Pacific Northwest
NATIONAL LABORATORY

PNNL-38772

Ranking Biological Features in Soil-Based Microbial Multi-Omics Data with Integration Modeling

December 2025

David J. Degnan
Ryan McClure
Daniel M. Claborne
Lisa M. Bramer
Javier E. Flores

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-38772

Ranking Biological Features in Soil-Based Microbial Multi-Omics Data with Integration Modeling

December 2025

David J. Degnan
Ryan McClure
Daniel M. Claborne
Lisa M. Bramer
Javier E. Flores

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Distinguishing the most important features (e.g. proteins, metabolites, etc.) per group (e.g. control and treatment) is a critical challenge in feature-rich multi-omics experiments, especially in soil data. Traditional feature identification and ranking approaches, such as differential expression, are based on single omics and thus not directly translatable to multi-omics experiments. Here, 5 multi-omics integration models (DIABLO, JACA, MOFA, MultiMLP, and SLIDE) that were not explicitly built for soil data applications were tested using a soil-based multi-omics experiment. The data were obtained from an experimental setup of an autoclaved soil system inoculated with 8 bacteria and using chitin as the carbon source and including samples collected at 0- (control), 4-, 8-, and 12-weeks post-inoculation. The omics data included metaproteomics, 16S rRNA sequencing, and LC-MS/MS metabolomics (in positive and negative mode). Each multi-omics integration model was implemented, and top features were compared to differential univariate statistics per omic type, demonstrating that integration approaches cut the potential number of top features from 2957 identified by differential statistics to 13-224 (a 99.6% to 92.4% reduction). Interestingly, most top features across integration models were not shared; though, scaling and averaging ranks across models shared similar patterns. This work highlights the usefulness of multi-omics integration models in soil-based microbial studies and the power of using multiple integration models together to interpret results.

Summary

This study explores five integration models to rank biological features in soil multi-omics data, enhancing environmental microbiology insights through a complementary approach.

Acknowledgments

This research was supported by the Earth and Biological Sciences Directorate Seed Projects, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

DIABLO, Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies; EM, expectation maximization; JACA, Joint Association and Classification Analysis; LC-MS/MS, liquid chromatography coupled with tandem mass spectrometry; MOFA, Multi-Omics Factor Analysis; MS, mass spectrometry; MultiMLP, Multiple Multi-Layer Perceptrons; Multi-Omics, Multiple Omics; SLIDE, Structural Learning and Integrative Decomposition

Contents

Abstract ii
Summary iii
Acknowledgments iv
Acronyms and Abbreviations v
1.0 Introduction 1
2.0 Methods 2
2.1 Data Acquisition 2
2.2 Data Filtering, Missing Value Imputation, and Scaling 2
2.3 Selection of Integration Models 3
2.4 Differential Expression and Abundance Statistics 4
2.5 Hyperparameter Tuning 4
2.6 Top Feature Selection 4
2.7 Top Feature Class Identification 5
3.0 Results 6
3.1 Omics Data Properties and the Impacts of Filtering Decisions 6
3.2 Selection of Hyperparameters, Components, and Top Features 7
3.3 Similarity of Top Features Across Integration Models 8
3.4 Top Feature Exploration 10
4.0 Discussion 14
5.0 References 15

Figures

Figure 1. (A) Log10 counts of features impacted by the missingness filter for metabolomics negative (left), metabolomics positive (middle), and metaproteomics (right) data. Log10 is used for visualization purposes. Black indicates the feature was maintained and red indicates a feature was filtered out. 16S data had no missingness. (B) Pearson correlation matrices of each omic before and (C) after expectation maximization imputation. Note that 16S data was not imputed but included in the plot for comparison purposes. 6
Figure 2. Feature absolute weights per integration model, with a cutoff determined by the kneedle algorithm (dashed line). Features are colored by omic type: 16S (red), metabolomics negative (light green), metabolomics positive (light blue), and metaproteomics (purple) 8
Figure 3. (A) Proportion of all possible features considered a “top feature” using differential statistics with a p-value cut-off of 0.05, and the integration models after kneedle top feature selection. (B) Proportions from A split by omics type. (C) Count of the number of top features shared across the

five integration models. (D) Shared features between integration models and differential expression. The numerator is the number of overlapping features between the models in the columns and rows, and the denominator in the number of features in the model listed at the bottom of the column. 8

Figure 4. Scaled ranks of top features, ordered from highest to lowest mean scaled rank. Features present in only one integration model or features without a specific identification (e.g. a class of metabolites as opposed to a specific metabolite) were removed. 11

Figure 5. Average scaled ranks of feature information, ordered from highest to lowest scaled mean rank across integration models for (A) CANOPUS categories of metabolomics data (combining both negative and positive ion mode), and (B) COG categories of protein functions. Mean ranks are reported, as long as the number of features in each category. 12

Tables

Table 1. Short descriptions of selected integration models with hyperparameters that were tuned. 3

Table 2. A description of integration models and selected hyperparameters. 7

Table 3. Number of features returned per integration model per omic (view). Differential expressed/abundant biomolecules are included for comparison. 9

1.0 Introduction

In traditional single omics bulk experiments, a sample (e.g. 1-10 grams of soil) is collected and biomolecules (e.g. DNA, proteins, lipids, metabolites) are extracted, undergo instrumentation (such as mass spectrometry analysis or sequencing), and are identified and quantified with computational annotation tools.[1] Typically, key biological features which distinguish sample groups (e.g. a control and an experimental group) are then determined using differential expression or abundance statistics[2-9], depending on the input data type. Features may then be ranked using p-values, fold changes, and other methods[10, 11] to identify those most important in distinguishing groups. Though useful, traditional single-omics statistical approaches are not well-suited for multi-omics experiments because p-values and fold changes cannot be directly compared across different omics datasets.

This discrepancy can be attributed to several factors, including differences in experimental design across omics (e.g., varying extraction techniques), instrumentation (e.g., mass spectrometry vs. sequencing), pre-processing choices, how expression values are measured (e.g., relative abundances in mass spectrometry vs. transcript counts in sequencing), and the generating mechanism and interpretation of missing values.[12-17] To address this gap, multi-omics integration models[18-23] have been developed to standardize disparate omics datasets, enabling easier comparison across omics datasets, often called views. Multi-omics integration models are variable in their design.[24] Most can be categorized into three integration types: early (concatenating features across datasets for high-dimensional analysis), middle (transforming datasets into a combined representation before analysis), or late (analyzing datasets separately and combining results through a model or algorithm).[24] An advantage of middle integration methods is that they address the “curse of dimensionality” problem (e.g. the increased risk of overfitting as the dimensions increase) of integration with reduction techniques, while also capturing both omic and inter-omic signals.[25] This is as opposed to early integration that inherently ignores the dimensionality problem and may be biased by the disparities in feature sizes across omics; and late integration methods which may overly emphasize omic-specific signals and are more limited in capturing trends persistent across omics.[25]

Here five middle integration models[18-23] were selected and compared using a previously published study on the biological response of eight soil bacteria on the breakdown of chitin with four omics datasets: 16S rRNA sequencing, metaproteomics, and LC-MS/MS metabolomics (one in positive ion mode and the other in negative ion mode).[26] To standardize hyperparameter tuning and feature ranking for ease of comparison, additional code and methods were developed. Differences in these feature rankings and their class representation (e.g. a metabolite class such as an organic acid), were compared and interpreted in the context of the original study and to univariate statistics (differential expression/abundance). This work highlights how integration models built for different biological contexts can be useful in soil-based research, and how a combination of multi-omics integration models together better resolves biological patterns than a single model on its own.

2.0 Methods

2.1 Data Acquisition

The experimental data has been published previously.[26, 27] Eight microbial strains (*Streptomyces* sp001905665 strain 001, *Neorhizobium tomejilense* strain 005, *Dyadobacter* sp. strain 007, *Sphingopyxis* sp. strain 008, *Ensifer adhaerens* strain 011, *Variovorax beijingensis* strain 012, *Sinorhizobium meliloti* strain 014, and *Rhodococcus* sp003130705 strain 016) with demonstrated roles in chitin breakdown in soil from the Washington State University Irrigated Agricultural Research and Extension Center (IAREC) in Prosser, WA, were selected, as explained elsewhere.[26, 27] Briefly, soil collected from IAREC was autoclaved, exposed to equal concentrations of each microbe, and underwent omics analysis (16S, metaproteomics, and LC-MS/MS metabolomics in positive and negative ion mode) captured at 0, 4, 8, and 12 weeks after inoculation of soil.

DNA was extracted using the Quick-DNA Fecal/Soil Microbe Microprep kit (ZYMO Research, Irvine, CA), underwent 16S rRNA sequencing with V4 forward (515F) and reverse (806R) primers (MiSeq Reagent Kits v2) with an Illumina MiSeq, and results were analyzed with QIIME2[28], DADA2[29], and the SILVA[30] database. Metabolites and peptides were isolated and separated with MPLEx.[31] Metabolites were analyzed with tandem LC-MS/MS using a Hypersil Gold C18 reverse-phase column and a UHPLC Waters Acquity (Waters, Milford, Massachusetts, USA) coupled to a high-resolution Q-Exactive HF-X Orbitrap mass spectrometer (HRMS) (Thermo Fisher Scientific, Waltham, MA). Instrumentation was conducted in Data Dependent Acquisition (DDA) mode in both positive and negative ion mode. Spectra were processed and metabolites identified with MZmine 3.2.3[32], SIRIUS 4[33], and MFAssignR.[34] LC-MS/MS metaproteomics was conducted with a Orbitrap Fusion Lumos, peptides were identified with MaxQuant[35] and rolled up to the protein level by summing the antilogs of log-transformed and normalized values, and proteins were filtered down to those with unique ties to specific bacteria.

2.2 Data Filtering, Missing Value Imputation, and Scaling

Unknown metabolites without at least class information (e.g. a metabolite class such as an organic acid) were filtered out to ensure interpretations can be made. To conduct some of the integration models (DIABLO[18], SLIDE[23]), missing values were not permitted and thus the multi-omics data required imputation. Features (e.g. metabolites, proteins, etc.) without enough representation per group (0 weeks (n = 3) and post-0 weeks (n = 9)) were filtered out before imputation. A maximum of 1 missing sample per feature (33% missingness) was permitted for group 0 weeks and 4 for group post-0 weeks (44% missingness). 16S data was log2 transformed. Proteomics data was log2-transformed and mean-centered normalized. Metabolomics data was processed as described previously.[26] Expectation maximization (EM) imputation[36, 37] was selected for its speed and performance in omics analysis, as compared to other imputation methods.[38, 39] Following imputation with the mvdalab[37] implementation of the EM algorithm, omics datasets were scaled with the scale[40] function.

2.3 Selection of Integration Models

All selected integration models[18-23] were approaches which use both unsupervised and supervised learning. First, an unsupervised step reduces an omics dataset (also called a view) into a lower dimensional space (also called a latent space) and then latent variables are calculated. Then, a supervised step separates groups (e.g. a control and an experimental group) based on those latent variable properties. Interestingly, MOFA[20] was the only approach where group information was not explicitly implemented into model parameters, as MOFA had better performances when groups were not specified. All models[18-23] were open-source, implemented in R or python, and either calculated or had a method to calculate feature rankings. More details on each integration model and the tuned hyperparameters are further explained in Table 1.

Table 1. Short descriptions of selected integration models with hyperparameters that were tuned.

Integration Model	Description	Tuned Hyperparameters
DIABLO[18]	**D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent variable approaches for **O**mics studies. A sparse partial least squares discriminant analysis (sPLS-DA) approach which uses a design matrix of correlations between omics. Latent variables are penalized with a LASSO.	the number of components, the design matrix. Hyperparameters were selected using BER (Balanced Error Rate).
JACA[19]	**J**oint **A**ssociation and **C**lassification **A**nalysis of multi-view data. Combines canonical correlation analysis (CCA) and linear discriminant analysis with (LDA).	alpha (weight between LDA and CCA), $lambda_d$ (regularization for sparsity level in each weight matrix), and rho (shrinkage for elastic net). Hyperparameters were selected using precision-recall area under the curve (PR-AUC).
MOFA[20]	**M**ulti-**O**mics **F**actor **A**nalysis. Incorporates Automatic Relevance Determination into Bayesian Group Factor Analysis to separate variation between multiple and single views. The MOFA+ version includes priors for more flexible regularization.	the number of components. Hyperparameters were selected using Evidence Lower Bound (ELBO).

MultiMLP[21, 22]	Averaging of **Multi**ple **M**ulti-**L**ayer **P**erceptrons. A non-variational adaption of the DeepIMV model. DeepIMV is a deep neural net information bottleneck approach built with PyTorch, originally designed with joint distributions to handle missing data. A slightly altered version of DeepIMV, which averages the marginal model predictions, is used here, as there was no missingness after imputation.	grid size of each omic (view). Hyperparameters were selected using cross-entropy loss.
SLIDE[23]	**S**tructural **L**earning and **I**ntegrative **De**composition. A linked component model that incorporates shared information between views which identifies the joint number of components in the latent space.	a structure matrix of shared components across views. Hyperparameters were selected with weighted Frobenius norm loss.

2.4 Differential Expression and Abundance Statistics

For comparison purposes, differential expression (16S) and differential abundance (metaproteomics, LC-MS/MS metabolomics in positive ion mode, LC-MS/MS metabolomics in negative ion mode) was conducted on the processed data between the two groups (0 weeks and post-0 weeks) using pmartR.[5, 6] To control the false discovery rate of significant biomolecules, *p*-values were adjusted using Benjamini-Hochberg.

2.5 Hyperparameter Tuning

Each model implementation had a tuning step to limit ambiguous hyperparameter selection. For ease of comparison, the same cross-validation splits were used with JACA[19], MOFA[20], MultiMLP[21, 22], and SLIDE.[23] DIABLO[18] did not have a simple method to implement cross-validation splits. DIABLO and JACA had built-in functions to conduct hyperparameter selection, and additional code was written to tune selected MOFA, MultiMLP, and SLIDE hyperparameters. Code was also written to extend the JACA hyperparameter tuning to incorporate cross-validation splits using a neural net.[41] Note that each model differs significantly in its approach and the tuning variables it uses (Table 1), such as cross-entropy loss or Evidence Lower Bound (ELBO).

2.6 Top Feature Selection

Each model yields one or more components, which are lower-dimensional representations of the high dimensional multi-omics data that capture the underlying structure. In cases where multiple components were returned for a specific integration model, the component with the best group separation was determined using the tidymodels[42] implementation of the neural net nnet[41] R package. Selected hyperparameters (the number of hidden units, the weight decay penalty, and the number of epochs) for the neural net (NN) were tuned with 3-cross fold validation repeated 5

times. The most important component was then determined with Shapley[43] values as calculated by the fastshap[44] R package.

Feature weights per top scoring component (if there were multiple) or the single component were then extracted for DIABLO[18], JACA[19], MOFA[20], and SLIDE.[23] Shapley values were calculated for MultiMLP[21, 22] using the shap[43] python library. Absolute weight values per feature had their algorithmic "knee", the point of diminishing returns in feature importance scores, detected by the kneedle[45] algorithm. All features above this threshold were considered "top features" per integration model and were subsequently compared for their omics type (e.g. 16S, metaproteomics, etc.) and shared feature importances.

Finally, to evaluate the similarity of model results, absolute weights from each model were extracted and a non-parametric Spearman correlation was calculated for all pairs of models. Euclidean distances between correlations were calculated and clustered with hierarchical clustering.

2.7 Top Feature Class Identification

Meta information on biomolecules was collected, specifically metabolite class information as determined by CANOPUS[46], protein function information as determined by COG[47, 48], and the bacteria which map to each protein, as previously described. Ranks were normalized by equation 1,

$$1 - \frac{1 - r}{\max(R)}$$

(Equation 1)

where "r" represents a specific rank in a list of ranks "R" for an integration model. Thus, if there were 5 features, ranks 1 through 5 would be 1, 0.8, 0.6, 0.4, and 0.2, respectively. For comparison purposes, average ranks per feature class were calculated, and feature classes sorted by highest to lowest rank in a plot. In cases where a feature class was not represented by an integration model, it was assigned a rank of 0. Counts of the number of features per class were also plotted.

3.0 Results

3.1 Omics Data Properties and the Impacts of Filtering Decisions

Before imputation, unknown metabolites were filtered out, removing 12,269 (81%) and 6,365 (78%) of the features in the positive and negative ion mode, respectively. Then, omics data was filtered to ensure enough feature representation per sample with at least 2 samples for the 0 week group (n = 3) and at least 5 samples for post-0 week group (weeks 4-12, n = 9). This resulted in the removal of 63.5%, 1.62%, and 4.83% of the metaproteomics, metabolomics positive, and metabolomics negative datasets post-unknown filtering, respectively (Fig. 1a). Pre-imputation, correlation heatmaps for the 16S and metabolomics negative datasets show a strong correlation within 0 timepoint samples, and within post-0 timepoint samples (Fig. 1b). Post-imputation, the separation of groups is clear for all omics views, even though group information is not explicitly specified in expectation maximization imputation (Fig. 1c). Note that 16S data had no missingness and thus did not have any imputation.

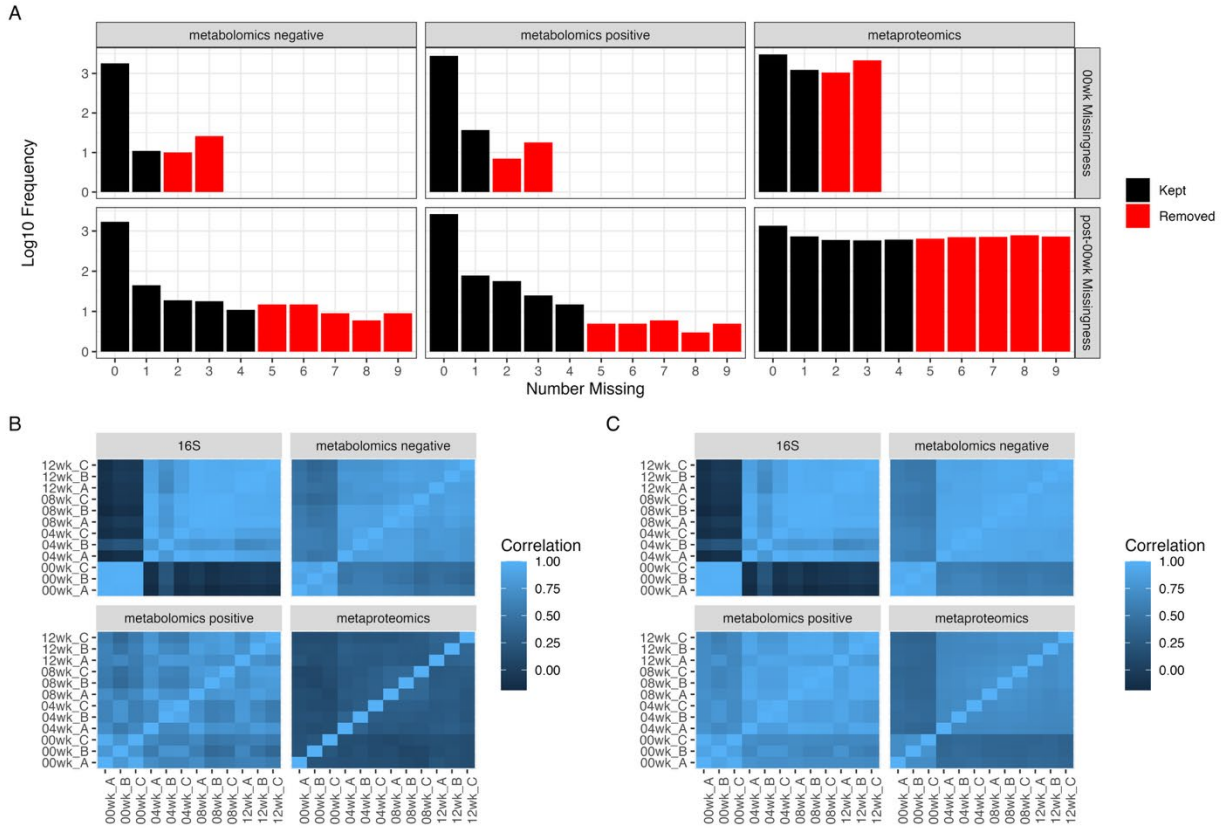

Figure 1. (A) Log10 counts of features impacted by the missingness filter for metabolomics negative (left), metabolomics positive (middle), and metaproteomics (right) data. Log10 is used for visualization purposes. Black indicates the feature was maintained and red indicates a feature was filtered out. 16S data had no missingness. (B) Pearson correlation matrices of each omic before and (C) after expectation maximization imputation. Note that 16S data was not imputed but included in the plot for comparison purposes.

3.2 Selection of Hyperparameters, Components, and Top Features

Table 2. A description of integration models and selected hyperparameters.

Integration Model	Hyperparameter Selection	Model Used
DIABLO[18]	Number of components: 1	diablo::tune.block.splsda()[18]
JACA[19]	alpha (weight between LDA and CCA): 0.5 $lambda_d$ (regularization for sparsity level in each weight matrix): 0.1 rho (shrinkage for elastic net): 0.01	jaca::jacaTrain()[19] Custom code[49] using tidymodels[42] and nnet::nnet()[41]
MOFA[20]	Number of components: 5	Custom code[49] using MOFA2::run_mofa()[20]
MultiMLP[21, 22]	Grid size: 2x original data size	Custom code[49]
SLIDE[23]	A structure matrix for the learned latent components, describing the extent that each component represents individual or multiple omics	Custom code[49]

Hyperparameter selection is summarized in Table 2. Briefly, default functions in packages were used wherever possible, and custom code was used to select hyperparameters for JACA[19], MOFA[20], MultiMLP[21, 22], and SLIDE.[23] After the selection of hyperparameters, integration models were run on the multi-omics dataset, and top components were selected for the only model that returned multiple components (i.e. MOFA). To determine the top component, a neural net[41] was fit to MOFA, and the component with the highest Shapley value returned (Component 1, with a Shapley value of 0.162, while all other values were below 0.001). There was clear separation between the 0 week and post-0 week groups for all models with components (DIABLO, MOFA, SLIDE). Interestingly, MOFA and SLIDE have some separation by the specific week the sample was collected (0, 4, 8, and 12 weeks), even though the integration model was not supplied that information.

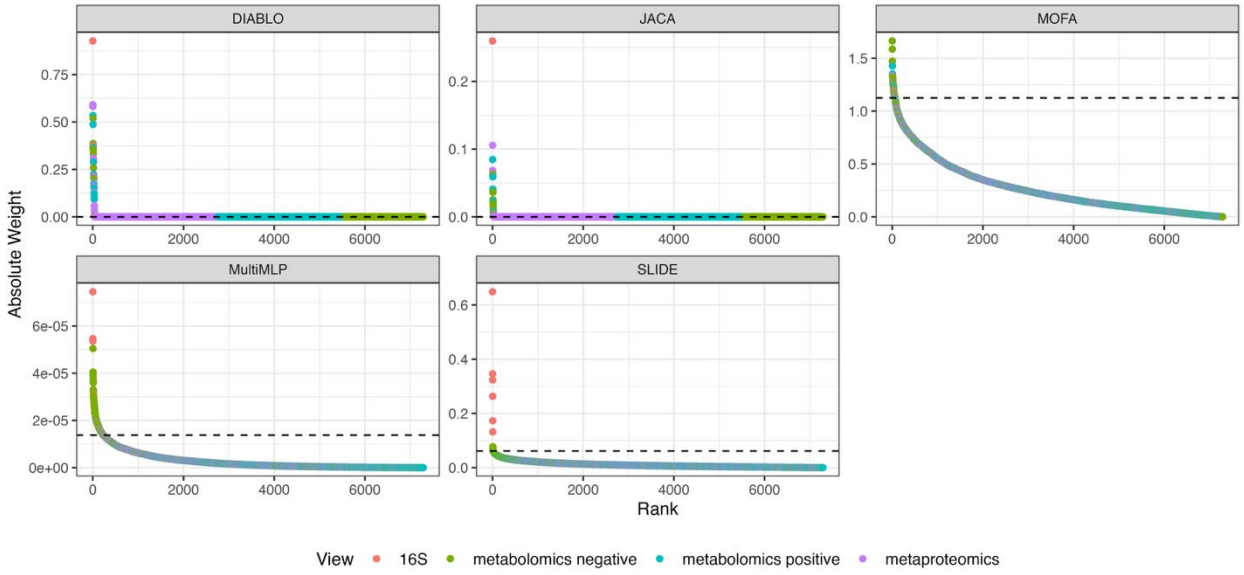

Figure 2. Feature absolute weights per integration model, with a cutoff determined by the kneedle algorithm (dashed line). Features are colored by omic type: 16S (red), metabolomics negative (light green), metabolomics positive (light blue), and metaproteomics (purple).

3.3 Similarity of Top Features Across Integration Models

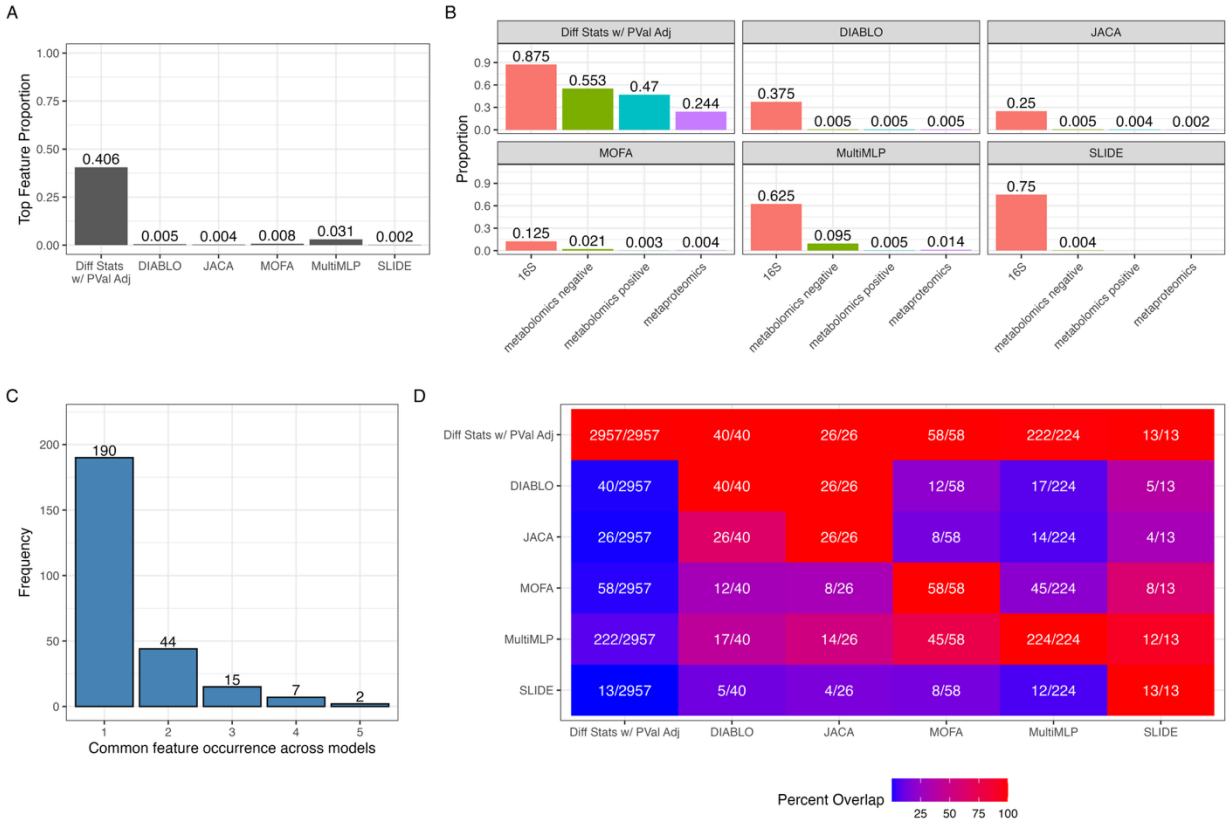

Figure 3. (A) Proportion of all possible features considered a “top feature” using differential statistics with a p-value cut-off of 0.05, and the integration models after kneedle top

feature selection. (B) Proportions from A split by omics type. (C) Count of the number of top features shared across the five integration models. (D) Shared features between integration models and differential expression. The numerator is the number of overlapping features between the models in the columns and rows, and the denominator in the number of features in the model listed at the bottom of the column.

As a base line, differential expression[3] and differential abundance[5, 6] statistics were calculated, and significant biomolecules ($\alpha \leq 0.05$) were concatenated together, resulting in 2957 potential top features (40.6% of the original dataset) to explain the difference between the control (0 week timepoint) and experimental (all other timepoints) conditions (Figure 3a). Each integration model returned far fewer top features, as MultiMLP had the largest count (224, 3.1% of the original dataset) and SLIDE had the lowest (13, 0.2%). Thus, each multi-omics integration model delivered on one of the goals of their design: to provide a smaller list of top targets explaining biological conditions than simply concatenating together significant biomolecules across omics datasets.

Four integration models (DIABLO, JACA, MOFA, MultiMLP) returned at least one feature from each omic dataset (view), while SLIDE only returned features from 16S and metabolomics negative (Figure 3b). The total number of possible features (p) per omic was 8, 1752, 2726, and 2800 for 16S, metabolomics negative, metaproteomics, and metabolomics positive, respectively; thus, high proportions of 16S data would reflect only a few features. Detailed counts of the number of features per omic are provided in Table 3. In terms of shared features across models, 44 features are shared across at least 2 models, 15 across 3, 7 across 4, and 2 across all 5 (Figure 3c). The 2 shared features across all models are a carboxylic acid and an amino acid with no further identification, making the exact shared interpretation of this feature difficult.

Table 3. Number of features returned per integration model per omic (view). Differential expressed/abundant biomolecules are included for comparison.

Integration Model	**16S**	**Metabolomics Negative**	**Metabolomics Positive**	**Metaproteomics**
DIABLO	3	9	14	14
JACA	2	9	10	5
MOFA	1	37	9	11
MultiMLP	5	167	13	39
SLIDE	6	7	0	0

Interestingly, all multi-omics integration models identified a subset of biomolecules that overlapped with those determined as significant through differential statistics, except for MultiMLP which selected two features with relatively close to significant values (*p*-values of 0.0520 and 0.0705) (Fig. 3d). Importantly, none of these integration models explicitly incorporate differential statistics; yet, the observed overlap provides validation for the

reliability of these models. JACA and DIABLO had significant overlap, where 100% of the features within JACA were also in DIABLO (Fig. 3d). To account for mismatched vector sizes when calculating overlap, we used the smaller vector size as the denominator in proportional comparisons. Based on this approach, MOFA and DIABLO exhibited the least overlap, sharing only 30% of their features, and the average proportion of shared features was 55.7%. Though the overlap was promising, it should not be missed that 190 features were not shared across models; thus, the selection of one individual integration model may lead to different downstream interpretations of the results.

Models were then clustered together by the Euclidean distances of the Spearman correlations of their absolute feature weights. DIABLO and JACA clustered together, which is not surprising given their shared top features (Figure 3) and how they assign weights of 0 to lowly important features (Figure 2). SLIDE, MOFA, and MultiMLP do not assign 0 weights (Figure 2), and formed a second cluster together, with a slight deviation from MultiMLP, which may be due to its multi-layer perceptron approach as opposed to a decomposition or factor analysis approach (Table 1).

3.4 Top Feature Exploration

Top feature ranks were then scaled (Equation 1) so that the highest-ranking features would be at or near 1, and the lowest top ranked features would be near 0. Any features that were not in the top rank were given a rank of 0. Average ranks were calculated for each feature, and any features present in at least 2 models or without a specific identification (e.g. a class of proteins as opposed to a specific protein) were removed, resulting in 22 features (Figure 4). The top two ranked features were *Sphingopyxis* and *Streptomyces*, which both have crucial rules in the initial breakdown (*Streptomyces, Neorhizobium*) and continued breakdown (*Sphingopyxis, Dyadobacter,* and *Variovorax*) of chitin.[26] The other bacteria of the eight species cohort, *Ensifer*, *Rhodococcus,* and *Sinorhizobium*, though important to the interspecies degradation of chitin, play less central and more specialized roles, as supported in the original study[26], and were not in the top features. Interestingly, the importance of these five bacteria was further emphasized in the protein features, where the mean scaled rank of top proteins in at least two models was, in order, *Sphingopyxis*, *Streptomyces, Variovorax, Dyadobacter,* and *Neorhizobium.*

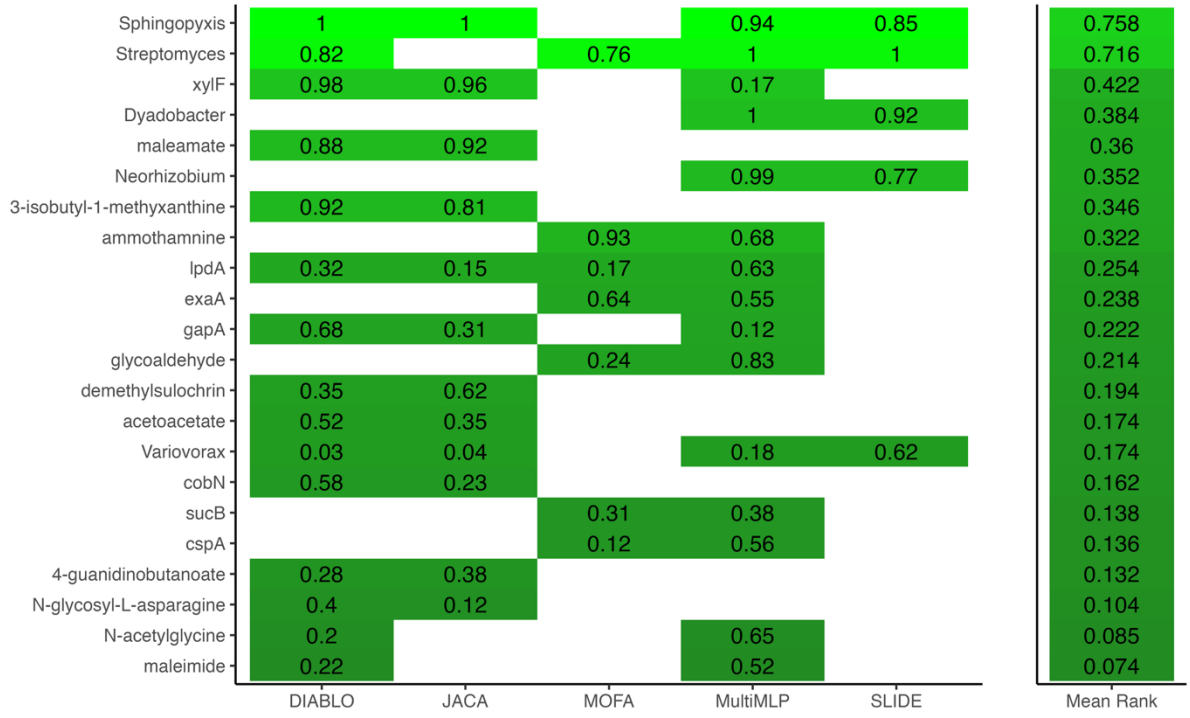

Figure 4. Scaled ranks of top features, ordered from highest to lowest mean scaled rank. Features present in only one integration model or features without a specific identification (e.g. a class of metabolites as opposed to a specific metabolite) were removed.

Though there are similar patterns between the highest ranked bacteria in both the 16S and metaproteomics, none of the top ranked proteins (xylF, lpdA, exaA, gapA, cobN, sucB, and cspA) exhibit any direct roles in the breakdown of chitin, suggesting that other pathways and processes of these species may respond more strongly to time than those that are involved in chitin degradation. Several *Streptomyces* proteins (xylF, gapA, cobN, sucB) had decreased abundance while exaA (*Variovorax*) and lpdA (*Dyadobacter*) had increased abundance, aligning with the overall changes in microbe populations as determined by the 16S data. Interestingly, the cold shock cspA protein (*Neorhizobium*) had a strong decrease in expression, likely because the samples were pre-incubated at 5°C and later incubated at 20°C, removing the need for cold shock proteins.

The metabolomics data had no specific ties to species and instead represent the abundances across all species, making their direct interpretation difficult. Ten specific metabolites were identified by at least two integration models. None had any direct ties to chitin metabolism. Some connections exist between top proteins and metabolites, such as acetoacetate which can be converted into acetoacetyl-CoA and enter the citric acid cycle (lpdA and sucB) and is involved in glycolysis (gapA). Interestingly, acetoacetate had decreased abundances, which could indicate a continued depletion as it feeds into the citric acid cycle or as it is converted into lipids and certain peptides. Although neither the metaproteomics nor the metabolomics data had a biomolecule associated with chitin metabolism as a top feature, several chitin metabolism features

were still used to separate groups, such as glmS and chitobiose in MOFA, MultiMLP, and SLIDE that had weights near the cutoff. The goal of each multi-omics integration model is to return an optimal set of all features, so exclusion from the “top features” does not indicate that the feature contains no value. Instead, it suggests a stronger signal to separate groups. A particular advantage of multi-omics integration models is that they allow for the identification of features with complementary information, such as the shared expression profiles of microbes in 16S and metaproteomics data, and the metabolic relationships between some top proteins and metabolites.

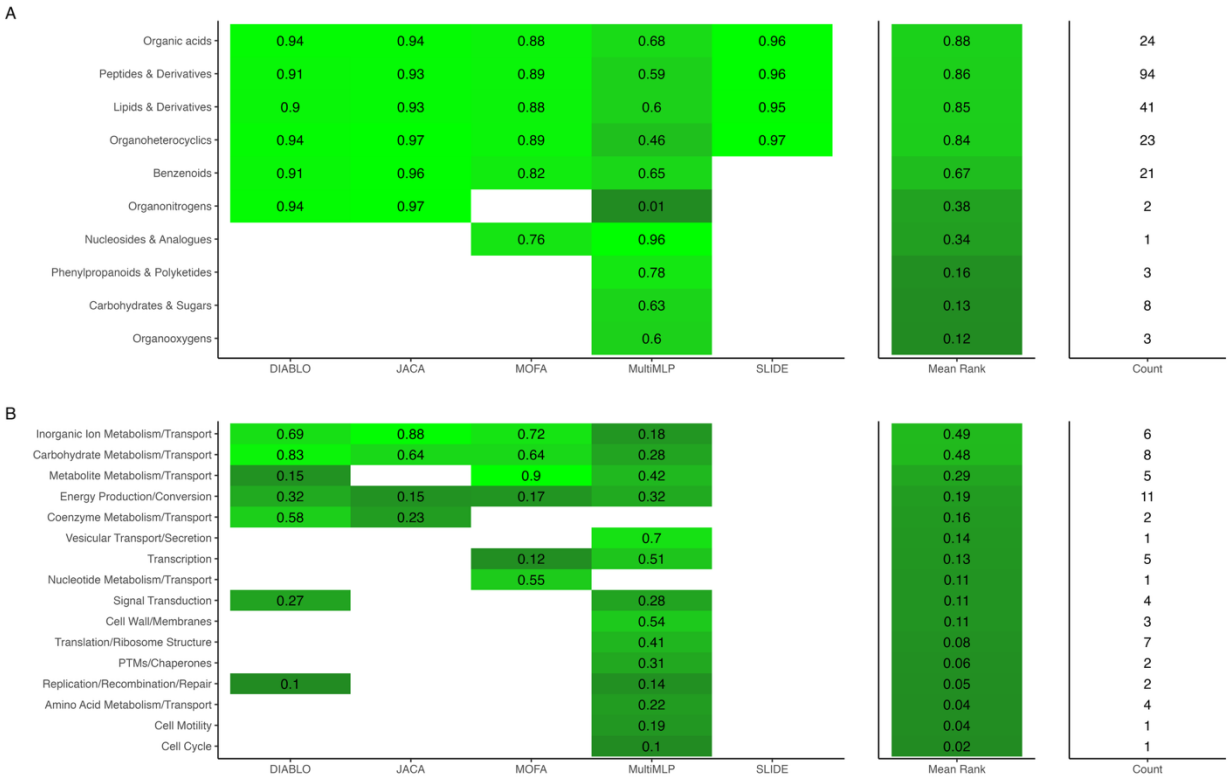

Figure 5. Average scaled ranks of feature information, ordered from highest to lowest scaled mean rank across integration models for (A) CANOPUS categories of metabolomics data (combining both negative and positive ion mode), and (B) COG categories of protein functions. Mean ranks are reported, as long as the number of features in each category.

To detect any patterns at biomolecule category levels (e.g. CANOPUS[46] categories for metabolites and COG[47, 48] categories for protein data), all top features were averaged at the category level, with reports of the number of biomolecules detected in each category. All integration models ranked organic acids, peptides & derivatives, lipids & derivatives, and organoheterocyclics as important classes of biomolecules for separating conditions (Figure 5). Differential abundance of these biomolecules shows an increased abundance, potentially indicating a focus on storage (as opposed to catabolism) of lipids as the main energy source (chitin) depletes. This may also explain the decreased abundance of acetoacetate, which could be stored in these lipids. The focus on anabolism of these specific energy-storing compounds may also be supported by protein COGs, where there

is a heavy focus on carbohydrate metabolism and energy conversion. These results demonstrate that incorporating additional meta information about features reveals additional information about the biological system of study.

4.0 Discussion

Each of the selected models has been demonstrated as a useful and informative approach for multi-omics analysis[18-23]; yet, on their own, no model provides a cohesive view of the most important features distinguishing conditions. For example, no model ranked all top bacteria in the 16S data. But, by scaling and averaging ranks across models, all the most important bacteria were seen in both the 16S and metaproteomics data. This demonstrates the value in utilizing several integration models, as opposed to just one, to get a more complete view of the top features, especially in cases where there are thousands of potential features. This work also shows the reliability of each of these approaches, as all models selected a subset of significant or near significant features without that information explicitly provided, and all models provided some top features which match previous work.[26] Limitations to this study include a small and imbalanced sample size (n = 3 for the control group, and n = 9 for experimental group), non-specific identifications in the metaproteomics and metabolomics data, high levels of missingness in the metaproteomics, and an inability to tie metabolomics to specific bacteria(essentially a “meta”-metabolomics approach). Future work in this area could focus on ensemble approaches and comparisons on larger datasets.

5.0 References

1. Overy, D. P.; Bell, M. A.; Habtewold, J.; Helgason, B. L.; Gregorich, E. G. “Omics” technologies for the study of soil carbon stabilization: a review. *Front. Environ. Sci.* **2021**, *9*, 617952. DOI: 10.3389/fenvs.2021.617952
2. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11* (10), R106. DOI: 10.1186/gb-2010-11-10-r106
3. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinform.* **2010**, *26* (1), 139-140. DOI: 10.1093/bioinformatics/btp616
4. Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **2015**, *43* (7), e47. DOI: 10.1093/nar/gkv007
5. Degnan, D. J.; Stratton, K. G.; Richardson, R.; Claborne, D.; Martin, E. A.; Johnson, N. A.; Leach, D.; Webb-Robertson, B. M.; Bramer, L. M. pmartR 2.0: A Quality Control, Visualization, and Statistics Pipeline for Multiple Omics Datatypes. *J. Proteome Res.* **2023**, *22* (2), 570-576. DOI: 10.1021/acs.jproteome.2c00610
6. Stratton, K. G.; Webb-Robertson, B. M.; McCue, L. A.; Stanfill, B.; Claborne, D.; Godinez, I.; Johansen, T.; Thompson, A. M.; Burnum-Johnson, K. E.; Waters, K. M.; et al. pmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *J. Proteome Res.* **2019**, *18* (3), 1418-1425. DOI: 10.1021/acs.jproteome.8b00760
7. McMurdie, P. J.; Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **2013**, *8* (4), e61217. DOI: 10.1371/journal.pone.0061217
8. Paulson, J. N.; Pop, M.; Bravo, H. C. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *R package version 1.50.0,* **2013.** https://www.bioconductor.org/packages/release/bioc/html/metagenomeSeq.html (accessed October 1, 2025).
9. Mandal, S.; Van Treuren, W.; White, R. A.; Eggesbø, M.; Knight, R.; Peddada, S. D. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **2015**, *26* (1), 27663. DOI: 10.3402/mehd.v26.27663
10. Li, Y.; Mansmann, U.; Du, S.; Hornung, R. Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinformatics* **2022**, *23* (1), 412. DOI: 10.1186/s12859-022-04962-x
11. Urkullu, A.; Pérez, A.; Calvo, B. Are the statistical tests the best way to deal with the biomarker selection problem? *Knowl. Inf. Systems* **2022**, *64* (6), 1549-1570. DOI: 10.1007/s10115-022-01677-6
12. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szczesniak, M. W.; Gaffney, D. J.; Elo, L. L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13. DOI: 10.1186/s13059-016-0881-8
13. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198-207. DOI: 10.1038/nature01511

14. Kammers, K.; Cole, R. N.; Tiengwe, C.; Ruczinski, I. Detecting Significant Changes in Protein Abundance. *EuPA Open Proteom.* **2015**, *7*, 11-19. DOI: 10.1016/j.euprot.2015.02.002
15. Love, M. I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15* (12), 550. DOI: 10.1186/s13059-014-0550-8
16. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15* (4), 1116-1125. DOI: 10.1021/acs.jproteome.5b00981
17. Silverman, J. D.; Roche, K.; Mukherjee, S.; David, L. A. Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2789-2798. DOI: 10.1016/j.csbj.2020.09.014
18. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13* (11), e1005752. DOI: 10.1371/journal.pcbi.1005752
19. Zhang, Y.; Gaynanova, I. Joint association and classification analysis of multi-view data. *Biometrics* **2022**, *78* (4), 1614-1625. DOI: 10.1111/biom.13536
20. Argelaguet, R.; Arnol, D.; Bredikhin, D.; Deloro, Y.; Velten, B.; Marioni, J. C.; Stegle, O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **2020**, *21* (1), 111. DOI: 10.1186/s13059-020-02015-1
21. Lee, C.; Van der Schaar, M. A variational information bottleneck approach to multi-omics data integration. *AISTATS*, **2021**, 1513-1521.
22. Claborne, D.; Flores, J.; Erwin, S.; Durell, L.; Richardson, R.; Fore, R.; Bramer, L. Consistency of Feature Attribution in Deep Learning Architectures for Multi-Omics. *arXiv* **2025.** DOI: 10.48850/arXiv.2507.22877
23. Gaynanova, I.; Li, G. Structural learning and integrative decomposition of multi-view data. *Biometrics* **2019**, *75* (4), 1121-1132. DOI: 10.1111/biom.13108
24. Flores, J. E.; Claborne, D. M.; Weller, Z. D.; Webb-Robertson, B. M.; Waters, K. M.; Bramer, L. M. Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front. Artif. Intell.* **2023**, *6*, 1098308. DOI: 10.3389/frai.2023.1098308
25. Ritchie, M. D.; Holzinger, E. R.; Li, R.; Pendergrass, S. A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **2015**, *16* (2), 85-97. DOI: 10.1038/nrg3868
26. McClure, R.; Rivas-Ubach, A.; Hixson, K. K.; Farris, Y.; Garcia, M.; Danczak, R.; Davison, M.; Paurus, V. L.; Jansson, J. K. Multi-omics of a model bacterial consortium deciphers details of chitin decomposition in soil. *mBio* **2025**, *16* (7), e0040425. DOI: 10.1128/mbio.00404-25
27. McClure, R.; Farris, Y.; Danczak, R.; Nelson, W.; Song, H. S.; Kessell, A.; Lee, J. Y.; Couvillion, S.; Henry, C.; Jansson, J. K.; et al. Interaction Networks Are Driven by Community-Responsive Phenotypes in a Chitin-Degrading Consortium of Soil Microbes. *mSystems* **2022**, *7* (5), e0037222. DOI: 10.1128/msystems.00372-22
28. Bolyen, E.; Rideout, J. R.; Dillon, M. R.; Bokulich, N. A.; Abnet, C. C.; Al-Ghalith, G. A.; Alexander, H.; Alm, E. J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37* (8), 852-857. DOI: 10.1038/s41587-019-0209-9

29. Verberkmoes, N. C.; Russell, A. L.; Shah, M.; Godzik, A.; Rosenquist, M.; Halfvarson, J.; Lefsrud, M. G.; Apajalahti, J.; Tysk, C.; Hettich, R. L.; et al. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **2009**, *3* (2), 179-189. DOI: 10.1038/ismej.2008.108

30. Yilmaz, P.; Parfrey, L. W.; Yarza, P.; Gerken, J.; Pruesse, E.; Quast, C.; Schweer, T.; Peplies, J.; Ludwig, W.; Glockner, F. O. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **2014**, *42*, D643-648. DOI: 10.1093/nar/gkt1209

31. Nakayasu, E. S.; Nicora, C. D.; Sims, A. C.; Burnum-Johnson, K. E.; Kim, Y. M.; Kyle, J. E.; Matzke, M. M.; Shukla, A. K.; Chu, R. K.; Schepmoes, A. A.; et al. MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses. *mSystems* **2016**, *1* (3). DOI: 10.1128/mSystems.00043-16

32. Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol.* **2023**, *41* (4), 447-449. DOI: 10.1038/s41587-023-01690-2

33. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16* (4), 299-302. DOI: 10.1038/s41592-019-0344-8

34. Schum, S. K.; Brown, L. E.; Mazzoleni, L. R. MFAssignR: Molecular formula assignment software for ultrahigh resolution mass spectrometry analysis of environmental complex mixtures. *Environ. Res.* **2020**, *191*, 110114. DOI: 10.1016/j.envres.2020.110114

35. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301-2319. DOI: 10.1038/nprot.2016.136

36. Chen, L. S.; Prentice, R. L.; Wang, P. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **2014**, *70* (2), 312-322. DOI: 10.1111/biom.12149

37. Afanador, N.; Tran, T.; Blanchet, L.; Baumgartner, R. Mvdalab: Multivariate data analysis laboratory. *R package version 1.7,* **2022.** https://cran.r-project.org/web/packages/mvdalab/index.html (accessed October 1, 2025).

38. Bramer, L. M.; Irvahn, J.; Piehowski, P. D.; Rodland, K. D.; Webb-Robertson, B. M. A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics. *J. Proteome Res.* **2021**, *20* (1), 1-13. DOI: 10.1021/acs.jproteome.0c00123

39. Webb-Robertson, B. J.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.* **2015**, *14* (5), 1993-2001. DOI: 10.1021/pr501138

40. R Core Team. *R: A language and environment for statistical computing;* R Foundation for Statistical Computing: Vienna, Austria, **2013.** https://www.R-project.org

41. Venables, W. N.; Ripley, B. D. *Modern applied statistics with S*; Springer Science & Business Media, **2013**.

42. Kuhn, M.; Silge, J. *Tidy modeling with R: A framework for modeling in the tidyverse*; O'Reilly Media, Inc., **2022**.
43. Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *NeurIPS* **2017**, *30*, 4768-4777. DOI: 10.5555/3295222.3295230
44. Greenwell, B. fastshap: fast approximate Shapley values. *R package version 0.1.1,* **2019.** https://cran.r-project.org/web/packages/fastshap/index.html (accessed October 1, 2025).
45. Satopaa, V.; Albrecht, J.; Irwin, D.; Raghavan, B. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. *ICDCSW* **2011,** 166-171. DOI: 10.1109/ICDCSW.2011.20
46. Duhrkop, K.; Nothias, L. F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **2021**, *39* (4), 462-471. DOI: 10.1038/s41587-020-0740-8
47. Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. A genomic perspective on protein families. *Science* **1997**, *278* (5338), 631-637. DOI: 10.1126/science.278.5338.631
48. Galperin, M. Y.; Vera Alvarez, R.; Karamycheva, S.; Makarova, K. S.; Wolf, Y. I.; Landsman, D.; Koonin, E. V. COG database update 2024. *Nucleic Acids Res.* **2025**, *53* (1), D356-D363. DOI: 10.1093/nar/gkae983
49. Flores, J; Degnan, D. Multi-omics Integration Tools. *Github,* **2025.** https://github.com/pnnl/multiomics_integration_tools (accessed October 1st, 2025).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov