84	0.90	0.92	0.93**	0.94*	0.93***
		0.91	0.88	0.91	0.91	0.91	0.91	0.91
		0.91	0.87	0.91	0.91	0.92	0.92	0.91
		0.96	0.77	0.94	0.94	0.97*	0.96***	0.96**
		0.92	0.84	0.93	0.93	0.93	0.93	0.93
		0.94	0.84	0.93	0.94	0.95	0.95	0.95
		0.57	0.59	0.59	0.59	0.61	0.62	0.63
		0.62	0.61	0.59	0.65***	0.65	0.67*	0.67**
		0.99	0.87	0.96	0.99	1.00[T]	0.99	0.99
		1.00[T]	0.89	0.97	1.00[T]	1.00[T]	1.00[T]	1.00[T]
		1.00[T]	0.91	1.00[T]	1.00[T]	1.00[T]	1.00[T]	1.00[T]
		0.69	0.67	0.70	0.73	0.75	0.75	0.75
		0.80	0.66	0.72	0.78	0.81**	0.80	0.81*
		0.79	0.68	0.79	0.79	0.81***	0.80	0.81
		0.65	0.61	0.60	0.63	0.69	0.69	0.70
		0.71	0.70	0.71	0.74	0.75***	0.75**	0.76*
		0.69	0.70	0.72	0.72	0.74	0.74	0.74
		0.55	0.58	0.56	0.56	0.55	0.58	0.61**
		0.58	0.58	0.61***	0.58	0.56	0.60	0.62*
		0.64	0.60	0.61	0.64	0.64	0.67	0.67
		0.66	0.65	0.62	0.67***	0.62	0.69**	0.69*
		0.70	0.59	0.58	0.70	0.70	0.70	0.72
		0.69	0.71	0.71	0.73	0.74	0.75	0.77
		0.64	0.78	0.76	0.76	0.79***	0.79**	0.80*

Model: Elastic Net, Random Forest, XGBoost

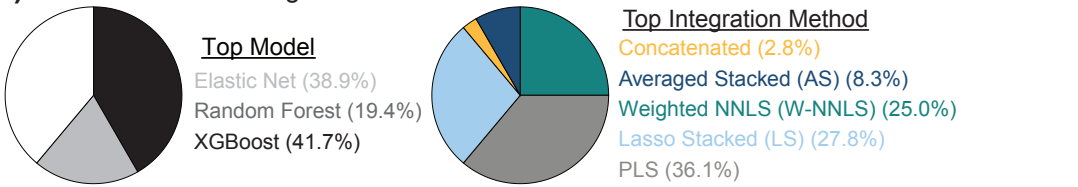
AUROC: 1, 0.5, 0

No Alcohol vs. Alcohol
Female vs. Male
Control vs. Gastrectomy
Control vs. Chrohn's Disease
Control vs. IBD
Control vs. Ulcerative Colitis
Female vs. Male
Control vs. Kidney Failure
Female vs. Male
Control vs. MP and CRC
Control vs. MP & Stage 0 CRC
Control vs. Stage 1 & 2 CRC
Control vs. Stage 3 & 4 CRC

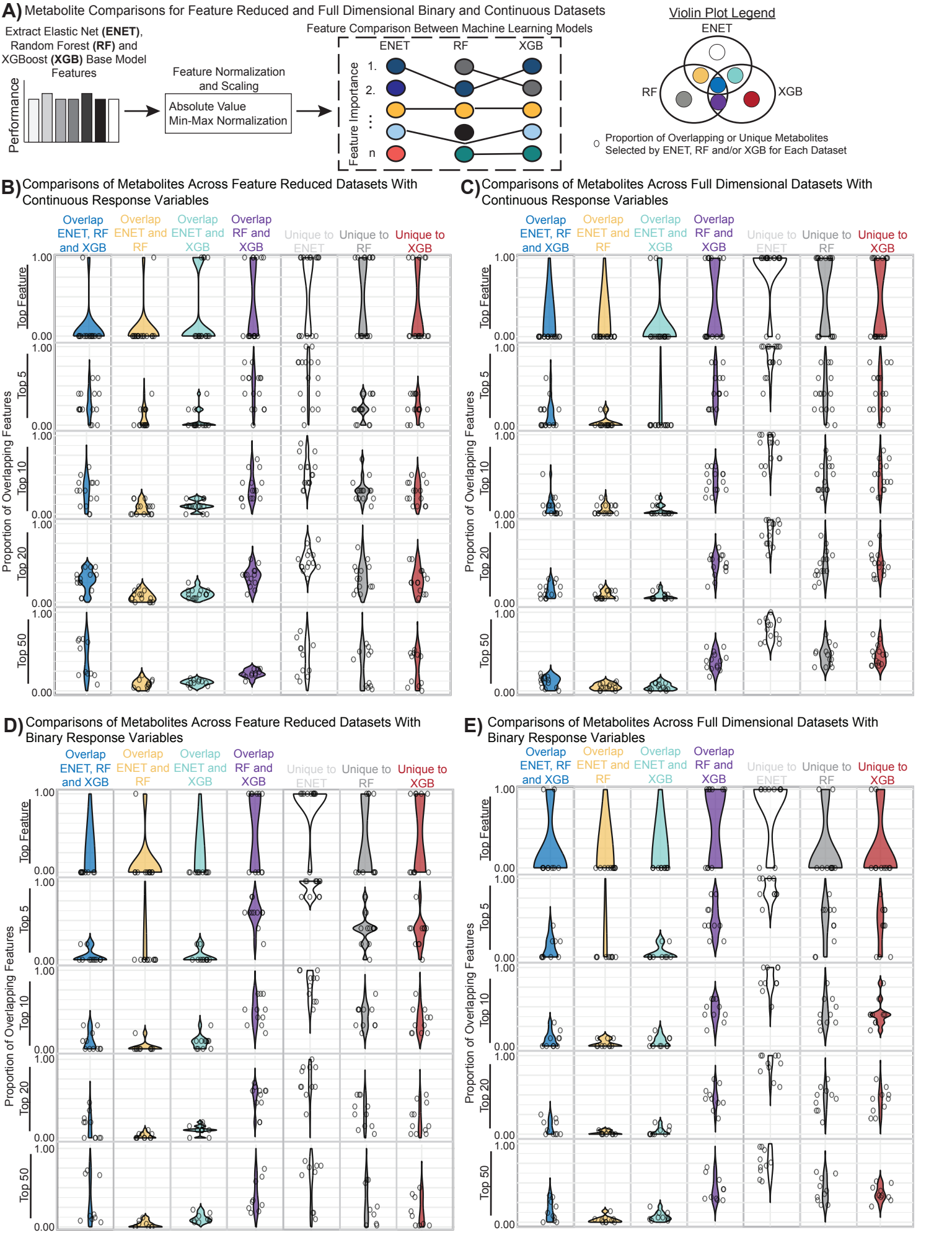
B) Feature Reduced Categorical Distributions with the Lowest AUROC



D) Full Dimensional Categorical Distributions with the Lowest AUROC

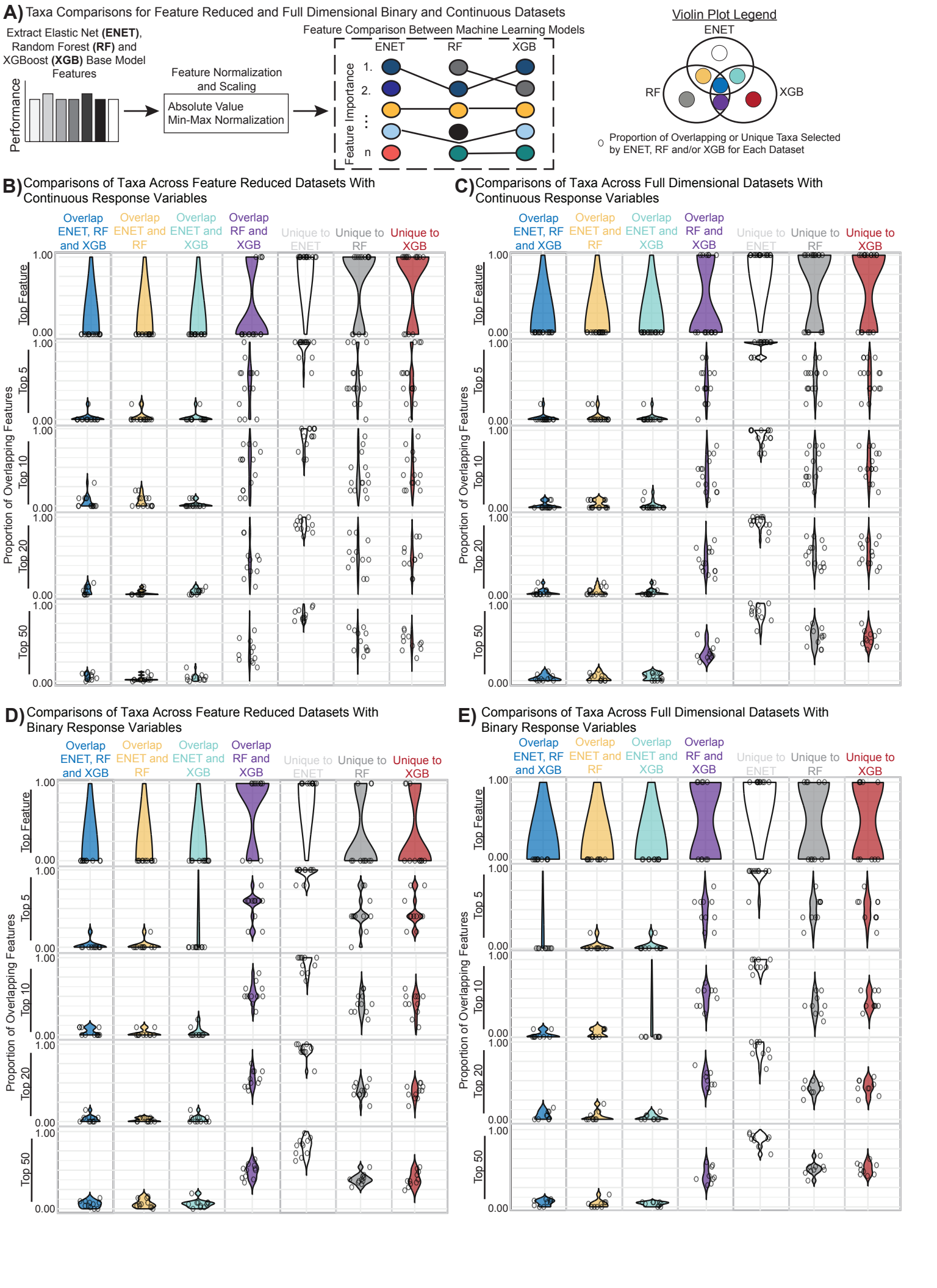


Supplemental Figure 2. Validation Performance Metrics Across Binary Feature Reduced and Full Dimensional Datasets. Validation performance for feature reduced matched microbiome and metabolomics datasets with binary response variables are represented by **A)**. The top performing machine learning models and top integration methods for each of the feature reduced datasets are represented as pie charts in **B)**. Testing performance for full dimensional matched microbiome and metabolomics datasets with binary response variables are represented by **C)**. The top performing machine learning models and top integration methods for each of the full dimensional datasets are represented as pie charts in **D)**. Performance is calculated using AUROC mean, which are shown in the heatmap. Response variables used for each model are shown on the right of each heatmap. The top performing models per response variable are represented by '*', where '*' is the best performing model, '**' is the second-best performing model and '***' is the third-best performing model. The proportions are calculated using the top three best performing models (*, **, ***) per response variable. 'T' represents ties in performance; response variables with majority 'T' were not used for pie chart calculations. Individual 'omics (Metabolomics and MSS) were not represented in the pie charts. '>' or 'T' represents increased or tied performance of an individual 'omics to one of the top performing integration models ('*', '**', and '***').


A) Metabolite Comparisons for Feature Reduced and Full Dimensional Binary and Continuous Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
ENET
RF
XGB
Feature Importance
1.
2.
n
Violin Plot Legend
ENET
RF
XGB
Proportion of Overlapping or Unique Metabolites Selected by ENET, RF and/or XGB for Each Dataset
B) Comparisons of Metabolites Across Feature Reduced Datasets With Continuous Response Variables
C) Comparisons of Metabolites Across Full Dimensional Datasets With Continuous Response Variables
D) Comparisons of Metabolites Across Feature Reduced Datasets With Binary Response Variables
E) Comparisons of Metabolites Across Full Dimensional Datasets With Binary Response Variables
Overlap ENET, RF and XGB
Overlap ENET and RF
Overlap ENET and XGB
Overlap RF and XGB
Unique to ENET
Unique to RF
Unique to XGB
Proportion of Overlapping Features
Top Feature
Top 5
Top 10
Top 20
Top 50
1.00
0.00

654 **Figure 4. Metabolite Feature Selection Comparisons across Elastic Net, Random Forest and XGBoost.**
655 The pipeline for extracting feature importance metrics and performing comparisons are represented in **A)**.
656 Feature Importance values for Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) were extracted for
657 each cross-validation step. The absolute value was taken for each value and min-max normalization was
658 performed. The overall average for each feature was then used as the metric for feature importance. See
659 **Methods** for more details. Comparisons for the metabolites selected across feature reduced datasets with
660 continuous response variables are represented by **B)**. Comparisons for the metabolites selected across full
661 dimensional datasets with continuous response variables are represented by **C)**. Comparisons for the
662 metabolites selected across feature reduced datasets with binary response variables are represented by **D)**.
663 Comparisons for the metabolites selected across full dimensional datasets with binary response variables are
664 represented by **E)**. For each analysis, top features for ENET, RF and XGB were binned into 5 levels – top feature
665 (i.e., features with highest feature importance values), top 5 features, top 10 feature, top 20 features and top 50
666 features. For each bin, comparisons for the features selected between ENET, RF and XGB were performed to
667 identify joint and unique features. The distribution of dataset proportions for overlapping and unique features are
668 represented by violin plots for panels **B-E**.

669


A) Taxa Comparisons for Feature Reduced and Full Dimensional Binary and Continuous Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
ENET
RF
XGB
Feature Importance
1.
2.
n
Violin Plot Legend
ENET
RF
XGB
Proportion of Overlapping or Unique Taxa Selected by ENET, RF and/or XGB for Each Dataset
B) Comparisons of Taxa Across Feature Reduced Datasets With Continuous Response Variables
C) Comparisons of Taxa Across Full Dimensional Datasets With Continuous Response Variables
D) Comparisons of Taxa Across Feature Reduced Datasets With Binary Response Variables
E) Comparisons of Taxa Across Full Dimensional Datasets With Binary Response Variables
Overlap ENET, RF and XGB
Overlap ENET and RF
Overlap ENET and XGB
Overlap RF and XGB
Unique to ENET
Unique to RF
Unique to XGB
Proportion of Overlapping Features
Top Feature
Top 5
Top 10
Top 20
Top 50
1.00
0.00

670 **Figure 5. Taxa Feature Selection Comparisons across Elastic Net, Random Forest and XGBoost.** The
671 pipeline for extracting feature importance metrics and performing comparisons are represented in **A)**. Feature
672 Importance values for Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) were extracted for each
673 cross-validation step. The absolute value was taken for each value and min-max normalization was performed.
674 The overall average for each feature was then used as the metric for feature importance. See **Methods** for more
675 details. Comparisons for the taxa selected across feature reduced datasets with continuous response variables
676 are represented by **B)**. Comparisons for the taxa selected across full dimensional datasets with continuous
677 response variables are represented by **C)**. Comparisons for the taxa selected across feature reduced datasets
678 with binary response variables are represented by **D)**. Comparisons for the taxa selected across full dimensional
679 datasets with binary response variables are represented by **E)**. For each analysis, top features for ENET, RF
680 and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance values), top 5
681 features, top 10 feature, top 20 features and top 50 features. For each bin, comparisons for the features selected
682 between ENET, RF and XGB were performed to identify joint and unique features. The distribution of dataset
683 proportions for overlapping and unique features are represented by violin plots for panels **B-E**.

684

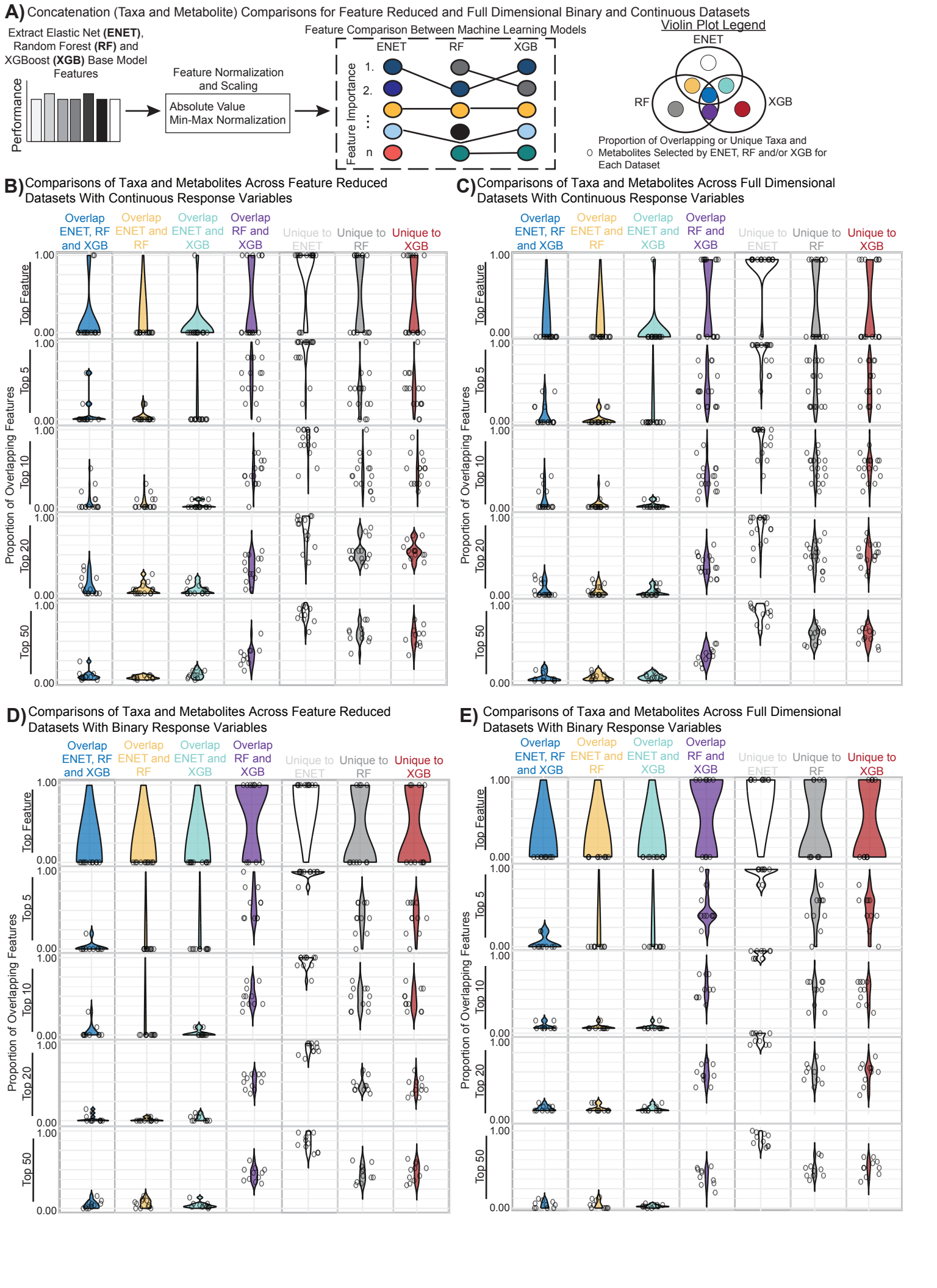

A) Concatenation (Taxa and Metabolite) Comparisons for Feature Reduced and Full Dimensional Binary and Continuous Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
ENET
RF
XGB
Feature Importance
1.
2.
n
Violin Plot Legend
ENET
RF
XGB
Proportion of Overlapping or Unique Taxa and Metabolites Selected by ENET, RF and/or XGB for Each Dataset
B) Comparisons of Taxa and Metabolites Across Feature Reduced Datasets With Continuous Response Variables
C) Comparisons of Taxa and Metabolites Across Full Dimensional Datasets With Continuous Response Variables
D) Comparisons of Taxa and Metabolites Across Feature Reduced Datasets With Binary Response Variables
E) Comparisons of Taxa and Metabolites Across Full Dimensional Datasets With Binary Response Variables
Overlap ENET, RF and XGB
Overlap ENET and RF
Overlap ENET and XGB
Overlap RF and XGB
Unique to ENET
Unique to RF
Unique to XGB
Proportion of Overlapping Features
Top Feature
Top 5
Top 10
Top 20
Top 50
1.00
0.00

Figure 6. Concatenation Feature Selection Comparisons across Elastic Net, Random Forest and XGBoost. The pipeline for extracting feature importance metrics and performing comparisons are represented in **A)**. Feature Importance values for Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) were extracted for each cross-validation step. The absolute value was taken for each value and min-max normalization was performed. The overall average for each feature was then used as the metric for feature importance. See **Methods** for more details. Comparisons for the metabolites and taxa selected across feature reduced datasets with continuous response variables are represented by **B)**. Comparisons for the metabolites and taxa selected across full dimensional datasets with continuous response variables are represented by **C)**. Comparisons for the metabolites and taxa selected across feature reduced datasets with binary response variables are represented by **D)**. Comparisons for the metabolites and taxa selected across full dimensional datasets with binary response variables are represented by **E)**. For each analysis, top features for ENET, RF and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance values), top 5 features, top 10 feature, top 20 features and top 50 features. For each bin, comparisons for the features selected between ENET, RF and XGB were performed to identify joint and unique features. The distribution of dataset proportions for overlapping and unique features are represented by violin plots for panels **B-E**.

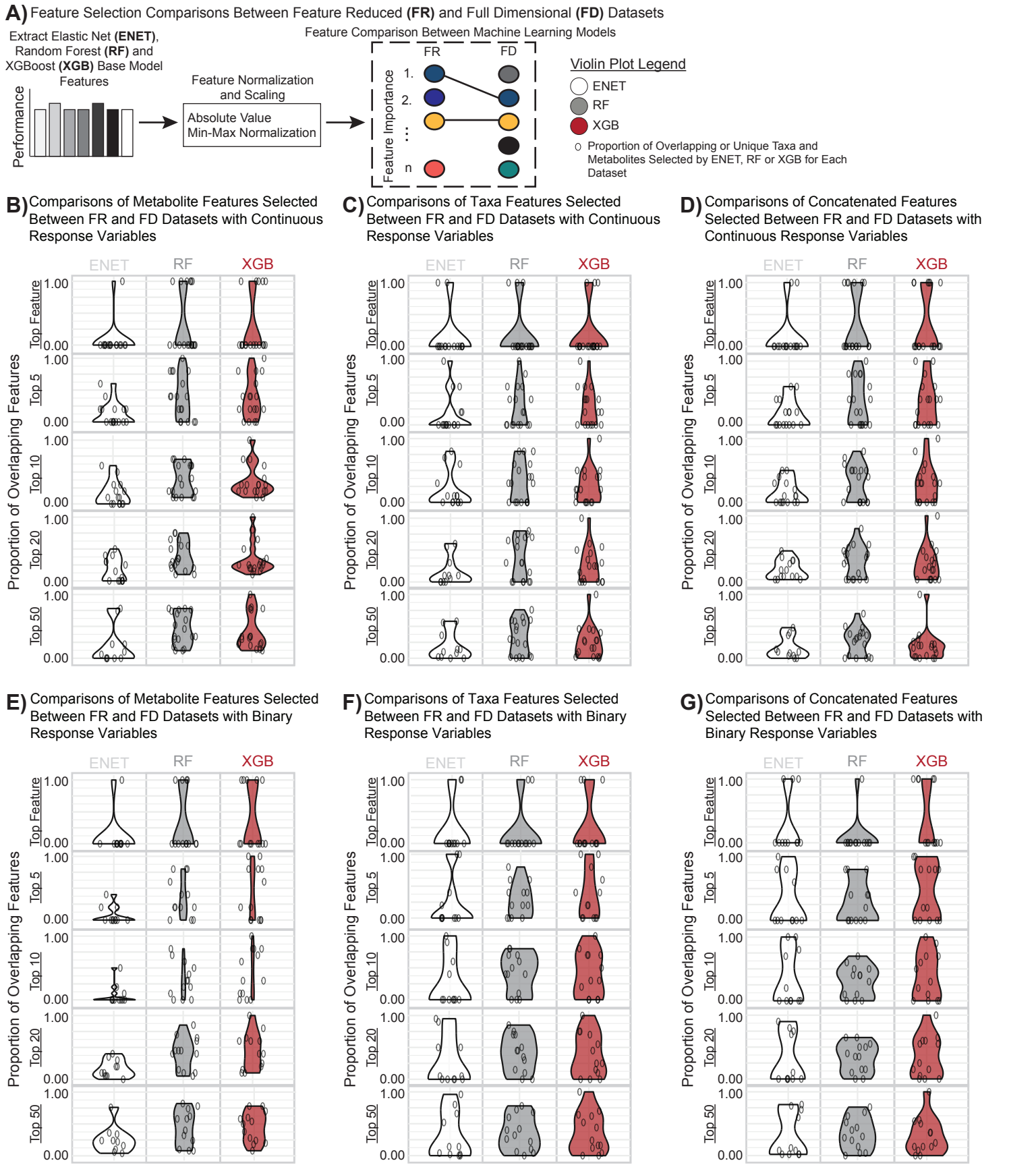
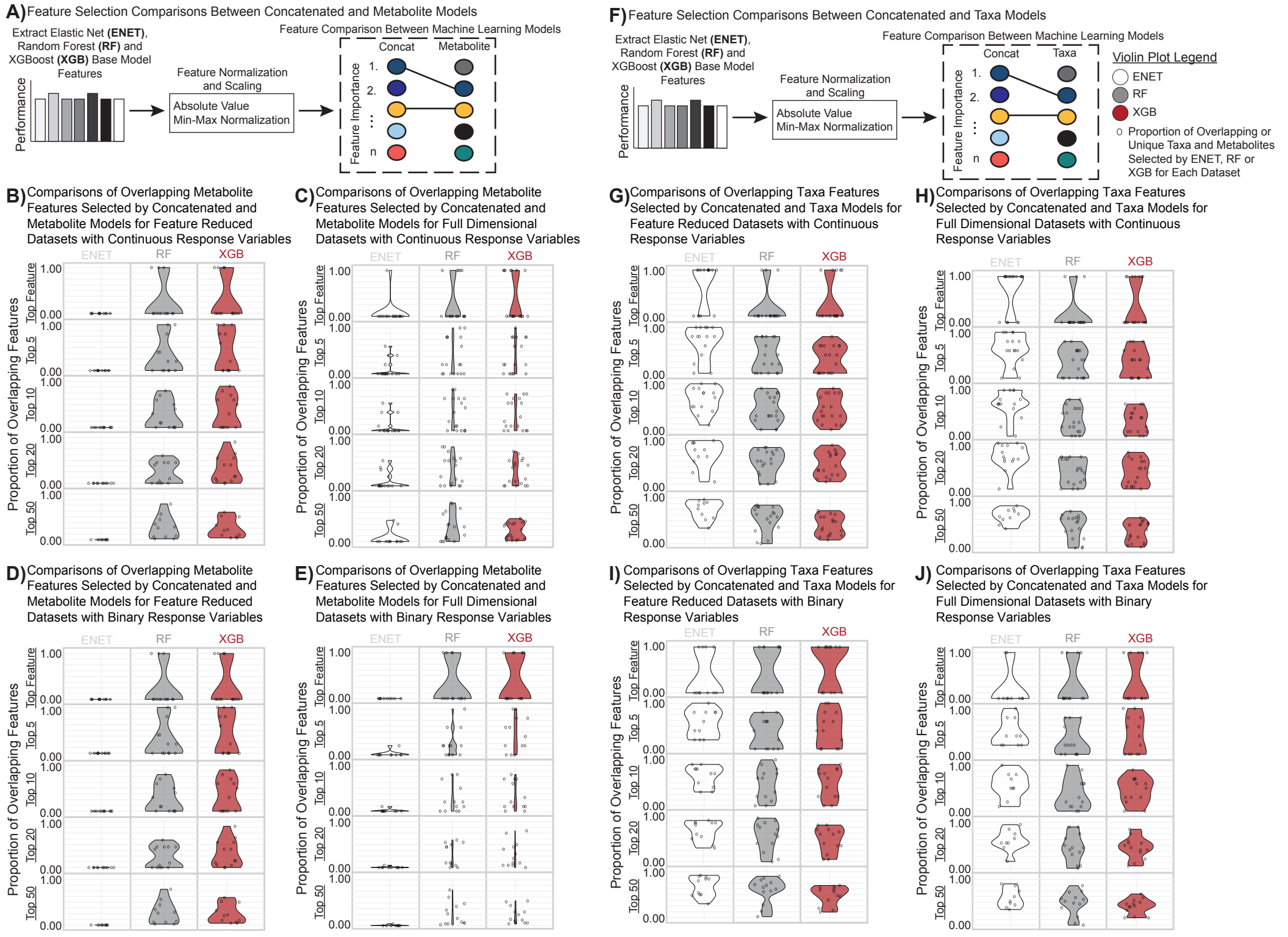
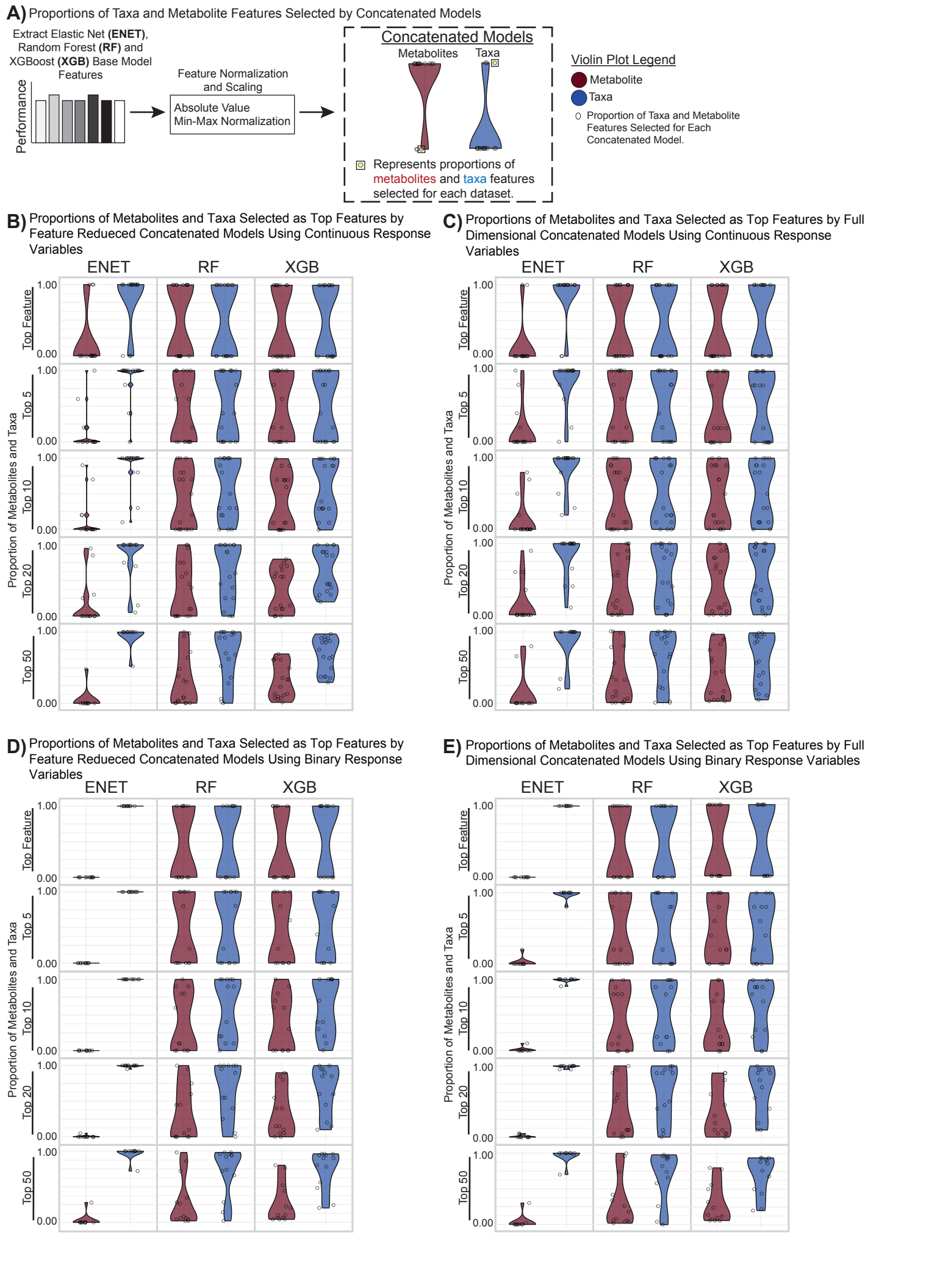

A) Feature Selection Comparisons Between Feature Reduced (FR) and Full Dimensional (FD) Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
FR
FD
Feature Importance
1.
2.
n
Violin Plot Legend
ENET
RF
XGB
Proportion of Overlapping or Unique Taxa and Metabolites Selected by ENET, RF or XGB for Each Dataset
B) Comparisons of Metabolite Features Selected Between FR and FD Datasets with Continuous Response Variables
C) Comparisons of Taxa Features Selected Between FR and FD Datasets with Continuous Response Variables
D) Comparisons of Concatenated Features Selected Between FR and FD Datasets with Continuous Response Variables
E) Comparisons of Metabolite Features Selected Between FR and FD Datasets with Binary Response Variables
F) Comparisons of Taxa Features Selected Between FR and FD Datasets with Binary Response Variables
G) Comparisons of Concatenated Features Selected Between FR and FD Datasets with Binary Response Variables
ENET
RF
XGB
Proportion of Overlapping Features
Top Feature
Top 5
Top 10
Top 20
Top 50
1.00
0.00

Figure 7. Feature Selection Comparisons Between Featured Reduced (FR) and Full Dimensional (FD) Datasets Across Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Models. The pipeline for extracting feature importance metrics and performing comparisons are represented in **A)**. Feature Importance values for ENET, RF and XGB were extracted for each cross-validation step. The absolute value was taken for each value and min-max normalization was performed. The overall average for each feature was then used as the metric for feature importance. See **Methods** for more details. Comparisons for the metabolites selected by feature reduced and full dimensional datasets with continuous response variables are represented by **B)**. Comparisons for the taxa selected by feature reduced and full dimensional datasets with continuous response variables are represented by **C)**. Comparisons for the taxa and metabolites selected by feature reduced and full dimensional datasets with continuous response variables are represented by **D)**. Comparisons for the metabolites selected by feature reduced and full dimensional datasets with binary response variables are represented by **E)**. Comparisons for the taxa selected by feature reduced and full dimensional datasets with binary response variables are represented by **F)**. Comparisons for the taxa and metabolites selected by feature reduced and full dimensional datasets with continuous response variables are represented by **G)**. For each analysis, top features for ENET, RF and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance values), top 5 features, top 10 feature, top 20 features and top 50 features. For each bin, comparisons for the features selected by feature reduced and full dimensional datasets for ENET, RF and XGB were performed to identify joint and unique features. The distribution of dataset proportions for overlapping and unique features are represented by violin plots for panels **B-G**.


A) Feature Selection Comparisons Between Concatenated and Metabolite Models
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
Concat
Metabolite
Feature Importance
1.
2.
n
B) Comparisons of Overlapping Metabolite Features Selected by Concatenated and Metabolite Models for Feature Reduced Datasets with Continuous Response Variables
C) Comparisons of Overlapping Metabolite Features Selected by Concatenated and Metabolite Models for Full Dimensional Datasets with Continuous Response Variables
D) Comparisons of Overlapping Metabolite Features Selected by Concatenated and Metabolite Models for Feature Reduced Datasets with Binary Response Variables
E) Comparisons of Overlapping Metabolite Features Selected by Concatenated and Metabolite Models for Full Dimensional Datasets with Binary Response Variables
F) Feature Selection Comparisons Between Concatenated and Taxa Models
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning Models
Concat
Taxa
Feature Importance
1.
2.
n
Violin Plot Legend
ENET
RF
XGB
Proportion of Overlapping or Unique Taxa and Metabolites Selected by ENET, RF or XGB for Each Dataset
G) Comparisons of Overlapping Taxa Features Selected by Concatenated and Taxa Models for Feature Reduced Datasets with Continuous Response Variables
H) Comparisons of Overlapping Taxa Features Selected by Concatenated and Taxa Models for Full Dimensional Datasets with Continuous Response Variables
I) Comparisons of Overlapping Taxa Features Selected by Concatenated and Taxa Models for Feature Reduced Datasets with Binary Response Variables
J) Comparisons of Overlapping Taxa Features Selected by Concatenated and Taxa Models for Full Dimensional Datasets with Binary Response Variables
ENET
RF
XGB
Proportion of Overlapping Features
Top Feature
Top 5
Top 10
Top 20
Top 50
1.00
0.00

721 **Figure 8. Feature Selection Comparisons Between Concatenated (Concat) and Individual**
722 **(Metabolomics/Taxa) Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Models.** The pipeline
723 for extracting feature importance metrics and performing comparisons between concatenated and metabolites
724 models are represented in **A)**. Feature Importance values for ENET, RF and XGB were extracted for each cross-
725 validation step. The absolute value was taken for each value and min-max normalization was performed. The
726 overall average for each feature was then used as the metric for feature importance. See **Methods** for more
727 details. Comparisons for the metabolites selected by concatenated and metabolomics models for feature
728 reduced datasets with continuous response variables are represented by **B)**. Comparisons for the metabolites
729 selected by concatenated and metabolomics models for full dimensional datasets with continuous response
730 variables are represented by **C)**. Comparisons for the metabolites selected by concatenated and metabolomics
731 models for feature reduced datasets with binary response variables are represented by **D)**. Comparisons for the
732 metabolites selected by concatenated and metabolomics models for full dimensional datasets with binary
733 response variables are represented by **E)**. The pipeline for extracting feature importance metrics and performing
734 comparisons between concatenated and metabolites models are represented in **F)** (See details above (**A**) and
735 in **Methods**). Comparisons for the taxa selected by concatenated and taxa models for feature reduced datasets
736 with continuous response variables are represented by **G)**. Comparisons for the taxa selected by concatenated
737 and taxa models for full dimensional datasets with continuous response variables are represented by **H)**.
738 Comparisons for the taxa selected by concatenated and taxa models for feature reduced datasets with binary
739 response variables are represented by **I)**. Comparisons for the taxa selected by concatenated and taxa models
740 for full dimensional datasets with binary response variables are represented by **J)**. For each analysis, top features
741 for ENET, RF and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance
742 values), top 5 features, top 10 feature, top 20 features and top 50 features. For each bin, comparisons for the
743 features selected by concatenated and individual (metabolomics and taxa) 'omics models for ENET, RF and
744 XGB were performed to identify joint and unique features. The distribution of dataset proportions for overlapping
745 and unique features are represented by violin plots for panels **B-E** and **G-J**.

746


A) Proportions of Taxa and Metabolite Features Selected by Concatenated Models
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Concatenated Models
Metabolites
Taxa
Represents proportions of metabolites and taxa features selected for each dataset.
Violin Plot Legend
Metabolite
Taxa
Proportion of Taxa and Metabolite Features Selected for Each Concatenated Model.
B) Proportions of Metabolites and Taxa Selected as Top Features by Feature Redueced Concatenated Models Using Continuous Response Variables
C) Proportions of Metabolites and Taxa Selected as Top Features by Full Dimensional Concatenated Models Using Continuous Response Variables
D) Proportions of Metabolites and Taxa Selected as Top Features by Feature Redueced Concatenated Models Using Binary Response Variables
E) Proportions of Metabolites and Taxa Selected as Top Features by Full Dimensional Concatenated Models Using Binary Response Variables
ENET
RF
XGB
Top Feature
Top 5
Top 10
Top 20
Top 50
Proportion of Metabolites and Taxa
1.00
0.00

Supplemental Figure 3: Proportions of Taxa and Metabolite Features Selected by Concatenated Models. The pipeline for extracting feature importance metrics and calculating proportions of the top features that are metabolites and taxa selected by the concatenated models are represented by **A)**. The Proportions of taxa and metabolite features selected by ENET, RF and XGB for concatenated feature reduced datasets with continuous response variable are shown in **B)**. The Proportions of taxa and metabolite features selected by ENET, RF and XGB for concatenated full dimensional datasets with continuous response variable are shown in **C)**. The Proportions of taxa and metabolite features selected by ENET, RF and XGB for concatenated feature reduced datasets with binary response variable are shown in **D)**. The Proportions of taxa and metabolite features selected by ENET, RF and XGB for concatenated full dimensional datasets with binary response variable are shown in **E)**. The distribution of proportions for each dataset are represented as violin plots. For each analysis, top features for ENET, RF and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance values), top 5 features, top 10 feature, top 20 features and top 50 features. The red violin plot represents metabolomics proportions for each dataset and the blue violin plot represents the taxa proportions for each dataset. These violin plot values are inverse proportional for each dataset, so that if the proportion of metabolites for Dataset A = 1, then the proportion of taxa for Dataset A = 0.

References

1. Gilbert JA, Quinn RA, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016 Jul 7;535(7610).
2. Joos R, Boucher K, Lavelle A, et al. Examining the healthy human microbiome concept. Nature Reviews Microbiology. 2024 Mar;23(3):192-205.
3. Walter J, Armet AM, Finlay BB, et al. Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human Microbiota-Associated Rodents. Cell. 2020 Jan 23;180(2):221-232.
4. Corander J, Hanage WP, Pensar J. Causal discovery for the microbiome. The Lancet Microbe. 2022 Nov;3(11):e881-e887.
5. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018 Jul;16(7):410-422.
6. Galloway-Peña J, Hanson B. Tools for Analysis of the Microbiome. Digestive Diseases and Sciences. 2020 Mar;65(3):674-685.
7. Song EJ, Lee ES, Nam YD. Progress of analytical tools and techniques for human gut microbiome research. Journal of Microbiology. 2018 Oct;56(10):693-705.
8. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. 2017 May;83(18).
9. Chetty A, Blekhman R. Multi-omic approaches for host-microbiome data integration. Gut Microbes. 2024 Jan-Dec;16(1):2297860.
10. Li P, Luo H, Ji B, et al. Machine learning for data integration in human gut microbiome. Microbial Cell Factories. 2022 Nov 23;21(1):241.
11. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nature Microbiology. 2019 Feb;4(2):293-305.
12. Li J, Zhao F, Wang Y, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017 Feb 1;5(1):14.
13. Gao B, Duan Y, Lang S, et al. Functional Microbiomics Reveals Alterations of the Gut Microbiome and Host Co-Metabolism in Patients With Alcoholic Hepatitis. Hepatology Communications. 2020 Jun 19;4(8):1168-1182.
14. Yachida S, Mizutani S, Shiroma H, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nature Medicine. 2019 Jun;25(6):968-976.
15. Ghaemi MS, DiGiulio DB, Contrepois K, et al. Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. Bioinformatics. 2019 Jan 1;35(1):95-103.
16. Mallick H, Porwal A, Saha S, et al. An integrated Bayesian framework for multi-omics prediction and classification. Statistics in Medicine. 2024 Feb 28;43(5):983-1002.
17. Liu YX, Qin Y, Chen T, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein & Cell. 2021 May;12(5):315-330.
18. Poussin C, Sierro N, Boué S, et al. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. Drug Discovery Today. 2018 Sep;23(9):1644-1657.
19. Huttenhower C, Finn RK, McHardy AC. Challenges and opportunities in sharing microbiome data and analyses. Nature Microbiology. 2023 Nov;8(11):1960-1970.
20. Santiago-Rodriguez TM, Hollister EB. Multi 'omic data integration: A review of concepts, considerations, and approaches. Seminars in Perinatology. 2021 Oct;45(6):151456.
21. Papoutsoglou G, Tarazona S, Lopes MB, et al. Machine learning approaches in microbiome research: challenges and best practices. Frontiers in Microbiology. 2023 Sep 22;14:1261889.
22. Mullen KM, van Stokkum IHM. nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS). https://rdrr.io/cran/nnls/ (27 June 2025, date last accessed).
23. Friedman J, Hastie T, Tibshirani R. et al. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010 Aug;33(1):1-22.
24. Mevik B-H, Wehrens R. The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software. 2007 Jan;18:1–23.
25. Erawijantari PP, Mizutani S, Shiroma H, et al. Influence of gastrectomy for gastric cancer treatment on faecal microbiome and metabolome profiles. Gut. 2020 Aug;69(8):1404-1415.
26. Wang X, Yang S, Li S, et al. Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and rodents. Gut. 2020 Dec;69(12):2131-2142.

27. Hernández Medina R, Kutuzova S, Nielsen KN, et al. Machine learning and deep learning applications in microbiome research. ISME Communications. 2022 Oct 6;2(1):98.
28. Kumar B, Lorusso E, Fosso B, et al. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. Frontiers in Microbiology. 2024 Feb 13;15:1343572.
29. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1;23(19):2507-17.
30. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006 Jan 6;7:3.
31. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019 Nov 28;20(1):257.
32. Lu J, Breitwieser FP, Thielen P, et al. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017 Jan;3:e104.
33. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 Apr 20;43(7):e47.
34. Karwowska Z, Aasmets O; Kosciolek T, et al. Effects of data transformation and model selection on feature importance in microbiome classification data. Microbiome. 2025 Jan 4;13(1):2.
35. Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019 Sep 1;35(17):3055-3062.
36. Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology. 2020 May 11;21(1):111.
37. Muller E, Algavi YM, Borenstein E. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. Npj Biofilms and Microbiomes. 2022 Oct 15;8(1):79.
38. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019 Nov; 4(43):1686.
39. Robin X, Turck N, Hainard A, et al. PROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12:77.
40. Gamer M, Lemon J, Fellows I, et al. irr: Various Coefficients of Interrater Reliability and Agreement. 2019. https://cran.r-project.org/web/packages/irr/index.html (27 June 2025, date last accessed).
41. Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists. Nature Reviews Molecular Cell Biology. 2022 Jan;23(1):40-55.
42. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Briefings in Bioinformatics. 2018 Mar 1;19(2):325-340.
43. Ghannam RB, Techtmann SM. et al. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Computational and Structural Biotechnology Journal. 2021 Jan 27;19:1092-1107.
44. Kühn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008 Nov;28(5):1-26.
45. Liaw A, Wiener MC. Classification and Regression by randomForest. R News. 2002 Dec;2:18-22.
46. Chen T, Guestrin C. XGBOOST: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug 13; 785-794.
47. Gu Z. Complex heatmap visualization. iMeta. 2022 Sept;3(1):e43.
48. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847-9.
49. Valero-Mora PM, Valero-Mora PM. Ggplot2: Elegant Graphics for Data Analysis. Journal of Statistical Software, Book of Reviews. 2010 Jul;35(1):1-3.