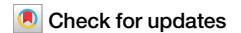


<https://doi.org/10.1038/s42003-025-08515-9>

A systematic benchmark of integrative strategies for microbiome-metabolome data



Loïc Mangnier^{1,6}, Antoine Bodein^{1,6}, Margaux Mariaz¹, Alban Mathieu¹, Marie-Pier Scott-Boyer¹, Neerja Vashist², Matthew S. Bramble^{3,4} & Arnaud Droit^{1,5} 

The rapid advancement of high-throughput sequencing technologies has enabled the integration of various omic layers into computational frameworks. Among these, metagenomics and metabolomics are increasingly studied for their roles in complex diseases. However, no standard currently exists for jointly integrating microbiome and metabolome datasets within statistical models. We benchmarked nineteen integrative methods to disentangle the relationships between microorganisms and metabolites. These methods address key research goals, including global associations, data summarization, individual associations, and feature selection. Through realistic simulations, we identified the best-performing methods and validated them on real gut microbiome datasets, revealing complementary biological processes across the two omic layers. Practical guidelines are provided for specific scientific questions and data types. This work establishes a foundation for research standards in metagenomics-metabolomics integration and supports future methodological developments, while also providing guidance for designing optimal analytical strategies tailored to specific integration questions.

The recent development of high-throughput sequencing technologies has permitted the generation of omics data at an exponential scale. Integrating these high-dimensional biological data holds great potential for elucidating the complex mechanisms underlying diverse biological systems^{1–3}. Among these, the interactions between microorganisms and metabolites have garnered growing interest due to their involvement in a wide range of biological functions. In human health, shifts in microbiome-metabolome relationships have been linked to conditions such as cardio-metabolic diseases⁴ and autism spectrum disorders⁵, potentially reflecting disruptions in microbiota-derived metabolic processes⁶. Beyond human health, microbiome-metabolome interactions also play key roles in plant productivity, soil nutrient cycling, and ecosystem resilience^{7–9}.

Microbiome data, generated through metagenomics sequencing, provides a broad overview of microbial community composition based on either functional gene content or taxonomic assignment¹⁰. However, it presents unique analytical challenges due to inherent properties such as over-dispersion, zero inflation, high collinearity between taxa, and its compositional nature^{11,12}. Proper handling of this compositionality, often through transformations like centered log-ratio (CLR) or isometric log-ratio

(ILR), is crucial for avoiding spurious results^{11–13}. Metabolomics, on the other hand, offers a comprehensive snapshot of the small molecules within a biological system, typically measured using techniques like LC-MS¹⁴. Similar to microbiome data, metabolomic profiles often exhibit over-dispersion and complex correlation structures. Integrating metabolomics data with pathway information alongside other omics, such as metagenomics, is vital for identifying disease-related biomarkers and metabolic signatures, with tools like MetaboAnalyst facilitating such analyses¹⁵.

The integration of microbiome and metabolome data into statistical frameworks provides key insights into how diet and lifestyle shape species-metabolite interactions. It also brings its own challenges¹⁶. The multiplicity of new available statistical models for the integration makes the choice of the right method a daunting task for many researchers as well as determining the most appropriate approach for the analysis.

Typical research questions include either the inference of interactions occurring between species and metabolites or in predicting one omic layer borrowing the biological information contained in the other omic layers^{17,18}. The inference of interactions focuses on understanding and identifying the complex relationships between microbial communities and the metabolites

¹Axe Endo-Nephro, Centre de recherche du CHU de Québec-Université Laval, Québec, QC, Canada. ²Department of Pathology and Laboratory Medicine, UCLA, Los Angeles, USA. ³Center for Genetic Medicine Research, Children's Research Institute, Children's National Hospital, Washington, DC, USA. ⁴Department of Genomics and Precision Medicine, The George Washington University of Medicine and Health Sciences, Washington, DC, USA. ⁵Département de médecine moléculaire, Faculté de médecine, Université Laval, Québec, QC, Canada. ⁶These authors contributed equally: Loïc Mangnier, Antoine Bodein.

 e-mail: Arnaud.Droit@crchudequebec.ulaval.ca

they influence. It aims to uncover causal dependencies and mechanisms underlying microbial-metabolite interactions in biological systems¹⁹. While advanced machine learning and deep learning methods are increasingly used for predictive tasks, they often lack the mechanistic interpretability needed for generating novel biological hypotheses and thus will not be the primary focus of this paper.

To elucidate the complex entanglement between microorganisms and metabolites, several strategies can be exploited, each exhibiting associations occurring at different scales. Consistent with a recent report from²⁰, traditional workflows include distinct types of analysis, addressing complementary biological questions², including the detection of global associations, data summarization, individual associations, and identification of the most relevant associated features across datasets.

Firstly, researchers often aim to determine the presence of an overall association between the two omic datasets. Multivariate methods like Procrustes analysis²¹, the Mantel test²², and MMiRKAT²³ are commonly employed for this purpose. This global assessment often comes before more specific analyses such as data summarization or the identification of the most relevant associated features across datasets².

Subsequently, mirroring approaches used in single-omics studies, summarizing the information within each dataset is a common objective, facilitating visualization and interpretation¹. The presence of two distinct omic layers allows for the exploration of both intra- and inter-correlations between their features. Data summarization methods like canonical correlation analysis (CCA)²⁴, Partial Least Squares (PLS)²⁵, redundancy analysis (RDA)²⁶, and MOFA2²⁷ help identify the most relevant associated features across datasets explaining significant data variability and have been successful in uncovering associations in diseases like Type 2 diabetes²⁸. However, these approaches often lack the resolution to pinpoint specific microorganism-metabolite relationships, which is crucial for identifying context-relevant core features.

To address this limitation, methods for detecting individual associations may prove relevant for the identification of metagenomic taxonomic features associated with diet-associated metabolites². One common strategy involves computing association measures (correlation or regression) for each metabolite-species pair. While straightforward, these methods face challenges due to the multiple testing burden. Alternative approaches include univariate or multivariate feature selection methods like LASSO²⁹, sparse CCA (sCCA)³⁰, and sparse PLS (sPLS)³¹, which aim to identify the most relevant associated features across datasets while addressing multicollinearity. However, the inherent complexities of microbiome and metabolome data can still limit the biological interpretability of results obtained from standard methods.

Thus, integrating microbiome and metabolome data requires particular attention. Approaches to deal with compositional data either as an outcome or explanatory variable have already been proposed^{12,32,33} covering applications of global association methods, data summarization, individual associations, or identification of core features. Conventional strategies involve applying standard statistical methods after appropriate transformations, such as cCLR or ILR for microbiome data, or log transformation for metabolomics data. Alternatively, purely compositional approaches, such as Dirichlet regression or balances derived from the ILR transformation, explicitly account for the compositional nature of microbiome data^{34–37}. Regarding the multiple possibilities of analysis methodologies, determining which strategy is the best, depending on the research question, remains an open problem with major implications for practitioners.

Despite recent efforts to evaluate statistical methods for joint microbiome and metabolome integration, these have often been limited to literature reviews or real-data applications lacking a known ground truth³⁸. This absence of unbiased comparisons is a significant gap in the field²⁰, hindering the establishment of best practices for result interpretability and reproducibility. For instance, while transformations like CLR and ILR are commonly used for microbiome data normalization, a clear consensus on their relative performance is lacking.

Therefore, in this paper, we present a comprehensive comparison of nineteen different integrative methods to decipher the complex relationships between microorganisms and metabolites.

Our selection of methods and analytical strategies is based on a recent review²⁰, covering common research aims such as global associations, data summarization, individual associations, and feature selection. Through extensive simulation studies, we provide valuable insights into the strengths and limitations of methods commonly used in practice. Subsequently, we apply the top-performing methods, identified based on power, robustness, and interpretability, to real gut microbiome and metabolome data from Konzo disease^{39,40}, revealing a complex multi-scale architecture between the two omic layers. Specifically, the best methods were selected based on their performance in: (i) global associations-detecting significant overall correlations while controlling false positives; (ii) data summarization-capturing and explaining shared variance; (iii) individual associations-detecting meaningful pairwise specie-metabolite relationships with high sensitivity and specificity; and (iv) feature selection-identifying stable and non-redundant most relevant associated features across datasets. Finally, we offer general and specific recommendations for researchers based on their data and research objectives, and we highlight potential avenues for future methodological development. A comprehensive user guide with all associated code is provided to facilitate the application of these methods in diverse contexts, promoting scientific replicability and reproducibility.

Methods

In this section, we describe the data simulation process, the methods tested for the four key analytical questions, and the evaluation metrics used to assess their performance.

Simulation setups

Microbiome and metabolome data were simulated using the *Normal to Anything* (NORTA) algorithm, which allows for generating data with arbitrary marginal distributions and correlation structures^{41–43}. Three real microbiome-metabolome datasets were used as templates:

- **Konzo dataset**³⁹: This high-dimensional dataset comprises 171 samples, 1,098 taxa, and 1,340 metabolites, collected from individuals affected by Konzo and matched controls. Microbiome data follow a negative binomial distribution, and metabolome data follow a Poisson distribution.
- **Adenomas dataset**⁴⁴: This intermediate-size dataset includes 240 samples, 500 taxa, and 463 metabolites, from patients with advanced adenomas and matched controls. Microbiome data exhibit zero-inflated negative binomial distributions, and metabolome data follow a log-normal distribution.
- **Autism spectrum disorder dataset**⁴⁵: This small dataset consists of 44 samples, 322 microbial taxa, and 61 metabolites, from children with autism spectrum disorder and neurotypical controls. Microbiome data exhibit zero-inflated negative binomial structures, and metabolome data follow a Poisson distribution.

To estimate the marginal distributions and correlation structures used in the simulations, we pooled all samples from each dataset regardless of study group, without explicitly modeling group-specific effects.

Correlation networks for species and metabolites were estimated using SpiecEasi⁴⁶, and normal distributions were converted into correlated distributions matching the original data structures. Simulations accounted for microbiome transformations (CLR, ILR, and alpha) and evaluated their impact on method performance (Table S1).

To assess Type-I error control, null datasets with no associations were generated. For alternative scenarios, we varied the number and strength of associations between microorganisms and metabolites. Methods were tested under three realistic scenarios with varying sample sizes, feature numbers, and data structures, with 1000 replicates per scenario. Simulation details and additional scenarios are provided in the supplementary, and key steps are illustrated in Fig. 1.

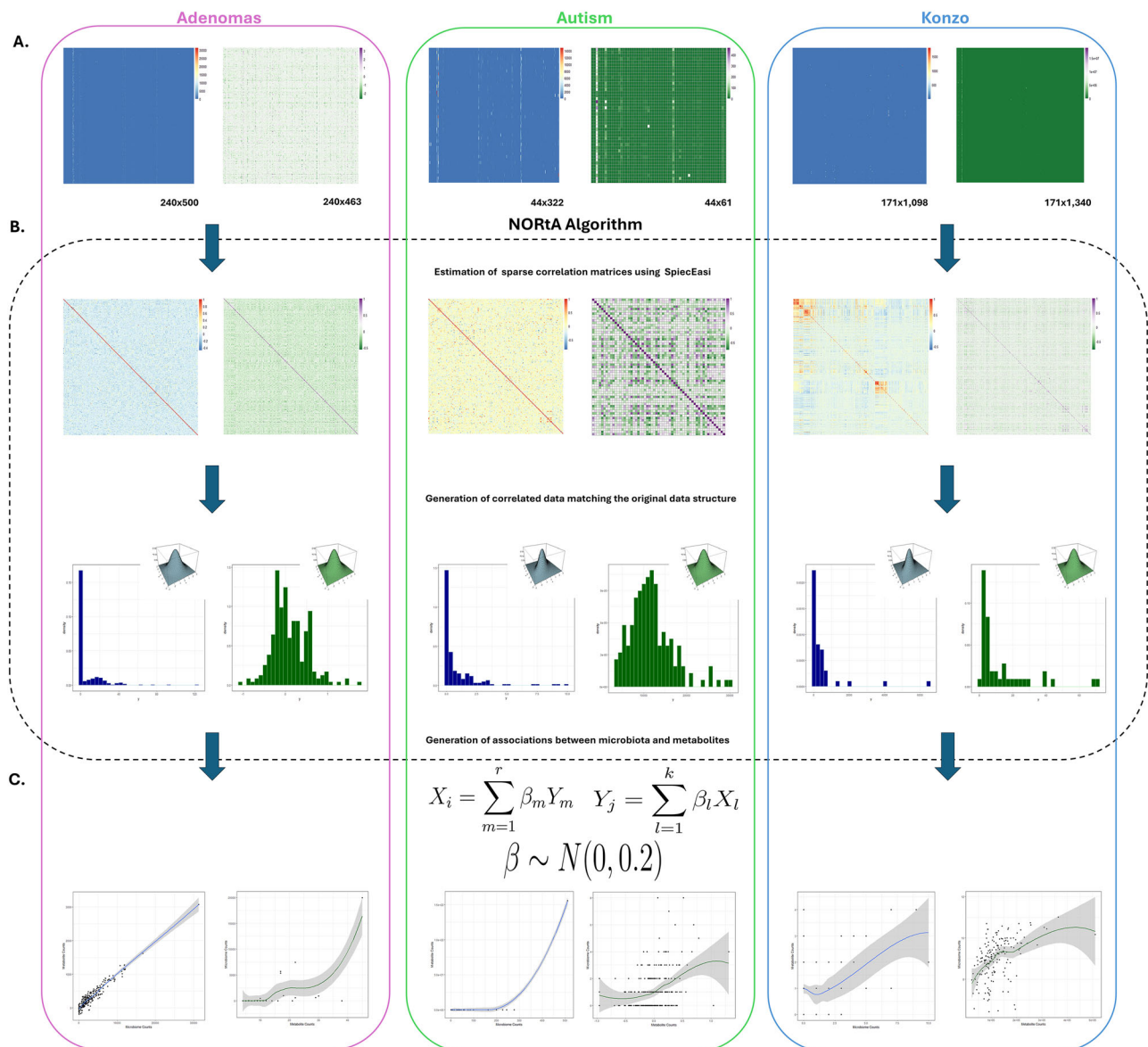


Fig. 1 | Overview of the simulation setup based on real datasets. A Three microbiome-metabolome datasets were selected, each exhibiting different data structures and correlations. We reported the sample size (N) and the number of features (P), as $N \times P$, for each dataset. **B** Realistic datasets were simulated using the “Normal-to-Anything” (NORTA) framework. First, we estimated sparse microbiome and metabolome correlation networks using SpiecEasi. Second, correlated multivariate Gaussian distributions were generated for both microbiome and metabolome

datasets using the correlation structures estimated in the previous step. Third, Gaussian distributions were converted into arbitrary distributions matching the original data structures. **C** Associations between species and metabolites were specified, mimicking the complex entanglement between the two omic layers. For each dataset, proportions of associated features vary between 1% and 10%, with association strengths randomly picked from a Gaussian distribution.

Additional scenarios were generated for methods requiring specific assumptions, detailed in the supplementary material “Additional simulation settings” section.

Statistical analyses

Let’s assume X and Y , a matrix of microbiome and metabolome, collected on the same set of samples, of size $n \times p$ and $n \times q$, where n is the number of samples, p the number of metagenomic features (i.e., species level assignment in our scenario), and q the number of metabolites, respectively. X_{ij} represents the j th microorganism in the i th sample, with $j = 1, 2, \dots, p$, while Y_{ik} is the k th metabolite in the i th sample, where $k = 1, 2, \dots, q$. For the sake of simplicity, we considered the case where $p = q$.

Prior to applying the methods, microbiome data were transformed using compositional data approaches, which are detailed in section 2.3.

The following sections describe the different categories of methods benchmarked in this study: (i) global association methods; (ii) data summarization methods; (iii) individual association methods; and (iv) feature selection methods. Since these methods rely on different statistical principles, their outputs are not directly comparable. We then used distinct performance metrics tailored to each category, which are described in their respective sections.

Data and distance kernel transformation

Most methods selected below require a normalization step or a distance-based transformation to handle compositional or over-dispersed data. We applied three microbiome normalizations and one metabolome transformation, selected based on the research objective.

To account for microbiome compositionality while preserving feature count, we used the CLR transformation^{47,48}, computed as:

$$CLR(X_j) = \log\left(\frac{X_j}{g(X)}\right)$$

where $g(X)$ is the geometric mean across microorganisms in a sample. CLR maintains interpretability and was used as the reference normalization for individual associations and feature selection. However, it does not ensure feature independence or sub-compositional coherence, limiting its suitability for distance-based methods.

For global associations and data summarization, we applied the ILR³⁵ and alpha transformations³⁴, which project data into an independent $D - 1$ quasi-orthogonal space. ILR is defined as:

$$ILR(X_j) = \sqrt{\frac{j}{j+1}} \log\left(\frac{\prod_{j=1}^{D-1} X_j}{X_j + 1}\right)$$

while the alpha transformation is a Box-Cox type transformation ensuring multivariate normality³⁴.

While we considered different normalization strategies for microbiome data, metabolome normalization can also significantly impact data analysis results. Various normalization approaches exist, including probabilistic quotient normalization, median fold change normalization, and total area normalization, each addressing specific biases in metabolomic measurements⁴⁹. However, in practice, log-transformation remains the most commonly used approach for metabolomic data, as it stabilizes variance and improves normality^{50–52}. Consequently and since metabolome data are log-normally distributed¹³, our evaluation focused on assessing the impact of applying or not applying a natural log-transformation on metabolomic data.

Additionally, for global and individual association analyses, we considered Euclidean, Canberra, and Manhattan distances on original and log-transformed metabolome data, and Euclidean distance on microbiome data. Notably, Euclidean distance on CLR-transformed data corresponds to Aitchison distance, which has been shown to outperform Bray–Curtis dissimilarity in stability and interpretability¹¹.

All transformations and distance kernels used per method are summarized in Table S1.

Global associations

We define global association methods as statistical approaches assessing overall correlations between microbiome and metabolome data. Following Deek et al. 2024, we evaluated the Mantel test²², MMiRKAT²³, and Procrustes Analysis²¹, using various data transformations and distance kernels.

Selected methods. The Mantel test²² measures global correlation between two datasets from the same samples, using distance or dissimilarity matrices. We applied it with Pearson's and Spearman's correlation on Euclidean, Canberra, and Manhattan distances for the original and log-transformed metabolome data. For microbiome data, Euclidean distance was used on a normalized matrix for better interpretability¹¹ (Table S1). P values were empirically derived from 1000 permutations. The analysis was conducted using the *vegan* R package⁵³.

MMiRKAT extends MiRKAT to assess global associations between a distance-transformed microbiome dataset (kernel function) and a low-dimensional multivariate phenotype²³. The model regresses the multivariate outcome on the non-parametric kernel-transformed microbiome with a Gaussian error term. Consistent with the Mantel test, we used Euclidean, Canberra, and Manhattan distances as kernel-based transformations for microbiome data, while the original or log-transformed metabolome matrix served as the outcome (Table S1). Significance was assessed using the exact Davies method. MMiRKAT was implemented via the *MiRKAT* R package⁵⁴.

Procrustes Analysis enables high-dimensional visualization by aligning datasets through linear translation, scaling, and rotation to minimize

Euclidean distance (Procrustes superimposition). Prior to alignment, both the microbiome and metabolome matrices were independently subjected to principal component analysis using the *rda* () function from the *vegan* R package, to obtain low-dimensional representations suitable for Procrustes comparison. Once aligned, inter-omics distances were computed for each sample, with smaller distances indicating stronger agreement between the two omic layers. Statistical significance was assessed via 999 permutations, comparing the observed sum of squared deviations to the resampled distribution. Procrustes Analysis was performed using the *vegan* R package⁵³.

Method evaluation. For global and individual association methods, we evaluated performance using Type-I error control and power, as these methods follow frequentist frameworks.

Type-I error control measures a method's ability to maintain the expected false positive rate (e.g., 5% at a 0.05 significance threshold). Values above this rate indicate a liberal method, while lower values suggest conservatism. This was assessed via quantile-quantile plots of $-\log_{10}$ p -values.

Power quantifies a method's ability to detect true associations while maintaining Type I error control. Ideally, methods should maximize power without inflating false positives.

For this category, the ground truth was defined at the global level, meaning that a structured dependence was introduced between microbiome and metabolome data without specifying individual species-metabolite pairs. A method was considered effective if it maintained proper Type-I error control while maximizing power.

Data summarization

Data summarization methods aim to capture shared variability between microbiome and metabolome data, facilitating dimensionality reduction and the identification of key underlying structures. We evaluated several approaches to extract latent components summarizing these complex relationships.

Selected methods. CCA²⁴ identifies linear combinations of two matrices that maximize correlation and was implemented using the *CCA* R package^{55,56}.

PLS maximizes covariance instead²⁵, with two variants: canonical PLS and regression PLS, applied via *mixOmics*.

RDA combines multivariate linear regression with PCA²⁶: regression is first applied to predict responses from predictors, followed by PCA on predicted values. RDA was performed using *vegan*.

MOFA2 is a Bayesian probabilistic framework that identifies shared sources of variability across omic layers²⁷, applied with default settings in MOFA2. For MOFA2, the number of latent factors was selected by the model, while for other methods, we retained all components up to the smallest feature set in either dataset.

Method evaluation. To ensure fair comparisons between data summarization methods, we considered the redundancy index initially proposed for CCA^{24,57}, and extended it to PLS²⁵. This index measures the proportion of variance in each dataset explained by the latent variables. In practice, we report the percentage of variance explained in each omic, a more interpretable metric across methods, and aligns with common practice in dimensionality reduction and data integration frameworks.

Individual associations between species and metabolites

This section evaluates methods designed to identify direct, pairwise associations between individual microbial taxa and metabolites. Each taxon-metabolite pair is tested independently to assess its association strength, aiming to recover true links introduced in the simulation while maintaining control over false positives. We additionally include pairwise Pearson and Spearman correlation tests as standard baseline methods for evaluating taxon-metabolite relationships.

Selected methods. We evaluated four methods, a log-linear regression on CLR-transformed microbiome data, a log-contrast model³², MiRKAT²³, and HALLA⁵⁸.

Table 1 | Confusion table

Predicted/ground truth	Non-associated	Associated
Non-associated	True negatives (TN): correctly identified as non-associated	False negatives (FN): missed associations
Associated	False positives (FP): incorrectly identified as associated	True positives (TP): correctly identified associations

The CLR-based log-linear model (CLR-lm) is defined as:

$$E(Y_{ik}^* | X_{ij}^*, \beta_j) = \beta_0 + X_{ij}^* \beta_j + \epsilon_i, \forall (j, k)$$

where Y^* is the log-transformed metabolome matrix and X^* the CLR-transformed microbiome data. While CLR accounts for compositionality, it does not preserve sub-compositionality³². The log-contrast model addresses this by enforcing $\sum_{j=1}^p \beta_j = 0$:

$$E(Y_{ik}^* | X_{i.}, \beta) = X_{i.} \beta + \epsilon_i, \quad \sum_{j=1}^p \beta_j = 0$$

To test for associations, we applied the global significance F-test via the `compositions` R package.

MiRKAT uses a semi-parametric kernel machine regression framework to summarize microbiome relationships²³. Unlike other methods, it allows multiple distance kernels, addressing uncertainty in kernel selection. We tested Euclidean, Canberra, and Manhattan distances on original and transformed microbiome data, with the original or log-transformed metabolome as the outcome, using the `MiRKAT` R package.

Finally, we applied HALLA, a hierarchical All-against-All statistical framework⁵⁸, designed for both homogeneous and heterogeneous datasets while controlling the False Discovery Rate. HALLA was implemented via the `halla` Python package with default parameters, using CLR-transformed and original microbiome data, along with log-transformed and original metabolome data.

Method evaluation. The performance of individual association methods was assessed using Type-I error rate and power, as described earlier for global association methods. Type-I error rate evaluates a method's ability to control false positives, while power measures its capacity to detect true associations.

For this category, the ground truth was defined as the set of true species-metabolite associations introduced during data simulation. A method was considered effective if it maximized power while maintaining proper Type-I error control.

To account for multiple testing in individual association methods, we combined p-values using the Aggregated Cauchy-based test (ACAT)⁵⁹ for CLR-lm, HALLA, and correlation-based approaches. ACAT, provided by the `ACAT` R package⁶⁰, aggregates p-values across microorganisms for each metabolite, reducing the multiplicity burden while controlling the Type-I error rate. Unlike Fisher's method, ACAT remains robust to p-value correlations and is computationally efficient for large datasets.

Feature selection: univariate

Univariate feature selection methods aim to identify relevant direct 1:1 associations between taxa and metabolites by evaluating their associations independently, across all possible pairs. In contrast to the previous section, which assesses each pair's significance, the focus here is on selecting the most informative features—those involved in at least one true association—based on individual association scores. Methods are evaluated using sensitivity, specificity, and sparsity to quantify their ability to prioritize relevant features.

While feature selection is often associated with supervised modeling, we use the term here to denote the identification of taxa or metabolites involved in true inter-omic associations, reflecting its broader usage in multi-omics studies^{1,19,61,62}.

Selected methods. Adapted from ref. 33, we evaluated two models incorporating compositional predictors when modeling metabolites as outcomes.

First, we applied CLR-LASSO, which performs a CLR transformation on microbiome data before fitting a univariate or multivariate LASSO log-linear regression²⁹. In the Results section, we refer to these models as LASSO and MLASSO. Formally, for a metabolite k , the LASSO log-linear model is:

$$\sum_{i=1}^n \left(Y_{ik}^* - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

where Y^* is the log-transformed metabolome matrix and X^* the CLR-transformed microbiome data. The penalty parameter λ was selected via 10-fold cross-validation over a grid from 0.01 to 1. Models were implemented using the `glmnet` R package⁶³.

We also applied coda-LASSO, a penalized log-contrast model ensuring scale invariance through a zero-sum constraint on regression coefficients. This approach, extending⁶⁴, selects the most contributive features by fitting a two-stage model on all possible log-ratios between species, leading to sparse solutions. The `coda4microbiome` R package⁶⁵ was used with default parameters.

Methods evaluation. The performance of univariate feature selection methods was assessed using sparsity, sensitivity, and specificity.

These metrics are used to evaluate how well the methods recover simulated associations. Specifically, sensitivity and specificity were calculated by comparing the selected features with the ground truth, which is the set of taxa or metabolites directly involved in the simulated associations (i.e., those with nonzero coefficients). While we do not include an explicit phenotype or classification task in our framework, the focus of the evaluation is on the ability of the methods to identify these predefined associations from the microbiome and metabolome data.

Specificity measures the proportion of true negatives among all non-associated features, while sensitivity quantifies the proportion of true positives among all associated features. Based on this, the confusion matrix (Table 1) is defined as follows:

They are computed as:

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}$$

For univariate feature selection methods, the ground truth corresponds to the set of taxa or metabolites directly involved in simulated 1:1 associations, meaning those with nonzero coefficients. Specificity and sensitivity are then evaluated based on the correct or incorrect identification of these associations.

Sparsity is also reported to quantify model parsimony, i.e., the proportion of selected features. While the ground truth is known, sparsity remains relevant to assess whether a method identifies only a minimal and informative set of features rather than an overly large or noisy subset.

Formally, the sparsity is given by:

$$\text{Sparsity} = \sum_{i=1}^p \frac{I(\beta_i \neq 0)}{p}$$

where β is either the coefficient for an association from a univariate feature selection method or corresponds to a loading in a multivariate feature selection. p is the total number of possible associations/features.

Feature selection: multivariate

Multivariate feature selection methods aim to identify subsets of taxa and metabolites that jointly contribute to the global covariance structure across omic layers. Unlike univariate approaches that test each pair independently, multivariate methods capture many-to-many relationships and select features that participate in shared latent structures, while ensuring model sparsity and interpretability.

Selected methods. sCCA³⁰ and sPLS³¹ extend CCA and PLS by incorporating penalization to perform both data summarization and feature selection.

For sCCA, an L1 penalty was applied to both datasets, retaining features contributing to the first two components. Optimal penalties were determined via 25 permutation-based samples with a 0.1-step grid search from 0.01 to 1. sCCA was implemented using the PMA R package⁶⁶.

Similarly, canonical and regression sPLS were tuned using 10-fold cross-validation with a 5-step grid search (5–25 for low-dimensional data, 10–50 for high-dimensional data). A maximum of two components was retained to select the most relevant features. sPLS was applied using the mixOmics R package¹.

For both sCCA and sPLS, only features with non-null loadings on the first two components were considered informative for performance evaluation.

Method evaluation. The performance of multivariate feature selection methods was assessed using sparsity, sensitivity, and specificity, as described earlier for univariate feature selection.

For multivariate feature selection methods, the ground truth is defined as the set of features contributing to the latent structures underlying the simulated associations, meaning those with nonzero loadings. Sensitivity and specificity are assessed in relation to these identified features.

Konzo data analysis workflow

Konzo is a paralytic neurological condition affecting children and women in sub-Saharan Africa, linked to chronic consumption of improperly processed cassava containing cyanogenic glucosides. While cyanide exposure is a key factor, not all consumers develop konzo, suggesting additional contributing factors.

Fecal samples from 171 individuals in Masi-Manimba ($n = 65$) and Kahemba ($n = 106$), Democratic Republic of the Congo, were analyzed using metagenomics and metabolomics, regardless of disease status. Shotgun metagenomic sequencing (~50 million reads per sample) was performed on DNA extracted from ~250mg of stool, using Kraken2 and Bracken for taxonomic classification³⁹. Metabolite profiling was conducted by Metabolon using a validated reference database. The dataset includes 1098 microorganisms and 1340 metabolites.

Microbiome data at the genus level were CLR-transformed, and metabolome data were log-transformed. The workflow included (1) global association, (2) data summarization, (3) univariate and multivariate feature selection, and (4) individual associations. Microorganisms were analyzed as both explanatory variables and outcomes. Given the high dimensionality, we applied the Mantel test for global associations and RDA to identify key contributors. Core microorganisms and metabolites were extracted using regression sPLS, selecting features with nonzero loadings on the first two components. Finally, CLR-lm regression and CODA-LASSO were used to assess microbial contributions to metabolite variations. The workflow is summarized in Fig. S13.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

We systematically evaluated the performance of microbiome-metabolome integration strategies by comparing different computational approaches across multiple simulated datasets. The following sections present the results

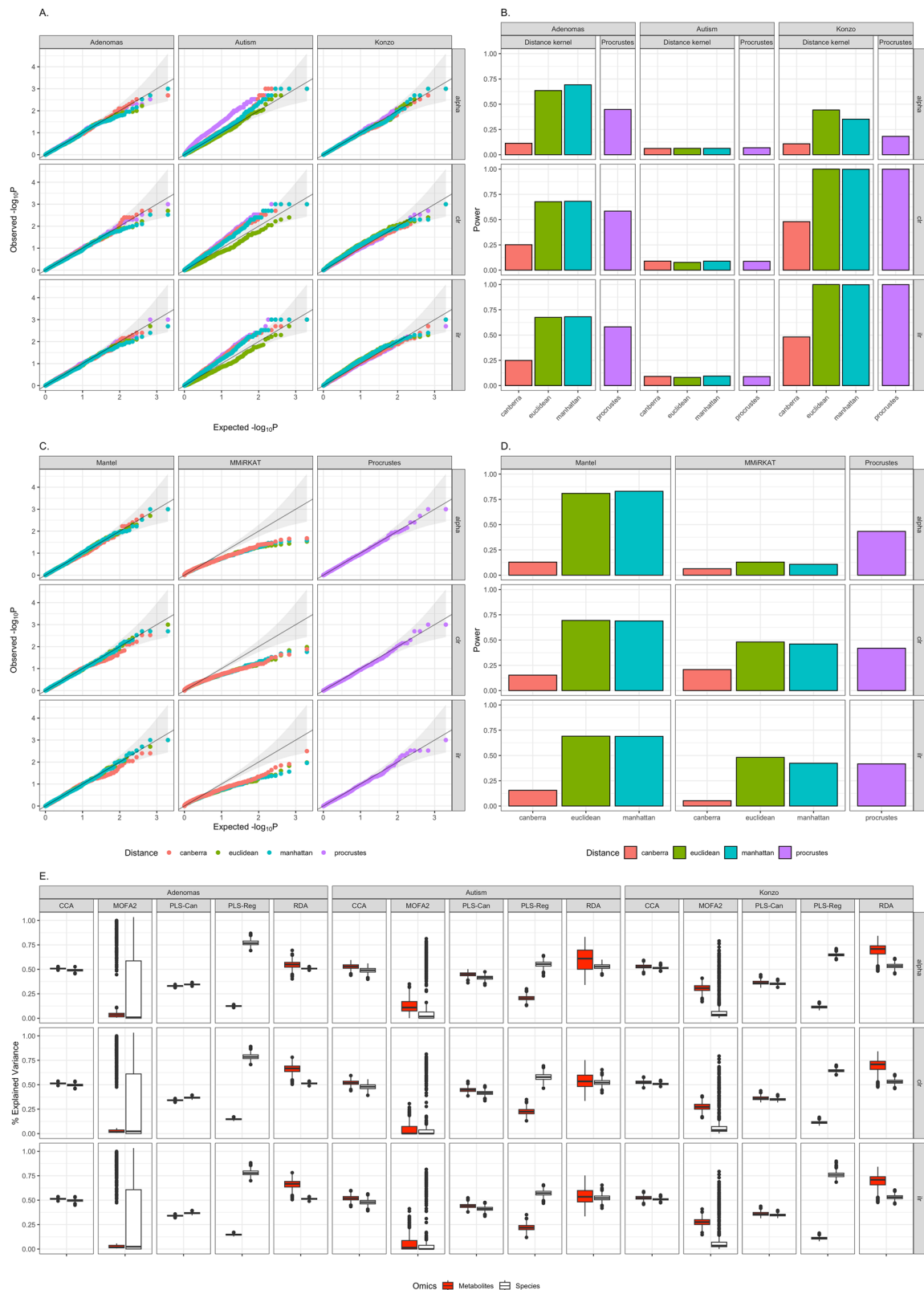
for each category of methods, using evaluation metrics specific to each approach to highlight their strengths and limitations.

Global associations

In our main setting, the Mantel test and Procrustes Analysis exhibited comparable results across microbiome data normalizations for controlling the false positive rate in the Adenomas and Konzo simulated-based scenarios (Fig. 2C). While in the Autism simulated datasets, the Mantel test considering the Euclidean distance provided the best control of Type-I error rate; the Procrustes analysis and other distance kernels showing inflation patterns (Fig. 2A). Nonetheless, methods highlighted variable powers depending on the scenario, microbiome normalizations and distance kernels considered (Fig. 2B). Indeed, in the Konzo simulated-based scenario, the alpha transformation could detect more than twice fewer associations compared to the CLR or ILR normalizations, whether the Mantel test or the Procrustes Analysis are considered (Fig. 2B). Additionally, the Canberra distance showed weakest powers across all the three settings, where no clear distinction between Euclidean and Manhattan distances could be observed (Fig. 2B). As an example, in the Adenomas simulation-based scenario, considering the CLR normalization, both the Manhattan and Euclidean distances offered up to 2.72-fold higher powers than the Canberra kernel (Power Manhattan = 68%; Power Euclidean = 67%; Power Canberra = 25%). In an additional scenario with a smaller number of features than individuals, the Mantel test was, on average, the best method across microbiome data normalizations and distance kernels, showing good control of the Type-I error rate and power (Fig. 2C, D). Briefly, when considering the Euclidean distance for the Mantel test, the method was minimally 1.5 times more powerful than MMiRKAT, considering the same distance kernel or the Procrustes Analysis (Fig. 2D). This result was consistent across other microbiome normalizations. Our results were confirmed in additional scenarios and considering other types of correlations (Figs. S1–4). Our findings point to different method capabilities to detect global associations depending on the data at hand. Importantly, our results suggest better performances for the Mantel test regarding both Type-I error rate and power when compared to MMiRKAT or the Procrustes analysis. Particularly, applying the Mantel test using the ILR normalization on microbiome data and the log transformation on metabolites was a robust strategy, offering the best results to detect global associations in a variety of scenarios.

Data summarization

As described in the Methods, we primarily focused on the first two components for all methods except MOFA2, where the number of components was inferred by the model. Regardless of the considered microbiome data normalization, method's performances were consistent in the three simulation setups. For example, RDA exhibited weak variations of the metabolome explained variances across data normalizations, with standard errors of 0.003, 0.003, and 0.002, for Adenomas, Autism, and Konzo simulation-based scenarios. Impacts of the choice of normalization on result variability for other methods were provided in Fig. S5. This result suggests that the normalization used for dealing with compositionality observed in metagenomics data does not affect the capability of methods to capture data variability. Importantly, we observed that no single method consistently outperformed others across all data summarization techniques and simulation scenarios, highlighting that performance is context-dependent and influenced by the specific characteristics of each dataset. However, RDA provided consistent performances compared to other methods in most scenarios, showing an average explained variance of 52% depending on the omics layer of interest (Fig. 2E). Moreover, MOFA2 exhibited the highest variable results, spanning from 0% to 100% of explained variances. This variability suggests that MOFA2's performance is highly sensitive to specific data properties, such as signal strength, noise level, and sample size. For regression-based methods, such as PLS-Reg or RDA, the integration directionality impacted the method performances, producing up to 6-fold higher explained variances (Fig. 2E). We confirmed our findings where considering the original metabolome (Fig. S6). Our results pointed to



RDA as the best trade-off to summarize data variability through latent factors, while PLS-Reg could be exploited when strong assumptions about the effect direction can be made. As a whole, our findings suggested that the RDA or PLS-Reg are versatile and robust under scenarios commonly encountered in practice, where the choice of the best method could be driven depending on the research question (See Discussion).

Individual associations between species and metabolites for compositional prediction

Studying the relationship between metabolites and microorganisms may represent an important challenge when accounting for the compositionality induced by microbiome datasets. Indeed, the perfect correlation brought by the compositionality makes the application of standard methods incorrect.

Fig. 2 | Performance of the multivariate methods for global associations or data summarization. **A** QQ-Plot of the Mantel test and the Procrustes Analysis across microbiome normalizations and distance kernels. For the Mantel test, we considered Spearman's method for computing the global association between the two datasets. *P* values for both the Mantel test and Procrustes Analysis were obtained empirically based on 1000 replicates. **B** Power of the Mantel test and the Procrustes Analysis across microbiome normalizations and distance kernels. For the Mantel test, we considered Spearman's method for computing the global association between the two datasets. *P* values for both the Mantel test and Procrustes Analysis were obtained empirically based on 1000 replicates. *P* values ≤ 0.05 were considered significant. **C** QQ-Plot of MMiRKAT, the Mantel test, and the Procrustes Analysis across microbiome normalizations and distance kernels. Points below the straight line refer to conservative behavior in the result section. To accommodate MMiRKAT (fewer

number of features than sample size), we considered scenarios with a smaller number of features in both omics layers than the number of individuals (See supplementary methods). **D** Power of MMiRKAT, the Mantel test, and the Procrustes Analysis across microbiome normalizations and distance kernels. To accommodate MMiRKAT (fewer number of features than sample size), we considered scenarios with a smaller number of features in both omics layers than the number of individuals (See supplementary methods). *P* values for both the Mantel test and Procrustes Analysis were obtained empirically based on 1000 replicates. *P* values ≤ 0.05 were considered significant. **E** Proportion of explained variance for the data summarization methods across different data structures and normalizations considering the log metabolome. Data summarization methods were compared considering scenarios with a number of features half the number of individuals (See supplementary methods).

This is particularly true when species are incorporated as covariates in statistical models analyzing microbiome-metabolome associations^{11–15,32}. To address this challenge while focusing on individual (1:1) associations, we evaluated four methods that test each microbe-metabolite pair independently: HALLA⁵⁸, the Log-contrast model³², MiRKAT²³, and a linear regression on CLR-transformed microbial abundances (referred to as CLR-lm). Overall, under the null hypothesis, across our three scenarios, methods exhibited different behaviors regarding the control of the Type-I error rate, showing an accurate control in the Adenomas simulation-based setting, or being either conservative or slightly liberal in the Autism and Konzo simulation-based scenarios (Fig. 3A). We observed a similar pattern when investigating Pearson's and Spearman's correlation with an accurate control or modest inflations of false positives depending on the considered setting (Fig. S7). These results highlight that Type I error control can be sensitive to dataset-specific characteristics, and no method maintained consistent performance across all settings. At the nominal 5% level, methods highlighted weak to modest powers, ranging from 7% to 35% depending on the method and simulated dataset considered, when these powers were drastically reduced after correcting for multiplicity (Figs. 3B and S8). These results are partially explained by the simulation setting, where a low signal-to-noise ratio was assumed, mimicking complex microbiome-metabolome relationships. In certain replicate, null powers were observed, while in others, a high percentage of significant associations (Fig. S9). Then, comparing models to Pearson's and Spearman's pairwise correlations (straight lines in Fig. 3B), we found no clear advantage of the Log-contrast, HALLA, MiRKAT, or CLR-lm over standard approaches, when higher powers were mainly due to uncalibrated false positive rates. Interestingly, MiRKAT was robust to microbiome normalization, offering consistent control of the Type-I error rate and power across CLR, ILR, or alpha transformations (Figs. 3A and S10). Collectively, our results align with a poor calibration of the Type-I error rate in our Konzo and Autism-based simulated scenarios for HALLA, the CLR-lm, or log-contrast regression, while MiRKAT exhibited consistent results across the considered scenarios. Across methods, well-calibrated QQ plots and reliable false positive control were only observed in the Adenomas-based simulations, indicating that dataset-specific features had a strong impact on method performance. Our evaluation suggested that MiRKAT offers the best false-positive power trade-off in all our simulation settings. The pros and cons of each strategy were further elaborated in the Discussion section.

Univariate feature-selection for compositional predictors

Feature selection methods have gained increasing interest from researchers in order to identify a subset of species associated with a variable of interest^{2,33,67,68}. However, due to the compositional nature of microbiome data, traditional methods have been shown to lead to results that may not accurately reflect the underlying biological structure, though they can still provide valuable insights in other contexts¹¹. Thus, we compared univariate feature selection methods accounting for compositional predictors, CODA-LASSO, CLR-LASSO, and CLR-MLASSO³³. First, we evaluated whether methods were able to provide sparse sets of microorganisms across our three scenarios. We found that

CLR-MLASSO exhibited higher sparsity than CLR-LASSO and CODA-LASSO, meaning that it retained a smaller subset of the original number of species (Fig. 4A). For example, in our Konzo-based simulation setting, we observed that CODA-LASSO offered the sparsest method with on average 0.2% of features selected ($sd = 0.001$) compared to 2% ($sd = 0.003$) and 41% ($sd = 0.09$), for the CLR-LASSO and CLR-MLASSO, respectively. This result was confirmed in the two other scenarios with different underlying data structures. Then, we assessed how accurate the methods are to find true associations, evaluating methods based on specificity and sensitivity (see Methods). Overall, methods exhibited different patterns of results, with weak to high sensitivities depending on the method and scenario considered (Fig. 4A). As an example, under our three simulation settings, CODA-LASSO showed sensitivities of 0.3%, 10%, and 30%, in the Adenomas-based, Autism-based, and the Konzo-based simulated scenario, respectively, suggesting method performance discrepancies under realistic data structures. We observed comparable results when considering CLR-LASSO or CLR-MLASSO. Additionally, because the great majority of features were non-associated, methods showed highly specific behaviors. Interestingly, CLR-MLASSO, due to the underlying feature selection process, could be highly sensitive at the price of lower sparsity or specificity compared to the two other methods. Collectively, our results highlight the importance of selecting univariate feature selection methods based on the data structure to ensure accurate and interpretable identification of key microbial signatures associated with metabolites. Specifically, methods incorporating compositional constraints, such as CODA-LASSO, demonstrated improved feature selection in microbiome data.

Multivariate feature-selection

Across our three scenarios, either sPLS-Can or sPLS-Reg provided lower sparsity scores compared to sCCA. While sPLS selected on average 12% of the total number of features, sCCA tends to keep either all or no variables, suggesting a poor method's performance to select features, as confirmed by the patterns of specificity and sensibility (Fig. 4B). This result is higher than the upper bound of true associations assumed by the simulation scenario (10%) suggesting that methods tend to provide a higher proportion of false associations than expected. Further investigation revealed that differences in sparsity could be attributed to the distribution of penalty parameters, which varied across simulation settings. In some cases, we observed strong or uniform penalization patterns, leading to overly sparse solutions (Fig. S11). Interestingly, both sPLS-Can and sPLS-Reg exhibited the same behavior, showing important levels of specificity and modest sensitivities, while sCCA showed inconsistent performances across our three simulation settings. For example, in the Konzo-based simulated scenario, the two methods offered on average 12% of signals detected as true when the signal is true (sensitivity), with much higher variability for sCCA as demonstrated by the shape of the boxplots in Fig. 4B. Finally, when focusing on sPLS-Reg, we found that directionality of integration did not offer discrepancy in results, suggesting robustness to the underlying outcome structure. These results were confirmed when considering metabolites on the original scale (Fig. S12). Collectively, our findings indicate that multivariate feature-selection methods

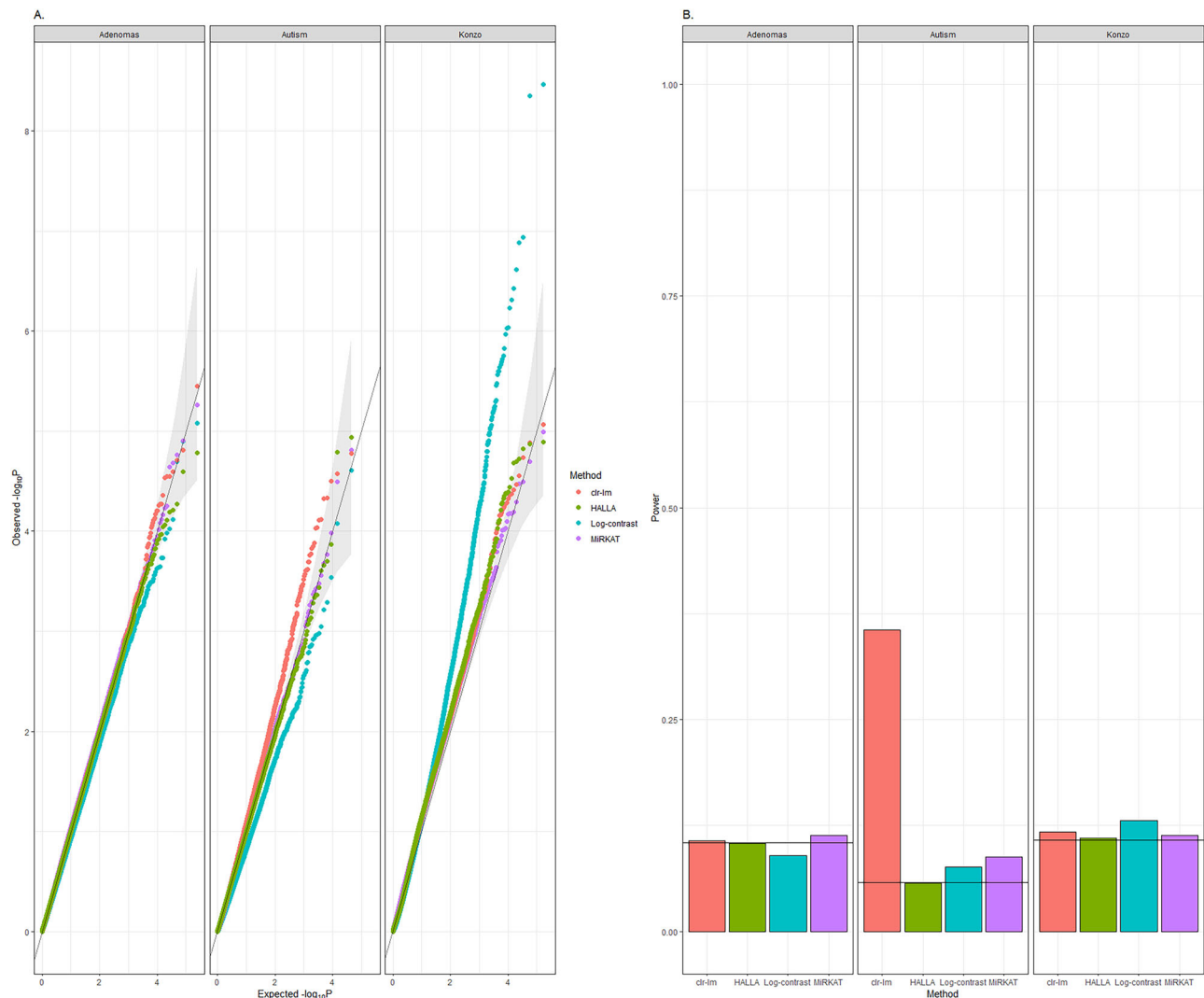


Fig. 3 | Performance of the individual association methods for compositional predictors. To accommodate long running times due to the number of pairs between species and metabolites, we considered scenarios with a number of features half the number of individuals (See supplementary methods) **A** QQplots of the individual association methods across our three simulation settings. **B** Power of the individual association methods across our two main scenarios. P values ≤ 0.05 were considered significant. For the CLR-lm method and HALLA, p -values were

combined using ACAT in order to provide similar comparisons with the log-contrast regression and MiRKAT (See Methods). For MiRKAT, we reported Type-I error rate and power using the ILR transformed microbiome data and the log transformed metabolites, while for HALLA, we considered the CLR transformed microbiome and the log metabolome. The straight line represents the background ACAT-combined power using Spearman's correlation on the CLR microbiome and the log metabolome. Powers were averaged over 1000 replicates.

have limited ability to distinguish informative from uninformative features when leveraging intra- and inter-correlations between omic layers.

Real-data application

Our systematic evaluation of strategies to jointly analyze microbiome and metabolome data has permitted the identification of the best methods depending on the research question. Thus, we illustrated the best approaches through an application on metabolomics and metagenomics data of the Konzo disease³⁹. We presented the exact workflow in the Konzo data analysis section and Fig. S13. Firstly, we asked whether there is a different pattern of global association between the two omics layer in cases and controls and found a stronger relationship in Konzo-affected individuals (Mantel statistic r : 0.4272; Spearman's permutation p value: 9.999×10^{-5}) than in healthy individuals (Mantel statistic r : 0.2838; Spearman's permutation p value: 0.0026997). Then we applied the RDA and found that the two first components explained roughly 26% of metabolome variability across the two conditions, while these proportions remained stable when considering microbiome (Fig. 5A). Moreover, the top-20 most contributing features in each omic on the two first RDA factors highlighted a large panel of

associations between species and metabolites with distinct patterns of correlations occurring in affected and unaffected subjects (Figs. 5B and S14). For example, in unaffected individuals, RDA identified *mevalonate* as strongly positively associated with species and *3-hydroxyisobutyrate* as exhibiting moderate negative association with microorganisms. These two metabolites have been shown to be linked to inflammatory- or oxidative-based processes potentially involved in Konzo^{69,70}. Consistently, the application of RDA has allowed the detection of 15 *Prevotella* species, which have a negative correlation with metabolites in unaffected samples, whereas they are positively correlated with metabolites in KONZO-affected samples. Moreover, *Bifidobacterium pseudocatenulatum*, *B. adolescentis*, and *B. angulatum* are positively correlated in unaffected samples, whereas *B. pseudocatenulatum*, *B. catenulatum*, *B. longum* are negative in affected patients. A high number of *Streptococcus* species have a positive correlation uniquely in Konzo-affected samples. This result points to distinct patterns of associations between species and metabolites in healthy and affected subjects. Subsequently we used the sPLS regression and identified 30, and 45 metabolites and 235, and 130 microorganisms significantly contributing to the two first components, in cases and controls, respectively (Fig. 5C).

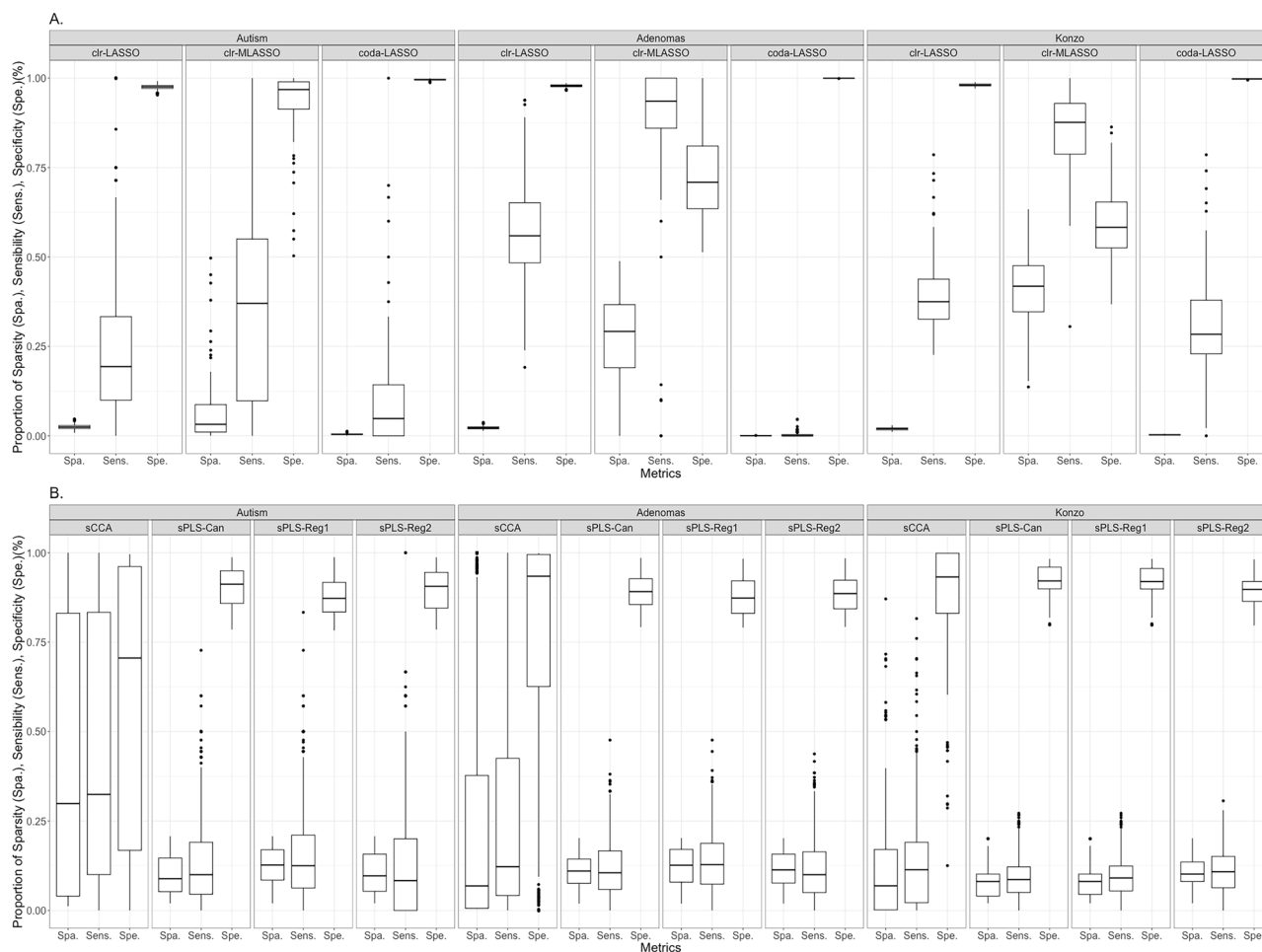


Fig. 4 | Performance (Sparsity (Spa.), Sensibility (Sens.), Specificity (Spe.)) of the feature selection methods for providing a sparse and reliable subset of elements across our two scenarios. A Performance of univariate feature selection methods considering microorganisms as covariates across our three settings. Metabolites were log-transformed before running the methods. Performances were calculated on 100 replicates. For CODA-LASSO in the Konzo scenario, we adapted the simulation

setting, selecting 300 species and 600 metabolites to accommodate running times of the method (See supplementary methods). **B** Performance of multivariate feature selection methods. Metabolites were log-transformed before running the methods. sPLS-Reg1 and sPLS-Reg2 correspond to the sPLS-Reg with X = microbiome and X = metabolome, respectively.

Interestingly, of the 130 species kept by the sPLS regression in healthy individuals, 64% were also found in affected individuals, while 46% of metabolites were preserved between the two conditions. Interestingly, both mevalonate and 3-hydroxyisobutyrate have been found to be contributing metabolites in Konzo-affected subjects. To investigate the implication of interactions between metabolites and species in Konzo, we sequentially applied MiRKAT and CODA-LASSO on the subset of unique metabolites found in affected individuals. Consequently, we found that *mevalonate* and *3-hydroxyisobutyrate* were significantly associated with 14 and 16 associated species, respectively, both exhibiting a large panel of associations (Fig. 5D, E). We identified species that could play a role in Konzo in affecting the oxidative response of the metabolism. For example, *Desulfovibrio desulfuricans* was positively associated with *mevalonate*, suggesting that increases in the microbial abundance are associated with an augmentation of metabolite levels. Also, *Clostridioides difficile* exhibited a consistent effect across *mevalonate* and *3-hydroxyisobutyrate*, suggesting common microbial dynamics between the two metabolites *Clostridioides difficile* has already been reported to have an impact in oxidative stress-related pathways, potentially involved in Konzo⁷¹. These associations were missed when applying the CLR-lm regression. We validated our findings at a larger scale by a systematic network analysis from the CLR-lm regression and CODA-LASSO (Figs. 5F and S15–17). Our results from metagenomics and metabolomics data from Konzo disease highlight distinct patterns of

interactions between microorganisms and metabolites occurring in both affected and unaffected individuals, where different microbial dynamics are involved.

Discussion

The integration of microbiome and metabolome datasets within statistical frameworks has become a valuable resource for researchers to comprehensively understand the underlying biological mechanisms involved in diseases. Indeed, recent studies in inflammatory bowel disease⁷² or cardio-metabolic traits⁴ have highlighted that pathoetiology may result in the disruption of the complex architecture occurring between the two omic layers. Understanding these interactions represents, therefore, a critical avenue for unraveling the biology of complex phenotypes. However, currently, there are no standards on how to integrate these two omic layers together, pushing researchers to constantly waste a lot of time in their decision-making process. Thus, deciding which method fits best for a specific biological question remains a daunting task, critically limiting the result interpretations and replicability. In this paper, we extensively benchmarked nineteen existing integrative methods to disentangle microbiome-metabolome interactions, covering most of the researchers' aims: global associations, data summarization, individual associations, and feature selection. Based on a comprehensive and realistic simulation study and a real data application, we highlighted the best methods depending on the

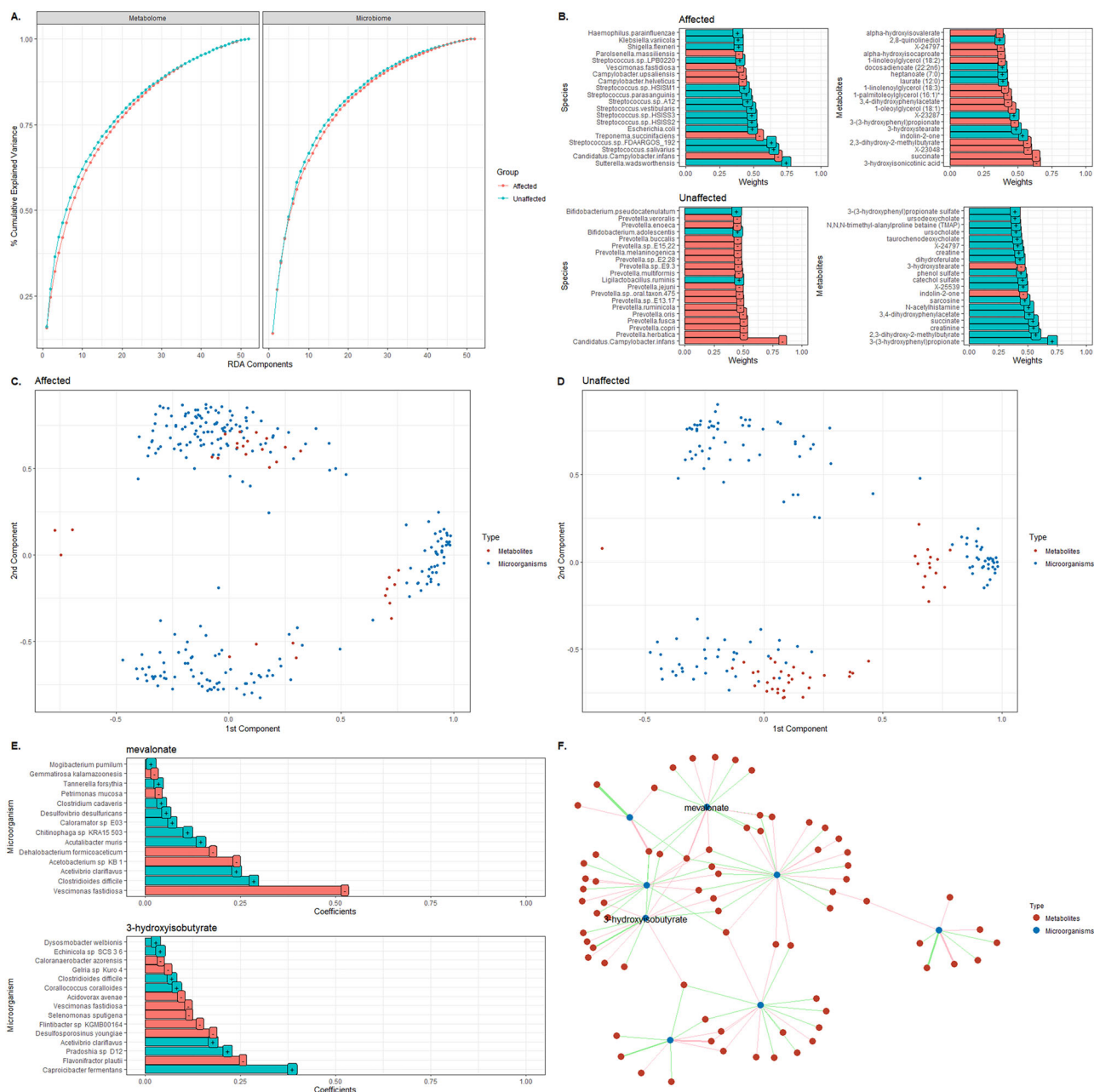


Fig. 5 | Application of best strategies highlights complementary biological interactions between microorganisms and metabolites in Konzo data.

A Proportion of cumulative explained variance in Metabolome and Microbiome datasets in both affected and unaffected individuals **B** Top-20 of the most contributing species and metabolites on the first RDA component in healthy and affected samples. Positive correlations were identified by a +, while negative correlations were identified with a - sign. Projection of metabolites (red) and microorganisms (blue) into the 2D regression SPLS space in **C** affected and **D** unaffected

individuals. Features with null loadings were removed from the analysis.

E Coefficients provided by the CODA-LASSO across mevalonate and 3-hydroxyisobutyrate were identified only in Konzo by the regression sPLS. Positive coefficients were identified by a +, while negative coefficients were identified with a - sign **F** Network between mevalonate and 3-hydroxyisobutyrate and their corresponding associated species found by CODA-LASSO. Positive associations were represented by green edges and negative associations by pink edges.

research question and data at hand, providing important insights about statistical good practices (Table 2) and avenues for future methodological developments (Table 3).

Despite its realism, our simulation framework has some limitations. In particular, it does not explicitly remove correlations driven by case-control group effects, meaning that some simulated associations may reflect group-level signals rather than true biological links. However, this limitation is mitigated in the context of a benchmark, where the primary goal is to compare methods under controlled yet realistic conditions, rather than to reproduce ground-truth biological mechanisms.

Furthermore, a user guide and all the scripts used in this work are available in the GitHub repository associated with the paper to facilitate reproducibility and further advancements in the field.

When evaluating global association methods, our results have pointed to important lessons for practitioners. Indeed, across our three realistic simulation scenarios, the Mantel test is the most powerful method to find associations occurring at a global scale compared to MMiRKAT and the Procrustes Analysis. Also, the method exhibits an adequate control of Type-I error rate, showing robustness across a variety of scenarios, considering several underlying data structures, normalizations, and distance kernels.

Table 2 | Summary of best methods depending on the research question

Scientific question	Research aim	Recommended method	Pros	Cons	Implementation
Is there any relationship between microorganisms and metabolites at a global level?	Global associations	Mantel test	Robust to data normalization and distance kernels, computationally efficient	Limited ability to adjust for covariates (possible through partial Mantel tests)	vegan R pkg.
Are microbiome and metabolome datasets summarizable through a limited number of components?	Data summarization	RDA	Robust to data normalization and distance kernels	Can only capture linear effects sensitive to directionality	vegan R pkg.
Can we identify associations between metabolites and species?	Individual associations	MIRKAT	Allow adjustment for covariates, Robust to microbiome normalization	Limited to a few families of generalized linear models	MIRKAT R pkg.
Can we identify core microorganisms and metabolites?	Feature selection (univariate)	CODA-LASSO (compositional covariates)	Compositional and sub-compositional coherent. No need for data transformation. Allow adjustment for covariates	Limited to a few families of generalized linear models, long running times in high-dimensional problems	coda4microbiome R pkg.

Table 3 | Overview of the research questions, corresponding methodological approaches, selected models, and performance metrics used in this benchmark

Scientific question	Research aim	Model selection	Selected models	Performance metrics
Is there any relationship between microorganisms and metabolites at a global level?	Global associations	Need to adjust for covariates	YES: Regression-based models	MMIRKAT
			NO: Bi-directional models	Mantel test, Procrustes analysis
Are microbiome and metabolome datasets summarizable through a limited number of components?	Data summarization	Does the directionality matter	YES: Regression-based models	PLS-Regression RDA
			NO: Canonical-based models	PLS-Canonical CCA MOFA2
Can we identify associations between metabolites and species?	Individual associations	Need to adjust for covariates	YES: Regression-based models	Log-Contrast MIRKAT
			NO: Correlation-based models	Pearson Spearman HALLA
Can we identify core microorganisms and metabolites?	Feature selection	Need to account for the between-within correlation	YES: Univariate models	CODA-LASSO CLR- (M) LASSO
			NO: Multivariate-models	sPLS-Regression sPLS-Canonical sCCA

Methods are categorized based on their ability to assess global associations, summarize data, identify individual associations, or perform feature selection.

This is an appealing feature in practice since choosing the right data transformation or distance metric may represent an important challenge for practitioners. However, by exploiting regression-based frameworks, MMRKAT, unlike the Mantel test, can adjust for confounding factors, making the correction for certain bias induced by individual characteristics, such as age, sex, lifestyle, or even batch effects, possibly²³. However, the method is unable to deal with scenarios with a larger number of features than individuals, limiting applications in most multi-omics scenarios, as pointed out by ref. 20. We therefore recommend using the Mantel test in most cases or MMRKAT filtering out features based on a feature selection approach when confounding is expected. The Procrustes Analysis could be exploited to have graphical representations, but shows no advantages over the Mantel test or MMRKAT in our benchmark. Importantly, when using the Mantel test, our results suggest that the Canberra distance on metabolome data is the poorest choice for detecting global associations across all our scenarios (Fig. 2B). The application of the Mantel test in our study identifies distinct patterns of global associations between microbiome and metabolome occurring in affected and unaffected individuals. Thus, our recommendation here is to apply Euclidean distance on the transformed microbiome data while applying Euclidean or Manhattan distances on metabolites in most cases.

Dimensionality reduction is often used by practitioners for summarizing information through a small number of components. Having an efficient method that recapitulates variability across two omic layers is critical for facilitating subsequent analyses, such as visualization or clustering. We considered five different methods exhibiting specific features to summarize omic layers information and found that in addition to being robust to data normalization, RDA is the most reliable method showing consistency across simulation scenarios and data normalizations (Fig. 2E). Interestingly, our results point to important impacts of directionality when applying regression-based methods, such as PLS-Reg or RDA, with performances drastically increasing depending on whether species or metabolites are considered as the outcome. This result could be explained by the underlying complex structure of microbiome data, strongly impacting the capability of methods to capture data variability. Consistent with this idea, we observed unpredictable performances for MOFA2 with explained variances going from 0% to 100% in all our scenarios (Fig. 2E). Although MOFA2 has been shown to capture complex relationships for gene expression or methylation⁷³, our results align with a distinct pattern of performance for metagenomics and metabolomics data. This behavior is explained by the underlying assumptions of MOFA2, where the use of the method could be restricted to cases where quasi-normality of data is expected, since other types of data require statistical approximations²⁷. We then applied RDA to our Konzo dataset and found distinct species-metabolites associations involved in healthy and affected samples (Fig. 5A, B). In unaffected patients, the metabolites 3-(4-hydroxyphenyl)lactate and 5-hydroxyindoleacetate are identified in positive correlations and may be linked to the fermentation process of the *Bifidobacterium* spp. positively correlated in unaffected samples and negatively correlated in affected samples.

In practice, another important question for researchers is to determine the relationship between microbial communities with a variable of interest^{39,74}. However, the underlying compositional structure of microbiome data is an important challenge for model performance. In this paper, we have compared four methods accounting for the compositionality of predictors with different strategies: a linear regression applied on the CLR transformed microbiome data, MMRKAT, the log-contrast model, and HALLA. Compared to correlations, these methods have not been shown to be significantly more powerful (Fig. 4B). As already pointed out in the result section, weak powers are mostly explained by the simulation scenario where we assumed a low signal-to-noise ratio. This behavior could therefore be expected in real-data applications. Thus, we recommend applying univariate methods to only a relevant subset of features, avoiding systematic applications across all metabolite-species pairs. Also, certain univariate methods suffer from inflation of the Type-I

error rate in our scenarios (Fig. 4A), suggesting that method performances are impacted by the underlying data structures. For example, the log-contrast model exhibits three different patterns of results across our three scenarios, showing well calibration, conservative, and liberal behaviors in our Adenomas-, Autism-, and Konzo- based scenarios (Fig. 4A). Our general recommendation to the reader is to use MMRKAT, since the method has shown accurate controls of Type-I error rate across microbiome normalizations, permitting correction for confounding factors. This is particularly important since the choice of the right normalization could be difficult. However, microbiome transformations could have critical impacts on result transferability and interpretability, where CLR provides still-correlated synthetic features and ILR or alpha reduce the dimensional space^{33,34}. This result highlights the need for new compositional data transformations, keeping the original number of features while linearly independent (Table 4).

Also, one important contribution of this work is to extensively evaluate feature selection methods. This is particularly crucial for researchers to accurately select metabolites and microorganisms involved in a specific biological context. Our results point to moderate performance of multivariate feature selection methods with inconsistent performances across scenarios and the data transformations considered (Figs. 5B and S12). The best performances are achieved for univariate feature-selection methods for compositional predictors, with CODA-LASSO as a good trade-off between sparsity and classification performances (Fig. 5A), while being sub-compositional coherent. However, all methods provide discrepancies in performance regarding the data structure. For example, CODA-LASSO is strongly affected by the proportion of zeros in microbiome data, as suggested by our results in the Adenomas-based scenario. Thus, we recommend in practice to use CODA-LASSO for scenarios with microbial predictors after removing taxa with a high proportion of zeros, to ensure the good performance of the method, whereas sPLS-Reg could be exploited to select features exploiting the inter-omics correlation. We therefore applied both regression sPLS and CODA-LASSO on the Konzo dataset. Regression sPLS has permitted the detection of a distinct set of metabolites and microorganisms in affected and unaffected individuals. From the subset of metabolites contributing to Konzo, CODA-LASSO has highlighted different microbial dynamics of effects (Fig. S17). This result is aligned with the model where microorganisms may be connected to a large set of metabolites. This complex microbiome-metabolome crosstalk has been shown to be associated with diseases⁷⁵. However, the identified associations may include artifact signals, as many feature selection methods benchmarked in this study exhibit limited sparsity and variability in their selected features, affecting result consistency. This result is aligned with previous reports where authors have shown poor performances of traditional feature selection models⁷⁶. Indeed, most penalized methods are built upon cross-validation, where small perturbations in data may yield drastic changes in results. Similarly to⁷⁶ extending sparse multivariate or univariate methods to the knockoff framework⁷⁷ or stability selection⁷⁸ should represent interesting avenues for improving both sparsity and reliability for compositional data⁷⁹ (Table 4).

Although our work is focused on evaluating methods for inferring associations between metagenomics and metabolomics data, additional work is still required to comprehensively compare methods in the context of prediction. Indeed, predicting metabolome levels from microbiome data is a flourishing research topic with critical implications for clinical applications. Indeed, addressing many challenges, such as integrating and analyzing diverse omics data types, dealing with high-dimensional data, addressing data heterogeneity, and developing robust computational models that capture complex relationships between different molecular layers, could be part of further investigation.

Although our simulation setup can realistically simulate microbiome and metabolome data, our framework is limited to the “Normal-to-Anything” framework. However, as discussed by⁴¹, simulating pure compositional data from a Dirichlet distribution induced only a small correlation between features, which is often unrealistic regarding the biology of the

Table 4 | Overview of avenues for future methodological developments to jointly analyze metagenomics and metabolomics data

Objective	Methods and limits	Methodological avenues
Data normalization	- CLR: provides still-correlated features in the original space - ILR; alpha are “black-box” transformations providing uncorrelated features in a restricted space	Data normalizations provide uncorrelated features in the original space for facilitating result interpretation.
Mechanistic interpretation	- Log-contrast; CODA-LASSO: captures statistical associations but does not infer causal or mechanistic relationships between microbiome and metabolome	Integration of causal inference frameworks (e.g., structural equation modeling, Bayesian networks) or dynamic modeling approaches to better characterize mechanistic links between microbial and metabolic changes
Feature selection	- CODA-LASSO; sPLS: lack of sparse solutions	Extension to the knockoff framework or stability selection for improving feature selection performances

Three research objectives with their corresponding methods were identified as important avenues for future methodological developments.

Box 1 | Recommendations of the best methods depending on the research objectives

Global Associations: We recommend using the Mantel test when evaluating the global associations between microbiome and metabolome data. To properly account for metagenomics data structure, we strongly recommend using the Euclidean Distance on the ILR normalization and log-transformed metabolites, when applicable. The ILR normalization preserves data isometry required for distance-based methods.

Data Summarization: If directionality is not of interest, we recommend using RDA on the ILR-transformed microbiome and log metabolites. However, depending on which omic layer is of interest, PLS-Reg is a great choice, particularly when studying the effect of microbiome on metabolome.

Individual Associations: We recommend using MiRKAT to obtain a well-controlled Type-I error rate and consistent power across a large

panel of data structures. Moreover, the method allows the incorporation of covariates to adjust for, which is an appealing feature in practice.

Univariate Feature Selection: We suggest using CODA-LASSO to identify core species associated with metabolites. In addition to being scale-invariant while preserving sub-compositionality, the method can automatically subset the most associated microorganisms. We recommend using the method after filtering out zero-inflated species to reduce the running times while maintaining good performance.

Multivariate Feature Selection: In practice, we recommend to use sPLS-Regression on the CLR microbiome and log metabolome. The method shows consistent results across scenarios, permitting the selection of most relevant associated features across datasets exploiting the inter-omics correlation.

microbial communities and metabolites. We therefore promote the correlation, zero-inflation, and overdispersion characteristics over a purely compositional structure, with no major impacts expected on the conclusions. Also, methods selected through the benchmark assumed a directionality of effects between species to metabolites; other types of approaches could be considered depending on the biological question. Indeed, in some contexts, linking metabolome to microbiome using models such as the Dirichlet regression could be achieved^{36,37}. We explored the performance of such approaches and found poor results. While we did not elucidate whether the results were explained by the methods themselves or the simulation setting, we decided to omit these approaches in the current version of the benchmark since they are not widely used in practice but could be explored in future works. Our illustration could identify biological species and metabolites involved in Konzo-related processes; our application is, however, limited to intersection-based analyses between affected and unaffected individuals. We are aware that explicit multi-omics models incorporating the disease information, such as DIABLO⁸⁰ or MDiNE⁴² are available. However, future evaluations of differential multi-omics integrative strategies are required to mechanistically link microbiome and metabolome to diseases from a dynamic perspective at a large scale (Table 4). We argue this aspect is particularly critical to pinpoint the underlying biological mechanisms, hence facilitating precision medicine applications^{81,82}.

To conclude, this study presents one of the most comprehensive benchmarks of purely statistical frameworks for integrating metagenomics and metabolomics data. We focus exclusively on purely statistical methods that naturally fall into four main categories, while acknowledging that other integrative approaches exist, including those leveraging external knowledge or addressing different research questions^{17,83,84}. Through extensive and realistic simulations, we systematically compared nineteen methods across key research questions, identifying the most effective approaches. Their applicability was demonstrated on the Konzo dataset, highlighting

complementary biological processes across different scales. This comprehensive evaluation provides practical guidance tailored to specific scientific questions and data characteristics. By providing a robust and replicable comparative framework, we aim to enhance research standards and foster the development of new integrative models. We anticipate that systematic applications of these methods to unified metagenomics-metabolomics resources⁸⁵ will be instrumental in identifying common and unique species-metabolite interactions, essential for understanding disease mechanisms.

Key Points

- We systematically evaluated nineteen statistical methods for jointly analyzing metagenomics and metabolomics data across a wide range of scientific questions, including global associations, data summarization, individual associations, and feature selection.
- We provided general guidelines for practitioners for properly analyzing microbiome and metabolome data together, facilitating result interpretation and replicability (Box 1).
- We illustrated the best methods through an application to metagenomics and metabolomics data in the gut for Konzo disease, highlighting complementary biological results.

Data availability

The data used in this study are available from the following sources: Public benchmark datasets for metagenomics and metabolomics (Autism and Adenomas) were obtained from the curated resource available at: <https://github.com/borenstein-lab/microbiome-metabolome-curated-data>. Data used to reproduce the analyses presented in the manuscript can be accessed at: <https://doi.org/10.6084/m9.figshare.25234915>. Simulated data were generated using the `simulate_data.R` script, provided in the associated GitHub repository. The metagenomics and metabolomics data for the Konzo disease case study are available upon request from Dr. Matthew S. Bramble.

Code availability

All code necessary to reproduce the analyses and figures presented in this manuscript is publicly available at: https://github.com/lmangnier/Benchmark_Integration_Metagenomics_Metabolomics. A permanent snapshot of the code has been archived on Zenodo and can be cited as ref. 86. The code was developed and tested using R version 4.2.2.

Received: 3 October 2024; Accepted: 11 July 2025;

Published online: 25 July 2025

References

- Rohart, F., Gautier, B., Singh, A. & Cao, K. A. L. mixOmics: an R package for omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
- Tang, Z. Z. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* **10**, 454 (2019).
- Vernocchi, P., Chierico, F. D. & Putignani, L. Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. *Front. Microbiol.* **7**, 1144 (2016).
- Fromentin, S. Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
- Dan, Z. Altered gut microbial profile is associated with abnormal metabolism activity of autism spectrum disorder. *Gut Microbes* **11**, 1246–1267 (2020).
- Lavelle, A. & Sokol, H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 223–237 (2020).
- Parray, J. A., Egamberdieva, D., Abd Allah, E. F. & Sayyed, R. Z. Editorial: soil microbiome metabolomics: a way forward to sustainable intensification. *Front. Sustain. Food Syst.* **7**, 1251054 (2023).
- Gopalakrishnan, S. Beyond correlation: Understanding the causal link between microbiome and plant health. *Rhizosphere* **32**, 100870 (2024).
- Huang, W., Sun, D., Chen, L. & An, Y. Integrative analysis of the microbiome and metabolome in understanding the causes of sugarcane bitterness. *Sci. Rep.* **11**, 6024 (2021).
- Pérez-Cobas, A. E., Gomez-Valero, L. & Buchrieser, C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Micro. Genom.* **6**, e000409 (2020).
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
- Greenacre, M. Compositional data analysis. *Annu. Rev. Stat. Its Appl.* **8**, 271–299 (2021).
- Xia, Y. & Sun, J. *Statistical data analysis of microbiomes and metabolomics*. American Chemical Society (2022).
- Zhou, B., Xiao, J. F., Tuli, L. & Ransom, H. W. LC-MS-based metabolomics. *Mol. Biosyst.* **8**, 470–481 (2012).
- Pang, Z. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Res.* **52**, W398–W406 (2024).
- Puig-Castellví, F. Advances in the integration of metabolomics and metagenomics for human gut microbiome and their clinical applications. *Trends Anal. Chem.* **167**, 117248 (2023).
- Noecker, C., Eng, A., Muller, E. & Borenstein, E. MIMOSA2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome metabolome data. *Bioinformatics* **38**, 1615–1623 (2022).
- Reiman, D., Layden, B. T. & Dai, Y. MiMeNet: exploring microbiome-metabolome relationships using neural networks. *PLoS Comput. Biol.* **17**, e1009021 (2021).
- Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).
- Deek, R. A., Ma, S., Lewis, J. & Li, H. Statistical and computational methods for integrating microbiome, host genomics, and metabolomics data. *Elife* **13**, e88956 (2024).
- Gower, J. C. Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975).
- Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
- Zhao, N. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96**, 797–807 (2015).
- Hotelling, H. Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
- Abdi, H. Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 97–106 (2010).
- Legendre, P. & Louis, J. *Numerical Ecology* 3rd Edn., Vol. 24 (Elsevier, 2012).
- Argelaguet, R. Multi Omics Factor Analysis—a framework for unsupervised integration of multi omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
- Al Bataineh, M. T. Uncovering the relationship between gut microbial dysbiosis, metabolomics, and dietary intake in type 2 diabetes mellitus and in healthy volunteers: a multi-omics analysis. *Sci. Rep.* **13**, 17943 (2023).
- Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
- Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).
- Chun, H. & Keleş, S. K. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 3–25 (2009).
- Aitchison, J. & Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–353 (1984).
- Susin, A., Wang, Y., Cao, K. A. L. & Calle, M. L. Variable selection in microbiome compositional data analysis. *NAR Genomics Bioinform.* **2**, lqaa029 (2020).
- Tsagris, M. Improved classification for compositional data using the α -transformation. *J. Classif.* **33**, 243–261 (2016).
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. Isometric Logratio transformations for compositional data analysis. *Math. Geol.* **35**, 279–300 (2003).
- Hijazi, R. H. & Jernigan, R. W. Modeling compositional data using dirichlet regression models. *J. Appl. Probab. Stat.* **4**, 77–91 (2009).
- Chen, J. & Li, H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7**, 418–442 (2013).
- Noecker, C., Chiu, H.-C., McNally, C. P. & Borenstein, E. Defining and evaluating microbial contributions to metabolite variation in microbiome-metabolome association studies. *MSystems* **4**, 10–1128 (2019).
- Bramble, M. S. The gut microbiome in konzo. *Nat. Commun.* **12**, 5371 (2021).
- Clos-Garcia, M. Multiomics signatures of type 1 diabetes with and without albuminuria. *Front. Endocrinol.* **13**, 1015557 (2022).
- Hawinkel, S., Mattiello, F., Bijmens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).
- McGregor, K., Labbe, A. & Greenwood, C. M. T. MDINE: a model to estimate differential co-occurrence networks in microbiome studies. *Bioinformatics* **36**, 1840–1847 (2020).
- Wang, Y. & Cao, K. A. L. PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. *Brief. Bioinform.* **24**, bbac622 (2023).

44. Kim, M. Fecal metabolomic signatures in colorectal adenoma patients are associated with gut microbiota and early events of colorectal cancer pathogenesis. *MBio* **11**, 10–1128 (2020).
45. Kang, D.-W. Differences in fecal microbial metabolites and microbiota of children with autism spectrum disorders. *Anaerobe* **49**, 121–131 (2018).
46. Kurtz, Z. D. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
47. Aitchison, J. *The Statistical Analysis of Compositional Data* (Monographs on Statistics and Applied Probability). (Chapman and Hall, 1986).
48. Tsagris, M. T., Preston, S. & Wood, A. T. A. A data-based power transformation for compositional data. *arXiv* <https://arxiv.org/abs/1106.1451> (2011).
49. Ejigu, A. et al. Normalization techniques in metabolomics: current challenges, recommendations, and future perspectives. *ACS Meas. Sci. Au.* <https://doi.org/10.1021/acsmeasuresci.4c00047> (2024).
50. van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142 (2006).
51. Katajamaa, M. & Orešič, M. Data processing for mass spectrometry-based metabolomics. *Nucleic Acids Res.* **1158**, 318–328 (2009).
52. Fernie, A. R., Aharoni, A. & Willmitzer, L. Metabolite profiling: from diagnostics to systems biology. *Plant Physiol.* **169**, 1727–1738 (2015).
53. Oksanen, J. et al. *vegan*: community ecology package (R package version 2.6-4). <https://CRAN.R-project.org/package=vegan> (2024).
54. Plantinga, A. et al. *MiRKAT*: microbiome regression-based kernel association tests. R package version 1.2.3, <https://CRAN.R-project.org/package=MIRKAT> (2023).
55. Hanson, B. A. *CCA*: canonical correlation analysis (R package version 1.2-2). <https://CRAN.R-project.org/package=CCA> (2024).
56. van den Boogaart, K. G. & Tolosana-Delgado, R. *Compositions*: a unified R package for compositional data analysis. *Comput. Geosci.* **34**, 1395–1411 (2008).
57. Stewart, D. & Love, W. A general canonical correlation index. *Psychol. Bull.* **70**, 160–163 (1968).
58. Ghazi, A. R. High-sensitivity pattern discovery in large, paired multiomic datasets. *Bioinformatics* **38**, i378–i385 (2022).
59. Liu, Y. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
60. Liu, Y. & Xie, W. ACAT: aggregated cauchy association test. R package version 1.6.0. <https://CRAN.R-project.org/package=ACAT> (2022).
61. Vahabi, N. & Michailidis, G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front. Genet.* **13**, 854752 (2022).
62. Kumar Singh, P. & Kumar, A. Identifying redundant features using unsupervised learning for high-dimensional data. *Expert Syst. Appl.* **66**, 218–225 (2016).
63. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
64. Bates, S. & Tibshirani, R. Log-ratio lasso: scalable, sparse estimation for log-ratio models. *Biometrics* **75**, 613–624 (2019).
65. Calle, M., Susin, T. & Pujolassos, M. *coda4microbiome*: compositional data analysis for microbiome studies (R package version 0.2.4). <https://CRAN.R-project.org/package=coda4microbiome> (2024).
66. Witten, D. M., Tibshirani, R. & Hastie, T. PMA: penalized multivariate analysis (R package version 1.3-1). <https://CRAN.R-project.org/package=PMA> (2024).
67. Scott, D. A. V. Bayesian compositional regression with microbiome features via variational inference. *BMC Bioinform.* **24**, 210 (2023).
68. Nguyen, Q. P. Associations between the gut microbiome and metabolome in early life. *BMC Microbiol.* **21**, 238 (2021).
69. Nilsen, M. S. 3-Hydroxyisobutyrate, a strong marker of insulin resistance in type 2 diabetes and obesity that modulates white and brown adipocyte metabolism. *Diabetes* **69**, 1903–1916 (2020).
70. Tricarico, P. M., Crovella, S. & Celsi, F. Mevalonate pathway blockade, mitochondrial dysfunction and autophagy: a possible link. *Int. J. Mol. Sci.* **16**, 16067–16084 (2015).
71. Weiss, A., Lopez, C. A., Beavers, W. N., Rodriguez, J. & Skaar, E. P. *Clostridioides difficile* strain-dependent and strain-independent adaptations to a microaerobic environment. *Micro. Genom.* **7**, 000738 (2021).
72. Ning, L. Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nat. Commun.* **14**, 7135 (2023).
73. Palou-Márquez, G., Subirana, I., Nonell, L., Fernández-Sanlés, A. & Elosua, R. DNA methylation and gene expression integration in cardiovascular disease. *Clin. Epigenet.* **13**, 75 (2021).
74. Yang, L. & Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* **10**, 130 (2022).
75. Lee-Sarwar, K. A., Lasky-Su, J., Kelly, R. S., Litonjua, A. A. & Weiss, S. T. Metabolome-microbiome crosstalk and human disease. *Metabolites* **10**, 181 (2020).
76. Hédou, J. Discovery of sparse, reliable omic biomarkers with Stabl. *Nat. Biotechnol.* **42**, 1581–1593 (2024).
77. Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: ‘model-x’ knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **80**, 551–577 (2018).
78. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 417–473 (2010).
79. Srinivasan, A., Xue, L. & Zhan, X. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics* **77**, 984–995 (2021).
80. Singh, A. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
81. Petrosino, J. F. The microbiome in precision medicine: the way forward. *Genome Med.* **10**, 12 (2018).
82. Talmor-Barkan, Y. Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease. *Nat. Med.* **28**, 295–302 (2022).
83. Shaffer, M. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. *BMC Bioinform.* **20**, 614 (2019).
84. Morton, J. T. Learning representations of microbe-metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).
85. Muller, E., Algavi, Y. M. & Borenstein, E. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *NPJ Biofilms Microbiomes* **8**, 79 (2022).
86. Mangnier, L. et al. *Benchmark Integration of Metagenomics and Metabolomics*. Zenodo. <https://doi.org/10.5281/zenodo.15602388>, (2024).

Acknowledgements

We would like to thank members of the Arnaud Droit Lab, particularly Louis-Maël Gueguen, Thomas Jeanne, and Tania Cuppens for their insightful comments on the manuscript.

Author contributions

L.M. and A.B. designed, conducted, performed the data analysis, and wrote the manuscript. M.M. performed the data analysis. A.M., M.P.S.B., and N.V. wrote the manuscript. A.B., M.P.S.B., M.S.B., and A.D. revised the manuscript. All authors read and approved the final version of the manuscript. L.M., A.B., M.P.S.B., and A.M. are research fellows in the

Arnaud Droit Lab affiliated with Laval University. M.M. is an intern in Biostatistics in the Arnaud Droit Lab affiliated with Laval University. N.V. is a postdoctoral fellow in the Arboleda Lab, in the Department of Pathology and Laboratory Medicine at UCLA. M.B. is an associate professor at The George Washington University School of Medicine and Health Sciences. A.D. is a full professor in the Department of Molecular Medicine at Laval University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-08515-9>.

Correspondence and requests for materials should be addressed to Arnaud Droit.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Silvio Waschina and Tobias Goris.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025