frontiers | Frontiers in Systems Biology

TYPE Specialty Grand Challenge
PUBLISHED 17 June 2024
DOI 10.3389/fsysb.2024.1432791

OPEN ACCESS

EDITED AND REVIEWED BY
Yoram Vodovotz,
University of Pittsburgh, United States

*CORRESPONDENCE
Umer Zeeshan Ijaz,
Umer.Ijaz@glasgow.ac.uk

RECEIVED 14 May 2024
ACCEPTED 17 May 2024
PUBLISHED 17 June 2024

CITATION
Ijaz UZ, Ameer A, Saleem F, Gul F, Keating C and Javed S (2024), Specialty grand challenge: how can we use integrative approaches to understand microbial community dynamics? *Front. Syst. Biol.* 4:1432791.
doi: 10.3389/fsysb.2024.1432791

Specialty grand challenge: how can we use integrative approaches to understand microbial community dynamics?

Umer Zeeshan Ijaz[1,2,3]*, Aqsa Ameer[1,4], Farrukh Saleem[4,5], Farzana Gul[4], Ciara Keating[6] and Sundus Javed[4]

[1]Water and Environment Research Group, University of Glasgow, Mazumdar-Shaw Advanced Research Centre, Glasgow, United Kingdom, [2]Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, United Kingdom, [3]National University of Ireland, Galway, Ireland, [4]Department of Biosciences, COMSATS University Islamabad, Islamabad, Pakistan, [5]National Veterinary Laboratories, Ministry of National Food Security and Research, Islamabad, Pakistan, [6]Department of Engineering, Durham University, Durham, United Kingdom

KEYWORDS

regression analysis, multivariate statistical analyses, integrative "omics", microbiology, ecology

Introduction

Microbiome studies have seen exponential growth since the advancement of next-generation sequencing (NGS) technologies (Qin, 2019)—albeit now old technologies! Sequencing approaches applied to mixed microbial populations involve either amplification of small genes such as the 16S rRNA gene (Johnson et al., 2019), or recovery of whole microbial genomes through shotgun metagenomics (Quince et al., 2017). The majority of observational and interventional studies are hypothesis-driven, with samples obtained either as case controls, spatially, or within temporal settings (Knight et al., 2018; Qian et al., 2020). Regardless of the methodology or study criteria, the analysis of the data through bioinformatics typically yields abundance/coverage tables that recover N (samples) x P (features) data within the chosen experimental or environmental context. Additional data (metadata) include parameters associated with the samples of interest. For environmental samples, these may include physicochemical parameters, and for human- or other host microbiome studies, additional data may include anthropometric measures and clinical data. Indeed, these data are essential to correlate treatments, conditions, or experimental variables with microbial community profiles.

The trend is increasingly geared toward collecting more and more metadata, such as the incorporation of metabolomics for metabolites (Bauermeister et al., 2021), metatranscriptomics for gene transcripts (Ojala et al., 2023), and metaproteomics for proteins (Armengaud, 2023). There are also commercial research services available such as Resistomap (https://www.resistomap.com/), which facilitates environmental monitoring of antibiotic resistance genes by offering a customizable target gene table using SmartChip qPCR. In host studies to unravel host-microbiome interactions, flow cytometry-based immunophenotyping is typically incorporated (Siebert et al., 2019). In clinical research, services such as Olink (https://olink.com) offer target platforms for protein biomarker analysis. This is based on a technology called Proximity Extension Assay, which uses labeled antibody pairs with DNA oligonucleotides that bind to the corresponding proteins in a sample. These oligonucleotides are then extended by DNA polymerase and are quantified through microfluidic qPCR. They offer different protein-associated panels/biomarkers with biological functions linked to cytokines, cardiovascular disease, immuno-oncology,

neurology, oncology, inflammation, and several other biological processes. Recently, there has been a focus on the study of microbial ecosystems in their entirety, with the buzzword being the "exposome", i.e., all the observable variations to which microbial communities are exposed (Gao et al., 2022; Gul et al., 2024). This plethora of additional data can then fill in the gaps of how the microbiome responds to the environment it is observed in. This will provide mechanistic insight into the function of microbial communities in a number of important contexts.

In this grand challenge review, we discuss numerous statistical approaches currently in use to find associations across multiple datasets sharing the same sampling space. Deducing discriminant features based on variations in the sampling space (case-control, spatial, temporal, etc.) and segregating them from features that remain fairly stable is a challenging issue. We also discuss challenges in their utility and where the gaps need to be filled.

Data that are not microbiome (sequence) data but provide further information about the microbiome samples are considered metadata. At the most basic level, metadata can be either categorical data (labeling of samples) or continuous data (for example, numerical data such as age, body mass index (BMI), anthropometric measures, and physicochemical parameters like pH and temperature). Additional metadata includes features recovered from other modalities, such as the metabolome (mainly continuous variables). While the goal is to capture as many perceived sources of variability within the confines of the environment in which the microbiome is observed, downstream statistical analyses become challenging and raise several questions:

- Of all the covariates that are captured, which ones should be included in the analyses?
- What about the confounders that are not captured?
- Do all covariates hold equal importance? Is there a way to rank them?
- If additional modalities generate a new set of features, where should the emphasis be in the downstream statistics once multiple feature tables are obtained? Finding discriminating features in the sampling space or finding correlating features across the datasets? Is there a trade-off?
- How do we translate an association between the covariate and an individual or subset of microbes into clinical or ecological relevance?
- Is linearity the best assumption to infer patterns of interest between the microbes and the covariates?
- How do we tackle heteroscedasticity and under-sampling particularly when there are more features than the number of samples P >> N?
- Which approach holds importance? Is a study-centric approach suitable where the emphasis is on features that remain stable or act discriminatory, or a taxa-centric approach, where given all observed variability, we can assess the ecological role of a particular microbe?
- Given the thousands of microbes that are detected, should we include all or some in the analyses? What is more important? Highly abundant microbes? Highly prevalent microbes? Highly interacting microbes?
- Is there any utility of rare biosphere in the analysis? How do we decide what is rare and is not a limitation of sequencing depth?
- How do we incorporate the inherent correlations that exist between samples, particularly in clinical studies, where a single subject has provided multiple samples?
- How can we assess the stability and complexity of microbial ecosystems in the wake of environmental perturbations?
- How can network reconstruction approaches be improved further? Given a network topology, how do we decide what the most influential species are?
- In spatial or temporal gradients, how do we compare datasets where the sampling time/space do not match?
- What is an appropriate normalization measure for different types of data?

Grand challenge: how can we find a relationship between a specific variable in a sea of variability and noise?

To assess the relationship between a single continuous outcome data and all measured independent variables we typically apply regression modeling. Regression models describe the relationship between one or more independent variables and a dependent variable. One of the most commonly used models is the linear regression model. Regression coefficients generally referred to as β-coefficients, are associated with each continuous parameter and several categorical parameters. The categorical parameters are often *dummified* into a numerical representation, typically one less than the total number of factors observed in a parameter, through a procedure called one-hot-encoding. The variable that is excluded becomes a reference variable often denoted as REF. The formula for linear regression is $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_k x_{k,i} + \varepsilon_i$. The equation can also be written in matrix form as $\boldsymbol{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. The signs of the β-coefficients give directionality with respect to the outcome $\boldsymbol{y}$ with the following interpretations: for continuous variables, a positive/negative coefficient is interpreted as "an increase/decrease in the covariate causes an increase/decrease in the outcome"; and for categorical variables, a positive/negative coefficient is interpreted as "as compared to the reference REF, the outcome Y is increasing/decreasing". There are several extensions to the linear regression model, the most popular of which are discussed below.

The Generalized Linear Model (GLM) is a popular model (Xiao et al., 2018; Koh et al., 2019) in microbiome studies. In the GLM model, an outcome $\boldsymbol{y}$ is assumed to be generated from a certain distribution; relevant distributions include normal, binomial, Poisson, and negative binomial distributions, among others. The regression model is then defined as $g(\boldsymbol{\mu}) = \mathbf{X\beta}$ where $\boldsymbol{\mu}$ is the conditional mean of the distribution, and $g(.)$ is the link function. The logistic regression is particularly important (i.e., the probability of an outcome with a specified variable), where the outcome is assumed to have a binomial distribution (the outcome variable takes values of 0 and 1) and the link function is the logit function $ln(p/(1-p))$. However, we are interested in the log-binomial regression model which also assumes a binomial distribution for a binary outcome but uses a log link function lnp. Fitting a log-binomial regression model with binomial errors and a log link to binary outcome data thus makes it possible to estimate risk ratios by taking the exponential of the beta coefficients

as e^{β_k}, e.g., in (Firew et al., 2020). It should be noted that the risk ratio is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group, and the log-binomial regression model was useful in the COVID-19 times to determine risks with occupational factors (Firew et al., 2020).

In many studies, the samples follow a case-control relationship, and therefore supplementing with a log-binomial regression model using all sources of variability fills in the gaps in our understanding. For outcome variables with more than two categories, *multinomial logistic regression* and *ordinal logistic regression* are recommended (Liang et al., 2020). For categorical outcomes, we obtain risk ratios from the models. On the other hand, for continuous outcomes, the typical strategy is to limit the number of variables (covariates) in the model, either through the *subset regression* approach applied to linear regression using R's leaps package (Lumley et al., 2013) or the Least Absolute Shrinkage and Selection Operator (LASSO) approach using R's glmnet package (Tay et al., 2023). In glmnet, "*The Relaxed LASSO*" is implemented which solves the following optimization problem, $\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1]$ over a range of values of the tuning parameter λ, with $l(.)$ being the negative log-likelihood contribution to the observation i. The method incorporates an *elastic net* penalty controlled by α, a compromise between lasso regression ($\alpha = 1$) and ridge regression ($\alpha = 0$). The LASSO constraint forces some of the beta coefficients to go to zero, acting as a variable selection approach. If the study has supplementary 'omics data as metadata, for example, targeted or untargeted metabolomics, then after appropriate normalization, e.g., probabilistic quotient normalization (Dieterle et al., 2006), these data can also be used in the LASSO regression. Furthermore, the approach is not just limited to continuous outcomes, as one can also use binary outcomes.

While the above models are mainly suitable for metadata, applying them to microbiome data is not straightforward. The read count data in 16S rRNA or metagenomic sequencing are typically summarized as a count table, and since the total sample read counts from an experiment, often referred to as the library size, are dependent on the sequencing technology, the absolute values are an artifact and present a challenge in how they are used in the regression models. Furthermore, depending on the depth of sequencing and the shape of the microbial community distribution (typically following a lognormal distribution), the table is highly sparse (50%–90% zero counts in the abundance table), often leading to overdispersion. These are the two main challenges that need to be addressed, i.e.,

- How do we effectively normalize the microbiome data when the samples do not have the same library size?
- In view of normalization procedures, how do we handle sparsity?

In the published literature, the above two questions are tackled in somewhat different ways, although there is no unifying framework. We list two recent regression approaches that address part of the problem with room for improvement.

1) In this case the microbial data are often taken to be compositional (i.e., the count table is converted to relative compositions constrained to 1). An example is the Compositional Decompositional Analysis (CODA)-LASSO approach (Calle et al., 2023) in which single binary/continuous outcome data from the meta table is regressed against the log abundances of microbes (Figure 1). While the approach offers variable selection by virtue of LASSO constraints, the challenging issue here is the log transform. The zero-count microbiome data cannot be log-transformed. A common practice is to add a pseudo-count, more commonly 1, 0.5 or even smaller values. There is no consensus on the appropriate choice of pseudo-count (Costea et al., 2014), and this remains an open problem to be solved despite recent attempts to address it (Hu et al., 2022).
2) In this case, abundance of individual microbes is regressed against covariates by fitting a distribution such as a Negative Binomial distribution that tackles overdispersion and sparsity. For example, using the *Generalized Linear Latent Variable Model* (GLLVM) approach (Niku et al., 2019) an extension of GLM, microbial abundances are regressed against all covariates including the latent variables (confounders that are not observed). The GLLVM approach (Figure 2) uses a link function g() similar to a GLM, fits a count distribution, and regresses against the covariates, where $\boldsymbol{\beta}_j$ are the coefficients of the microbes associated with individual covariates. After estimating a 95% confidence interval for these coefficients, there are three possibilities for a given beta coefficient: the 95% confidence interval is all positive (an increase in the covariate causes an increase in the abundance of the given microbe); the 95% confidence interval is all negative (an increase in the covariate causes a decrease in the abundance of the given microbe); and where the 95% confidence crosses the zero threshold (the covariate is insignificant). For scenarios where the covariate is categorical in nature, it is dummified (converted to 0s and 1s) with one factor acting as a reference. The interpretation is similar to the explanation given above for continuous covariates, except that now the interpretation is with respect to the reference factor: the 95% confidence interval is all positive (as compared to the reference factor, there is an increase in the abundance of the given microbe); the 95% confidence interval is all negative (as compared to the reference factor, there is a decrease in the abundance of the given microbe); and the 95% confidence interval crosses the 0 boundary (the covariate is not significant). While $\boldsymbol{\beta}_j$ are the coefficients associated with covariates $\boldsymbol{x}_i$, $\boldsymbol{\theta}_j$ are the corresponding coefficients associated with latent variable $\boldsymbol{u}_i$. β_{0j} are the intercepts and α_i are optional sample effects that can be chosen as either fixed effects or random effects. In addition, the residual covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ of the $\boldsymbol{\theta}_j$ coefficients stores correlations between microbes where $\boldsymbol{\Gamma} = [\theta_1 \ldots \theta_m]$ for m latent variables. This residual covariance matrix can then indicate co-occurrence relationships between microbes that are not explained by the observed covariates. However, fitting GLLVM against exhaustive sources of variability when there are thousands of taxa, significantly more than the number of samples, is computationally challenging and impractical for larger datasets. Including all of them may not result in the convergence of the likelihood function. The open-ended

10.3389/fsysb.2024.1432791

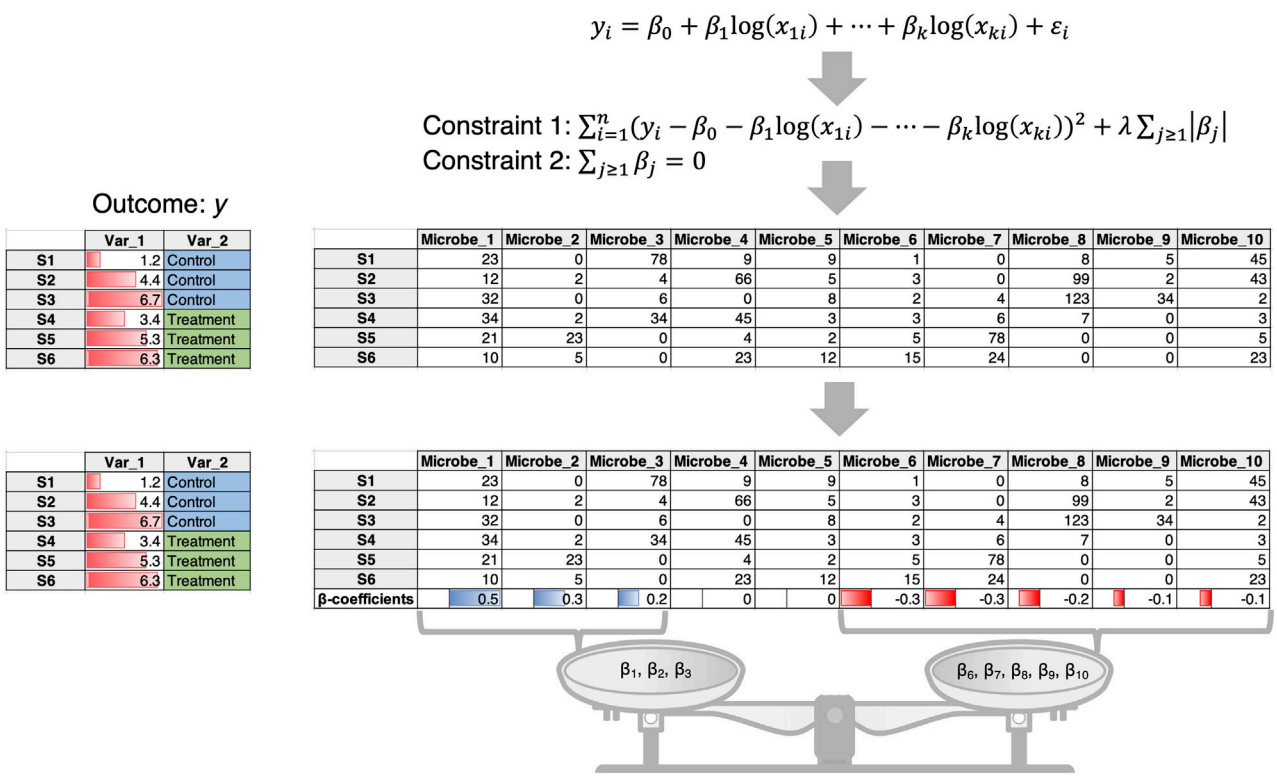


FIGURE 1
The CODA-LASSO approach extends the linear regression approach by taking log abundances of microbiome data and considering two additional constraints. Constraint 1 is the LASSO constraint that forces some of the beta coefficients to go to zero allowing for variable selection, while constraint 2 ensures that there are two subsets of beta coefficients of microbes, those that are positively associated with the outcome, and those that are negatively associated with the outcome.

questions are then: Within the framework of GLLVM, and other regression models in general, should we incorporate a subset of microbes? What should be the criteria for the inclusion of a microbe?

To limit the number of microbes in the regression model, the filtering criteria is often to ignore low abundance or low prevalence microbes. While this may have worked in several studies where the dominant role was played by the abundant or prevalent taxa, however, there is growing literature that emphasizes the importance of the rare biosphere (Lynch and Neufeld, 2015), which may have ecological, taxonomic, and functional potential. Furthermore, a recent study proposes that keystone interacting microbial species (Layeghifard et al., 2019) have far more clinical relevance than the abundant or prevalent ones. They have associated hubs from the network topology of interacting species with clinical covariates, demonstrating this approach to be better. However, more research is needed to identify keystone interacting species from microbiome datasets, particularly to formulate network-wide statistics that are not only robust against biases but also offer biological relevance. The *Integrated Value of Influence* (Salavaty et al., 2020) is a good starting point for this as a reasonable measure to identify keystone microbial species.

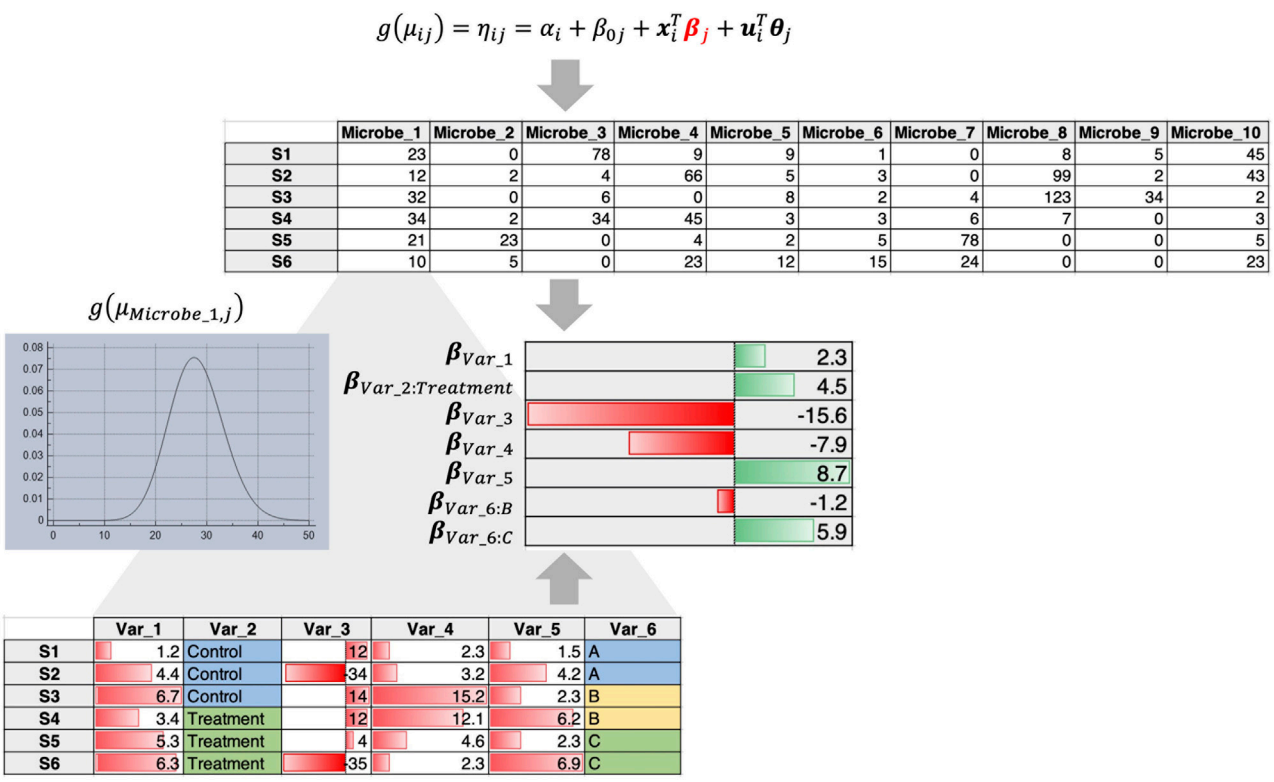
Grand challenge: how can we unravel the mediating role of microbes?

While most microbiome studies focus on observing changes in microbes in a case-control setting emphasizing their differential abundances, there are trait or performance data (outcomes) that are not explicitly incorporated. Therefore, a few challenges that arise include:

- Can we identify microbes that play a mediating role between the treatments and an outcome of interest?
- What is the nature of mediation? Is it local or global? Is there no mediation at all?
- Can the mediating microbes be potential targets for the development of therapeutic or clinical interventions?

Although still in its infancy, a recently proposed framework (Yue and Hu, 2022), simultaneously links Treatment T, microbial mediators $M = (M_1, ., M_J)$, outcome O, and confounding covariates Z. The classical model for multiple mediators (Van Der weele & Vansteelandt, 2013) is a double regression problem, where for a continuous outcome and J mediators, we have

10.3389/fsysb.2024.1432791



	Microbe_1	Microbe_2	Microbe_3	Microbe_4	Microbe_5	Microbe_6	Microbe_7	Microbe_8	Microbe_9	Microbe_10
S1	23	0	78	9	9	1	0	8	5	45
S2	12	2	4	66	5	3	0	99	2	43
S3	32	0	6	0	8	2	4	123	34	2
S4	34	2	34	45	3	3	6	7	0	3
S5	21	23	0	4	2	5	78	0	0	5
S6	10	5	0	23	12	15	24	0	0	23

	Var_1	Var_2	Var_3	Var_4	Var_5	Var_6
S1	1.2	Control	12	2.3	1.5	A
S2	4.4	Control	34	3.2	4.2	A
S3	6.7	Control	14	15.2	2.3	B
S4	3.4	Treatment	12	12.1	6.2	B
S5	5.3	Treatment	4	4.6	2.3	C
S6	6.3	Treatment	35	2.3	6.9	C

FIGURE 2
GLLVM procedure that fits a distribution for each microbe and regresses against all covariates to obtain beta coefficients that reveal positive or negative relationships.

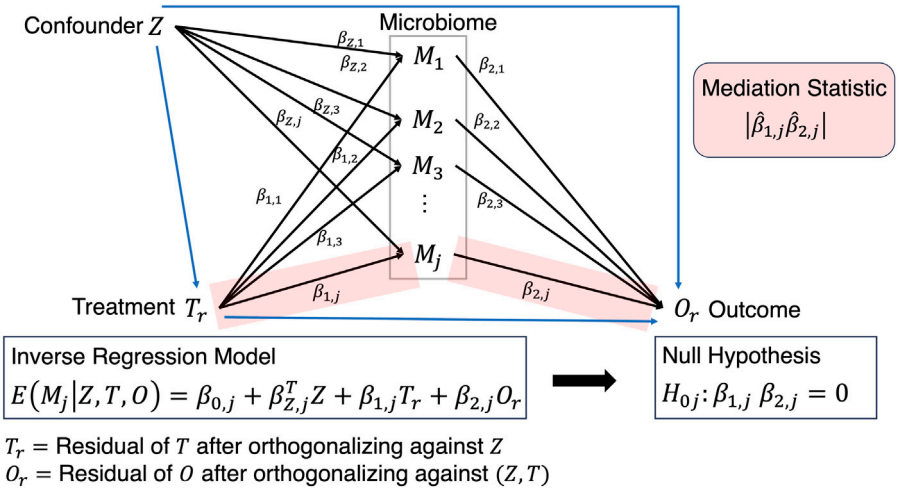


FIGURE 3
Mediation analysis where the path from *Treatment* to *Outcome* is established after solving the inverse regression problem and testing the corresponding beta coefficients that exist on that path.

$$E(M_j \mid Z,T) = \alpha_{0,j} + \alpha_{Z,j}^T Z + \alpha_{1,j} T$$

$$E(O|Z,T,M_1,\ldots,M_J) = \theta_0 + \theta_Z^T Z + \theta_1 T + \sum_{j=1}^{J} \theta_{2,j} M_j$$

where the total mediation effect through $(M_1,\ldots,M_J)$ microbial mediators take the form $\sum_{j=1}^{J} \alpha_{1,j}\theta_{2,j}$ with $\alpha_{1,j}$ characterizing the effect of T on M_j given Z, and $\theta_{2,j}$ characterizes the effect of M_j on O given Z and T and all other M_j s. Testing $\alpha_{1,j}\theta_{2,j} = 0$ for an individual mediator achieves the purpose as the non-zero value indicates the contribution of M_j to the overall mediation effect. $E(O|Z,T,M_1,...,M_J)$ called forward outcome model is difficult to solve as there are more mediators than the number of samples, and

therefore an inverse regression model is proposed (Figure 3) where orthogonalized versions of the residual of treatment T against Z defined as T_r, and orthogonalized versions of the residual of O against (Z, T) defined as O_r are used, and the two equations are merged into a single equation,

$$E\left(M_j \mid Z, T, O\right) = \beta_{0,j} + \beta_{Z,j}^{T} Z + \beta_{1,j} T_r + \beta_{2,j} O_r$$

such that $\beta_{1,j}$ corresponds to $\alpha_{1,j}$, and $\beta_{2,j}$ corresponds to $\theta_{2,j}$ with the test becoming $\beta_{1,j}\beta_{2,j} = 0$ for an individual mediator.

Further variations include causal mediation methods specifically designed to handle high-dimensional and compositional microbiome data (Wang et al., 2020). This rigorous Sparse Microbial Causal Mediation Model (Sparse MCMM) applies a linear log-contrast regression model and Dirichlet regression model to estimate the causal direct effect of treatment and microbiome mediation effects at both the community and individual taxon levels. PhyloMed (Hong et al., 2023) on the other hand discovers mediation signals by analyzing sub -compositions defined on the phylogenetic tree. While these mediation approaches offer a deeper understanding of the causal mediation effect of the microbiome and have a growing number of applications in microbiome studies, they are often context-dependent, and cater to those situations where a sizeable proportion of the microbial community changes. Where the changes are small or localized, the above mediation approaches do not appear to work well, and thus there is room for improvement.

Grand challenge: how can we determine whether variables cause a shift in microbial community diversity?

Microbial diversity measures provide insight into the structure and dynamics of microbial communities. Differences within treatments (alpha diversity) and between treatments (beta diversity) can highlight responses in microbial populations to environmental or other conditions. Diversity measures themselves do not provide statistics, only trends. How can we statistically determine whether a variable is linked to patterns in diversity? First, we need to apply an algorithm to establish whether the variable (covariate) is causing a change in beta diversity between groups. In this regard, Permutational Multivariate Analysis of Variance (PERMANOVA) is a very useful tool as it employs any of the beta diversity dissimilarity metrics suitable for microbiome data (traditional ones include Bray-Curtis distance and UniFrac metrics) and can be applied to a wide range of complex models. PERMANOVA is a permutation test that uses an F test to assess whether the variances of two populations are equal by comparing groups of objects with the null hypothesis being that centroids and dispersions are equivalent. For each of the covariates, the test returns an R^2 value which, if significant, is the percentage of variability in the microbiome explained by that covariate. As PERMANOVA is sensitive to the order of variables, it is often combined with a filtering process such as *Redundancy Analysis (RDA) with forward selection* (Vass et al., 2020). An alternate method is the *Fuzzy Set Ordination* (FSO) method (Roberts, 2009). Similar to PERMANOVA it uses dissimilarity metrics and metadata, but it is based on the principle of fuzzy set theory and reports correlation R as a quality-of-fit metric. Moving forward, two challenges need to be addressed:

1. All of these methods rely on distances that use all the measured microbiome count without incorporating individual covariances, i.e., the importance of individual microbial species is lost. Emerging approaches (Satten et al., 2017; Andries and Nikzad-Langerodi, 2022) do offer a bit of reprieve, but a concerted effort is required to develop this direction.
2. Another issue is that the majority of methods require multivariate uniformity of variability (homoscedasticity) and balanced sample sizes. In particular, PERMANOVA suffers from loss of power and type 1 error inflation (Alekseyenko, 2016). Therefore, there is a need to develop new robust methods that can ensure correct data analysis. The W_d^* test (Hamidi et al., 2019) may be a good advancement in this direction, but there is a lack of a unified framework to tackle heteroscedasticity.

Perhaps an alternative to PERMANOVA could be to knock out unnecessary covariates through the approach by (Clarke and Ainsworth, 1993) which presents algorithms that allow for the comparison of beta diversity distances between two sets of data that have either samples or features in common. This approach facilitates the exploration of environmental variables (or clinical parameters) that best correlate with sample similarities in the biological community (microbiome). In the procedure (termed BIOENV), the similarity matrix of the community is fixed, while subsets of the environmental variables are used in the calculation of the environmental similarity matrix. A correlation coefficient is then calculated between the two matrices and the best subset of environmental variables can then be identified and further subjected to a permutation test to determine significance. R's vegan package (Dixon, 2003) implements the bioenv() function, where the similarity matrix of environmental data is assumed to be based on normalized Euclidean distances (Figure 4). This makes sense with environmental data where one normalizes the data to remove the effect of differing scales between parameters. For the microbiome, the Bray-Curtis Similarity Index is commonly used due to its non-parametric nature.

Grand challenge: how can we visualize the relationship between covariates and microbial diversity patterns?

Constrained ordination approaches can be used to visualize the variation in microbial communities that can be explained by external environmental variables or constraints. There are different types of constrained ordination approaches such as Canonical Correspondence Analysis (CCA) and Redundancy Analysis (RDA) (Legendre and Legendre, 2012) (Figure 5). In the case of CCA, a Chi-squared transformed microbial abundance table is subjected to weighted linear regression on the constraining variable. The fitted values are then subjected to correspondence analysis using Singular Value Decomposition (SVD). RDA on the

10.3389/fsysb.2024.1432791

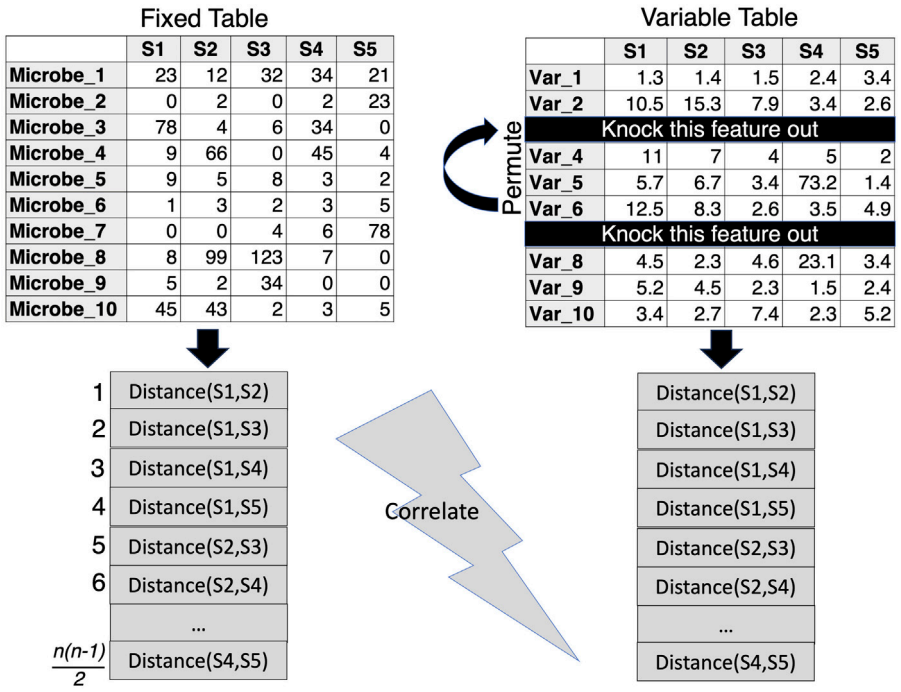


FIGURE 4
BIOENV approach, where the similarity distance is calculated for a fixed matrix (microbiome) and the features in the variable table are permuted to calculate a variable similarity distance in such a way that those subsets are retained where the correlation between the distances of the two matrices is optimized.

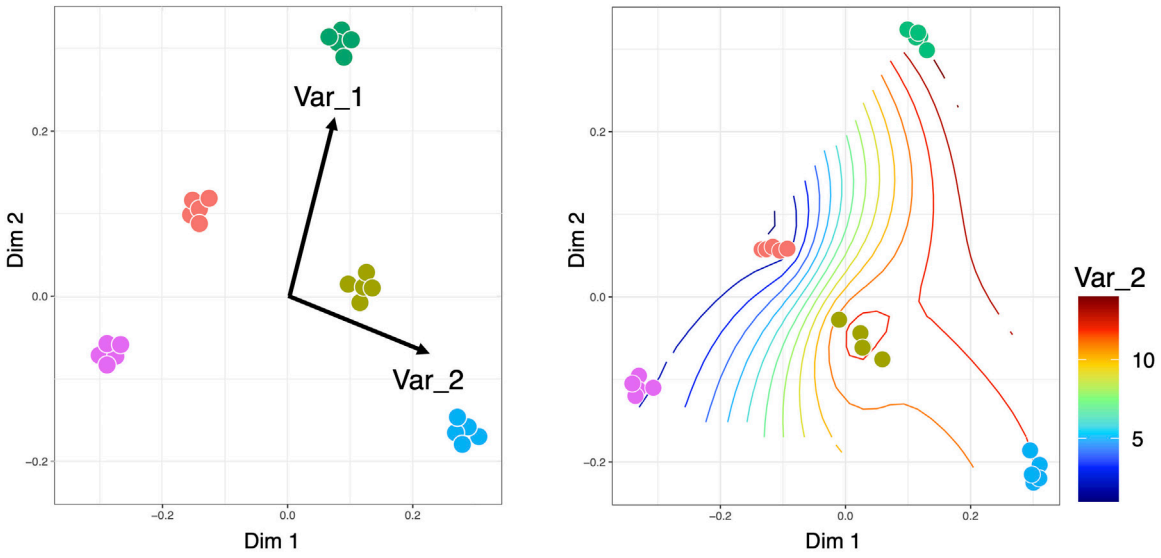


FIGURE 5
Constrained ordinations, where the left side visualization is obtained for CCA/RDA with the length of the arrows pointing to the direction of increase of the continuous covariates, while the right side is the smooth contour of a continuous covariate on the reduced ordination space. It should be noted that for CCA/RDA categorical variables can also be used where we get separate vectors for each factor of a categorical variable.

other hand uses ordinary unweighted linear regression and unweighted SVD. Additionally, there is a distance-based Redundancy Analysis (dbRDA) that allows non-Euclidean dissimilarity indices such as the Bray-Curtis distance. There are two versions of dbRDA implemented in R's vegan package: the capscale() function based on (Anderson and Legendre, 1999); and the dbrda() function based on (McArdle and Anderson, 2001). The two methods differ in how dissimilarities are handled but they essentially do the same thing. To facilitate stepwise model building for constrained ordination methods, one can use the ordistep() function (Blanchet et al., 2008) from R's Vegan package, which can perform forward, backward, and stepwise model selection using a permutation test. Alternatively, the ordisurf() function from R's Vegan package can fit smooth surfaces using penalized splines (Wood, 2003) in the *Generalized Additive Model* (GAM). The method uses a single continuous metadata and regresses it against the smooth values of the scores obtained from any of the ordination techniques such as Non-Metric Distance Scaling (NMDS) or Principal Coordinate Analysis (PCoA). Nevertheless, there are challenges that need to be addressed in constrained ordination approaches. One of the shortcomings is the assumption of linearity in approximating the response of microbial species to environmental gradients (which typically follow a log-linear relationship). Although (Makarenkov and Legendre, 2002) provided a non-linear approach based on polynomial regression, more research is required in this area. Another problem is that CCA does not work well when there is a large variability in the library sizes of microbial samples (typical of metagenomics datasets), often leading to inflated Type 1 errors (Ter Braak and Te Beest, 2022). Further research is needed on the choice of test statistics associated with CCA to address this issue (Ter Braak, 2022).

Grand challenge: how can we assess stability and complexity in microbiome studies?

Many studies today are concerned with assessing the stability of an observed microbial ecosystem against environmental perturbations (Mills et al., 2023) and whether the structure of the microbiome offers some sort of resilience. Typically, it is assumed that higher taxonomic diversity leads to higher functional redundancy which may provide an advantage when individual taxa are displaced or knocked out (the author's paper above suggests otherwise). The challenge here is to come up with easy-to-use metrics to assess how stable an ecosystem is. Stability can be defined in terms of functional stability. For example, in (Eng and Borenstein, 2018a), artificial perturbations in taxonomic composition are created, and function $f = \frac{1}{e^a} t^b$ is fitted between the taxonomic difference t (using Weighted UniFrac) and the functional difference f (cosine dissimilarity between the original and perturbed functional profiles) of these perturbations to give *Attenuation* a and *Buffering* b coefficients. On a Buffering-Attenuation plot, these authors have compared different environments, showing gut communities to be more robust while vaginal communities to be unstable. However, to apply this procedure, the predicted functional profiles of individual observed taxa need to be known in advance which may be impractical for 16S rRNA studies unless there is an improvement in the database development of metabolic prediction software such as PICRUSt2 (Douglas et al., 2020).

Other approaches stem from May's stability theory (May 1972), which states that the stability of n interacting microbial species is determined by the interacting community matrix M (also called the "adjacency matrix" obtained by network inference). The complexity is defined as $\alpha^2 nC$, where α^2 and C are the variance and density of the non-zero off-diagonal components of M. With the scaled diagonal components as −1, the ecosystem is stable as long as it satisfies the stability criterion: $\alpha\sqrt{nC} < 1$. There are issues associated with the application of this theory to microbial interaction networks:

a) It is not possible to accurately reconstruct an interaction network from abundance data, as this would require high-quality time series data, and well-designed control experiments.
b) The current network reconstruction approaches (e.g., SPIEC-EASI (Kurtz et al., 2015), SparCC (Friedman and Alm, 2012), Phi statistics (Lovell et al., 2015), Probabilistic co-occurrence (Veech, 2013), MENA (Deng et al., 2012), etc.) that infer the community matrix M as a co-occurrence network are not really useful because they do not encode causal relationships.

The second problem can be addressed by estimating the effective connectance D^2 after fitting a regression model to samples that overlap in terms of the species they share and the sample dissimilarities (Yonatan et al., 2022). This avoids the need to infer a co-occurrence relationship explicitly, leading to D^2 serving as a proxy for stability. D^2 is then obtained by the slope of the regression fitted to the dissimilarity-overlap plot on the 25% top overlap values for the paired-wise dissimilarity/overlap values for N samples in a given category from a total of $N(N-1)/2$ paired-wise values. However, this approach requires at least 35 biological replicates which may be impractical for most studies.

For temporal datasets, a community-level measure of stability can be the *Local Contribution to Beta Diversity* (LCBD) measure (Legendre and De Cáceres, 2013). Any deviation from the mean LCBD can serve as a means to assess the stability of the system (Ijaz et al., 2018). Another advantage is that LCBD is a unidimensional measure that can be used in the regression approaches discussed previously and can be studied in the presence of covariates.

Moving on from community-level stability metrics to the identification of subcommunities that remain stable is another challenging issue. Not much work has been done in this direction, and it is still in its infancy. An important development in finding a minimal subset of microbes that either remain stable or change with respect to a continuous covariate of interest, is the Ensemble Quotient Optimization (EQO) approach by (Shan et al., 2023) (Figure 6). The approach uses a relative abundance table, called the community matrix $\boldsymbol{M}$ (m microbes over n samples), where the goal is to obtain a vector $\boldsymbol{x} \in (0,1)^P$ where the ith position in the vector is either 0 or 1, i.e., a subset of microbes where values of 1 belong to an ensemble that we are interested in recovering. This ensemble is recovered in the context of a phenotype/predictor variable $\boldsymbol{y}$ by optimizing an *Ensemble Quotient* $EQ = \frac{\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x}}$, through a genetic algorithm (an optimization algorithm), where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are algebraic transformations of the community matrix that capture

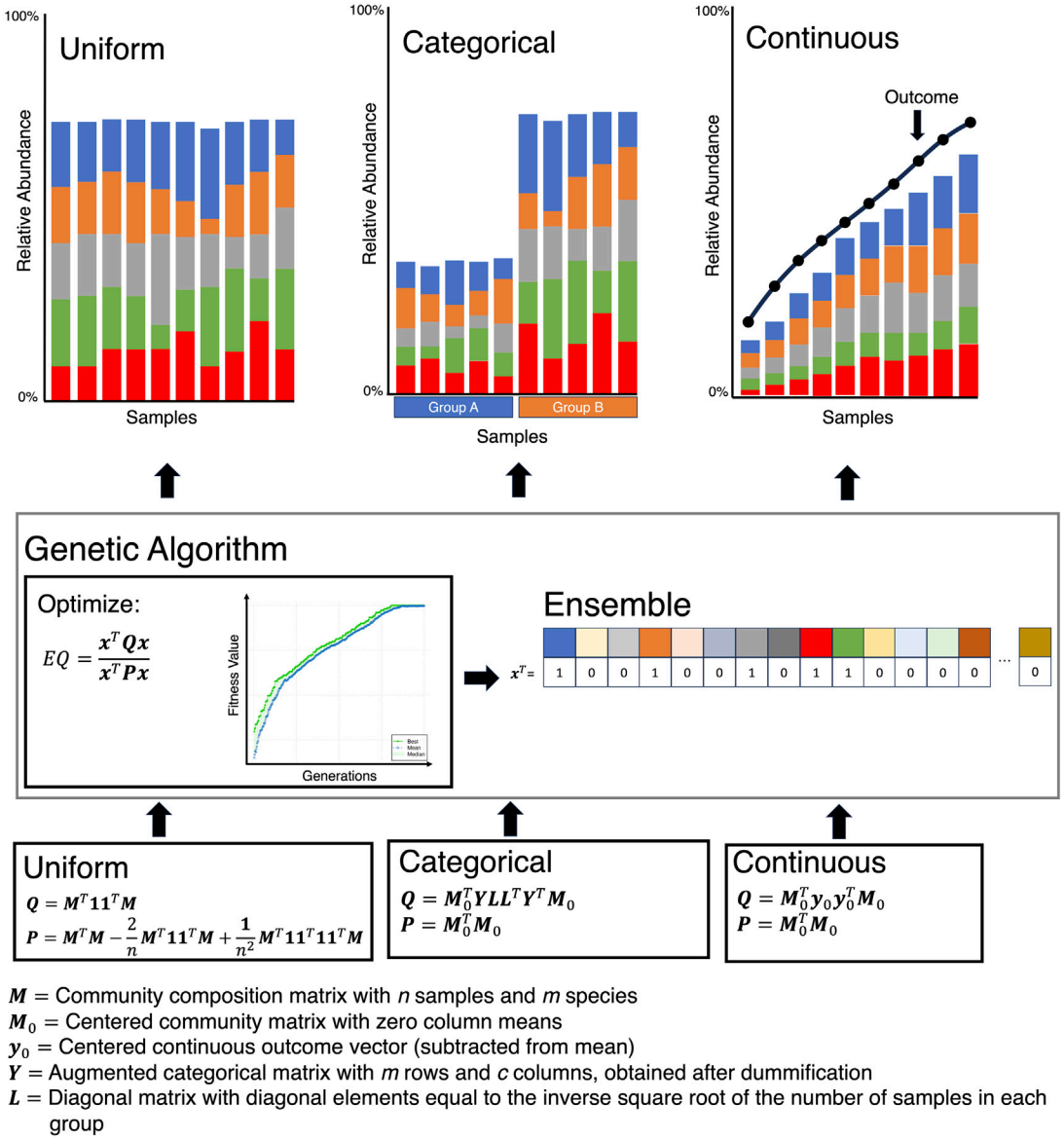


FIGURE 6
The Ensemble Quotient optimization algorithm works by finding an ensemble (a subset of microbes) that either remains stable (Uniform), has a stratified response (Categorical), or correlates with a covariate of interest (Continuous).

the covariance between microbes, and the covariance between microbes and y. The choice of y dictates which ensemble one can recover, and it can be used in three cases: a) If the interest lies in an ensemble of microbes *that remains stable* for a set of samples, then y is considered uniform, i.e., consisting of 1s; b) If the interest lies in an ensemble of species whose cumulative abundance correlates with a continuous physico-chemical parameter y, then the algorithm is optimized with a centered community matrix M_0, and a centered continuous parameter y_0; and c) if the interest lies in an ensemble of species whose cumulative abundance is stratified across different categories, then an augmented Y matrix is considered that captures the categorical information as 1s or 0s after applying dummification, and uses M_0. In the context of temporal data, case (a) can be used to see which subset of microbes does not change over the whole time span (quality of fit is returned as the Coefficient of Variation CV), while case (b) can be used to see which subset of microbes has a relationship with the performance parameters (quality of fit is returned as the correlation coefficient between the continuous outcome and the cumulative abundance of the ensemble). In case (c) of a stratified response, the quality of fit is established by the Coefficient of Determination CD. To optimize the EQ to obtain x, the genetic algorithm optimization code is located at https://github.com/Xiaoyu2425/Ensemble-Quotient-Optimization.

In summary, the following questions need to be addressed:

- How can we construct microbial networks that can accurately capture microbial interactions including causality, and that too with reduced sample numbers?

- Can we construct easy-to-use metrics that can describe the complexity and stability of an ecosystem at both temporal and spatial scales, and that can seamlessly integrate with metadata?
- How can we numerically assess the resilience of an observed microbial community?
- What is the role of a stable subcommunity in the functioning of a microbial ecosystem?
- While the existing literature focuses on taxonomic stability, how can we recover stable functions from microbial ecosystems and relate them to the covariates (metadata)?

Grand challenge: should more emphasis be given to taxa-centric approaches as opposed to study-centric ones?

The taxa-centric approach differs from the study-centric approach in that the emphasis is on elucidating how a particular microbe behaves in a variety of environments. In ecology, one of the important methods is to assess which niches microbes occupy, and whether there is a degree of overlap between them. For such an assessment, it is important to consider all possible sets of environments (dictated by biotic or abiotic variation), with the total number of environments serving as a parameter in the model. To identify the roles of microbes in the context of these environments, R's MicroNiche package (Finn et al., 2020) is useful. It facilitates the identification of *generalist* (which should exist in the majority of the environments) and *specialist* (which should exist in some environments) microbial species in addition to the environment-dependent positive/negative association of microbial species with the continuous covariates observed in the data. Why is this important? There is growing literature suggesting that generalist and specialist species impact the microbial community dynamics differently (Sriswasdi et al., 2017), with generalists in particular playing a key role in maintaining taxonomic diversity. In fact, an author's recent work (Mills, 2023) suggests that different microbial communities are disproportionately impacted by environmental disturbances, and stable environments lead to the proliferation of generalist species. Therefore, there is a need to consider the ecological roles of microbial species, and to distinguish between different categories of microbial species when studying a microbial ecosystem (Xu et al., 2022). How to identify distinct roles remains a challenge. Here, we discuss the recent taxa-centric approaches.

Before making distinctions in the MicroNiche framework, as a pre-processing step, microbes are first selected using the limit of quantification (LOQ) approach. Briefly, LOQ filters out microbes that fall below a "decision boundary", calculated from the distribution of microbes with 95% confidence that these microbes will fall within a null distribution where the mean microbial abundance is zero. To calculate the standard deviation of the null distribution, the lognormal rank distribution of the microbes with the dataset is fitted with $S(R) = S_0 e^{-a^2 R^2}$ where the log abundance of the microbe S at rank R is dependent on the coefficient a and rank R calculated as $a = \sqrt{\frac{\ln S_0}{S_m}/R^2}$ where S_m is the lowest taxon abundance of S. To calculate the LOQ, we fit the above log-normal model to the data, and the LOQ is then determined as the overlap between the null hypothesis (i.e., a microbe's mean abundance is zero) and where the microbe falls within 1 standard deviation of the above model. After filtering out the microbes, we then calculated the niche breadth as Levins' $B_N = \frac{1}{R}\sum_{i=1} p_i^2$, where p_i is the proportional abundance of a microbe in the i-th environment, with the total number of environments being R. If B_N approaches 1 for a given microbe, then it is considered a "generalist", while if it approaches $1/R$, then it can be considered a "specialist". To derive the p-value for Levins' B_N i.e., whether we can call a microbe a generalist or a specialist with a high degree of certainty, a null modeling approach is used, where a random normal distribution of 999 possible niche breadths are produced for a microbe, and allows a p-value to be assigned depending on whether a microbe's B_N is greater or lower than the mean of the null model. As per the author's recommendation, after applying null modeling, the fifth Quantile and 95th Quantile are obtained to tag the microbes as specialists if its B_N < fifth Quantile, and generalists if its B_N > 95th Quantile. Those that fell in the inter-range were tagged as undecided.

In the second step, the overlap of these specialist or generalist microbes is calculated using Levins' Overlap formula $\mathrm{LO}_{i,j} = \frac{\sum_{i,j=1}(p_{ir})(p_{jr})}{\sum_{i=1}(p_{ir}^2)}$, where p_i is the proportional abundance of the microbe i in the r-th environment, and p_j is the abundance of the microbe j in the r-th environment. In addition to Levins' B_N, one can calculate Hurlbert's B_N, where an additional covariate r_i observed for the environment i is incorporated in the formula (Figure 7). The model yields a value between 0 and 1 for each microbe and corresponding covariate, indicating whether there is an inverse (~0) or a positive relationship (~1), with 0.5 indicating no relationship to the covariate. To determine positive and negative relationships (potentially symbiosis and antagonism) between the microbes, Proportional Overlap $\mathrm{PO}_{i,j}$ is used. The Proportional Overlap $\mathrm{PO}_{i,j}$ is a Jaccard similarity coefficient that approaches 0 for microbe pairs that are inversely related to each other and approaches 1 for microbe pairs that are positively related to each other. Similar to the above approach is the development of the *Social Niche Breadth* score (von Meijenfeldt et al., 2023) which revealed across ~22,000 environments that social generalists have a diverse pan-genome, are mainly opportunistic, and dominate local communities. Social specialists, on the other hand, exhibit mixed behavior, i.e., they are stable but low in abundance, and their genome sizes change with the diversity of their environments.

Another way of imparting distinction is to tag microbial species as either *specific* (existing within a narrow range of a particular covariate) *or cosmopolitan* (existing in a broader range of a particular covariate). In this regard, R's Specificity Package (Darcy et al., 2022) is an important development (see Figure 8). It calculates Rao's Quadratic Entropy (RQE) as $RQE = \sum_{i=1}^{s-1}\sum_{j=i+1}^{s} D_{ij}p_i p_j$ where microbial abundance $p_i p_j$ is the multiplication of the abundance of a specific microbe in samples i and j, respectively, each weighted by the difference in the covariate value D_{ij}. A null modeling procedure is then applied (statistical effect size) where 999 random permutations are obtained for the abundance table, and RQE values are then obtained for these random permutations. The deviation of the original RQE from

10.3389/fsysb.2024.1432791

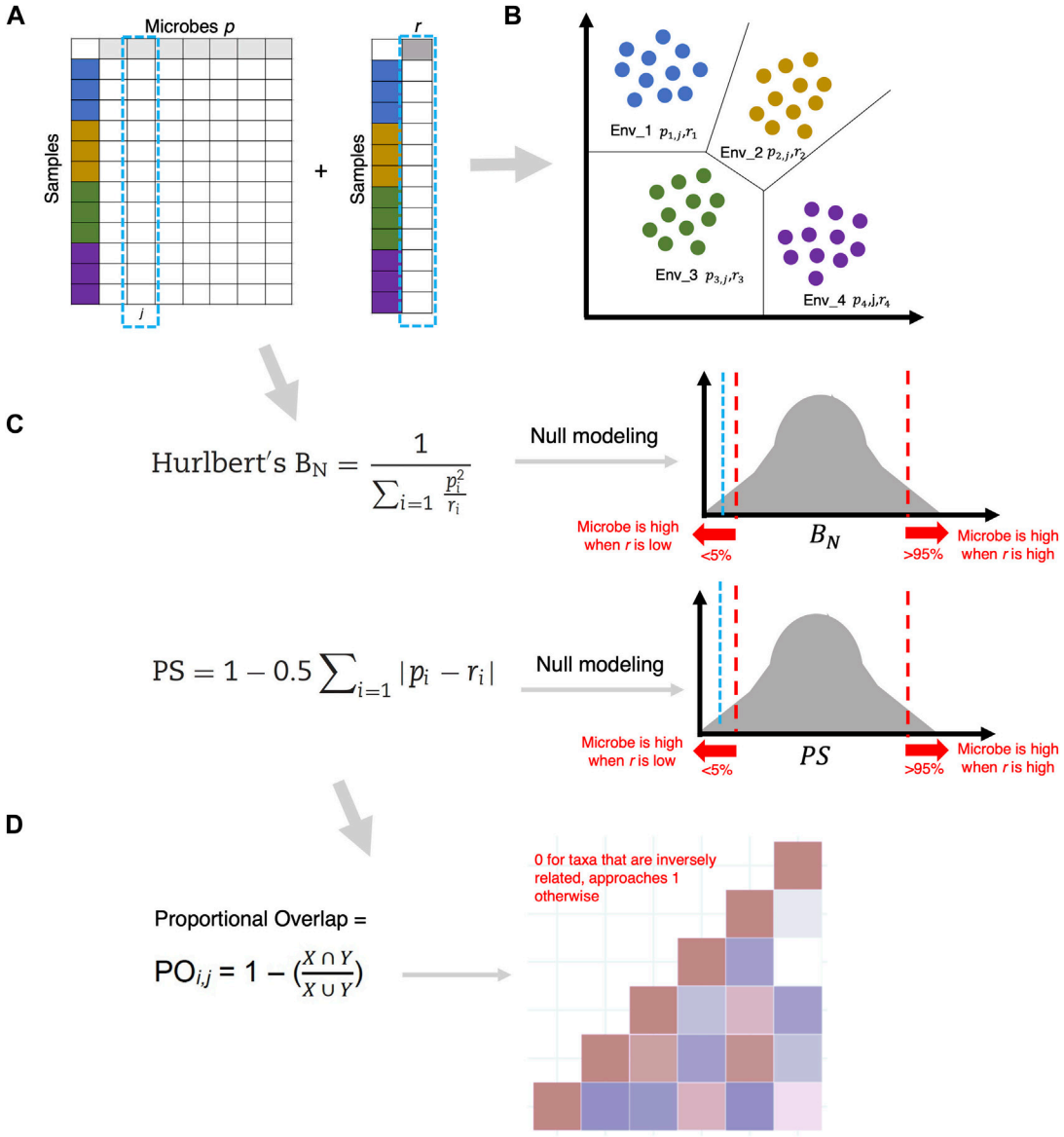


FIGURE 7
(A) Given p_i, the proportional abundance of a microbe in the i-th environment, r_i an environmental property (continuous covariate) is associated with each environment. **(B)** In the total number of environments being R (stratification based on some commonality among samples), we can assume for p_i, and r_i to be fairly stable within the environment and vary more across the environments. **(C)** Hurlbert's B_N and Feinsinger's PS statistic indicates a negative or positive relationship between a microbe and an environmental property. A null modeling procedure is considered by generating a random normal distribution with 999 possible estimates, and by labeling a microbe as "negative" if the metric <fifth Quantile, and "positive", if the metric >95th Quantile. **(D)** For Proportional Overlap $PO_{i,j}$, X (for microbe i) and Y (for microbe j) are Feinsinger's PS.

the average of the RQEs of these random permutations then returns a "Spec" number, ranging from −1 to +1, with 0 as the null hypothesis that the genus weights are randomly ordered with respect to sample identity, with perfect *specificity* when Spec approaches −1 and perfect *cosmopolitanism* when Spec approaches +1, and with the null modeling procedure providing additional *p*-values for significance.

Another popular approach to identifying ecological classes is to fit a neutral model (Burns et al., 2016) to the observed microbial abundance-occupancy relationships. This allows the separation of microbial community members into three subsets: a) those that satisfy the 95% confidence interval of the fitted neutral model, and are driven by stochastic processes; b) those that fall above the 95% confidence of the neutral model and are selected by the environment; and c) those that fall below the model, and are driven by the dispersal limitation process (Figure 9).

In summary, the challenges are as follows:

- With the expansion of public databases with deposited sequences, there is a need to understand microbial niche breadth in a meta-analysis setting and to revisit the definition of stability, particularly in light of how niche theory unravels eco-evolutionary processes.

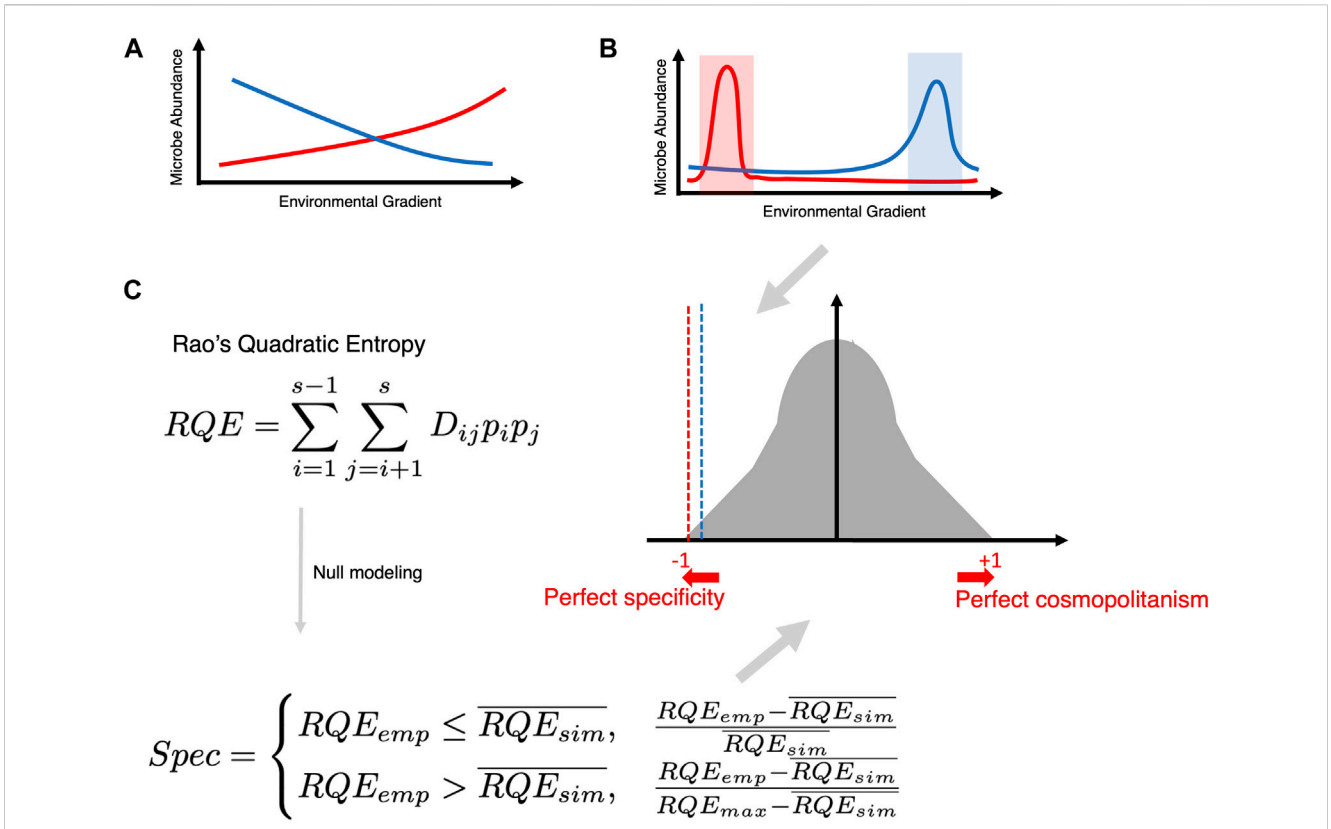


FIGURE 8
(A) Traditional way of finding an association (whether negative or positive) between microbial abundance and a continuous covariate of interest. **(B)** The goal of specificity analysis is to find microbes that become specific within a narrow range of a covariate of interest. **(C)** Creating a null distribution utilizing Rao's Quadratic Entropy formula (which incorporates the difference of the covariate across two samples *i* and *j* as D_{ij}, multiplied by the relative abundance of the microbe across two samples as p_i and p_j). These are calculated for all pairwise combinations of samples ($\frac{s(s-1)}{2}$). The empirical value is then compared with the average of the 999 null distribution values to tag a microbe as specific (Spec ~ −1) or cosmopolitan (Spec ~ +1).

- New metrics need to be developed that identify the ecological roles of microbes for better mechanistic understanding rather than simply identifying microbial species that are up- or downregulated.
- A taxa-centric framework needs to be developed that also incorporates covariates (metadata) into the modeling process based on the hypothesized distinctions.

Grand challenge: how do you identify a signature microbiome when each ecosystem differs in terms of variability?

One of the less explored areas is the recovery of the core/signature microbiome, which is shared by the majority of samples (or individuals in clinical settings). Traditionally, a core microbiome has been defined as a subset of microbes with high prevalence (typically 50% or 85%) across all samples (Shetty et al., 2017). This prevalence goes down to 30% where core membership is considered stable given individual variabilities (Ainsworth et al., 2015). There is no real consensus on what is an appropriate threshold. Also, different ecosystems have different levels of inter-subject variability. For example, while gut microbial communities may be more similar, vaginal microbial communities show much more variability (Eng and Borenstein, 2018a). Therefore, the challenge is to develop a unified framework where the crisp prevalence threshold is avoided, and the core membership of microbes is dynamically learned from the data. There exists one such recent dynamic strategy (Shade and Stopnisek, 2019) for inferring the core microbiome. The strategy considers (Figure 9) the sample occupancy of microbes at different sites (whether in space or time) along with the replicate information, and then dynamically calculates the minimum occupancy threshold by learning from the data. The ranking of microbes is done using a combination of two metrics: *Site-specific occupancy* (the proportion of microbes within a given site; and *Replicate Consistency* (the consistency of microbes across replicates within a site). After ranking the microbes using the two metrics, the subset of core taxa is constructed by iteratively adding one microbe at a time to the core set of microbes, i.e., from the high-ranked microbes to the low-ranked ones. The contribution of the core subset to beta diversity is then calculated every time a new microbe becomes a member of the core set using the Bray-Curtis contribution, $C = 1 - \frac{BC_{core}}{BC_{all}}$. There are two stopping criteria used in (Shade and Stopnisek, 2019), of which a relaxed criterion for inclusion of a microbe in the core microbiome subset is recommended:

10.3389/fsysb.2024.1432791

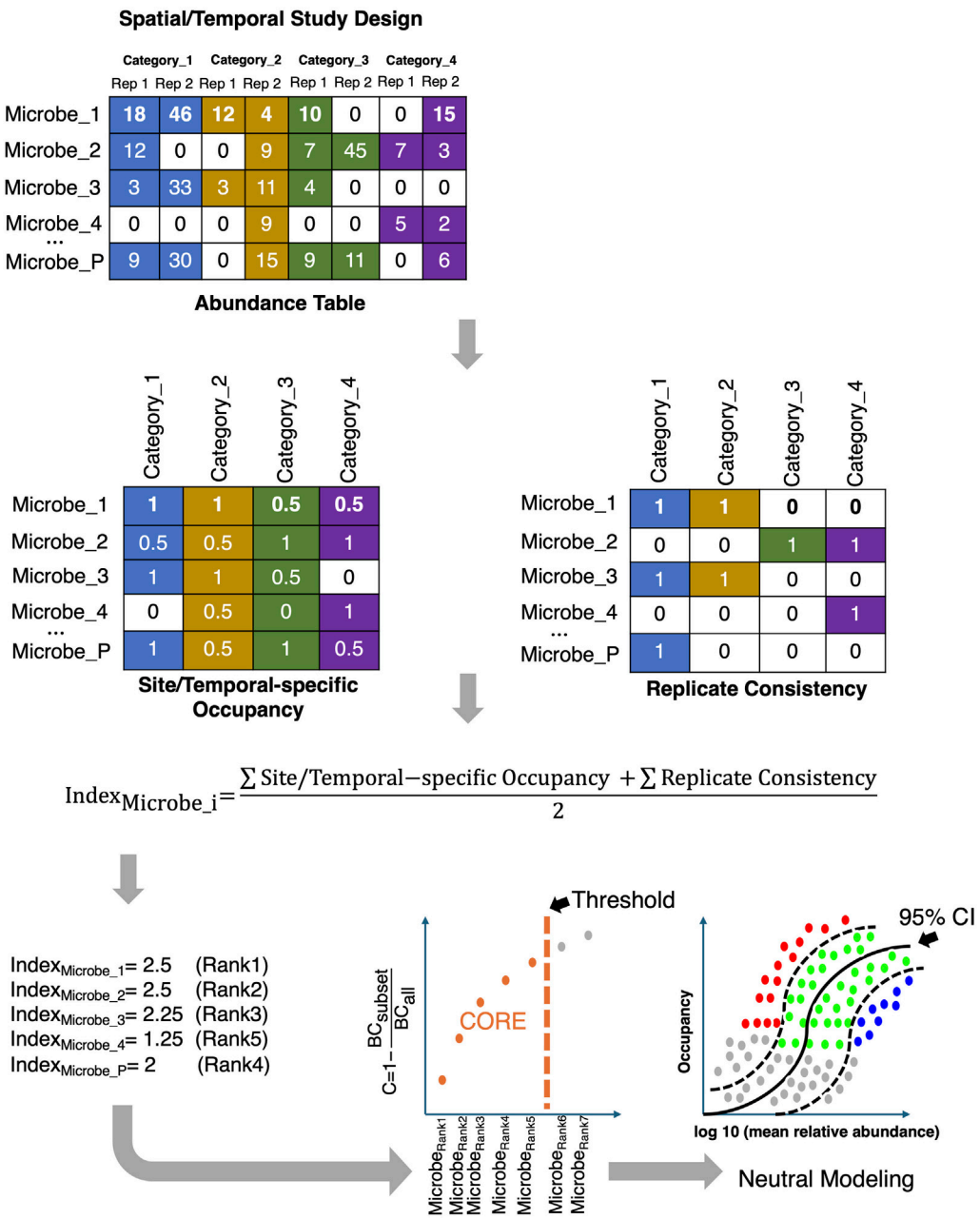
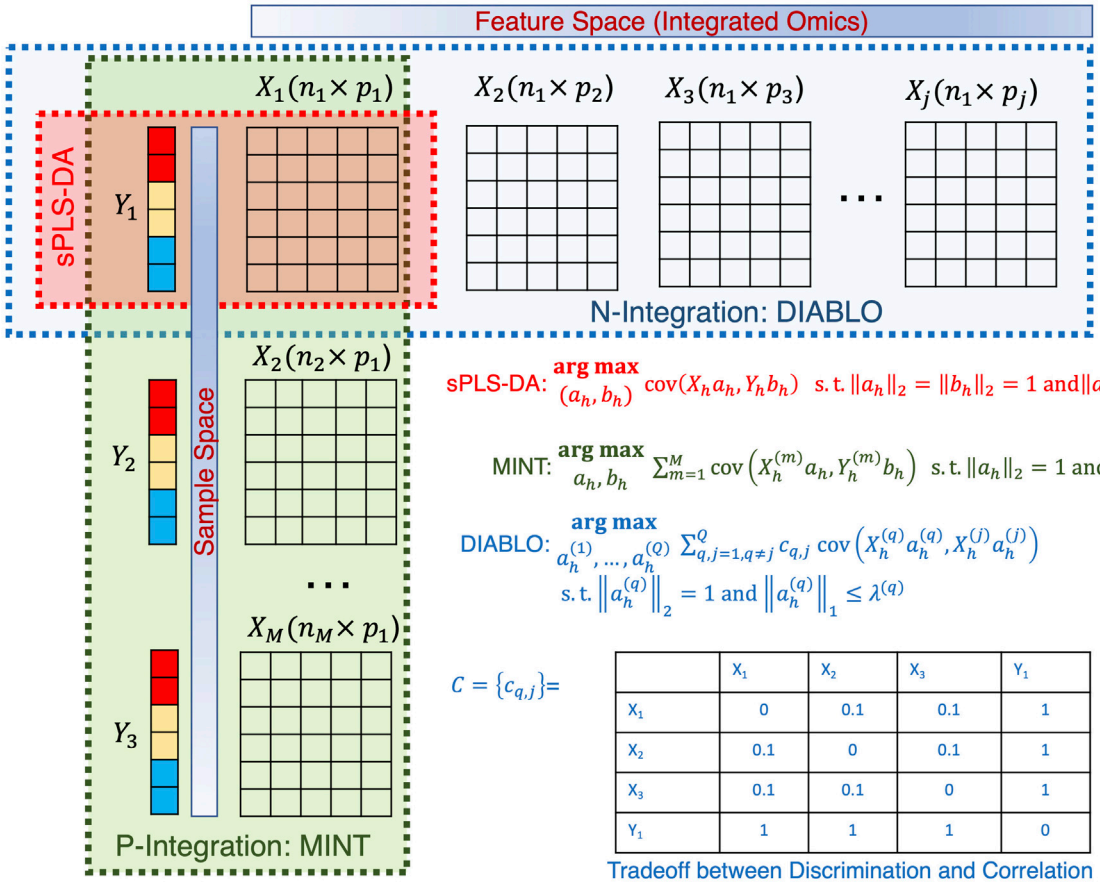


FIGURE 9
The core microbiome is established by first ranking the microbes based on a combination of two criteria, *Site-specific Occupancy* and *Replicate Consistency*, and then using the Bray-Curtis (BC) contribution of the core subset to the whole community; the core subset is iteratively constructed until the stopping criteria are reached. The core subset is then further discretized to those that: fall above the neutral model (selected by host/environment and represented by red dots in the bottom right figure); fall below the neutral model (selected by dispersal limitation and represented by blue dots); and those that fit the 95% CI of the neutral model (represented by green dots).

inclusion of an additional microbe does not cause more than a 2% increase in the explanatory value by Bray-Curtis distance. Although not extensively tested on a variety of datasets, this strategy seems promising (in the absence of alternatives) and needs to be explored further including through the development of different metrics to the Bray-Curtis contribution *C*, and other stopping criteria that may be ecologically inspired. The identified core microbiome can then be further regressed against all sources of variation, for example, using the GLLVM framework or can be fitted with a neutral model.

Grand challenge: how can we integrate additional datasets?

With the multitude of complex 'omics datasets that can be attached to the microbiome samples of interest, the process of integrating additional datasets is of great interest. The challenge with 'omics data is the sheer volume of data (with many features), high noise, sparsity, and potentially missing data points. To overcome these issues and integrate the datasets (often with differing sample numbers), we need to apply tools that reduce



	X_1	X_2	X_3	Y_1
X_1	0	0.1	0.1	1
X_2	0.1	0	0.1	1
X_3	0.1	0.1	0	1
Y_1	1	1	1	0

FIGURE 10
Illustration of the sPLSDA algorithm and its extension. The optimization algorithms are shown for all cases along with the weighting strategy for DIABLO shown in a table.

the dimensionality of the datasets. One tool commonly used (and variations thereof) is Partial Least-Squares Discriminant Analysis (PLS-DA) (Worley and Powers, 2012). Arguably one of the most crucial developments in microbiome data integration has been the development of 'mixOmics' an R package that holds a repository of functions for multivariate analysis of biological data such as dimensionality reduction, and visualization and includes multiple integration tools (Rohart et al., 2017b). This powerful resource allows the integration of microbiome and 'omics datasets. This package focuses on dimensionality reduction by statistically integrating several datasets using *Projection to Latent Structure* models and their multigroup extensions. The multivariate approaches project the sample matrix X into H latent components giving scores of samples on these components $(t_1, t_2, \ldots, t_H)$ which are defined as a linear combination of the original predictors. The weights of each of the predictors are given by the loading vectors on these components as $(a_1, a_2, \ldots, a_H)$. The matrix $X = (X_1, X_2, ..., X_P)$ is then represented in the first latent component as $t_1 = Xa_1 = X_1 a_1^1 + \ldots + X_P a_1^P$. For each loading vector a_h, there is one latent component t_H with dimension $H \ll P$. To enable variable selection, the optimization algorithm maximizes the covariance between the scores of two data matrices X_i and X_j as $cov(t_h^i, t_h^j)$ which are subject to LASSO constraints imposed on the loading vectors a_h^i and a_h^j. This forces some of the components of the loading vector to go to zero, thus enabling discrimination. The approach is referred to as *sparse Projection to Latent Structure Discriminant Analysis* (sPLSDA). Two of its extensions are widely used: a) the Multivariate INTegrative (MINT) algorithm (Rohart et al., 2017a) called the P-Integration algorithm where the matrices X_j share the same features in a multifactorial design; and b)the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) algorithm (Singh et al., 2019) called the N-Integration algorithm where the matrices X_j originate from multiple modalities each with different features but on the same samples. The optimization strategy is similar to sPLSDA where now the covariance of the scores between multiple matrices is simultaneously optimized either as a simple sum of covariances (MINT) or as a weighted sum of covariances(DIABLO). The weights give DIABLO a trade-off between correlation and discrimination (see Figure 10). There are several challenges associated with the applicability of such approaches:

- The main challenge with these approaches is the appropriate type of normalization model. For datasets where the features are count data, TSS + CLR (Total Sum Scaling followed by Centralized Log Ratio) may suffice. For other types of datasets such as those originating from flow cytometry or

metabolomics, there is no real consensus on what is an appropriate normalization measure.
- The P- and N-integration approaches involve optimizing the additive sum of weighted covariances across multiple datasets. Identification of the correct weights that offer reasonable trade-offs between discrimination and correlation is a largely unexplored topic.
- While extensions such as timeOmics (Bodein et al., 2022) have been proposed for temporal datasets that primarily return clustering of time series data across different datasets, the method shows poor performance when there is high inter-replicate variability. Normalization strategies, interpolation strategies (when time points do not match), and pre-processing strategies still need to be further explored.

Concluding remarks

While there are numerous ways to divulge patterns of interest in microbiome data, and associate them with covariates of interest (metadata), we have discussed those methods that have gained importance in recent years, and this list is by no means exhaustive. Statistical approaches that offer multivariate data integration are few and far between, and those that are used in routine practice have very strong assumptions of linearity. There is room for improvement in these techniques and guided by our experience, we have highlighted the challenges associated with some of these approaches. There is an ever-increasing pressure to utilize analytical techniques that lead to a mechanistic understanding of the ecosystem under study, and perhaps inspired by the work done in ecology, there may be a way. Integration algorithms in recent years have also gained popularity as there is a shift toward incorporating multiple omics technologies in microbiome surveys each offering complementarity. However, we are still far from being able to infer direct causality, although correlation and association inference are well explored. Also, there is a need to develop analytical strategies, that can address nonlinearity, low sample numbers, and unbalanced study designs. Analytical techniques now offer more insight than ever before, but verifying the patterns in the lab or *in situ* is still required to not only yield mechanistic insights, but also to aid in tool development.

Author contributions

UI: Writing–original draft, Funding acquisition, Conceptualization. AA: Writing–review and editing. FS: Writing–review and editing. FG: Writing–review and editing. CK: Writing–review and editing. SJ: Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. UI is funded by EPSRC (EP/W037475/1 and EP/V030515/1) and BBSRC (BB/T010657/1).

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Ainsworth, T. D., Krause, L., Bridge, T., Torda, G., Raina, J. B., Zakrzewski, M., et al. (2015). The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *ISME J.* 9 (10), 2261–2274. doi:10.1038/ISMEJ.2015.39

Alekseyenko, A. V. (2016). Multivariate Welch t-test on distances. *Bioinforma. Oxf. Engl.* 32 (23), 3552–3558. doi:10.1093/BIOINFORMATICS/BTW524

Anderson, M. J., and Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Stat. Comput. Simul.* 62 (3), 271–303. doi:10.1080/00949659908811936

Andries, E., and Nikzad-Langerodi, R. (2022). Dual-constrained and primal-constrained principal component analysis. *J. Chemom.* 36 (5), e3403. doi:10.1002/CEM.3403

Armengaud, J. (2023). Metaproteomics to understand how microbiota function: the crystal ball predicts a promising future. *Environ. Microbiol.* 25 (1), 115–125. doi:10.1111/1462-2920.16238

Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K., and Dorrestein, P. C. (2021). Mass spectrometry-based metabolomics in microbiome investigations. *Nat. Rev. Microbiol.* 20, 143–160. doi:10.1038/s41579-021-00621-9

Blanchet, F. G., Legendre, P., and Borcard, D. (2008). FORWARD SELECTION OF EXPLANATORY VARIABLES. *Ecology* 89 (9), 2623–2632. doi:10.1890/07-0986.1

Bodein, A., Scott-Boyer, M. P., Perin, O., Lê Cao, K. A., and Droit, A. (2022). timeOmics: an R package for longitudinal multi-omics data integration. *Bioinforma. Oxf. Engl.* 38 (2), 577–579. doi:10.1093/BIOINFORMATICS/BTAB664

Burns, A. R., Stephens, W. Z., Stagaman, K., Wong, S., Rawls, J. F., Guillemin, K., et al. (2016). Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* 10, 655–664. doi:10.1038/ismej.2015.142

Calle, M. L., Pujolassos, M., and Susin, A. (2023). coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinforma.* 24 (1), 82–19. doi:10.1186/s12859-023-05205-3

Clarke, K. R., and Ainsworth, M. (1993). A method of linking multivariate community structure to environmental variables. *Mar. Ecol. Prog. Ser.* 92 (3), 205–219. doi:10.3354/meps092205

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nat. Methods 2014* 11, 359. doi:10.1038/nmeth.2897

Darcy, J. L., Amend, A. S., Swift, S. O. I., Sommers, P. S., and Lozupone, C. A. (2022). specificity: an R package for analysis of feature specificity to environmental and higher dimensional variables, applied to microbiome species data. *BioRxiv* 2021, 467582. doi:10.1101/2021.11.06.467582

Deng, Ye, Jiang, Y.-H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinforma.* 13 (1), 113. doi:10.1186/1471-2105-13-113

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* 78 (13), 4281–4290. doi:10.1021/AC051632C

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14 (6), 927–930. doi:10.1111/J.1654-1103.2003.TB02228.X

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi:10.1038/s41587-020-0548-6

Eng, A., and Borenstein, E. (2018a). Taxa-function robustness in microbial communities. *Microbiome* 6 (1), 45–19. doi:10.1186/s40168-018-0425-4

Finn, D. R., Yu, J., Ilhan, Z. E., Fernandes, V. M. C., Penton, C. R., Krajmalnik-Brown, R., et al. (2020). MicroNiche: an R package for assessing microbial niche breadth and overlap from amplicon sequencing data. *FEMS Microbiol. Ecol.* 96 (8), fiaa131. doi:10.1093/FEMSEC/FIAA131

Firew, T., Sano, E. D., Lee, J. W., Flores, S., Lang, K., Salman, K., et al. (2020). Protecting the front line: a cross-sectional survey analysis of the occupational factors contributing to healthcare workers' infection and psychological distress during the COVID-19 pandemic in the USA. *BMJ Open* 10 (10), e042752. doi:10.1136/BMJOPEN-2020-042752

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLOS Comput. Biol.* 8 (9), e1002687. doi:10.1371/JOURNAL.PCBI.1002687

Gao, P., Shen, X., Zhang, X., Jiang, C., Zhang, S., Zhou, X., et al. (2022). Precision environmental health monitoring by longitudinal exposome and multi-omics profiling. *Genome Res.* 32 (6), 1199–1214. doi:10.1101/GR.276521.121

Gul, F., Herrema, H., Davids, M., Keating, C., Nasir, A., Ijaz, U. Z., et al. (2024). Gut microbial ecology and exposome of a healthy Pakistani cohort. *Gut Pathog.* 16 (1), 5–18. doi:10.1186/s13099-024-00596-x

Hamidi, B., Wallace, K., Vasu, C., and Alekseyenko, A. V. (2019). W*d -test: robust distance-based multivariate analysis of variance. *Microbiome* 7 (1), 51. doi:10.1186/S40168-019-0659-9

Hong, Q., Chen, G., and Tang, Z. Z. (2023). PhyloMed: a phylogeny-based test of mediation effect in microbiome. *Genome Biol.* 24 (1), 72–21. doi:10.1186/s13059-023-02902-3

Hu, Y., Satten, G. A., and Hu, Y. J. (2022). LOCOM: a logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. *Proc. Natl. Acad. Sci. U. S. A.* 119 (30), e2122788119. doi:10.1073/pnas.2122788119

Ijaz, U. Z., Sivaloganathan, L., McKenna, A., Richmond, A., Kelly, C., Linton, M., et al. (2018). Comprehensive longitudinal microbiome analysis of the chicken cecum reveals a shift from competitive to environmental drivers and a window of opportunity for Campylobacter. *Front. Microbiol.* 9 (OCT), 2452. doi:10.3389/fmicb.2018.02452

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10, 5029. doi:10.1038/s41467-019-13036-1

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422. doi:10.1038/s41579-018-0029-9

Koh, H., Li, Y., Zhan, X., Chen, J., and Zhao, N. (2019). A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. *Front. Genet.* 10 (MAY), 453444. doi:10.3389/fgene.2019.00458

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11 (5), e1004226. doi:10.1371/JOURNAL.PCBI.1004226

Layeghifard, M., Li, H., Wang, P. W., Donaldson, S. L., Coburn, B., Clark, S. T., et al. (2019). Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. *Npj Biofilms Microbiomes 2019* 5, 4–12. doi:10.1038/s41522-018-0077-y

Legendre, P., and De Cáceres, M. (2013). Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecol. Lett.* 16 (8), 951–963. doi:10.1111/ELE.12141

Legendre, P., and Legendre, L. (2012). Numerical ecology. Available at: https://books.google.com/books/about/Numerical_Ecology.html?id=6ZBOA-iDviQC.

Liang, J., Bi, G., and Zhan, C. (2020). Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. *Ann. Transl. Med.* 8 (16), 982. doi:10.21037/ATM-2020-57

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* 11 (3), e1004075. doi:10.1371/JOURNAL.PCBI.1004075

Lumley, A. T., Miller, A., Regression, D., Suggests, D., Gpl, L., Lumley, M. T., et al. (2013) *Package ' leaps*.

Lynch, M. D. J., and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* 13 (4), 217–229. doi:10.1038/NRMICRO3400

Makarenkov, V., and Legendre, P. (2002). Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. *Ecology* 83 (4), 1146–1161. doi:10.1890/0012-9658(2002)083[1146:NRAACC]2.0.CO;2

May, R. M. (1972). Will a large complex system be stable? *Nature* 238, 413–414. doi:10.1038/238413a0

McArdle, B. H., and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82 (1), 290–297. doi:10.1890/0012-9658(2001)082[0290:FMMTCD]2.0.CO;2

Mills, S. (2023). Environmental stability promotes generalist taxa which increase resistance to environmental shocks in methanogenic microbial communities. 1–21.

Mills, S., Yen Nguyen, T. P., Ijaz, U. Z., and Lens, P. N. L. (2023). Process stability in expanded granular sludge bed bioreactors enhances resistance to organic load shocks. *J. Environ. Manag.* 342, 118271. doi:10.1016/J.JENVMAN.2023.118271

Niku, J., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019). gllvm: fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods Ecol. Evol.* 10 (12), 2173–2182. doi:10.1111/2041-210X.13303

Ojala, T., Kankuri, E., and Kankainen, M. (2023). Understanding human health through metatranscriptomics. *Trends Mol. Med.* 29 (5), 376–389. doi:10.1016/J.MOLMED.2023.02.002

Qian, X. B., Chen, T., Xu, Y. P., Chen, L., Sun, F. X., Lu, M. P., et al. (2020). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chin. Med. J.* 133 (15), 1844–1855. doi:10.1097/CM9.0000000000000871

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biol. Med.* 16 (1), 4–10. doi:10.20892/J.ISSN.2095-3941.2018.0055

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi:10.1038/nbt.3935

Roberts, D. W. (2009). Comparison of multidimensional fuzzy set ordination with CCA and DB-RDA. *Ecology* 90 (9), 2622–2634. doi:10.1890/07-1673.1

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Lê Cao, K. A. (2017a). MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* 18 (1), 128. doi:10.1186/s12859-017-1553-8

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017b). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13 (11), e1005752. doi:10.1371/journal.pcbi.1005752

Salavaty, A., Ramialison, M., and Currie, P. D. (2020). Integrated value of influence: an integrative method for the identification of the most influential nodes within networks. *Patterns (New York, N.Y.)* 1 (5), 100052. doi:10.1016/J.PATTER.2020.100052

Satten, G. A., Tyx, R. E., Rivera, A. J., and Stanfill, S. (2017). Restoring the duality between principal components of a distance matrix and linear combinations of predictors, with application to studies of the microbiome. *PLOS ONE* 12 (1), e0168131. doi:10.1371/JOURNAL.PONE.0168131

Shade, A., and Stopnisek, N. (2019). Abundance-occupancy distributions to prioritize plant core microbiome membership. *Curr. Opin. Microbiol.* 49, 50–58. doi:10.1016/J.MIB.2019.09.008

Shan, X., Goyal, A., Gregor, R., and Cordero, O. X. (2023). Annotation-free discovery of functional groups in microbial communities. *Nat. Ecol. Evol.* 7, 716–724. doi:10.1038/s41559-023-02021-z

Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H., and de Vos, W. M. (2017). Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* 41 (2), 182–199. doi:10.1093/FEMSRE/FUW045

Siebert, J. C., Görg, C., Palmer, B., and Lozupone, C. (2019). Visualizing microbiome-immune system interplay. *Immunotherapy* 11 (2), 63–67. doi:10.2217/imt-2018-0138

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 35 (17), 3055–3062. doi:10.1093/BIOINFORMATICS/BTY1054

Sriswasdi, S., Yang, C. C., and Iwasaki, W. (2017). Generalist species drive microbial dispersion and evolution. *Nat. Commun.* 8, 1162–1168. doi:10.1038/s41467-017-01265-1

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* 106, 1. doi:10.18637/JSS.V106.I01

Ter Braak, C. J. F. (2022). Predictor versus response permutation for significance testing in weighted regression and redundancy analysis. *J. Stat. Comput. Simul.* 92 (10), 2041–2059. doi:10.1080/00949655.2021.2019256

Ter Braak, C. J. F., and Te Beest, D. E. (2022). Testing environmental effects on taxonomic composition with canonical correspondence analysis: alternative permutation tests are not equal. *Environ. Ecol. Statistics* 29 (4), 849–868. doi:10.1007/S10651-022-00545-4

Van Der weele, T., and Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiol. Methods* 2 (1), 95–115. doi:10.1515/em-2012-0010

Vass, M., Székely, A. J., Lindström, E. S., and Langenheder, S. (2020). Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. *Sci. Rep. 2020* 10, 2455–2513. doi:10.1038/s41598-020-59182-1

Veech, J. A. (2013). A probabilistic model for analysing species co-occurrence. *Glob. Ecol. Biogeogr.* 22 (2), 252–260. doi:10.1111/J.1466-8238.2012.00789.X

von Meijenfeldt, F. A. B., Hogeweg, P., and Dutilh, B. E. (2023). A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat. Ecol. Evol.* 7, 768–781. doi:10.1038/s41559-023-02027-7

Wang, C., Hu, J., Blaser, M. J., and Li, H. (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics* 36 (2), 347–355. doi:10.1093/BIOINFORMATICS/BTZ565

Wood, S. N. (2003). Thin Plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 65 (1), 95–114. doi:10.1111/1467-9868.00374

Worley, B., and Powers, R. (2012). Multivariate analysis in metabolomics. *Curr. Metabolomics* 1 (1), 92–107. doi:10.2174/2213235X11301010092

Xiao, J., Chen, L., Johnson, S., Yu, Y., Zhang, X., and Chen, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Front. Microbiol.* 9 (JUN), 1391. doi:10.3389/fmicb.2018.01391

Xu, Q., Vandenkoornhuyse, P., Li, L., Guo, J., Zhu, C., Guo, S., et al. (2022). Microbial generalists and specialists differently contribute to the community diversity in farmland soils. *J. Adv. Res.* 40, 17–27. doi:10.1016/J.JARE.2021.12.003

Yonatan, Y., Amit, G., Friedman, J., and Bashan, A. (2022). Complexity–stability trade-off in empirical microbial ecosystems. *Nat. Ecol. Evol.* 6, 693–700. doi:10.1038/s41559-022-01745-8

Yue, Y., and Hu, Y. J. (2022). A new approach to testing mediation of the microbiome at both the community and individual taxon levels. *Bioinformatics* 38 (12), 3173–3180. doi:10.1093/BIOINFORMATICS/BTAC310