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 Suzette N. Palmer or Xiaowei Zhan.

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. 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) and 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

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

69 **Results**

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

71 We collected and processed four multi-omics datasets consisting of matched metagenomic shotgun sequencing
72 and metabolomics data **(Table 1-2)**[11,14,25,26], and we evaluated 18 continuous and 13 binary matching response
73 variables. We then developed an integrative analysis pipeline to assess the performance of three
74 machine-learning algorithms (Elastic Net, Random Forest, XGBoost) in combination with five integration
75 approaches (Concatenated, Averaged Stacked, Weighted NNLS, Lasso Stacked, and PLS). For concatenation,
76 the metabolomics and metagenomic abundances for each matched sample were appended to form a single long
77 feature vector that served as input for the analysis. The remaining four approaches relied on stacked
78 generalization, where separate base models were first trained on the metabolomics block and on the
79 metagenomics block. The out-of-fold predictions from the two base models were stacked to form two meta-
80 features, which were combined by a meta-learner using all four strategies: a simple mean that assigns equal
81 weight to each prediction (Averaged Stacking), weighted non-negative least square (NNLS) regression, lasso-
82 penalized regression (Lasso stacking), and partial least squares (PLS). We evaluated each stacking approach
83 on feature reduced and full dimensional datasets to assess how well these models performed under different
84 feature scenarios (**Table 1-2**). Each analysis incorporated a grid search with 3 repeated 5-fold cross-validation
85 to optimize the base models, and performance was assessed using root-mean-squared error (RMSE) for
86 continuous outcomes and area under the receiver-operating-characteristic curve (AUROC) for binary outcomes.
87 Uniform grid parameters were applied across all analyses, and features (or, for stacked models, base model-
88 level importances) were extracted for downstream selection analysis (see **Methods** for more details).

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

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

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

<u>Random Forest Maintains Dominance in Binary Outcome Datasets</u>

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**).

<u>Uncovering Trends in Feature Selection Across Elastic Net, Random Forest and XGBoost</u>

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

To assess feature selection across models, we first extracted feature importance metrics from each cross-validation step (see **Methods** for more details). For each run, we computed the absolute value of each feature's importance score and applied Min-Max Normalization. This normalization ensures that feature importance values are comparable across models, preventing any single model from dominating due to differences in scale. We then used the Mean Importance to identify the top overall features for each model. For

146 this analysis, we grouped features into bins based on their importance scores within each model, selecting those
147 with the highest feature importance. To ensure rigor, we used predefined bin sizes, which represent the top 1,
148 5, 10, 20 and 50 features. These discrete bins create a manageable list of high-priority candidates that can be
149 feasibly validated through peer review, or experimentally through *in vivo* and *in vitro* studies. These predefined
150 lists sizes allow us to compare the most influential features across models without the impractical task of vetting
151 each variable across every dataset.

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

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

159 We examined metabolite feature selection patterns from Elastic Net, Random Forest, and XGBoost
160 across feature reduced and full dimensional datasets with binary and continuous response variables (**Figure**
161 **4A**). Across all conditions, Elastic Net consistently identified a large proportion of unique features, regardless of
162 dataset dimensionality, response variable type or bin size (1, 5, 10, 20 or 50) (**Figure 4B-E**).

163 Overlap among the top-ranked features selected by all three models was minimal (**Figure 4B-E**). In
164 nearly all cases, the top feature selected by Elastic Net, Random Forest and XGBoost differed, irrespective of
165 whether the response variable was binary or continuous, and whether the dataset was full dimensional or feature
166 reduced. Only one dataset with a continuous response variable and reduced features (**Figure 4B**) and two
167 datasets with binary response with full dimensional features (**Figure 4E**) showed any overlap in the top ranked
168 feature across all three models.

169 Another consistent trend across all conditions were that overlap between Elastic Net and the tree-based
170 models were limited (**Figure 4**). For feature bins ranging from 5 to 50, overlapping features between Elastic Net
171 and either, Random Forest or XGBoost accounted for less than 25% of the selected features across all dataset
172 types. Two exceptions to the trend occurred in the feature-reduced continuous dataset (**Figure 4B**), where these

173 comparisons showed approximately 40% overlap in the top 5 features for both Elastic Net and Random Forest
174 and Elastic Net and XGBoost.

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

184 Tree-Models Reveal Partial Concordance in Taxa Selection While Elastic Net Remains Distinct

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

193 Patterns of top feature overlap between the tree-based models varied depending on response type and
194 dimensionality. For continuous response variables, Random Forest and XGBoost exhibited greater agreement
195 in their top features when applied to full-dimensional datasets compared to feature reduced datasets (**Figure**
196 **5B-E**). This trend was reversed for binary response variables, where tree-based models showed higher
197 concordance in their top selected features in feature reduced datasets than in full dimensional datasets (**Figure**
198 **5D-E**).

199 Violin plot distributions of feature overlap between Random Forest and XGBoost for bins 5-50 maintained
200 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 high agreement for their top selected features across binary and continuous datasets, with approximately half of the proportion of features being shared (**Figure 6B-E**). Elastic Net consistently selected distinct top features for continuous datasets, with the exception of two feature reduced datasets and one full dimensional dataset (**Figure 6B-C**).

When examining the top selected features (bins 5-50, which represent the top 5 through 50 features selected per dataset), this trend of limited agreement between Elastic Net and tree-based models continued. The proportion of overlapping features among all three models, as well as between Elastic Net and either tree-based methods remained low across most datasets. Overlapping proportions exceeding 25% were rare, and the

majority of distributions were centered near zero, further confirming that Elastic Net largely selected features distinct from those chosen by Random Forest and XGBoost (**Figure 6B-E**). Similar to the taxa-only models, Random Forest and XGBoost demonstrated greater consistency in their feature selections for the concatenated datasets (**Figure 5-6**). Across bins 5-50, approximately 50% of the selected features were shared between the two tree-based model, while the remaining features were largely unique to each method (**Figure 6B-E**). These results reinforce the pattern that Elastic Net tends to prioritize a separate set of features, while Random Forest and XGBoost show more concordance in their selections, even when analyzing integrated taxonomic and metabolomics data.

Dimensionality Reduction Influences Feature Selection Stability in Machine Learning Models

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

Across all data types and response variables, Elastic Net consistently demonstrated minimal agreement in top feature selection for the full dimensional and feature reduced models (**Figure 7B-G**). Regardless of whether the dataset involved metabolomics, taxonomy or concatenated or the response type, only 1-2 datasets showed overlap in the top feature selected by Elastic Net (**Figure 7B-G**). In contrast, the tree-based models, with greater emphasis on Random Forest, showed stronger consistency in the top features selected for the metabolomics datasets (**Figure 7B, Figure 7E**). Approximately half of the Random Forest models identified the same top feature in both the full and feature reduced datasets for both binary and continuous response variables. XGBoost followed a similar trend for metabolomics data, with slightly lower proportions of agreement in top features (**Figure 7B, Figure 7E**).

In contrast, for taxa and concatenated datasets, top feature agreement between full dimensional and reduced models was rare, even for the tree-based methods (**Figure 7C-F**, **Figure 7D-G**). Across these analyses, Random Forest and XGBoost typically shared the same top feature in only 1-4 datasets indicating that

256 dimensionality reduction substantially influences which features are prioritized for these data types (**Figure 7C-**
257 **F**, **Figure 7D-G**).

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

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

273 Trends observed in the concatenated models closely mirrored these seen in the taxa data (**Figure 7D**,
274 **Figure 7G**). However, distributions for bins 5-50 were generally shifted closer to zero across all three modeling
275 approaches, indicating even lower agreement in selected features between full dimensional and feature reduced
276 models (**Figure 7D**, **Figure 7G**). This suggests that when combining multiple data types, feature selection
277 becomes increasingly sensitive to dimensionality, especially for Elastic Net, which consistently demonstrated the
278 least concordance between full and feature reduced models.

279 <u>Differences in Concatenated and Single 'Omics Feature Selection</u>

280 To determine whether concatenated models incorporating both metabolomics and taxa data could
281 recover features identified by single 'omics models, we compared feature selection patterns across Elastic Net,
282 Random Forest and XGBoost for both full dimensional and feature reduced datasets with continuous and binary
283 response types (**Figure 8A**, **Figure 8F**). For metabolomics data, Elastic Net models trained on feature reduced

concatenated datasets showed no overlap in features with the corresponding single ‘omics metabolomics models across bins 1-50, regardless of response type (**Figure 8B-E**). In the full dimensional datasets, there was limited agreement between the concatenated and metabolomics models for continuous outcomes, with most overlap proportions falling below 0.4 and the highest density occurring near zero **(Figure 8C**). For full dimensional data with binary outcomes, consensus remained minimal with many datasets showing no overlapping features between the concatenated and metabolomics models (**Figure 8D-E**).

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

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

312 in every bin originating from the taxa data, regardless of dimensionality or response type (**Supplemental Figure**
313 **3B-E**). This suggests Elastic Net may be more sensitive to high dimensional features within concatenated
314 models. In contrast, Random Forest and XGBoost selected a more balanced mix of metabolites and taxa features
315 across bins, indicating a more even consideration of both data modalities in feature importance rankings
316 (**Supplemental Figure 3B-E**).

317 **Discussion**

318 This study presents a comprehensive evaluation of machine learning performance and feature selection
319 stability across multi-omics microbiome and metabolomics datasets using Elastic Net, Random Forest and
320 XGBoost. By systematically comparing these models across both continuous and binary response variables,
321 full dimensional and feature reduced datasets, and integration strategies including concatenation and ensemble-
322 based methods, we aimed to uncover generalizable trends in model behavior and feature selection stability
323 (**Figure 1**). One of the most consistent findings was the robust performance of Random Forest, particularly when
324 combined with Non-Negative Least Squares (NNLS) integration, for datasets with continuous outcomes (**Figure**
325 **2**). This is consistent with prior studies highlighting the capacity of Random Forest to handle noisy, sparse,
326 nonlinear and high dimensional data, such as microbiome data, effectively[5,27,28]. NNLS further enhanced model
327 performance, reinforcing recent evidence that integration methods leveraging weighted combinations of
328 predictions can outperform simpler or more complex stacking strategies for microbiome multi-omics data[16].

329 Across binary outcomes, Random Forest also outperformed other models, though the advantage was
330 less pronounced with more comparable performances across the stacked integration methods (**Figure 3**).
331 Notably, several individual 'omics models outperformed integrative models for both binary and continuous
332 responses, particularly metabolomics models. This suggests that in certain cases, single 'omics models may
333 identify strong predictive features and/or relationships that are lost during integration. We also observed that
334 model performance results were largely consistent between full dimensional and feature reduced datasets, with
335 Random Forest maintaining top performance across both response types (**Figure 2-3**). This suggests that,
336 despite substantial reductions in input features, overall predictive performance can be preserved when robust
337 models and integration strategies are applied[29].

338 In terms of feature selection, a key trend that emerged across all datasets is that Elastic Net consistently
339 identified a distinct and largely non-overlapping set of features compared to Random Forest and XGBoost

340 (**Figure 4-6**). This divergence is likely attributable to the linear and sparsity-inducing nature of Elastic Net, which
341 tends to select a small subset of features that are linearly associated with the outcome and includes groups of
342 correlated predictors. Alternatively, tree-based models are more adept at capturing nonlinear, higher-order
343 interactions and distributed signals[30].

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

354 When comparing feature selection between single 'omics and concatenated models, tree-based methods
355 retained a relatively high degree of consistency, while Elastic Net rarely recovered the same features—especially
356 in metabolomics analyses (**Figure 8**). Interestingly, for taxonomic features, Elastic Net showed much higher
357 agreement between concatenated and single 'omics models. This pattern likely reflects higher dimensionality in
358 the concatenated datasets, where taxa features outnumber metabolite features. As a result, Elastic Net appears
359 biased towards selecting taxa features, while Random Forest and XGBoost maintained a more balanced
360 selection across both data types. These results are consistent with previous findings suggesting that tree-based
361 models are more resilient to greater numbers of features and sparsity in high-dimensional data[30].

362 While this work provides several new insights into model behavior and feature selection for microbiome
363 multi-omics data, there are several limitations. First, the study was conducted using four multi-omics microbiome
364 datasets[11,14,25,26], limiting generalizability. Broader applications across a wider range of microbiome-related
365 diseases would help validate these findings. Additionally, taxonomic profiling was performed using Kracken

366 v2.1.1 and Bracken v2.8, which may produce different results when compared to other classification tools[31,32].
367 We also only selected one method for feature reduction analyses. We used Limma[33] for metabolomics data and
368 Wilcoxon signed-rank test for the taxonomic relative abundance data. These might not be the best method for
369 feature reduction and/or may not accurately reflect feature signal structure across all datasets. Exploration into
370 other methods of microbiome data transformation and its effect on multi-omics data integration should be
371 explored[34]. Furthermore, the concatenated datasets suffered from greater dimensionality, with taxa feature
372 outnumbering metabolomics for all datasets. This imbalance likely contributed to Elastic Net's consistent
373 selection of taxa-only features. Finally, a uniform hyperparameter grid was used across all models, which
374 provided standardization but may not have captured optimal tuning for individual datasets. Future works should
375 incorporate additional datasets exploring diverse diseases, consider additional integration methods such as
376 DIABLO[35] or MOFA+[36], explore alternative dimensionality reduction strategies and taxa classification strategies
377 and include refined hyperparameter optimization. Additionally, biological validation of selected features, through
378 literature mining or experimental follow-up, is crucial to ensure biological relevance. Overall, this study offers a
379 robust framework for evaluating machine learning performance and feature selection in multi-omics microbiome
380 data, highlighting the strengths and limitations of current modeling strategies and laying the groundwork for more
381 interpretable and scalable integrative analysis.

382 **Methods**

383 Data Collection and Processing

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

388 Data Processing for the Integrated Analysis

389 Processed data files for metabolomics, metadata, species level relative abundance and genus level relative
390 abundance were further processed for downstream analysis. Taxonomic data at the genus and species levels
391 were concatenated for each dataset, and taxa with an overall abundance across samples that was less than
392 0.1% were removed. For metabolite differential intensity analysis, Limma 3.64.1[33] was applied, which employs
393 linear modeling and empirical Bayes methods to identify significant differences while adjusting for multiple testing

using Benjamini-Hochberg procedure. For taxonomic differential analysis, the Wilcoxon rank-sum test was used to compare taxa abundance across experimental groups (More details for data processing can be found in **Supplemental File 2**).

Overview of Multi-omics Integrative Analysis

The integrative analysis was developed using a custom in-house script that is readily available at https://github.com/suziepalmer10/Multiomics-Integrative-Pipeline/tree/main. The packages tidyverse 2.0.0 and argparse 2.2.3 were used to parse and pre-process the data for this integrative analysis[38]. For continuous outcomes, performance was assessed with Root Mean Squared Error (RMSE), and binary classification was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC). For the integrative 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, $\sum \beta_j^2$

421 represents the L_2 penalty, λ determines the overall regularization strength, and α controls the balance between
422 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
423 **Supplemental File 2** for more details).

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

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

434 Data integration using direct concatenation: To integrate taxa and functional data using direct concatenation, we
435 will horizontally concatenate the two matrices, where A represents taxonomy data and B represents functional
436 data. Both matrices share the same number of rows (m), corresponding to the samples and disease phenotype
437 that is used as the response variable. However, these matrices differ in the number of columns, n_A, which
438 represents the number of taxa and, n_B, which represents the number of functional data (i.e., metabolites).
439 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 $C =$
440 $[A|B]$. This creates a unified matrix, where $C \in \mathbb{R}^{m\times(n_A+n_B)}$, which enables the integration of the distinct datatypes
441 while preserving the sample structure. We then use the combined matrix, C, for machine learning model
442 development using Elastic Net, XGBoost and Random Forest algorithms.

443 Data integration using averaged stacking of the predictors: Data integration using averaged stacking combines
444 predictions from individual machine learning models (Random Forest, XGBoost or Elastic Net) for each dataset
445 (taxonomy and metabolomics) into a single, averaged prediction for each sample. For averaged stacking, let
446 $P_A \in \mathbb{R}^{m\times 1}$ represent predictions from a machine learning model trained on taxonomy data, A and let $P_B \in \mathbb{R}^{m\times 1}$

447 represent predictions from a machine learning model trained on functional data. Let m represent the number of
448 samples and response variables that are shared across both datasets. The averaged stacking prediction, P_{avg},
449 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
450 vector across all samples.

451 <u>Data integration using weighted non-negative least squares (NNLS):</u> Data integration using weighted nnls
452 combines predictions from each model by assigning weights to minimize the error between the combined
453 predictions and the observed values, with the constraint that weights are non-negative. As previously described
454 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,
455 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
456 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
457 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
458 package was used for non-negative least squares analysis[22].

459 <u>Data integration using least absolute shrinkage and selection operator (LASSO) stacked generalization:</u> Data
460 integration using LASSO combines predictions by assigning weights to minimize the error between the combined
461 prediction and the observed values while enforcing sparsity in the rights through L_1 regularization. As previously
462 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
463 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
464 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
465 w_A and w_B are determined by solving the following optimization problem using L_1 regularization:
466 $\min_{w_A, w_B \geq 0} \|y - (w_A P_A + w_B P_B)\|^2 + \lambda(|w_A| + |w_B|)$, where λ is the weight of the L_1 regularization term and controls
467 solution sparsity. The glmnet 4.1-6 package was using for LASSO[23].

468 <u>Data integration using partial least squares (PLS):</u> PLS is a regression method that combines the predictors from
469 separate machine learning models by projecting the predictors into a lower dimensional space to maximize the
470 covariance between the predictors and the response variables. This integrative method can handle
471 multicollinearity and captures the shared variation between different datatypes. PLS operates in three stages.
472 The first stage is latent component extraction, where PLS identifies latent components, T, by projecting the
473 combined predictor matrix, P, onto a set of weight vectors that maximize the covariance between the latent

474 components, T, and the response variable, y. Mathematically, $T = PW$, and as previously described, let $P_A \in$
475 $\mathbb{R}^{m\times1}$ and $P_B \in \mathbb{R}^{m\times1}$, represent predictions from a machine learning model trained on taxa data, A, and
476 functional data, B, and let $y \in \mathbb{R}^{m\times1}$ represent response variables. Let $P = [P_A|P_B] \in \mathbb{R}^{m\times2}$, represent the
477 concatenated matrix of predictors. Let $T \in \mathbb{R}^{m\times k}$ represent the latent variables extracted by P, where k
478 represents the number of components. Let $W \in \mathbb{R}^{2\times k}$ represent a weight matrix that defines the linear
479 combination between P_A and P_B. The next stage of PLS is to create a linear regression of the response variable.
480 This is accomplished following extraction of latent components, T, and the response variable, y, are then
481 modeled as a linear function of T. Mathematically, $y = TC + \varepsilon$, where $C \in \mathbb{R}^{k\times1}$ represent regression coefficients
482 for T to predict y and ε represents the residual error term. The final stage is to combine the projection and the
483 regression stages, which can be mathematically represented as $y_{PLS} = PWC$. The pls 2.8-4 package[24] was used
484 for partial least squares model generation.

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

494 Performance Pie Charts: The R package ggplot2 3.4.1 was used to create the performance heatmaps for Figures
495 2-3 and Supplemental Figures 1-2[49]. These metrics were calculated by quantifying the number of models for
496 each dataset that were part of the top 3 performing models (represented as '*', "**" and "***") (described above
497 for Performance Heatmap Generation).

498 Feature Extraction:

499 Feature importance values were extracted from Elastic Net, XGBoost and Random Forest across all 15 cross-
500 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 and repeats from *5*-fold cross validation. Let $|S_{i,j}|$ represent the absolute value for the importance score of feature i in model j. Next min-max normalization, $N_{i,j}$, for feature i for model j is calculated by $N_{i,j} = \frac{|S_{i,j}| - min_k(|S_{k,j}|)}{max_k(|S_{k,j}|) - min_k(|S_{k,j}|)}$, where k indexes over all features in model j. I next compute the averaged normalized 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 features in each model.

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

Data Availability

Data is free and available through: https://github.com/borenstein-lab/microbiome-metabolome-curated-data/wiki/Data-overview#datasets-included[33].

Code Availability

527 Code for this integrative analysis and associated analysis are free and available through:
528 https://github.com/suziepalmer10/Multiomics-Integrative-Pipeline.

529 **Acknowledgements**

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

535 **Competing Interests**

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

538

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

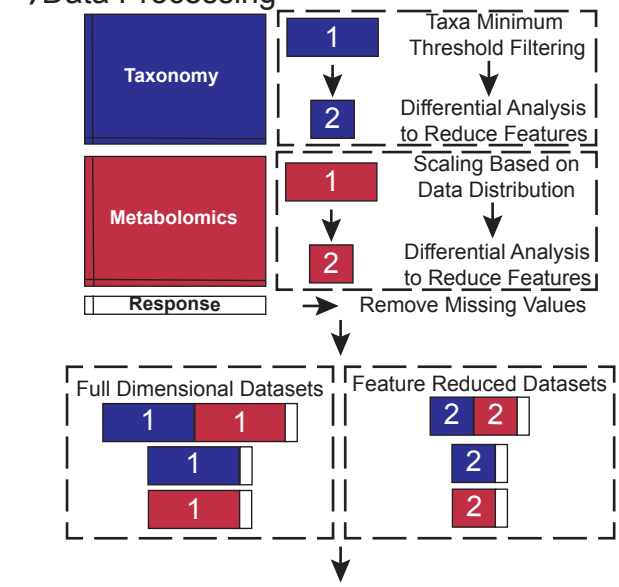
'-' represents models that failed to complete.

Table 2: Details for Data With Binary Response Variables

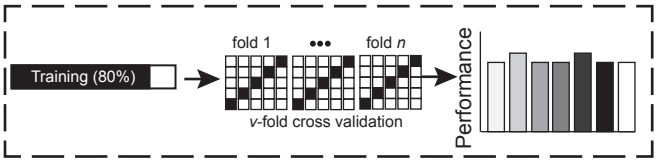
			Featured Reduced Details					Full Dimensional Details				
					Run Time (HH:MM)					Run Time (HH:MM)		
Study Name (Subset)	Response Variable	Sample Size	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost	Number of Metabolites	Number of Taxa	Elastic Net	Random Forest	XGBoost
Erawijantari	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

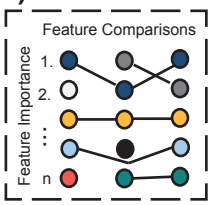


B) Model Training and Testing



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.

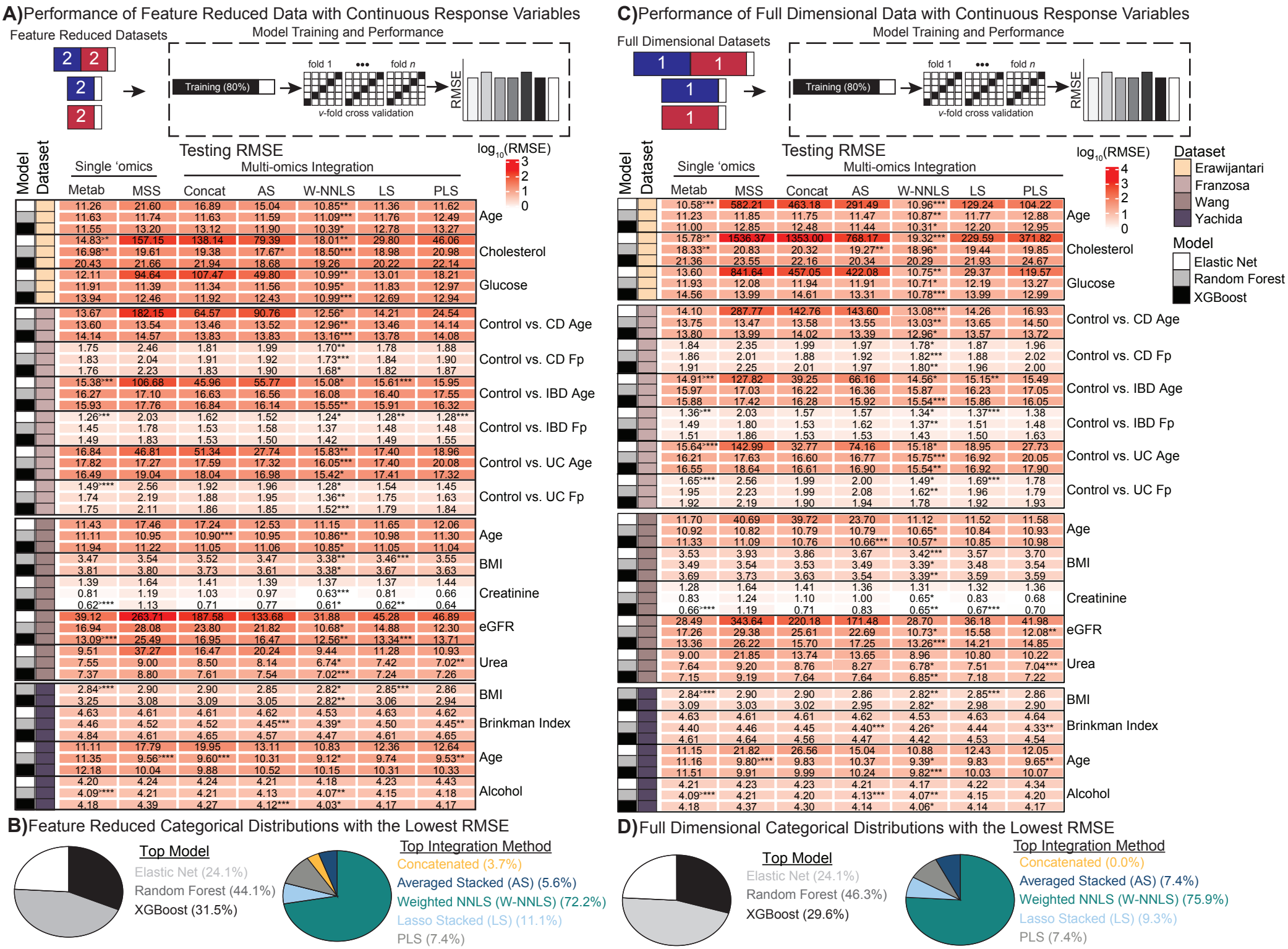
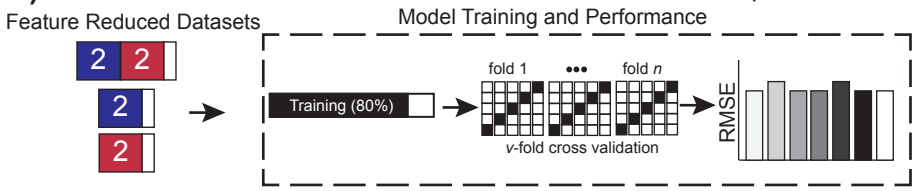

A) Performance of Feature Reduced Data with Continuous Response Variables
Feature Reduced Datasets
Model Training and Performance
Training (80%)
fold 1
fold n
v-fold cross validation
RMSE
Testing RMSE
log10(RMSE)
Model
Dataset
Single 'omics
Multi-omics Integration
Metab
MSS
Concat
AS
W-NNLS
LS
PLS
Age
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 (24.1%)
Random Forest (44.1%)
XGBoost (31.5%)
Top Integration Method
Concatenated (3.7%)
Averaged Stacked (AS) (5.6%)
Weighted NNLS (W-NNLS) (72.2%)
Lasso Stacked (LS) (11.1%)
PLS (7.4%)
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
Testing RMSE
log10(RMSE)
Dataset
Erawijantari
Franzosa
Wang
Yachida
Model
Elastic Net
Random Forest
XGBoost
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 (24.1%)
Random Forest (46.3%)
XGBoost (29.6%)
Top Integration Method
Concatenated (0.0%)
Averaged Stacked (AS) (7.4%)
Weighted NNLS (W-NNLS) (75.9%)
Lasso Stacked (LS) (9.3%)
PLS (7.4%)

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

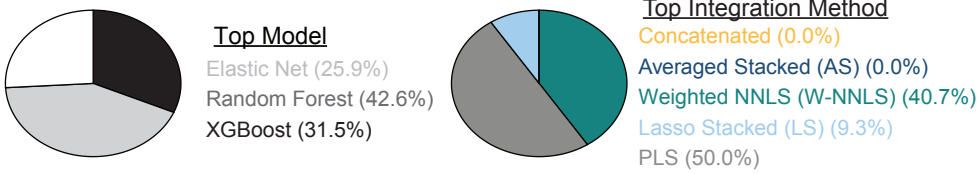


Validation RMSE

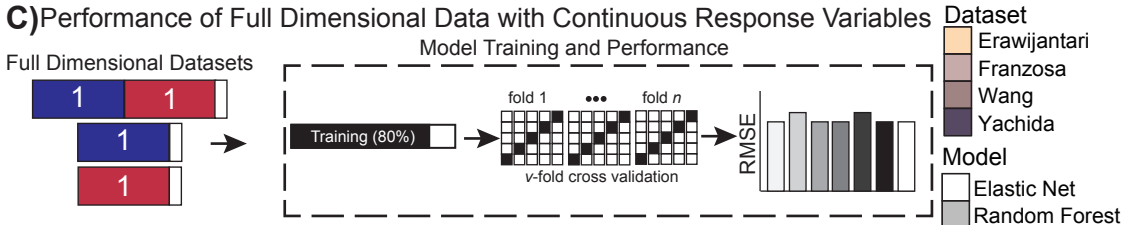
$\log_{10}$(RMSE): 3, 2, 1, 0

Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
	Age	11.65	44.24	60.47	25.15	10.81	11.19	10.60
		11.28	10.43	10.56	10.73	9.80**	10.24	9.69*
		13.12	10.82	10.96	11.36	10.04	10.50	9.83***
	Cholesterol	36.22	267.05	214.72	140.03	33.26	32.95	32.09*
		35.87	34.65	34.73	34.95	32.75***	33.80	32.15**
		38.25	37.82	37.17	35.97	34.28	34.78	33.16
	Glucose	16.11	171.46	131.20	89.34	14.62	15.77	14.38***
		16.25	16.32	16.36	16.19	14.49	15.68	13.97*
		19.38	18.25	18.29	18.03	14.68	17.39	13.98**
	Control vs CD Age	11.80	97.81	65.35	51.43	11.18**	11.64	11.08*
		12.28	12.95	12.58	12.53	11.43***	12.13	11.43
		12.34	14.19	12.47	12.60	11.69	11.85	11.61
	Control vs CD Fp	1.66	9.85	1.84	5.59	1.53*	1.63***	1.56**
		1.83	2.12	1.95	1.95	1.64	1.81	1.65
		1.79	2.22	1.81	1.91	1.74	1.75	1.75
	Control vs IBD Age	14.39	224.80	58.62	114.04	13.96	14.11	14.04
		14.92	15.55	15.13	15.10	13.83*	14.81	13.86**
		14.60	16.11	15.01	14.63	13.90	14.23	13.88***
	Control vs IBD Fp	1.47	2.09	1.62	1.65	1.41**	1.44	1.41*
		1.63	2.01	1.71	1.79	1.46	1.60	1.48
		1.50	2.18	1.52	1.71	1.43***	1.47	1.44
	Control vs UC Age	15.73	326.83	63.97	167.74	15.12**	15.43	14.99*
		16.82	16.95	16.81	16.65	15.23	16.29	15.15***
		17.45	18.16	17.72	17.18	16.27	16.40	16.01
	Control vs UC Fp	1.52	2.15	1.70	1.70	1.47**	1.48***	1.45*
		1.63	1.98	1.68	1.76	1.56	1.60	1.56
		1.67	2.08	1.78	1.74	1.55	1.58	1.52
	Age	12.21	197.51	192.76	100.12	11.37	11.91	11.26
		11.68	10.91	10.90	11.20	10.56*	10.88	10.57**
		12.92	11.13	10.90	11.43	10.88	10.94	10.84***
	BMI	4.06	4.13	4.12	4.06	3.98**	4.03	3.93*
		4.31	4.38	4.36	4.18	4.02***	4.11	4.00
	Creatinine	1.77	1.69	1.54	1.50	1.38	1.41	1.30
		0.88	1.31	1.13	1.06	0.66	0.87	0.66
		0.67	1.26	0.74	0.82	0.63*	0.65***	0.63**
	eGFR	45.11	289.11	214.20	149.46	30.94	30.18	30.48
		21.12	33.00	28.10	26.10	15.46	19.09	15.54
		13.74	31.84	14.14	19.27	12.81*	13.10***	12.90**
	Urea	10.48	74.72	12.82	38.42	8.46	9.46	8.34
		6.79	8.16	7.59	7.30	6.13*	6.72	6.16***
		6.41	7.95	6.48	6.56	6.16**	6.22	6.17
	BMI	3.13	3.14	3.12	3.11	3.08	3.08	3.06***
		3.29	3.26	3.29	3.13	3.04**	3.09	3.03*
	Brinkman Index	4.36	4.35	4.35	4.35	4.26	4.33	4.25
		4.18	4.18	4.19	4.14	4.00**	4.09	3.99*
		4.22	4.30	4.34	4.08	4.01	4.05	4.01***
	Age	9.92	17.89	21.62	12.94	9.38	9.40	9.37
		9.76	9.08	9.06	9.25	8.82*	9.03***	8.86**
		10.11	9.56	9.31	9.27	9.17	9.17	9.13
	Alcohol	4.17	4.14	4.14	4.15	4.11	4.13	4.09
		4.03	4.00	3.98	3.98	3.87*	3.95	3.87**
		4.20	4.07	4.02	3.97	3.91***	3.92	3.91

B) Feature Reduced Categorical Distributions with the Lowest RMSE



C) Performance of Full Dimensional Data with Continuous Response Variables

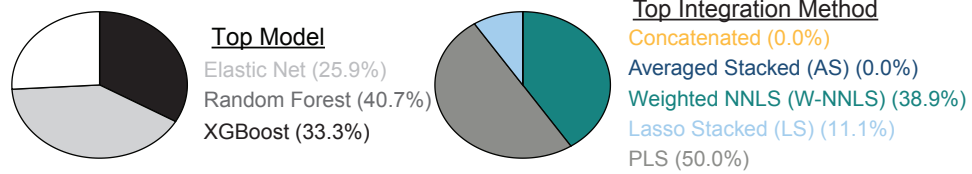


Validation RMSE

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

Model	Dataset	Single 'omics: Metab	Single 'omics: MSS	Multi-omics Integration: Concat	AS	W-NNLS	LS	PLS
	Age	12.10	61.54	59.29	34.06	10.77	11.35	10.62
		11.50	10.58	10.52	10.94	9.74***	10.38	9.60*
		12.72	10.74	11.18	11.27	9.87	10.37	9.66**
	Cholesterol	35.46	496.33	247.94	257.49	34.25	34.05	32.51*
		35.21	35.53	35.21	35.18	33.42	34.40	32.61**
		37.51	37.04	36.97	35.83	34.51	34.33	33.19***
	Glucose	16.24	126.11	110.19	66.41	14.93	15.81	14.74
		15.67	15.95	16.03	15.72	14.30***	15.19	13.92*
		18.55	16.33	16.83	16.75	14.37	15.61	14.13**
	Control vs. CD Age	11.04	131.82	54.76	67.39	10.57**	10.86***	10.44*
		12.15	12.96	12.40	12.47	11.53	12.05	11.44
		12.47	14.14	12.87	12.75	11.86	12.14	11.82
	Control vs. CD Fp	1.67	8.61	1.81	4.99	1.56**	1.64***	1.55*
		1.90	2.14	1.96	2.00	1.68	1.88	1.68
		1.83	2.29	1.88	1.96	1.70	1.78	1.69
	Control vs. IBD Age	14.45	562.61	161.93	282.27	14.11	14.24	14.16
		15.05	15.45	15.06	15.14	13.81**	14.86	13.75*
		14.89	15.85	14.76	14.65	14.12	14.41	14.07***
	Control vs. IBD Fp	1.52	2.09	1.62	1.68	1.45**	1.48***	1.45*
		1.70	2.00	1.73	1.82	1.52	1.67	1.52
		1.62	2.09	1.67	1.73	1.51	1.57	1.51
	Control vs. UC Age	15.10	650.60	330.01	327.41	14.95***	14.69**	14.53*
		16.86	16.80	16.71	16.68	15.14	16.41	15.12
		16.52	18.03	16.51	16.49	15.70	15.97	15.50
	Control vs. UC Fp	1.55	2.43	1.87	1.82	1.46*	1.49	1.50
		1.69	2.01	1.71	1.82	1.55	1.66	1.53
		1.58	2.14	1.64	1.75	1.48***	1.52	1.48**
	Age	12.69	448.48	432.12	226.59	11.40	12.40	11.29
		11.50	10.87	10.91	11.10	10.46*	10.84	10.48**
		12.23	10.91	11.04	11.04	10.63	10.69	10.61***
	BMI	4.18	4.42	4.42	4.23	4.14	4.14	4.11
		4.07	4.13	4.12	4.08	3.99**	4.05	3.97*
		4.37	4.27	4.31	4.17	4.04	4.12	4.03***
	Creatinine	1.76	1.69	1.54	1.50	1.38	1.40	1.33
		0.88	1.35	1.21	1.09	0.66***	0.88	0.66
		0.69	1.27	0.76	0.85	0.64*	0.68	0.65**
	eGFR	35.84	279.15	215.54	142.06	32.70	33.08	33.09
		20.95	34.00	29.95	26.59	14.84	19.02	14.91
		13.08	30.91	15.19	18.82	11.99*	12.40***	12.11**
	Urea	9.82	127.22	10.34	65.66	8.51	8.69	8.27
		6.82	8.36	7.91	7.42	6.03*	6.75	6.03**
		6.44	8.19	6.67	6.75	6.13***	6.23	6.15
	BMI	3.12	3.14	3.14	3.11	3.08	3.07	3.06***
		3.25	3.28	3.23	3.14	3.02**	3.08	3.01*
	Brinkman Index	4.36	4.35	4.35	4.35	4.26	4.33	4.26
		4.17	4.19	4.17	4.14	3.99*	4.10	4.00**
		4.21	4.30	4.26	4.08	4.03	4.03	4.02***
	Age	11.60	45.47	63.43	26.05	9.58	9.64	9.55
		9.83	9.05	9.05	9.31	8.82*	9.03***	8.83**
		10.22	9.34	9.45	9.29	9.09	9.12	9.09
	Alcohol	4.11	4.17	4.17	4.12	4.07	4.10	4.06
		3.98	3.98	3.97	3.95	3.81*	3.93	3.81**
		4.09	4.06	4.00	3.91	3.86***	3.88	3.86

D) Full Dimensional Categorical Distributions with the Lowest RMSE



Supplemental Figure 1. **Validation Performance Metrics Across Continuous Feature Reduced and Full Dimensional Datasets.** Validation 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)**. Validation 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 calculated 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 '***').

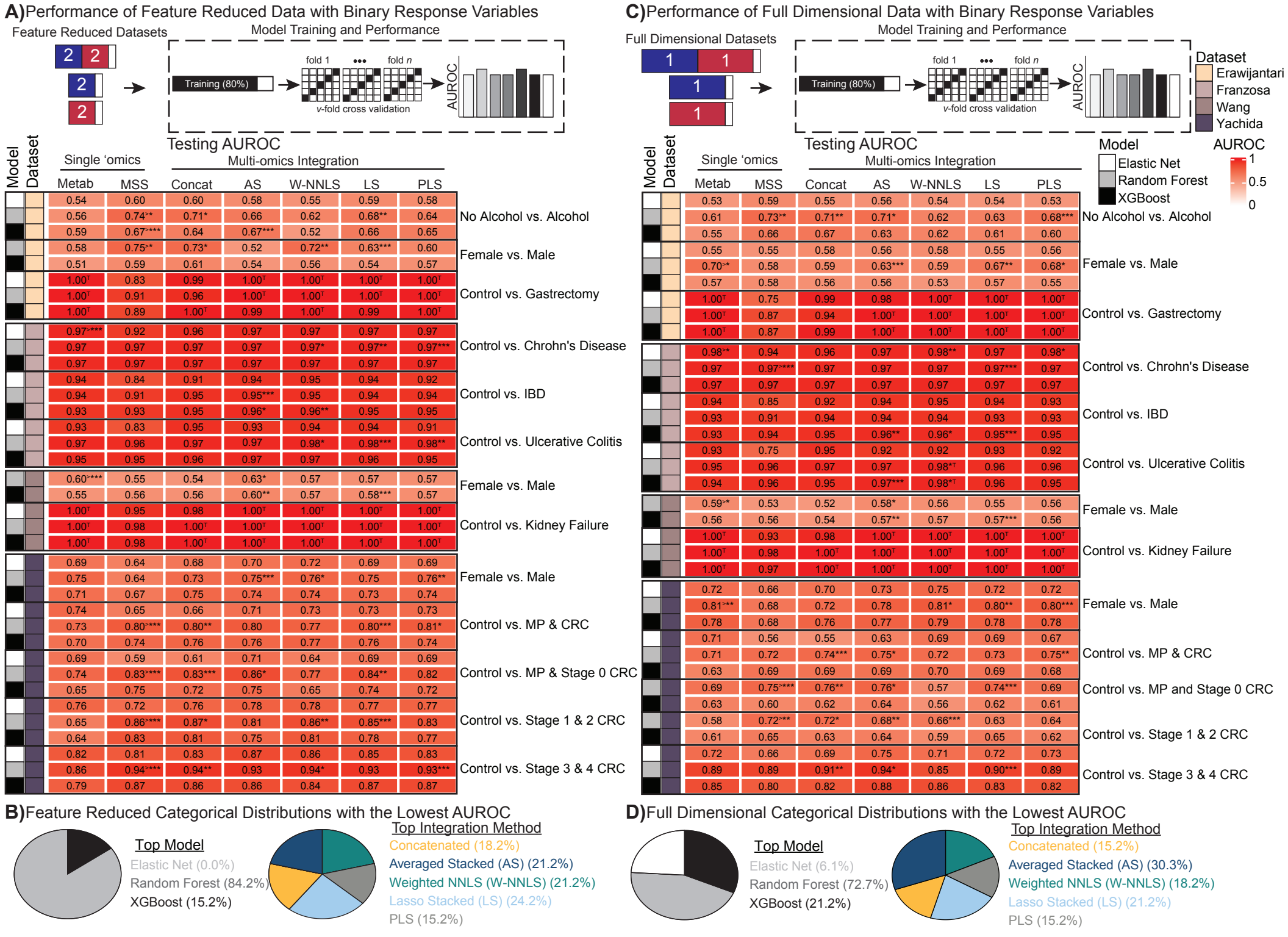
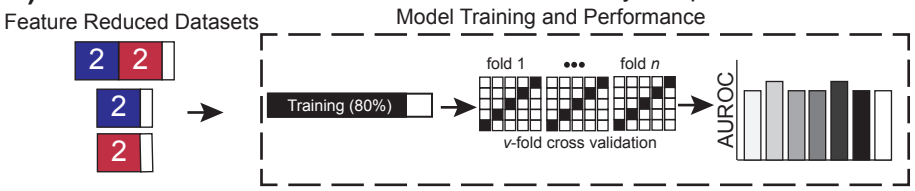

A) Performance of Feature Reduced Data with Binary Response Variables
Feature Reduced Datasets
Model Training and Performance
Training (80%)
fold 1
fold n
v-fold cross validation
AUROC
Testing AUROC
Model
Dataset
Single 'omics
Multi-omics Integration
Metab
MSS
Concat
AS
W-NNLS
LS
PLS
No Alcohol vs. Alcohol
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
B) Feature Reduced Categorical Distributions with the Lowest AUROC
Top Model
Elastic Net (0.0%)
Random Forest (84.2%)
XGBoost (15.2%)
Top Integration Method
Concatenated (18.2%)
Averaged Stacked (AS) (21.2%)
Weighted NNLS (W-NNLS) (21.2%)
Lasso Stacked (LS) (24.2%)
PLS (15.2%)
C) Performance of Full Dimensional Data with Binary Response Variables
Full Dimensional Datasets
Model Training and Performance
Training (80%)
fold 1
fold n
v-fold cross validation
AUROC
Dataset
Erawijantari
Franzosa
Wang
Yachida
Testing AUROC
Model
Elastic Net
Random Forest
XGBoost
AUROC
1
0.5
0
Single 'omics
Multi-omics Integration
Metab
MSS
Concat
AS
W-NNLS
LS
PLS
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 and Stage 0 CRC
Control vs. Stage 1 & 2 CRC
Control vs. Stage 3 & 4 CRC
D) Full Dimensional Categorical Distributions with the Lowest AUROC
Top Model
Elastic Net (6.1%)
Random Forest (72.7%)
XGBoost (21.2%)
Top Integration Method
Concatenated (15.2%)
Averaged Stacked (AS) (30.3%)
Weighted NNLS (W-NNLS) (18.2%)
Lasso Stacked (LS) (21.2%)
PLS (15.2%)

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 '***').

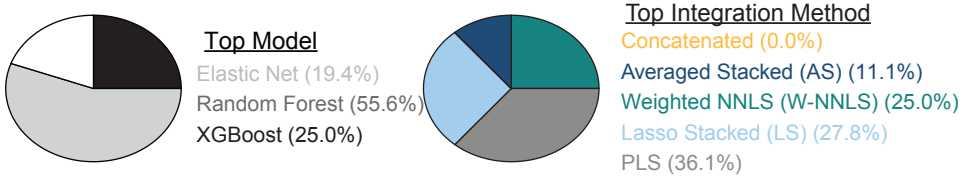
A) Performance of Feature Reduced Data with Binary Response Variables



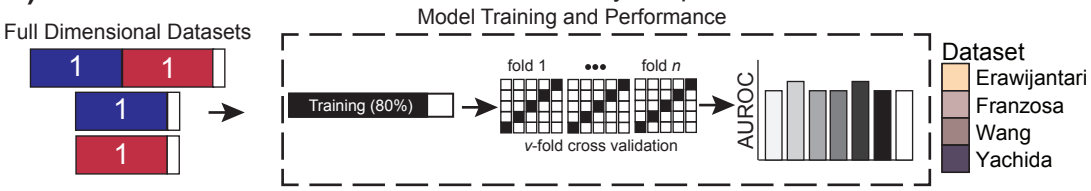
Validation AUROC

Model	Dataset	Metab (Single 'omics)	MSS (Single 'omics)	Concat (Multi-omics Integration)	AS (Multi-omics Integration)	W-NNLS (Multi-omics Integration)	LS (Multi-omics Integration)	PLS (Multi-omics Integration)	
Elastic Net		0.59	0.60	0.62	0.60	0.59	0.59	0.65	No Alcohol vs. Alcohol
Random Forest		0.63	0.64	0.63	0.67***	0.60	0.70*	0.69**	
XGBoost		0.54	0.57	0.58	0.60	0.53	0.60	0.63	
Random Forest		0.72	0.71	0.69	0.76***	0.75	0.79**	0.80*	Female vs. Male
XGBoost		0.64	0.62	0.68	0.70	0.65	0.71	0.72	
Elastic Net		0.97	0.76	0.87	0.93	0.97	0.96	0.97	Control vs. Gastrectomy
Random Forest		0.95	0.90	0.93	0.98**	0.98*	0.98	0.98***	
XGBoost		0.92	0.87	0.94	0.96	0.98	0.96	0.97	
Elastic Net		0.94	0.87	0.93	0.93	0.94	0.94	0.94	Control vs. Chrohn's Disease
Random Forest		0.93	0.94	0.94	0.94	0.94	0.94	0.94	
XGBoost		0.94	0.93	0.93	0.94	0.95***	0.95**	0.95**	
Elastic Net		0.95	0.83	0.90	0.94	0.95***	0.95**	0.95*	Control vs. IBD
Random Forest		0.91	0.88	0.91	0.91	0.91	0.92	0.91	
XGBoost		0.92	0.88	0.92	0.92	0.93	0.93	0.92	
Elastic Net		0.96	0.75	0.93	0.93	0.96***	0.96	0.96	Control vs. Ulcerative Colitis
Random Forest		0.94	0.85	0.94	0.94	0.94	0.95	0.94	
XGBoost		0.96	0.85	0.95	0.95	0.96*	0.96	0.96**	
Random Forest		0.57	0.63	0.62	0.60	0.61	0.64***	0.65*	Female vs. Male
XGBoost		0.60	0.62	0.60	0.64	0.63	0.64**	0.63	
Elastic Net		1.00^{T}	0.89	0.97	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	Control vs. Kidney Failure
Random Forest		1.00^{T}	0.90	0.98	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	
XGBoost		1.00^{T}	0.92	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	
Elastic Net		0.70	0.71	0.73	0.75	0.76	0.76	0.76	Female vs. Male
Random Forest		0.77	0.64	0.74	0.77	0.79**	0.78	0.79*	
XGBoost		0.77	0.64	0.76	0.77	0.79	0.79	0.79***	
Elastic Net		0.68	0.65	0.69	0.68	0.72	0.73	0.72	Control vs. MP & CRC
Random Forest		0.70	0.77	0.77	0.78***	0.78	0.79**	0.79*	
XGBoost		0.64	0.75	0.74	0.73	0.76	0.75	0.76	
Elastic Net		0.68	0.51	0.54	0.66	0.66	0.67	0.68***	Control vs. MP & Stage 0 CRC
Random Forest		0.56	0.67	0.65	0.67	0.67	0.68**	0.67	
XGBoost		0.59	0.67	0.65	0.65	0.65	0.68	0.69*	
Elastic Net		0.80	0.74	0.78	0.80	0.83***	0.83**	0.83	Control vs. Stage 1 & 2 CRC
Random Forest		0.73	0.80	0.80	0.81	0.83*	0.82	0.82	
XGBoost		0.70	0.77	0.77	0.78	0.78	0.79	0.79	
Elastic Net		0.77	0.69	0.68	0.76	0.78	0.79	0.79	Control vs. Stage 3 & 4 CRC
Random Forest		0.69	0.82	0.83	0.80	0.83**	0.84*	0.83	
XGBoost		0.64	0.83	0.83	0.80	0.81	0.82	0.83***	

B) Feature Reduced Categorical Distributions with the Lowest AUROC



C) Performance of Full Dimensional Data with Binary Response Variables



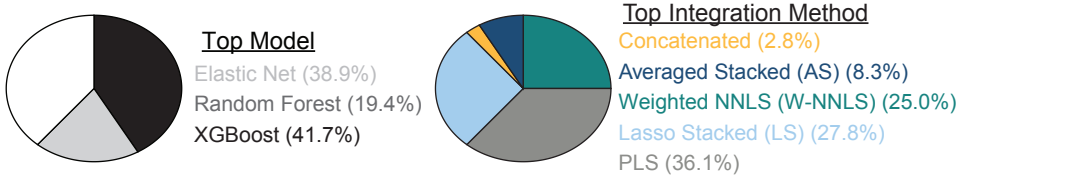
Validation AUROC

Model	Dataset	Metab (Single 'omics)	MSS (Single 'omics)	Concat (Multi-omics Integration)	AS (Multi-omics Integration)	W-NNLS (Multi-omics Integration)	LS (Multi-omics Integration)	PLS (Multi-omics Integration)	
Elastic Net		0.66	0.60	0.63	0.61	0.63	0.67	0.69	No Alcohol vs. Alcohol
Random Forest		0.64	0.60	0.60	0.61	0.67	0.67	0.70*^{T}	
XGBoost		0.68	0.60	0.59	0.66	0.66	0.69***	0.70*^{T}	
Elastic Net		0.73^{T}***	0.54	0.69	0.74**	0.73	0.73***	0.75*	Female vs. Male
Random Forest		0.68	0.61	0.64	0.66	0.65	0.68	0.70	
XGBoost		0.65	0.62	0.60	0.64	0.59	0.66	0.68	
Elastic Net		0.97	0.68	0.88	0.90	0.98*	0.97**	0.97***	Control vs. Gastrectomy
Random Forest		0.93	0.87	0.92	0.96	0.97	0.95	0.97	
XGBoost		0.94	0.84	0.91	0.94	0.96	0.96	0.96	
Elastic Net		0.94	0.89	0.93	0.93	0.94**	0.94***	0.94	Control vs. Chrohn's Disease
Random Forest		0.93	0.93	0.94	0.94	0.94	0.94	0.94	
XGBoost		0.94	0.93	0.94	0.94	0.95*	0.94	0.94	
Elastic Net		0.93***	0.84	0.90	0.92	0.93**	0.94*	0.93***	Control vs. IBD
Random Forest		0.91	0.88	0.91	0.91	0.91	0.91	0.91	
XGBoost		0.91	0.87	0.91	0.91	0.92	0.92	0.91	
Elastic Net		0.96	0.77	0.94	0.94	0.97*	0.96***	0.96**	Control vs. Ulcerative Colitis
Random Forest		0.92	0.84	0.93	0.93	0.93	0.93	0.93	
XGBoost		0.94	0.84	0.93	0.94	0.95	0.95	0.95	
Random Forest		0.57	0.59	0.59	0.59	0.61	0.62	0.63	Female vs. Male
XGBoost		0.62	0.61	0.59	0.65***	0.65	0.67*	0.67**	
Elastic Net		0.99	0.87	0.96	0.99	1.00^{T}	0.99	0.99	Control vs. Kidney Failure
Random Forest		1.00^{T}	0.89	0.97	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	
XGBoost		1.00^{T}	0.91	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	1.00^{T}	
Random Forest		0.69	0.67	0.70	0.73	0.75	0.75	0.75	Female vs. Male
XGBoost		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	
Random Forest		0.65	0.61	0.60	0.63	0.69	0.69	0.70	Control vs. MP and CRC
XGBoost		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	
Random Forest		0.55	0.58	0.56	0.56	0.55	0.58	0.61**	Control vs. MP & Stage 0 CRC
XGBoost		0.58	0.58	0.61***	0.58	0.56	0.60	0.62*	
Random Forest		0.64	0.60	0.61	0.64	0.64	0.67	0.67	Control vs. Stage 1 & 2 CRC
XGBoost		0.66	0.65	0.62	0.67***	0.62	0.69**	0.69*	
Random Forest		0.70	0.59	0.58	0.70	0.70	0.70	0.72	Control vs. Stage 3 & 4 CRC
XGBoost		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*	

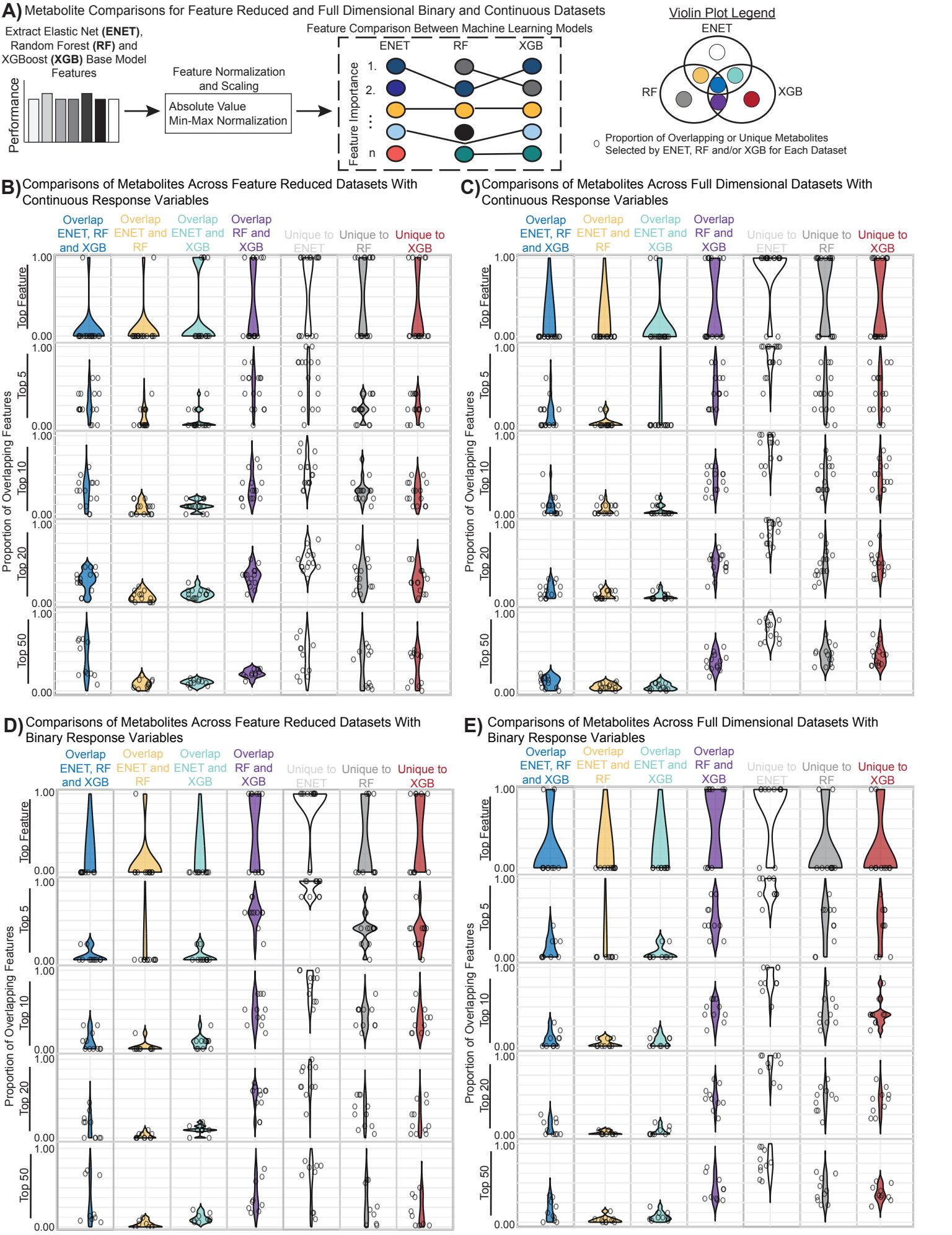
Model: Elastic Net, Random Forest, XGBoost

AUROC: 1, 0.5, 0

D) Full Dimensional Categorical Distributions with the Lowest AUROC



610 **Supplemental Figure 2. Validation Performance Metrics Across Binary Feature Reduced and Full**
611 **Dimensional Datasets.** Validation performance for feature reduced matched microbiome and metabolomics
612 datasets with binary response variables are represented by **A)**. The top performing machine learning models
613 and top integration methods for each of the feature reduced datasets are represented as pie charts in **B)**. Testing
614 performance for full dimensional matched microbiome and metabolomics datasets with binary response
615 variables are represented by **C)**. The top performing machine learning models and top integration methods for
616 each of the full dimensional datasets are represented as pie charts in **D)**. Performance is calculated using
617 AUROC mean, which are shown in the heatmap. Response variables used for each model are shown on the
618 right of each heatmap. The top performing models per response variable are represented by '*', where '*' is the
619 best performing model, '**' is the second-best performing model and '***' is the third-best performing model. The
620 proportions are calculated using the top three best performing models (*, **, ***) per response variable. 'T'
621 represents ties in performance; response variables with majority 'T' were not used for pie chart calculations.
622 Individual 'omics (Metabolomics and MSS) were not represented in the pie charts. '>' or 'T' represents increased
623 or tied performance of an individual 'omics to one of the top performing integration models ('*', '**', and '***').
624


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

625 **Figure 4. Metabolite Feature Selection Comparisons across Elastic Net, Random Forest and XGBoost.**
626 The pipeline for extracting feature importance metrics and performing comparisons are represented in **A)**.
627 Feature Importance values for Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) were extracted for
628 each cross-validation step. The absolute value was taken for each value and min-max normalization was
629 performed. The overall average for each feature was then used as the metric for feature importance. See
630 **Methods** for more details. Comparisons for the metabolites selected across feature reduced datasets with
631 continuous response variables are represented by **B)**. Comparisons for the metabolites selected across full
632 dimensional datasets with continuous response variables are represented by **C)**. Comparisons for the
633 metabolites selected across feature reduced datasets with binary response variables are represented by **D)**.
634 Comparisons for the metabolites selected across full dimensional datasets with binary response variables are
635 represented by **E)**. For each analysis, top features for ENET, RF and XGB were binned into 5 levels – top feature
636 (i.e., features with highest feature importance values), top 5 features, top 10 feature, top 20 features and top 50
637 features. For each bin, comparisons for the features selected between ENET, RF and XGB were performed to
638 identify joint and unique features. The distribution of dataset proportions for overlapping and unique features are
639 represented by violin plots for panels **B-E**.

640

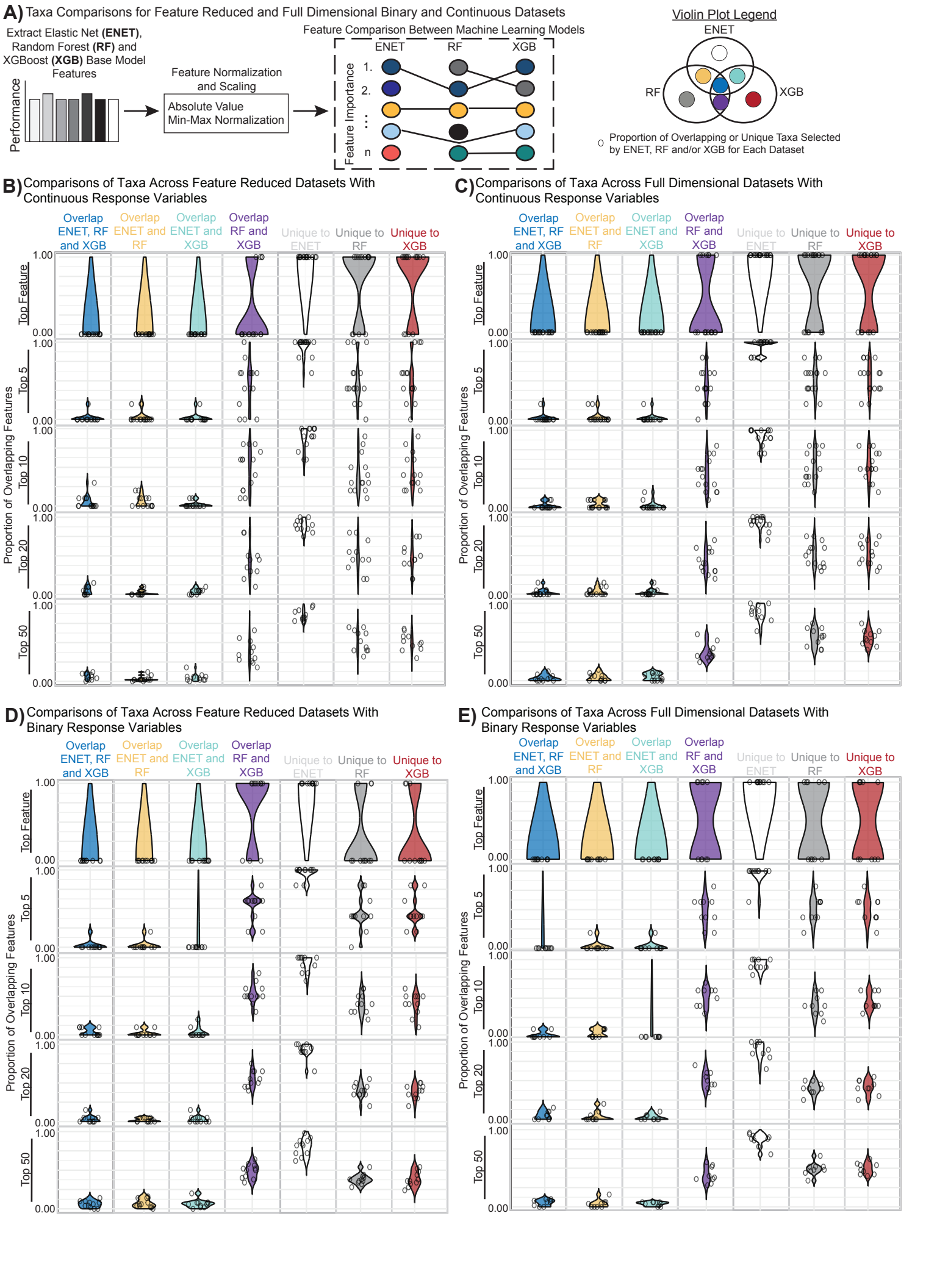
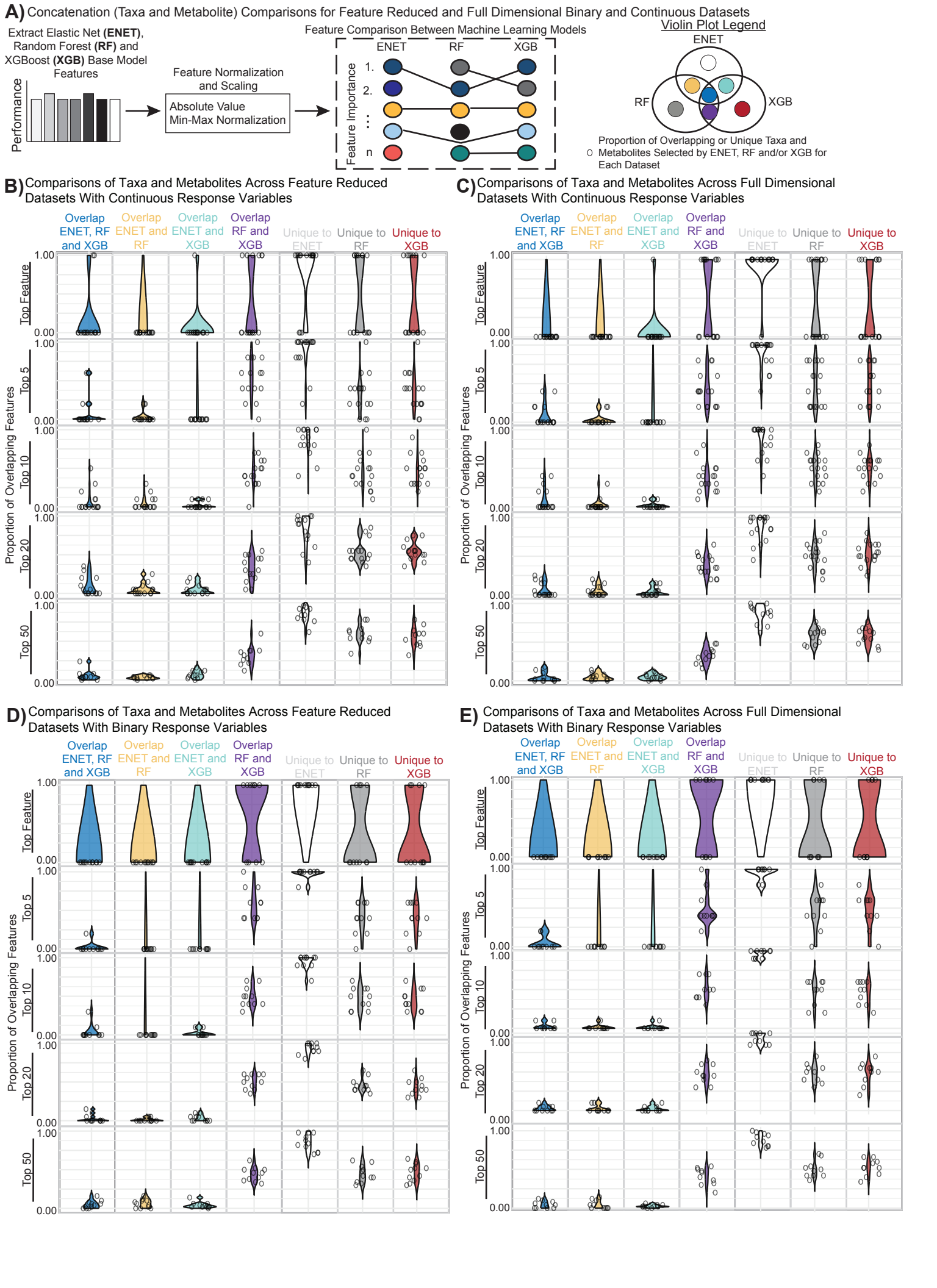

A) Taxa Comparisons for Feature Reduced and Full Dimensional Binary and Continuous Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning 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

Figure 5. Taxa 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 taxa selected across feature reduced datasets with continuous response variables are represented by **B)**. Comparisons for the taxa selected across full dimensional datasets with continuous response variables are represented by **C)**. Comparisons for the taxa selected across feature reduced datasets with binary response variables are represented by **D)**. Comparisons for the 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) 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

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

671

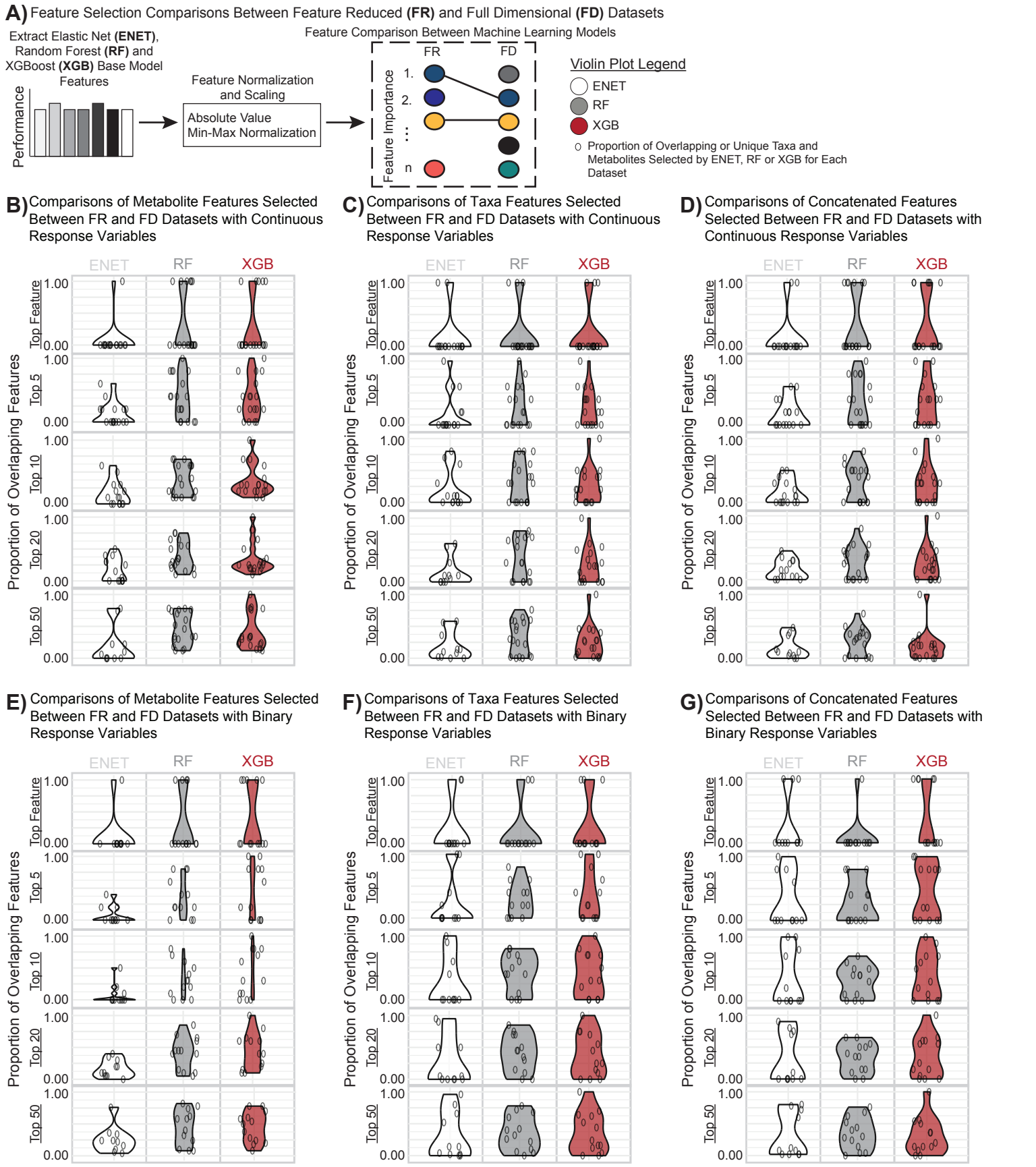
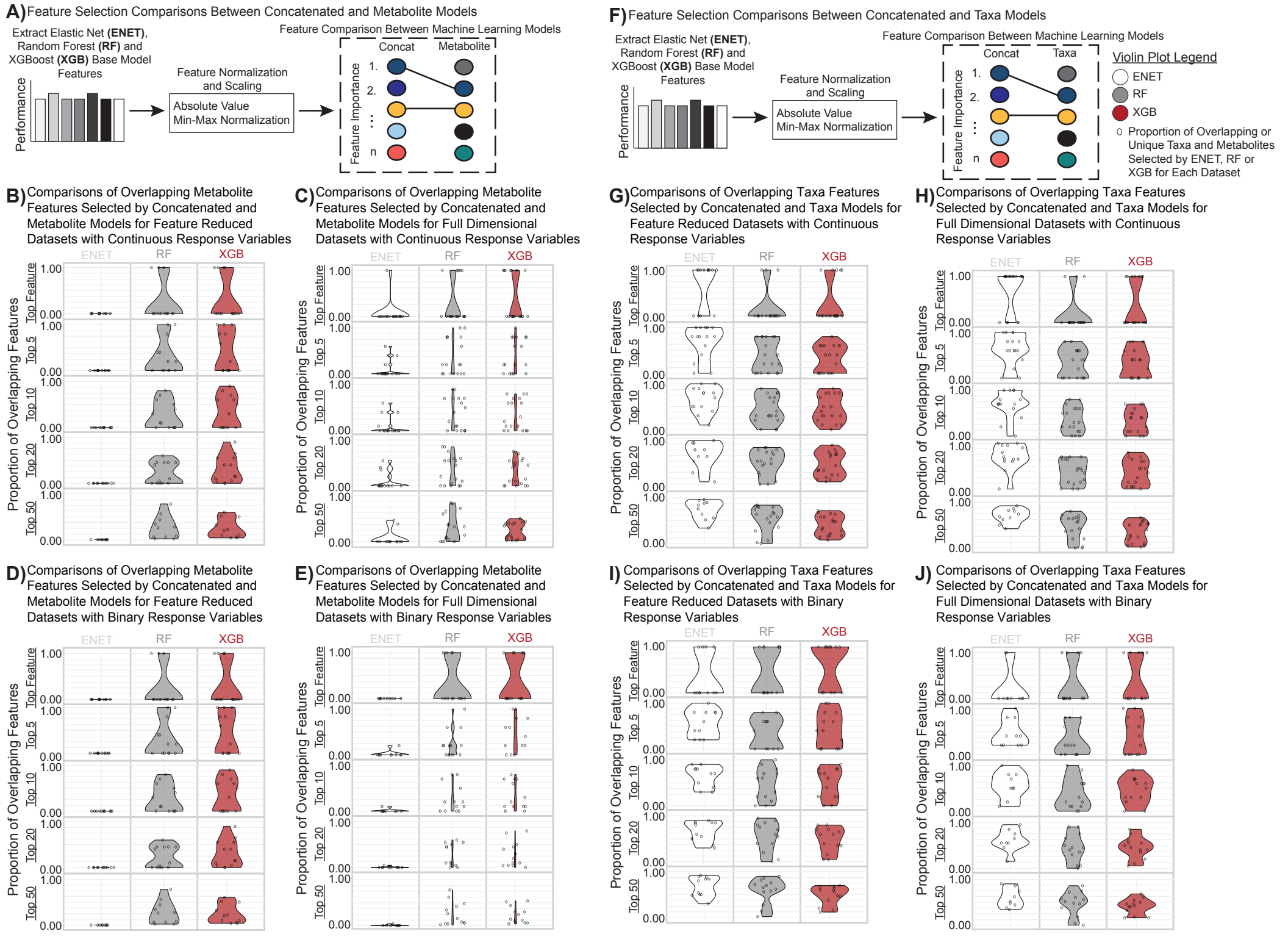
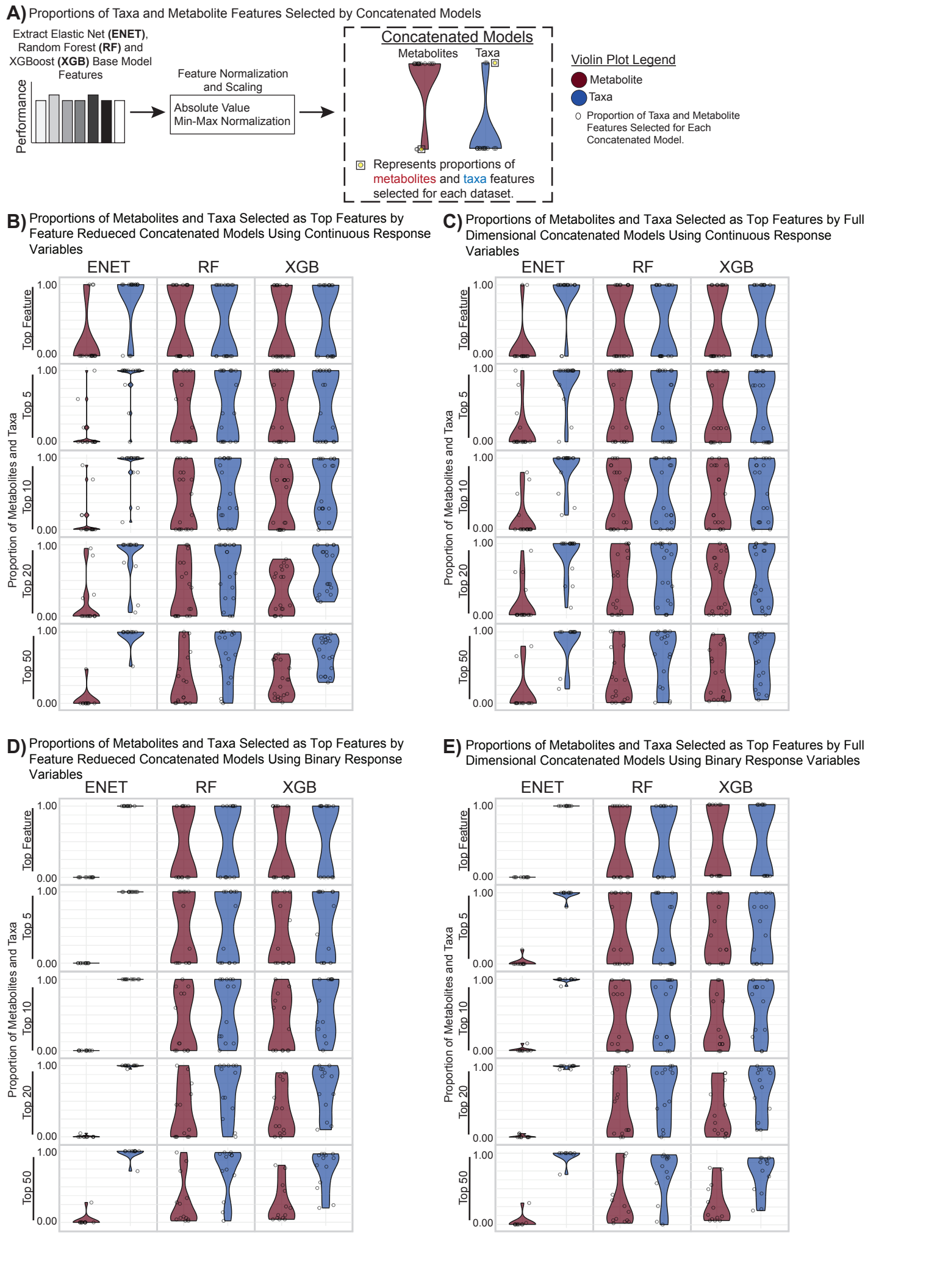

A) Feature Selection Comparisons Between Feature Reduced (FR) and Full Dimensional (FD) Datasets
Extract Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Base Model Features
Performance
Feature Normalization and Scaling
Absolute Value
Min-Max Normalization
Feature Comparison Between Machine Learning 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

692 **Figure 8. Feature Selection Comparisons Between Concatenated (Concat) and Individual**
693 **(Metabolomics/Taxa) Elastic Net (ENET), Random Forest (RF) and XGBoost (XGB) Models.** The pipeline
694 for extracting feature importance metrics and performing comparisons between concatenated and metabolites
695 models are represented in **A)**. Feature Importance values for ENET, RF and XGB were extracted for each cross-
696 validation step. The absolute value was taken for each value and min-max normalization was performed. The
697 overall average for each feature was then used as the metric for feature importance. See **Methods** for more
698 details. Comparisons for the metabolites selected by concatenated and metabolomics models for feature
699 reduced datasets with continuous response variables are represented by **B)**. Comparisons for the metabolites
700 selected by concatenated and metabolomics models for full dimensional datasets with continuous response
701 variables are represented by **C)**. Comparisons for the metabolites selected by concatenated and metabolomics
702 models for feature reduced datasets with binary response variables are represented by **D)**. Comparisons for the
703 metabolites selected by concatenated and metabolomics models for full dimensional datasets with binary
704 response variables are represented by **E)**. The pipeline for extracting feature importance metrics and performing
705 comparisons between concatenated and metabolites models are represented in **F)** (See details above (**A**) and
706 in **Methods**). Comparisons for the taxa selected by concatenated and taxa models for feature reduced datasets
707 with continuous response variables are represented by **G)**. Comparisons for the taxa selected by concatenated
708 and taxa models for full dimensional datasets with continuous response variables are represented by **H)**.
709 Comparisons for the taxa selected by concatenated and taxa models for feature reduced datasets with binary
710 response variables are represented by **I)**. Comparisons for the taxa selected by concatenated and taxa models
711 for full dimensional datasets with binary response variables are represented by **J)**. For each analysis, top features
712 for ENET, RF and XGB were binned into 5 levels – top feature (i.e., features with highest feature importance
713 values), top 5 features, top 10 feature, top 20 features and top 50 features. For each bin, comparisons for the
714 features selected by concatenated and individual (metabolomics and taxa) 'omics models for ENET, RF and
715 XGB were performed to identify joint and unique features. The distribution of dataset proportions for overlapping
716 and unique features are represented by violin plots for panels **B-E** and **G-J**.

717


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

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

733 **References**

734 1. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*
735 (2016) doi:10.1038/nature18850.
736 2. Joos, R. *et al.* Examining the healthy human microbiome concept. *Nature Reviews Microbiology* (2024)
737 doi:10.1038/s41579-024-01107-0.
738 3. Walter, J. *et al.* Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human
739 Microbiota-Associated Rodents. *Cell* (2020) doi:10.1016/j.cell.2019.12.025.
740 4. Corander, J. *et al.* Causal discovery for the microbiome. *The Lancet microbe* (2022) doi:10.1016/s2666-
741 5247(22)00186-0.
742 5. Knight, R. *et al.* Best practices for analysing microbiomes. *Nature Reviews Microbiology* (2018)
743 doi:10.1038/s41579-018-0029-9.
744 6. Galloway-Peña, J. *et al.* Tools for Analysis of the Microbiome. *Digestive Diseases and Sciences* (2020)
745 doi:10.1007/s10620-020-06091-y.
746 7. Song, E.-J. *et al.* Progress of analytical tools and techniques for human gut microbiome research. *Journal*
747 *of Microbiology* (2018) doi:10.1007/s12275-018-8238-5.
748 8. Hasin, Y. *et al.* Multi-omics approaches to disease. *Genome Biology* (2017) doi:10.1186/s13059-017-1215-
749 1.
750 9. Chetty, A. & Blekhman, R. Multi-omic approaches for host-microbiome data integration. *Gut microbes*
751 (2024) doi:10.1080/19490976.2023.2297860.
752 10. Li, P. *et al.* Machine learning for data integration in human gut microbiome. *Microbial Cell Factories* (2022)
753 doi:10.1186/s12934-022-01973-4.
754 11. Franzosa, E. A. *et al.* Author Correction: Gut microbiome structure and metabolic activity in inflammatory
755 bowel disease. *Nature microbiology* (2019) doi:10.1038/s41564-019-0442-5.
756 12. Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* (2017)
757 doi:10.1186/s40168-016-0222-x.
758 13. Gao, B. *et al.* Functional Microbiomics Reveals Alterations of the Gut Microbiome and Host Co-Metabolism
759 in Patients With Alcoholic Hepatitis. *Hepatology Communications* (2020) doi:10.1002/hep4.1537.

760 14. Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the
761 gut microbiota in colorectal cancer. *Nature Medicine* (2019) doi:10.1038/s41591-019-0458-7.
762 15. Ghaemi, M. S. *et al.* Multiomics modeling of the immunome, transcriptome, microbiome, proteome and
763 metabolome adaptations during human pregnancy. *Bioinformatics* (2018)
764 doi:10.1093/bioinformatics/bty537.
765 16. Mallick, H. *et al.* An integrated Bayesian framework for multi-omics prediction and classification. *Stat*
766 *Medicine* (2023) doi:10.1002/sim.9953.
767 17. Liu, Y.-X. *et al.* A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell*
768 (2020) doi:10.1007/s13238-020-00724-8.
769 18. Poussin, C. *et al.* Interrogating the microbiome: experimental and computational considerations in support
770 of study reproducibility. *Drug Discovery Today* (2018) doi:10.1016/j.drudis.2018.06.005.
771 19. Huttenhower, C., Finn, R. K. & McHardy, A. C. Challenges and opportunities in sharing microbiome data
772 and analyses. *Nature Microbiology* (2023) doi:10.1038/s41564-023-01484-x.
773 20. Santiago-Rodríguez, T. M., Santiago-Rodriguez, T. M., Hollister, E. B. & Hollister, E. B. Multi 'omic data
774 integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology* (2021)
775 doi:10.1016/j.semperi.2021.151456.
776 21. Papoutsoglov, G. *et al.* Machine learning approaches in microbiome research: challenges and best
777 practices. *Frontiers in Microbiology* (2023) doi:10.3389/fmicb.2023.1261889.
778 22. Mullen, K. M. & Stokkum, I. H. M. van. *Nnls: The Lawson-Hanson Algorithm for Non-Negative Least*
779 *Squares (NNLS)*. (2024).
780 23. Friedman, J. H. *et al.* Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal*
781 *of Statistical Software* (2010) doi:10.18637/jss.v033.i01.
782 24. Mevik, B.-H. & Wehrens, R. The pls Package: Principal Component and Partial Least Squares Regression
783 in R. *Journal of Statistical Software* **18**, 1–23 (2007).
784 25. Erawijantari, P. P. *et al.* Influence of gastrectomy for gastric cancer treatment on faecal microbiome and
785 metabolome profiles. *Gut* (2020) doi:10.1136/gutjnl-2019-319188.
786 26. Wang, X. *et al.* Aberrant gut microbiota alters host metabolome and impacts renal failure in humans and
787 rodents. *Gut* (2020) doi:10.1136/gutjnl-2019-319766.

27. Medina, R. H. *et al.* Machine learning and deep learning applications in microbiome research. *ISME Communications* (2022) doi:10.1038/s43705-022-00182-9.
28. Kumar, B., Lorusso, E., Fosso, B. & Pesole, G. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in Microbiology* (2024) doi:10.3389/fmicb.2024.1343572.
29. Saeys, Y. *et al.* A review of feature selection techniques in bioinformatics. *Bioinformatics* (2007) doi:10.1093/bioinformatics/btm344.
30. Díaz-Uriarte, R., Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* (2006) doi:10.1186/1471-2105-7-3.
31. Wood, D. E. *et al.* Improved metagenomic analysis with Kraken 2. *Genome Biology* (2019) doi:10.1186/s13059-019-1891-0.
32. Lu, J. *et al.* Bracken: estimating species abundance in metagenomics data. *PeerJ* (2017) doi:10.7717/peerj-cs.104.
33. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* (2015) doi:10.1093/nar/gkv007.
34. Karwowska, Z., Aasmets, O., Kosciólek, T. & Org, E. Effects of data transformation and model selection on feature importance in microbiome classification data. *Microbiome* (2025) doi:10.1186/s40168-024-01996-6.
35. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* (2019) doi:10.1093/bioinformatics/bty1054.
36. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* (2020) doi:10.1186/s13059-020-02015-1.
37. Muller, E. *et al.* The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *npj biofilms and microbiomes* (2022) doi:10.1038/s41522-022-00345-5.
38. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* (2019) doi:10.21105/joss.01686.
39. Robin, X. *et al.* PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* (2011) doi:10.1186/1471-2105-12-77.

40. Gamer, M., Lemon, J. & Singh <puspendra.pusp22@gmail.com>, I. F. P. *Irr: Various Coefficients of Interrater Reliability and Agreement*. (2019).
41. Greener, J. G. *et al.* A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology* (2021) doi:10.1038/s41580-021-00407-0.
42. Li, Y. *et al.* A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics* (2016) doi:10.1093/bib/bbw113.
43. Ghannam, R. B., Ghannam, R. B., Techtmann, S. M. & Techtmann, S. M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and structural biotechnology journal* (2021) doi:10.1016/j.csbj.2021.01.028.
44. Kühn, M. & Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* (2008) doi:10.18637/jss.v028.i05.
45. Liaw, A., Liaw, A., Wiener, M. C. & Wiener, M. C. Classification and Regression by randomForest. *null* (2007) doi:null.
46. Chen, T. *et al. Xgboost: Extreme Gradient Boosting*. (2025).
47. Gu, Z. & Gu, Z. Complex heatmap visualization. *iMeta* (2022) doi:10.1002/imt2.43.
48. Gu, Z. *et al.* Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw313.
49. Valero-Mora, P. M. & Valero-Mora, P. M. Ggplot2: Elegant Graphics for Data Analysis. *Journal of Statistical Software* (2010) doi:10.18637/jss.v035.b01.