Predictive modeling of microbial data with interaction effects

Mara Stadler[1,2], Jacob Bien[3], Christian L. Müller[1,2,4]

1 Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
2 Department of Statistics, Ludwig Maximilians University Munich, 80539 Munich, Germany
3 Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA
4 Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA

* Correspondence: mara.stadler@helmholtz-munich.de

The code for reproducing the analysis and figures is accessible on the following GitHub public repository: `https://github.com/marastadler/Microbial-Interactions`

Abstract

Microbial interactions are of fundamental importance for the functioning and the maintenance of microbial communities. Deciphering these interactions from observational data or controlled lab experiments remains a formidable challenge due to their context-dependent nature, i.e., their dependence on (a)biotic factors, host characteristics, and overall community composition. Here, we present a statistical regression framework for microbial data that allows the inclusion and parsimonious estimation of species interaction effects for an outcome of interest. We adapt the penalized quadratic interaction model to accommodate common microbial data types as predictors, including microbial presence-absence data, relative (or compositional) abundance data from microbiome surveys, and quantitative (absolute abundance) microbiome data. We study the effect of including hierarchical interaction constraints and stability-based model selection on model performance and propose novel interaction model formulations for compositional data. To illustrate our framework's versatility, we consider prediction tasks across a wide range of microbial datasets and ecosystems, including metabolite production in model communities in designed experiments and environmental covariate prediction from marine microbiome data. While we generally observe superior predictive performance of our interaction models, we also assess limits of these models in presence of extreme data sparsity and with respect to data type. On a large-scale gut microbiome cohort data, we identify sparse family-level interaction models that accurately predict the abundance of antimicrobial resistance genes, enabling the formulation of novel biological hypotheses about microbial community interactions and antimicrobial resistance.

1 Introduction

A fundamental objective in microbial ecology is to elucidate how species compositions and species-species interactions are related to the maintenance and functioning of a microbial community [1]. Interactions between microbial species come in many forms, including cross-feeding interactions through metabolite exchange, bacteriocin-induced growth-inhibitory interactions, and exchange of genetic material for genotype selection [2, 3]. Conceptually, microbial interactions can be described in terms of their net positive, negative, or neutral effect on their interaction partner, resulting in broad categories such as mutualistic, commensal, amensal, predatory/parasitic/exploitative, antagonistic or competitive interactions [4, 5, 6, 2]. Experimentally identifying and verifying such interactions within natural communities has remained a difficult task, owing to the sheer complexity of microbial ecosystems and limited technical capabilities to dissect such communities.

With the emergence of large-scale microbial survey data, computational approaches have become popular that use statistical regression and correlation methods to estimate sparse species-species association and co-occurrence networks from microbial abundance measurements [5, 7, 8, 9, 10, 11]. While these networks do not necessarily reflect true ecological relationships [12], they can provide valuable insights into the global structure of microbial communities across ecosystems [13, 14]. However, none of these methods allow to relate species-species associations or "interactions" to a community functional outcome of interest or to concomitant environmental or host-related covariates. Furthermore, most network approaches deliver context-independent (or averaged) pairwise associations, thus potentially missing species-species interactions that are relevant for a specific function of the community.

In this contribution, we provide a statistical regression framework for microbial data that allows the parsimonious inclusion of microbial interactions for predicting an outcome of interest, such as butyrate [15] or a concomitantly measured covariate [16]. Using the generic quadratic interaction regression model as a starting point, we adapt the model to accommodate all common microbial abundance data modalities (see Fig. 1 for an illustration). Important examples include data from designed *in-vitro* experimental studies on model microbial communities where microbial abundance comes in form of presence-absence (binary) data or absolute abundances, i.e., non-negative count or continuous data [17]. The majority of microbiome survey data, however, quantify taxon abundances by amplicon sequencing, thus providing primarily relative abundance (or compositional) data [18] in form of Operational Taxonomic Units (OTUs) or Amplicon Sequencing Variants (ASVs) [19]. Moreover, recent quantitative microbiome profiling techniques [20, 21, 22] combine absolute cell count measurements and relative amplicon data, thus providing absolute microbial abundance information, albeit with potential biases [23].

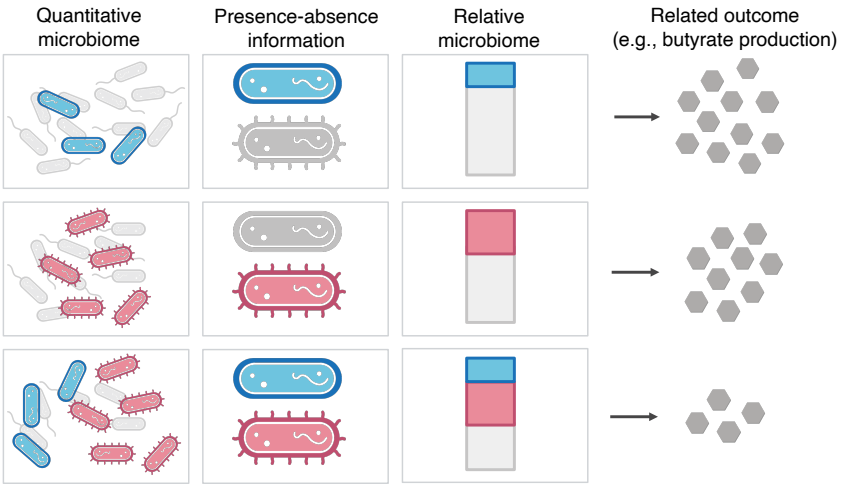

Fig 1. Illustration of three data modalities in microbiome analysis and their combinatorial behavior with respect to an outcome, e.g., a community function (created with BioRender.com). The sketch depicts three distinct data modalities in the columns: (i) quantitative microbiome data, representing absolute counts; (ii) presence-absence information of microbial species; and (iii) relative abundance data, also known as compositions. Each row illustrates a simplified scenario. In the first scenario, blue microbes are present while red are absent, resulting in a large (production of) outcome (e.g., butyrate). In the second scenario, red microbes are present while blue are absent, leading to another large (production of) outcome. In the third scenario, both blue and red groups of microbes are present, yet only minimal amounts of the outcome are produced, indicating an antagonistic combinatorial effect between the two groups.

Our framework unifies and generalizes several seemingly disjoint approaches in the literature of microbial ecology and microbiome data science. For example, several recent studies in microbial ecology use presence-absence and absolute abundance data from designed experiments on small microbial communities to predict community functions [24, 15], such as, e.g., butyrate production, or overall host fitness [25] using the quadratic (and higher-order) interaction model. Our framework is readily available for such studies and gives statistical guidelines how to choose model complexity, how hierarchical constraints can increase model interpretability, and how to analyze higher dimensional datasets.

On the other hand, for regression tasks based on high-dimensional large-scale amplicon sequencing data, many statistical approaches consider the *linear* log-contrast model [26], which is the standard linear model for compositional data, as the baseline model. To deal with the high dimensionality (where typically the number of features p is larger than number of samples n), penalized and structured regression models have been proposed [27, 28, 29, 30, 31, 16]. Here, we extend these linear (main effects) model to include species interaction effects. Specifically, starting with Aitchison's (low- dimensional) proposal [26], we introduce three models with quadratic interactions that work with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic

log-ratio model. To achieve parsimonious models in the high-dimensional setting, we employ ℓ_1 penalization and illustrate via semi-synthetic data simulations when quadratic interactions are identifiable, given the excess sparsity of typical microbiome data.

To achieve stable and interpretable interaction models [32], we follow [33] and incorporate hierarchical interaction modeling [34, 35, 36] and stability-based model selection [37, 38] into our framework. The hierarchy assumption enforces constraints on interaction features, requiring that they can only be included in the model if both features (strong hierarchy) or at least one feature (weak hierarchy) are already present as main effects. Stability-based model selection ensures that interactions are only included if they can be consistently and reproducibly identified across different subsets of the data which will likely help reduce the number of testable biological hypotheses.

We demonstrate the versatility of our framework by analyzing datasets that encompass all three data modalities across various ecosystems, including synthetic microbial communities, human gut microbiomes, and marine microbial ecosystems. Notably, our application of the quadratic interaction model on a quantitative microbiome data from the Metacardis study [39] reveals its effectiveness in accurately estimating the abundance of antimicrobial resistance genes (ARGs) from microbial taxa abundances. Furthermore, our analysis of a microbiome dataset containing presence-absence information rediscovers a stable interaction effect, specifically the inhibitory role of *D. piger* on the butyrate producer *A. caccae* [24]. For the newly introduced *sparse quadratic log-contrast model* tailored for relative microbiome data, we provide both semi-synthetic data simulations to demonstrate the model's ability to accurately detect interaction effects and, following [16], re-analyze Tara ocean data [40], highlighting superior predictive performance of sparse interaction modeling compared to their linear counterparts. We conclude by providing a comparative analysis of the quadratic interaction models across the three data modalities using the Metacardis ARG prediction task, illustrating commonalities and differences across the resulting predictive models. The latter analysis gives further guidance for the practitioner regarding merits and pitfalls of quadratic interaction models. Our framework for quadratic interaction modeling is freely available as reproducible R code at `https://github.com/marastadler/Microbial-Interactions`.

2 Methods

2.1 Interaction modeling strategy

Given the abundance information of p microbial taxa $X = (X_1, ..., X_p)$, the baseline model for uncovering (joint) additive effects of the microbial taxa on an outcome $Y \in \mathbb{R}^n$ (e.g., butyrate production), is the linear model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{1}$$

where β_0 is the intercept term, β_j is the effect of taxon j on Y, and ϵ models the technical and biological noise term.

In many prediction tasks, relying on a linear (main effect) model alone is insufficient to capture the complexity of dynamics within microbial communities. A common approach to introduce a more intricate yet interpretable model is the inclusion of quadratic terms. Here, we extend the baseline model by introducing a generic quadratic interaction model, incorporating all pairwise interactions between microbial taxa, namely

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \epsilon, \tag{2}$$

where $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$ is a symmetric matrix of interactions. We assume $\Theta_{jj} = 0$ in this model formulation. However, the general principles still apply if this constraint is removed.

In the following section, we instantiate the interaction model to accommodate distinct data types and denote the microbial abundance information by A for count information (absolute or relative) and B for presence-absence information (see Fig. 1).

2.1.1 Interaction model for quantitative microbiome data

Whenever microbial abundance information is given as absolute counts, the model is equal to the generic model 2 and does not require further transformation of the input data or any constraints on the model coefficients. Throughout this work, we denote the absolute count input data by $A \in \mathbb{R}_+^{n \times p}$. Assuming that Y depends on the actual amounts of taxon abundances, the quadratic interaction model is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j A_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} A_j A_k + \epsilon, \tag{3}$$

where the model parameters follow the description provided in model 2.

2.1.2 Interaction model for presence-absence microbiome data

If the microbial abundance information is represented as presence-absence data, given by a binary matrix, we denote the microbial abundance information as $B \in \{0, 1\}^{n \times p}$, where 1 indicates the presence of a microbial taxon, and 0 indicates its absence. One common alternative encoding is $B \in \{-1, 1\}^{n \times p}$, where the absence is encoded as -1. The choice of encoding does not affect the ability of the model to fit the data, it only changes the interpretation of the coefficient. Assuming that Y depends on the presence-absence information

of microbial taxa, the quadratic interaction model is given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j B_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} B_j B_k + \epsilon, \tag{4}$$

where the model parameters follow the description provided in model 2. For $B \in \{0, 1\}$, β_0 is the baseline effect when all features, here referring to microbial taxa, are absent, and β_j for $j = 1, ..., p$ represents the effect of the presence of B_j when all other taxa are absent. The interaction term, Θ_{jk}, accounts for the additional effect when both features B_j and B_k are present. For $B \in \{-1, 1\}$, β_0 signifies the overall mean (assuming a completely balanced design). For more details on the interpretation and the linear transformations of model coefficients between these two encodings, see the Supplementary Material. When describing Y as a fitness or phenotypic landscape, the different encodings in the interaction model are often associated with Fourier and Taylor expansions, allowing the parameters to describe landscape properties, like ruggedness [15, 41, 42].

2.1.3 Interaction modeling for relative microbiome data

When the microbial count information in X is provided as (sparse) compositions rather than as absolute counts, one way of modeling interaction effects between microbial taxa includes converting X to a binary matrix $B = \mathbb{1}_{\{X>0\}}$ that carries the presence-absence information of microbial taxa.

However, the compositional information might hold valuable insights beyond that provided by presence-absence data. We introduce three methods for modeling quadratic interactions with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic log-ratio model. The three models differ in terms of interpretability, dimensionality, and optimization, and the choice of which model to use depends on the underlying data and the biological question.

Alr transformed quadratic model While comparing the relative count information in compositional data is not biologically meaningful, describing the response as a linear combination of log-ratios derived from the original compositions is a valid comparison. One popular way of building log-ratios is by choosing a common reference feature p, such that the transformed count is given by $C_j = \log(A_j/A_p),\ j = 1, ..., p-1$ [26]. This transformation is known as the additive log-ratio transformation (alr)-transformation. The alr transformation allows modeling an outcome Y based on the $(p-1)$-dimensional compositional input data, assuming that Y depends on the composition of X, not on the actual amount, by a model linear in the features

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \epsilon \tag{5}$$

and an extension to interaction effects, given by

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j=1}^{p-1} \sum_{k=1}^{p-1} \Theta_{jk} C_j C_k + \epsilon. \tag{6}$$

For $p := p - 1$ this formulation is equal to the generic model 2 (with Θ not being symmetric). While this very general model formulation allows for the interpretation of the effects with respect to a specific reference feature p, extensions to expressions in the p-dimensional space [26] and log-ratio models that allow pairwise comparisons between features [30] have been proposed. Both approaches can be translated back to the model formulation in 5 and 6, respectively, and will be discussed in the following two paragraphs.

Constrained quadratic log-contrast model As shown in [26], a more convenient symmetric expression of the linear alr transformed model 5, that does not require a reference feature, can be derived by reformulating the equation as a p-dimensional problem including a zero-sum constraint, given by

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \epsilon, \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j = 0, \tag{7}$$

where the main (log) effect coefficients $\beta_j, j = 1, ..., p$ sum up to zero. As illustrated in [26], the extension to the quadratic log-contrast model can be represented as

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j \log(A_j) + \frac{1}{2} \sum_{j \neq p} \Theta_{jk} \log(A_j / A_k)^2, \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j = 0, \tag{8}$$

where the main (log) effect coefficients $\beta_j, j = 1, ..., p$ sum up to zero, with $\beta \in \mathbb{R}^p$, and the interaction effect coefficients Θ_{jk} correspond to the quadratic (log-ratio) interaction effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$. In the Supplementary information we show how to formulate the alr transformed model as constrained quadratic log-contrast model.

Quadratic log-ratio model Another way of accounting for compositionality in regression models is to build log-ratios between all possible pairs of features in $A \in \mathbb{R}_+^{n \times p}$. This approach is referred to the (all-pairs) log-ratio model [30], which is given by

$$Y = \beta_0 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j / A_k) + \epsilon, \tag{9}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k.

In the same way as in 8 the log-ratio model can be extended to a quadratic version, the quadratic log-ratio interaction model (qlr), namely,

$$Y = \beta_0 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j/A_k) + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log(A_j/A_k)^2 + \epsilon, \tag{10}$$

where the main effect coefficient $\beta_{j,k}$ corresponds to the pairwise (log-ratio) effect of A_j and A_k, with $\beta \in \mathbb{R}^{p(p-1)/2}$ and the interaction effect coefficient Θ_{jk} corresponds to the quadratic (log-ratio) effect of A_j and A_k, with $\Theta = \Theta^T \in \mathbb{R}^{p \times p}$. There exists a linear transformation between the main effect coefficients β_j in model 7 and model 8 and the main effects coefficients $\beta_{j,k}$ in model 9 and model 10, $\beta_j = -\sum_{k=1}^{j-1} \beta_{k,j} + \sum_{k=j+1}^{p} \beta_{j,k}$, implying that the zero-sum constraint on $\beta \in \mathbb{R}^p$ is inherently met in the linear and quadratic log-ratio model. While the models are mathematically equivalent, their interpretations are different and the choice might depend on the particular data application.

2.2 Penalized model estimation

Microbial datasets typically include a large number of features p and interactions between features $p(p-1)/2$ compared to the number of observations n. Moreover, even in scenarios where $n > p(p1)/2$, we assume that a parsimonious model is most appropriate, focusing only on the selection of few features and interactions that are relevant for the outcome.
To facilitate penalized model estimation, we employ regularized maximum-likelihood estimation incorporating ℓ_1-norm (lasso) penalization for both linear and interaction coefficients, as proposed by [43]. We introduce a generic optimization problem, consisting of an objective function $\rho(l, \beta_0, \beta, \Theta)$ and a (potential) constraint set on the model parameters $c(\beta_0, \beta, \Theta)$ that facilitates parameter estimation for all (linear and interaction) models introduced in Section 2.1. The objective function takes the general form

$$\rho(l, \beta_0, \beta, \Theta) = l(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1 . \tag{11}$$

Here, $\lambda > 0$ serves as a tuning parameter, regulating the sparsity levels of the coefficients β and Θ, respectively. The loss function $l(\beta_0, \beta, \Theta)$ is specific to each model. Consequently, the generic optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{12}$$

This optimization problem is subsequently instantiated by specific loss functions and constraints.

Sparse quadratic interaction model for quantitative and presence-absence microbiome data The loss function $l(\beta_0, \beta, \Theta)$ for the sparse quadratic interaction model, also all-pairs lasso, for the interaction models for absolute count data or presence-absence data, introduced in 3 and 4, is defined as

$$l^{\mathrm{qi}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k \right\|_2^2,$$

with $X := A \in \mathbb{R}_+^{n \times p}$ for absolute count data and $X := B \in \{0, 1\}^{n \times p}$ (or $B \in \{-1, 1\}^{n \times p}$) for presence-absence data. This model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\mathrm{qi}}, \beta_0, \beta, \Theta). \tag{13}$$

In the linear model case the loss function in the optimization problem reduces to $l(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right\|_2^2$.

Sparse alr transformed quadratic model Given $A \in \mathbb{R}^{n \times p}$ is a matrix containing the relative abundance information of p microbial taxa, the loss function in 11 for the sparse alr transformed quadratic model, introduced in 6, is defined as

$$l^{\mathrm{qalr}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j + \frac{1}{2} \sum_{j=1}^{p-1} \sum_{k=1}^{p-1} \Theta_{jk} C_j C_k \right\|_2^2,$$

with $C_j = \log(A_j / A_p)$, $j = 1, ..., p-1$. The model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. Consequently, the optimization problem is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\mathrm{qalr}}, \beta_0, \beta, \Theta). \tag{14}$$

In the linear model case the loss function in the optimization problem reduces to $l^{\mathrm{alr}}(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p-1} \beta_j C_j \right\|_2^2$.

Sparse quadratic log-contrast model The linear log-contrast model has been extended to the high-dimensional setting [31, 27, 28, 16], and is also known as the sparse log-contrast model. While this model has been used in various microbiome data analysis applications, the concept of introducing interactions in the log-contrast model has been defined in [26], but has not been extended to the high-dimensional setting or used in practical applications. Here, we translate the interaction model proposed in [26] to the high-dimensional setting. The loss function for the sparse quadratic log-contrast model (qlc) corresponding to the interaction

model for compositional data, introduced in 8, is defined as

$$l^{\mathrm{qlc}}(\beta_0, \beta, \Theta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) - \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log(A_j/A_k)^2 \right\|_2^2.$$

As this model comes with a zero-sum constraint on the main effect coefficients, the constraint set in 12 is given by

$$c(\beta_0, \beta, \Theta) = \left\{ \sum_{j=1}^{p} \beta_j = 0 \right\}.$$

Thus, the optimization problem for the sparse quadratic log-contrast model is given by

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\mathrm{qlc}}, \beta_0, \beta, \Theta) \text{ s.t. } c(\beta_0, \beta, \Theta). \tag{15}$$

In the linear sparse log-contrast model defined in 7, the loss function reduces to $l^{\mathrm{lc}}(\beta_0, \beta) = \left\| Y - \beta_0 - \sum_{j=1}^{p} \beta_j \log(A_j) \right\|_2^2$.
The main effect covariates, denoted by A_j for $j = 1, ..., p$ typically remain unscaled under the zero-sum constraint. However, the interaction features $\log(A_j/A_k)^2$ are not subject to the zero-sum constraint and we scale them. The ℓ_2-norm of these interaction features tends to increase with the ℓ_2-norm of their associated main effects (more specifically, the ℓ_2-norm of the main effects after transforming them with the centered log-ratio (clr) transformation). The clr divides each compositional part by the geometric mean of all parts, namely

$$\mathrm{clr}(A) = \left(\log \frac{A_i}{g(A_i)} \right)_{i=1,\ldots,n} \text{ with } g(A_i) = \exp\left(\frac{1}{p} \sum_{j=1}^{p} \log(A_{ij}) \right).$$

Here, we introduce a way of scaling the interaction features that ensures equal penalization of the interaction features. Moreover, we adjust the scale of the interaction features to align with the norm of the average clr transformed ℓ_2-norms of all main effects. Mathematically, this can be expressed as follows: We denote each column of the interaction feature matrix as $A^I_{\cdot jk} = \log(A_j/A_k)^2$, with $A^I \in \mathbb{R}^{n \times p(p-1)/2}$, and its scaled version is given by

$$A^I_{\cdot jk} \left(\left\| A^I_{\cdot jk} \right\|_2 \right)^{-1} \frac{1}{p} \sum_{k=1}^{p} \left\| A^{\mathrm{clr}}_k \right\|_2,$$

where $A^{\mathrm{clr}} = \mathrm{clr}(A) \in \mathbb{R}^{n \times p}$ is the clr transformed main effects matrix A and $\left\| A^{\mathrm{clr}}_k \right\|_2$ is the ℓ_2-norm of the k-th column of A^{clr}.

Sparse quadratic log-ratio model The loss function of the sparse quadratic log-ratio (qlr) model corresponding to the interaction model for compositional data, introduced in 10,

is defined as

$$l^{\mathrm{qlr}}(\beta_0, \beta, \Theta) = \frac{1}{2}\left\| Y - \beta_0 - \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j/A_k) - \frac{1}{2} \sum_{j \neq k} \Theta_{jk} \log(A_j/A_k)^2 \right\|_2^2.$$

This model does not require further constraints on the model parameters, such that $c(\beta_0, \beta, \Theta) = \emptyset$. The optimization problem for the sparse quadratic log-ratio model is therefore given as

$$\underset{\beta_0, \beta, \Theta}{\text{minimize}}\, \rho(l^{\mathrm{qlr}}, \beta_0, \beta, \Theta). \tag{16}$$

In the sparse log-ratio model, that is linear in the features and corresponds to the model in 9, the loss function reduces to $l^{\mathrm{lr}}(\beta_0, \beta) = \frac{1}{2}\left\| Y - \beta_0 - \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \beta_{j,k} \log(A_j/A_k) \right\|_2^2$. The $p(p-1)/2$-dimensional sparse log-ratio problem has been shown to be equivalent to the sparse log-contrast model problem for $\lambda^{\mathrm{qlr}} = 2\lambda^{\mathrm{qlc}}$ [30]. This equality can be directly translated to their quadratic extensions. As the dimensionality of the predictor space in the $p(p-1)/2$-dimensional log-ratio model becomes computationally inefficient for large p, the authors in [30] propose a two-stage procedure that involves a pre-selection step for covariates to reduce the predictor space before applying the log-ratio lasso. This two-step procedure can be directly applied to the $2 \cdot p(p-1)/2$-dimensional quadratic log-ratio lasso, introduced in 10, in scenarios where p is large.

2.3 Modeling hierarchical interactions

The quadratic interaction models, introduced before, can enhance predictive performance compared to models that are linear in the features, but they may not detect robust microbial interaction effects suitable for further functional analysis. To enhance model interpretability, we introduce the statistical concept of hierarchy in the context of quadratic models for microbiome data. The concept of hierarchy permits the inclusion of an interaction Θ_{jk} in the model *only if* both associated main effects are also present in the model (strong hierarchy), or if at least one of the associated main effects is included (weak hierarchy) [see [34], and references therein]. This hierarchy can be implemented by imposing constraints on the interaction effects $\Theta_j \in \mathbb{R}^p$ for $j = 1, ..., p$ namely

$$c(\beta_0, \beta, \Theta) = \left\{ \Theta = \Theta^T,\ \left\|\Theta_j\right\|_1 \leq |\beta_j| \right\}. \tag{17}$$

By eliminating the symmetry constraint on Θ, the resulting model relaxes to weak hierarchy on the interaction features. While 17 is non-convex, we follow [34] who proposed a convex relaxation of the problem and provided an efficient implementation in the corresponding R package `hierNet` [44] (v1.9). The hierarchical constraint can be imposed within the generic optimization problem described in 12 and allows a direct application under the

convex relaxation for (i) quantitative microbiome data (ii) or presence-absence information of microbial species with 13; and (iii) relative microbiome data, after performing the alr transformation with 14.

2.4 Model selection

An essential challenge in high-dimensional penalized regression is the selection of the regularization parameter λ. This parameter balances the sparsity of model coefficients with out-of-sample predictive performance [45, 46]. Standard methods for main effects and interaction models often involve techniques such as cross-validation [34] or Information Criteria like the Akaike (AIC) and the Bayesian Information Criterion (BIC) [47]. However, these methods tend to select more predictors and interactions than necessary [47]. If the main aim lies in detecting robust effects, one way of accounting for the potential limitations caused by cross-validation in penalized regression models is the concept of stability selection [37] for identifying a set of predictive features and interactions in microbial data. Stability selection has shown effectiveness across various scientific domains, ranging from network learning [48, 49] to data-driven partial differential equation identification [50, 51]. In the context of regression, stability selection involves iteratively learning sparse regression models from subsamples of the data (e.g., $n_s = \lfloor n/2 \rfloor$), recording the frequency of selected predictors across models, and selecting the most frequent predictors for the final model. A variant of stability selection, complementary pairs stability selection (CPSS) [38], is particularly advantageous for handling unbalanced experimental designs, as it ensures that individual subsamples are independent of each other. CPSS draws b subsamples as complementary pairs $\{(a_{2l-1}, a_{2l}) : l = 1, ..., b\}$, with $a_{2l-1} \cap a_{2l} = \emptyset$ from samples $\{1, ...n\}$ of size $\lfloor n/2 \rfloor$. Applying a variable selection procedure S (for instance choosing the k first predictors entering the penalized model in the regularization path or cross-validation) to each subsample allows defining a feature specific selection probability $\hat{\pi}_i$ for $i = 1, ..., p + p(p-1)/2$ that is given by

$$\hat{\pi}_i = \frac{1}{2b} \sum_{l=1}^{2b} \mathbb{1}_{\{i \in \hat{S}(a_l)\}}. \tag{18}$$

The final selection set, denoted as $\hat{S}^{\text{CPSS}}$, consists of predictors i for which the estimated selection probability $\hat{\pi}_i$ exceeds a predefined threshold π_{thr}, that represents the minimum selection frequency required for a predictor to be included in the final set. We employ the `stabs` R package [52] (v0.6-4), which offers an efficient implementation of the CPSS procedure. This approach involves defining several hyperparameters, including the set of regularization parameters Λ, the threshold $\pi_{\text{thr}} \in [0, 1]$, the number of initial predictors k entering the sparse model, and the number of complementary splits b. The CPSS procedure in [52] can be directly applied to linear models. As CPSS does not make a distinction between main and interaction effects, it can be directly applied to quadratic models. An integration of the

CPSS procedure within the hierarchical interaction modeling framework has been introduced in [33].

3 Results

3.1 Applications in quantitative and presence-absence microbial data

Antimicrobial resistance genes (ARGs) play a crucial role in the survival and evolution of individual microbial species. The extensive use of antimicrobials has increased the development of resistance in pathogens, leading to an increased presence of ARGs. The number of ARGs might be associated with the composition and abundance of certain microbes in the human gut [53].

To get a better understanding of how community composition and interactions might be related to ARGs, we use quantitative microbial count information derived as mOTUs (metagenomic operational taxonomic units) from quantitative microbiome profiling for a subset of $n = 690$ individuals. Specifically, we take the abundance data from the MetaCardis cohort [39] for which metadata information is available. We aggregate the mOTUs on genus level and illustrate the modeling strategy by considering the 30 most abundant genera in our model. We denote the underlying data by $A \in \mathbb{R}^{n \times p}$ and the number of ARGs by $y \in \mathbb{R}^n$. Given that the microbial counts in A are quantitative, we fit the sparse interaction model for absolute count data defined in 3 and 13, respectively, for 10 train test splits by using 5-fold cross-validation (CV). Our results suggest that some genera and interactions between them can explain the prevalence of ARGs (see Fig. 2a, right panel). In Fig. 2a (left panel), we visualize the estimated coefficients that exhibit a non-zero median over 10 train test splits. Next to some minor main and interaction effects, Bacteroides and Escherichia show a substantial effect on the increase of the number of ARGs, while Prevotella shows a decrease. Moreover, we identify a positive interaction effect between Prevotella and Faecalibacterium, which is contrary to their individual negative effects, indicating an antagonistic association. These findings suggest that the presence and co-presence of certain bacterial species in the gut microbiota can influence the prevalence of ARGs.

In a second example we investigate the contribution of certain bacteria as well as their pairwise interplay on butyrate production, a short-chain fatty acid beneficial to human health, within an in-vitro community given the presence-absence information of bacteria within a synthetic community from [24]. Certain bacteria, known as butyrate-producers, have the ability to ferment dietary fibers into butyrate, contributing to gut health, immune function, and energy metabolism [54]. Understanding how bacteria interact in this context is essential for understanding the complexity of this process. Following [15], we use the presence-absence information, denoted by $B \in \{0, 1\}^{n \times p}$ of $p = 25$ bacteria in $n = 1561$ experiments, to fit the

sparse quadratic interaction model defined in 4 and 13, respectively, for 10 train test splits by using 5-fold cross-validation, to the butyrate production, denoted by $y \in \mathbb{R}^n$.

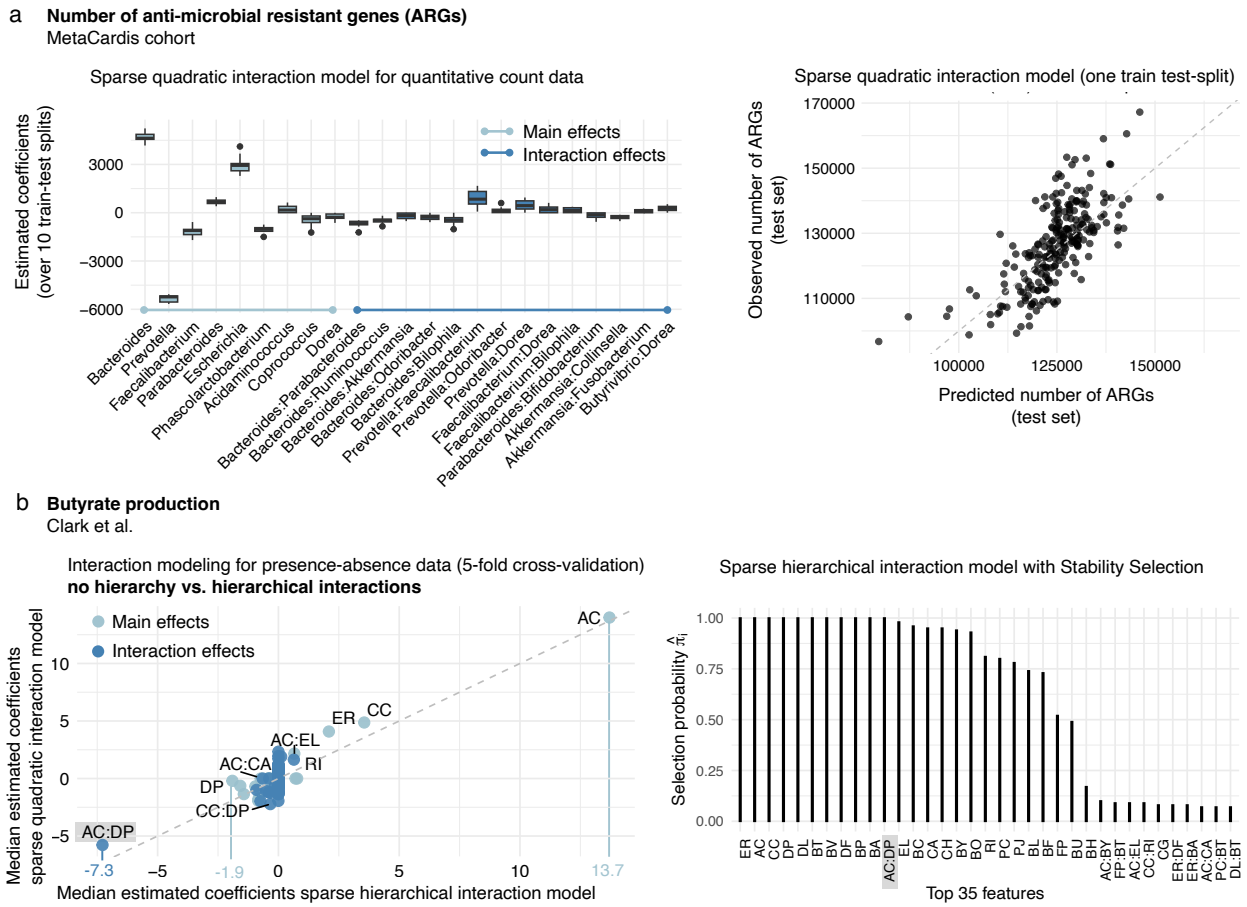

Fig 2. **a**, Prediction of the number of anti microbial resistance genes (ARGs) from absolute abundance information of microbial genera from the MetaCardis cohort. Left: distribution of estimated coefficients over 10 train test splits in the sparse quadratic interaction model 13. Only coefficients with a non-zero median estimated coefficient are shown; right: Scatterplot comparing the observed and predicted number of ARGs on a test data set for the sparse quadratic interaction model. **b**, Prediction of butyrate production from the abundance information of microbial species from [24]. Left: comparison of median estimated coefficients (over 10 train test splits) between the sparse quadratic interaction model (y axis) and the sparse hierarchical interaction model 17 with weak hierarchy (x axis); right: Top 35 selection probabilities from complementary pairs stability selection (CPSS) in the sparse hierarchical interaction model.

The model identifies a strong positive effect of *A. caccae* (AC) on butyrate production. AC is known to be a butyrate producer [55]. However, this effect experiences notable inhibition when AC is combined with *D. piger* (DP), reflected by a strong negative interaction effect (AC:DP) in our model, while DP itself shows only a modest negative impact on the butyrate

production. The inhibiting effect of DP on AC with respect to butyrate production has been shown in tri-cultures with *E. hallii* before [56] as well as in the context of hydrogen sulfide production by DP [24]. The model identifies a multitude of further minor main and interaction effects, including *R. intestinalis* (RI), *E. rectale* (ER), AC and *E. lenta* (AC:EL), and AC and *C. aerofaciens* (AC:CA). Under the assumption that at least one bacterium from each pair contributing to a pairwise interaction influences butyrate production individually, we apply the sparse hierarchical interaction model with weak hierarchy on the same data.
The model with hierarchy yields a substantially reduced set of selected effects while maintaining a similarly strong predictive performance (test set $R^2 = 0.72$ with weak hierarchy versus $R^2 = 0.78$ without hierarchy) (see Fig. 2b, left panel). While the effects of AC or the interaction between AC and DP stand out as clearly important predictors of butyrate production, regardless whether we fit the model with or without hierarchy or the choice of model selection procedure, the stability of smaller effects in the model, like ER or AC:ER, remains unclear. To further investigate, we combine the sparse hierarchical interaction model with stability selection, specifically complementary pairs stability selection (CPSS). The selection probabilities $\hat{\pi}_i$, where $i = 1, ..., p + p(p-1)/2$, indicate that some of the small effects are robust such as ER, *C. comes* (CC), and *D. longicatena* (DL) ($\hat{\pi}_i > 0.7$). However, all other interaction effects, such as AC:EL or AC:CA, fail to demonstrate stability across multiple subsamples.

3.2 Application of interaction modeling on relative microbiome data

3.2.1 Feasibility study of accurate interaction detection on semi-synthetic relative microbiome data

In this section, we generate semi-synthetic data to demonstrate the ability of the sparse quadratic log-contrast model (sparse qlc), defined in 8 and 15, to accurately detect interaction effects in relative microbiome data.
We elucidate the conditions under which accurate estimation is feasible by varying the degree of sparsity of the interaction features and the level of noise present in the data. In the semi-synthetic scenario, we leverage real-world relative microbial count data derived from 16S rRNA sequencing, $A \in \mathbb{R}_+^{n \times p}$, to generate S synthetic outcomes $y_s \in \mathbb{R}^n$ for $s = 1, \ldots, S$. The generation of synthetic outcomes from real data ensures that all inherent distribution properties of the dataset are retained.
In the first simulation, our goal is to understand how the sparsity level of an interaction feature affects the accuracy of the estimates. We use a subset of the data of the American Gut Project [57], processed in [16], comprising a selection of $p = 50$ OTUs, ranging from dense to sparse, and $n = 300$ subsamples (see Fig. 3b). We define $S = 5$ semi-synthetic scenarios, where $y_s \in \mathbb{R}^n$ for $s = 1, ..., S$ is given as the sum over a sparse linear combination

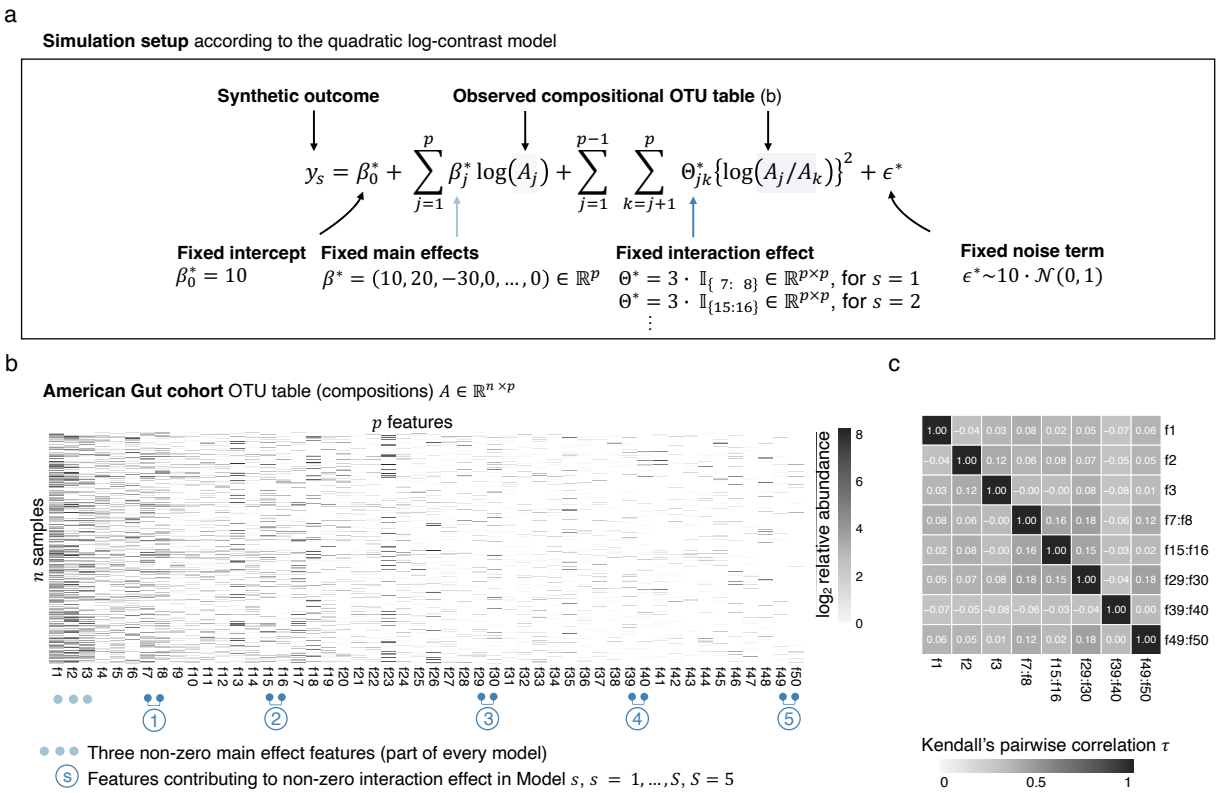

Fig 3. Semi-synthetic simulation setup for varying feature sparsity levels. **a**, Simulation setup for generating a synthetic outcome y_s for $s = 1, \ldots, S$ based on the quadratic log-contrast model formulation. **b**, Heatmap of the OTU table carrying compositional information for a subset of $p = 50$ OTUs from the American Gut cohort sorted by sparsity in descending order. Non-zero main effects contributing to each of the $S = 5$ semi-synthetic scenarios (light blue) and features contributing to the non-zero interaction effect in model scenario s for $s = 1, \ldots, S$ (dark blue) are highlighted. **c**, Kendall's pairwise correlations τ between features that have non-zero effects in the models $s = 1, \ldots, S$. They should be as uncorrelated as possible ($|\tau| < .2$) to eliminate effects of correlated features.

of main and interaction effects in $A \in \mathbb{R}_+^{n \times p}$ according to model formulation 8 (see Fig. 3a). To account for the zeros in the data when building log-ratios we add a pseudo count of one to each entry in A, such that $A := A + 1$. Each outcome is characterized by a fixed intercept term $\beta_0^* = 10$, three non-zero main effects ($\beta_1 = 10$, $\beta_2 = 20$, and $\beta_3 = -30$), and a noise term $\epsilon^* = c_1 \cdot \mathcal{N}(0, 1)$, with $c_1 = 10$. Moreover, each outcome y_s for $s = 1, \ldots, S$ is characterized by a unique non-zero interaction effect $\Theta_{jk}^* = 3$ between two features A_j and A_k for $j = 1, \ldots, p-1$ and $k = j+1, \ldots, p$ which varies across interaction features of different sparsity levels, namely, 36% (f7:f8, $s = 1$), 52% (f15:f16, $s = 2$), 67% (f29:f30, $s = 3$), 74% (f39:f40, $s = 4$), and 88% (f49:f50, $s = 5$) zero entries (see Fig. 3b). To avoid undesired correlation effects between the predictive features in the model, we ensure that the interaction features are uncorrelated (absolute Kendall's pairwise correlation $|\tau| < 0.2$) with the main effects (see Fig. 3c). By

fitting the sparse qlc model to the semi-synthetic data, we demonstrate that as interaction features become sparser, the accuracy of feature estimates with comparably small effect sizes (here $\Theta^*_{jk} = 3$) diminishes (see Tab. 1). Our simulations show that, while very sparse features with small true nonzero estimates are still selected by the sparse quadratic log-contrast model, they tend to be underestimated and are accompanied by many spuriously selected features with estimates of similar magnitude (see also Fig. S1).

Table 1. Median and Variance of the estimation error of the interaction coefficient $\sqrt{(\Theta^*_{jk} - \hat{\Theta}_{jk})^2}$ for $S = 5$ semi-synthetic scenarios.

	Sparsity interaction feature				
	36% (f7:f8)	52% (f15:f16)	67% (f29:f30)	74% (f39:f40)	88% (f49:f50)
Median estimation error	0.15	0.30	0.82	0.87	1.78
Variance estimation error	0.04	0.09	1.15	0.21	1.22

In a second simulation, we investigate the impact of noise in the data on the accurate estimation of main and interaction effects. Again, we utilize a subset of the data from the American Gut Project [57], preprocessed as described in [16]. In contrast to the previous scenario, our objective is to evaluate the influence of noise while mitigating the effects of sparsity. To achieve this, we aggregate the data at the phylum level which is the highest taxonomic level with $p = 10$ phyla. We generate $L = 4$ synthetic outcomes $y_l,\ l = 1, ..., L$ according to model formulation 8 by fixing main and interaction effects, while allowing the noise levels to vary. We fix the intercept term at $\beta^*_0 = 10$, assign six non-zero main effects that sum up to zero to the first six features as $\beta^* = (10, 20, 30, -10, -20, -30, 0, \ldots, 0) \in \mathbb{R}^p$, and introduce three non-zero interaction effects (between A_1 and A_3, A_8 and A_{10}, and A_9 and A_{10}) as $\Theta^* = 10 \cdot \mathbb{1}_{1:3,\ 8:10,\ 9:10} \in \mathbb{R}^{p \times p}$. The noise terms ϵ^*_l follow a normal distribution with mean 0 and variance 1 with that is multiplied by a constant factor c_l, $l = 1, \ldots, L$, given by $c = (10, 100, 200, 500)^T$, that varies the noise level. Thus, $\epsilon^*_l = c_l \cdot \mathcal{N}(0, 1)$. Consequently, the synthetic outcomes y_l, $l = 1, ..., L$ are given by

$$y_l = \beta^*_0 + \sum_{j=1}^{p} \beta^*_j \log(A_j) + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \Theta^*_{jk} \log(A_j/A_k)^2 + \epsilon^*_l. \tag{19}$$

To relate the noise terms to the signal part, we translate the noise levels to signal-to-noise ratios (SNR) denoted by snr_l for $l = 1, ..., L$ given by $\text{snr} = (178.01, 1.78, 0.45, 0.07)^T$. We fit both models, the sparse linear log-contrast model (sparse lc) and the sparse quadratic log-contrast (sparse qlc), to the newly generated data. Notably, features f2, f4, f5, and f6 show no contribution to interaction effects in our synthetic example and are accurately estimated by the (misspecified) sparse lc model (see Fig. 4a and b). However, for the features f1 and f3,

which both have non-zero main effects as well as a common interaction effect, the sparse lc model accommodates the positive interaction effect between f1 and f3 within their main effect estimates, leading to an overestimation of the true positive main effect of f1 ($\beta_1^* = 10$) and an underestimation of the true positive main effect of f3 ($\beta_3^* = 30$). Moreover, the sparse lc model selects f8, f9 and f10 as relevant main effect features despite their lack of true nonzero main effects, in order to integrate their true underlying interaction effects. In contrast, the sparse qlc model captures the coefficients accurately. For both models, the overall performance deteriorates as the SNR decreases (see R^2 values in Fig. 4c and estimation error summary statistics in Tab. 2). Our simulations indicate that the sparse qlc model outperforms the sparse lc model when true interaction effects are present, provided that the noise level is not excessively high or the SNR is not too low (see Fig. 4c). In scenarios where noise levels are exceptionally high, the interaction model tends to overfit the data (see Fig. 4c, right plot).

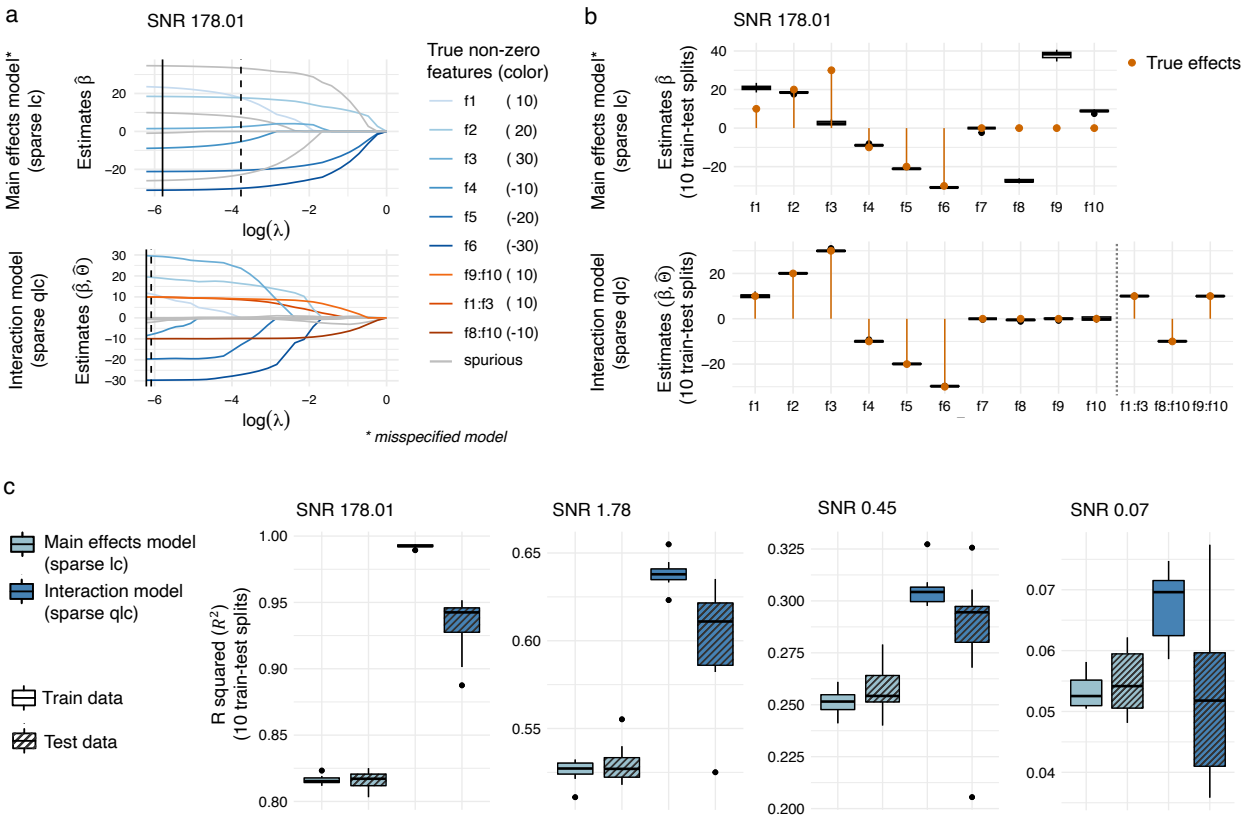

Fig 4. Influence of model misspecification and noise in semi-synthetic scenarios. **a**, Solution path for the misspecified main effects model (sparse lc) and the interaction model (sparse qlc) for the synthetic scenario $l = 1$ with a signal-to-noise ratio (SNR) of 178.01 for one train test split. **b**, Estimated coefficients distributions over 10 train test splits corresponding to the solution paths in a. For the interaction model only three non-zero interaction features are shown for visualization purposes. **c**, Comparison of model performance via R squared (R^2) for the main effects model and the interaction effects model on train and test data for four different SNRs.

Table 2. Median and Variance of the estimation error of the interaction coefficient $\sqrt{([\beta^*, \Theta^*] - [\hat{\beta}, \hat{\Theta}])^2}$ for $L = 4$ semi-synthetic scenarios with different Signal-to-noise ratios (SNR).

	SNR: 178.01	SNR: 1.78	SNR: 0.45	SNR: 0.07
Median	3.89	38.44	66.96	138.78
Variance	1.74	57.44	362.86	30.35

3.2.2 Interaction modeling on real-world relative microbiome data

Here, we perform sparse interaction modeling on environmental rather than experimental or host-associated microbiome data, highlighting salinity as a crucial factor in marine ecosystems. Variations in salinity play a critical role in shaping microbial community diversity and functionality in marine ecosystems, thereby influencing nutrient dynamics and ecosystem health [58, 59]. Using the marine data collection from Tara Oceans [60], which includes relative microbial abundance information derived as metagenomic OTUs (mOTUs) of ocean surface water and associated environmental covariates, we illustrate how sparse interaction modeling can substantially improve the predictive performance compared to models linear in the features. We aggregate the data on family level and learn a sparse quadratic log-contrast model (sparse qlc) as defined in 15 for ocean salinity from $n = 136$ samples and a subset of $p = 30$ most abundant families [61] and compare this to the corresponding sparse log contrast model that is linear in the features (sparse lc). Our comparison over 10 train test splits shows that modeling interaction effects rather than only main effects not only improves predictive accuracy (see Fig. 5a), but also allows predicting salinity concentrations beyond the interval $[33.8, 36.6]$ (see Fig. 5b). The model selects various features that are not consistently selected among all train test splits and no large main effects. However, it also identifies strong negative interaction effects between a family without annotation (f55) that belongs to the order SAR11 clade and the family Sphingomonadaceae as well as between two families without annotation (f13 and f96) that both belong to the order SAR11 clade. The negative interactions indicate that the quadratic effect of these groups (high abundances of both groups) comes with reduced salinity levels. For the families f8 and f60, that are also both part of SAR11 clade, the model identifies a positive interaction effect for all train test splits, indicating that the quadratic effect of these families comes with higher salinity levels. Several studies indicate a potential link between salinity in the ocean and the abundance of the SAR11 clade [62, 63, 64, 65]. Our findings suggest that interactions among subtypes of the SAR11 clade and between SAR11 clade families and other families may play a role in relation to the salinity level in the ocean.

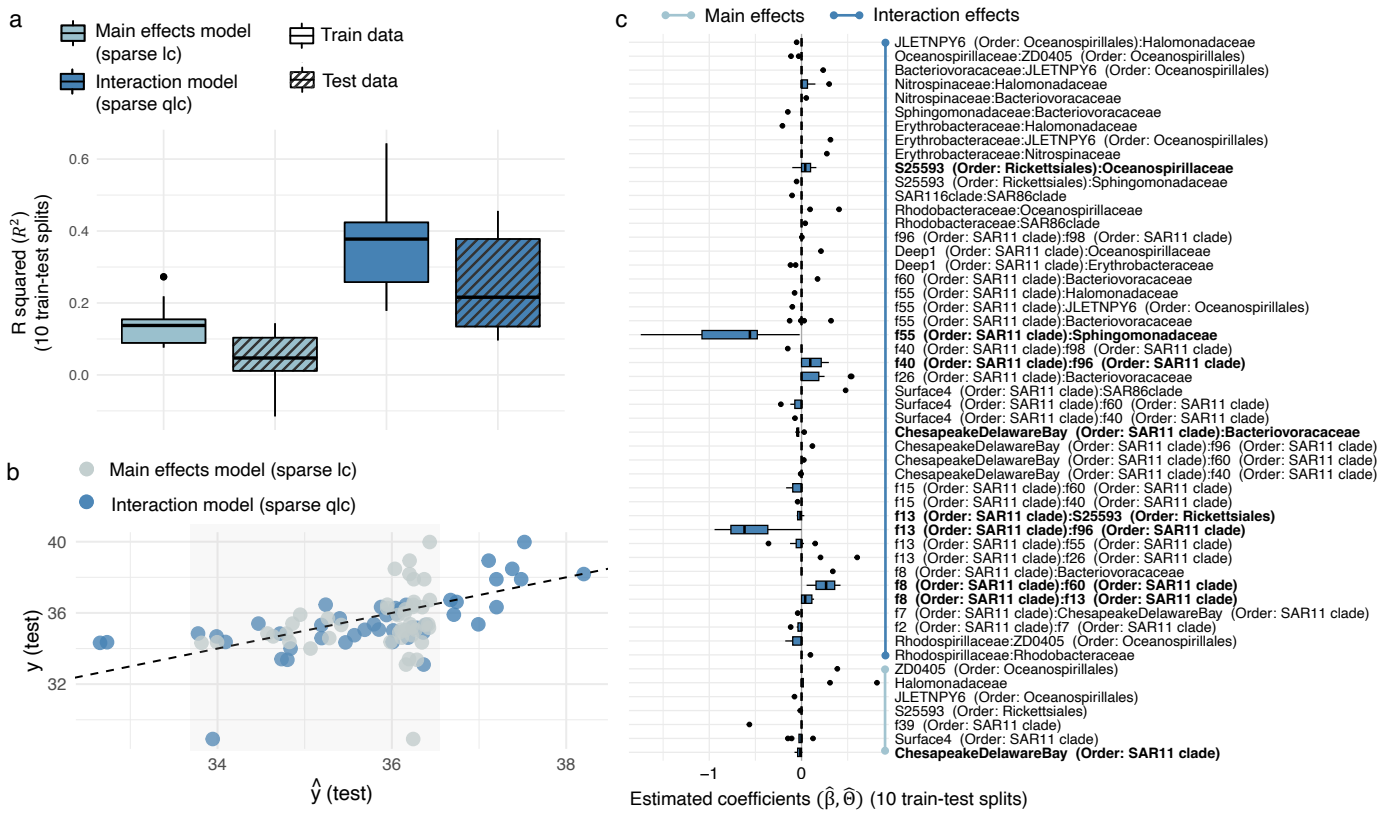

Fig 5. Summary plots of the Tara ocean data on family level for salinity prediction. **a,** Train and test set R squared R^2 distribution over 10 train test splits for the main effects sparse log contrast model (SLC) and the all-pairs log-contrast interaction lasso (SLC + int.). **b,** Scatterplot between the observed test set outcome y (salinity) and the prediction $\hat{y}$ from the main effects sparse log-contrast model (SLC) and the all-pairs log-contrast interaction lasso (SLC + int.). **c,** Distribution of estimated main effect and interaction effect coefficients in the the all-pairs log-contrast interaction lasso (SLC + int.) over 10 train test splits. Only features (main or interaction features) with a non-zero mean coefficient are shown. Features with a non-zero median are bold.

3.3 Abundance information can hold valuable insights beyond presence-absence information

The interpretation of the relationship between an outcome variable Y and species abundance information varies with the different data modalities discussed in this work. When using absolute counts, the effect of the actual abundance value can be derived, whereas presence-absence data provide insights into the effects of existence. Compositional data offer information on the impact of proportions between species abundances. While these modalities are typically analyzed based on availability, it remains unclear how effectively each data modality truly explains an outcome of interest. Furthermore, it is uncertain whether the same microbial taxa and interactions between taxa would be identified if another modality was used in the

prediction task.

To illustrate this, we revisit the example on quantitative microbial abundance information from the MetaCardis cohort as discussed in Section 3.1. This data can be transformed into compositions or presence-absence information using $\boldsymbol{B} = \mathbb{1}_{A>0}$, allowing for a comparative analysis across the three data modalities for predicting the number of antimicrobial resistance genes (ARGs). We apply the sparse quadratic interaction model for absolute count data (defined in 3), the sparse quadratic interaction model for presence-absence data (defined in 4), the sparse quadratic log-contrast model (defined in 8) to the three versions of the abundance data. In this comparison, we employ the quadratic log-contrast model to analyze the relative information, as it offers the most straightforward interpretation for such comparisons.

We observe that transforming counts to relative abundances does not substantially affect the predictive performance, underscoring the importance of both absolute values and their proportions in describing the number of antimicrobial-resistant genes (ARGs), as shown in Fig. 6a. However, the presence-absence information alone does not adequately explain the number of ARGs. The genera, Prevotella, Bacteroides, and Escherichia, are identified as important predictors with consistent signs for the number of ARGs in both absolute counts and relative count information, as shown in Fig. 6b. This indicates that both the absolute abundance and the proportions relative to other taxa are significant. Escherichia, which is also part of the Proteobacteria phylum known to harbor a variety of ARGs [53, 66], and Bacteroides fragilis, a species within the Bacteroides genus, exhibits high antimicrobial resistance rates and possesses numerous mechanisms related to antimicrobial resistance [67]. These factors could explain the effects detected by our model. However, the presence-absence data do not select any of the main effects. Instead, this model frequently selects a multitude of interaction effects involving Prevotella. Overall, there is almost no agreement in the interaction effects identified across the different data modalities. In summary, our findings suggests that abundance information is overall more informative than presence-absence information and that whether we identify an interaction between two taxa is highly dependent on the underlying data modality.

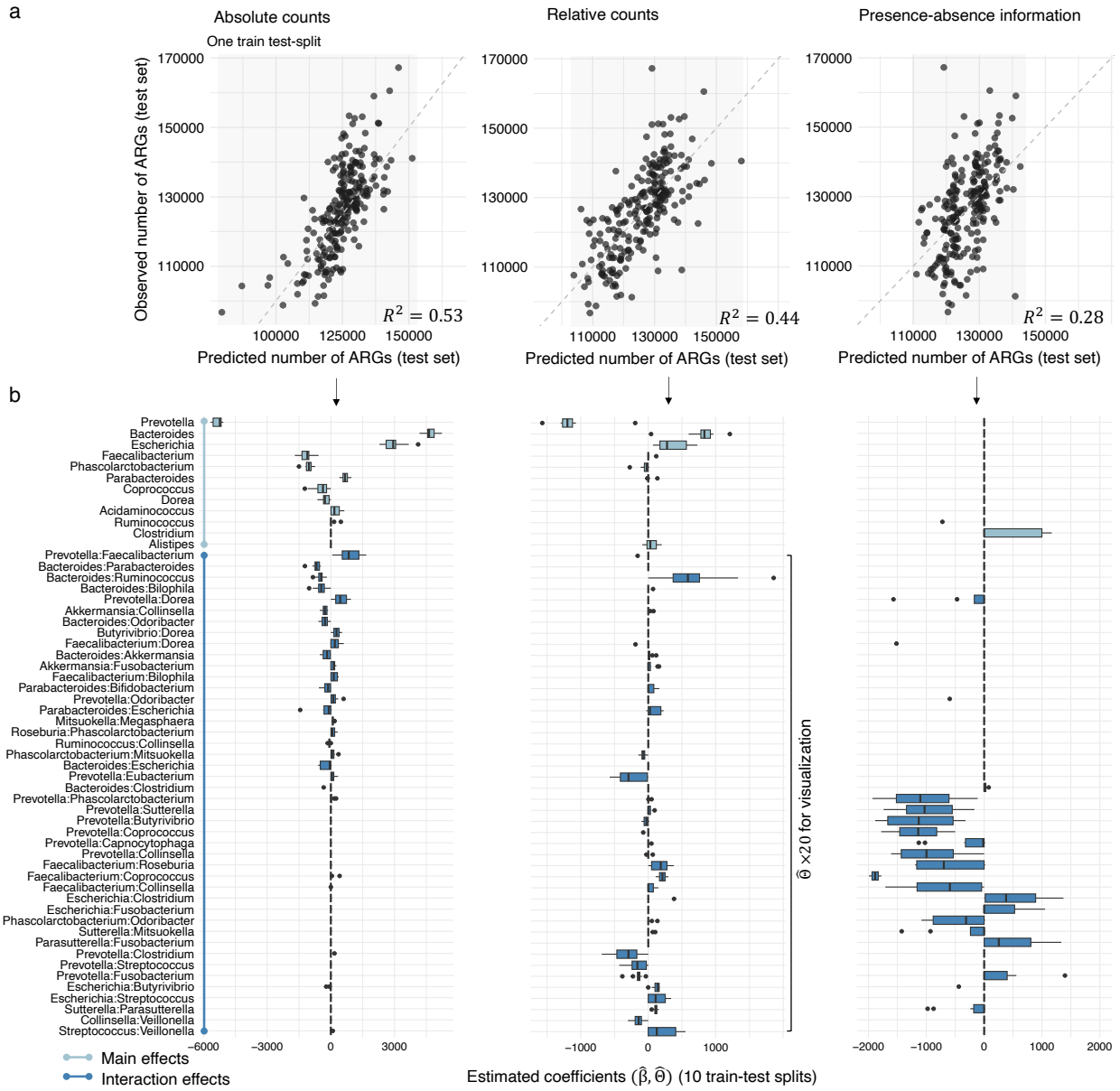

Fig 6. **a**, Scatterplots (one train test split) and test set R squared R^2 (average over 10 train test splits) for the comparison of the predicted number of anti microbial resistance genes (ARGs) and the observed number of ARGs on a test dataset based on the absolute count information of genera, the relative count information of genera and the presence-absence information of genera of the MetaCardis cohort. **b**, Distribution of coefficients over 10 train test splits for the superset of coefficients that are non-zero in one of the three data modalities.

4 Discussion

Identifying predictive and interpretable main and interaction effects between microbial taxa that can be related to ecological, host-associated or environmental features is a cornerstone of statistical data analysis of microbiome data. To this end, we have introduced a generic modeling approach for quadratic interactions, that is further instantiated to accommodate three distinct data types, in which microbial abundance information typically appears: (i) quantitative microbiome, (ii) presence-absence information, or (iii) relative count information. For these three data modalities we introduce a generic optimization problem, that allows penalized model estimation to estimate parsimonious models, where the number of features p (here the microbial taxa) and pairwise interactions often exceeds the number of measurements n. Our interaction modeling strategy combines existing approaches for sparse interaction modeling and introduces novel ways of modeling interaction effects in the compositional setting. In the realm of interpretability, we combine the quadratic interaction modeling framework with the concept of hierarchical interactions [34] and stability selection [38, 37] as optional extensions. We introduce three mathematically equivalent ways of modeling quadratic interactions with relative input data: (a) the alr transformed quadratic model, (b) the quadratic log-contrast model, and (c) the quadratic log-ratio model. The three models differ in terms of interpretability, dimensionality, and optimization. In the $\tilde{p} + \tilde{p}(\tilde{p} - 1)/2$-dimensional alr transformed quadratic model, where $\tilde{p} = p - 1$, the interpretation of coefficients is consistently tied to a reference feature, which is enforced to be part of the model in the regularized model case. This model can be optimized without constraints on the main effects and can be flexibly extended to hierarchical interactions. The $p + p(p - 1)/2$-dimensional quadratic log-contrast model requires a zero-sum constraint on the main effects but offers a more convenient expression that does not require the assignment of a reference feature. In this model, each main effect is interpreted with respect to all other features. The $2 \cdot p(p - 1)/2$-dimensional quadratic log-ratio model does not require a constraint on the main effect coefficients in the optimization, as this property is automatically met, and allows for the interpretation of the relative effects between all pairs of features. Notably, the interpretation of the interaction coefficients in both the quadratic log-contrast model and the quadratic log-ratio model are identical. Although the log-contrast model in the linear case enjoys the clear advantage of lower dimensionality compared to the log-ratio model, this distinction becomes less relevant when introducing interactions. Hence, the primary criterion for selecting between the models should be based on the preferred interpretation.

We demonstrate the broad applicability of our framework by analyzing microbial data covering all three data modalities across various ecosystems, including synthetic microbial communities, human gut microbiomes, and marine microbial ecosystems. We show how quadratic models can improve the predictive performance compared to linear models on real-world data and illustrate how main and interaction effects can be robustly estimated by integrating hierarchical interaction modeling and stability selection. Notably, using a

synthetic community dataset from [24], we identified a strong inhibition of butyrate production by *A. caccae* when *D. piger* is present. We demonstrate that this is the only robust and consequently relevant combinatorial effect within this community. Additionally, we generate semi-synthetic data to demonstrate the ability of the sparse quadratic interaction model for relative microbiome data to accurately detect interaction effects. Based on this, we showcase how sparse an interaction feature can be in order to achieve an accurate estimation in our model. Further, we demonstrate, for varying noise levels, how a misspecified main effects model tends to inaccurately estimate effects when true interactions are present. Finally, we perform a comparative analysis of the quadratic interaction models for the three data modalities discussed in this study, demonstrating that absolute microbial counts and compositions achieve similarly strong predictive performance when predicting the number of antimicrobial resistance genes (ARGs). In contrast, transforming abundance data into presence-absence information results in a decline in predictive accuracy. Our analysis suggests that the identified (main and) interaction effects highly depend on the underlying data type. Moreover, the identified effects of individual genera and interactions between them are an interesting finding in themselves. Many of the relevant features correspond to the definition of enterotypes [68], as they were defined for this specific data in [69], suggesting a potential link between enterotypes and ARGs.

As more and larger data sets become available, our models can be extended to higher-order interactions or more complex, arbitrary non-linear effects. Moreover, our results suggest that different data modalities carry different information, and it might be meaningful to include multiple layers of information in statistical prediction tasks with interaction effects.

In summary, we believe that our framework and its implementation in R provide a valuable tool to study robust interaction effects in microbiome data of different modalities and distinct ecosystems, ranging from synthetic communities and community function studies in microbial ecology to observational microbiome data linking the microbiome to the host or environment.

References

[1] A. Konopka. "What is microbial community ecology?" In: *The ISME journal* 3.11 (2009), pp. 1223–1230.

[2] E. J. Culp and A. L. Goodman. "Cross-feeding in the gut microbiome: Ecology and mechanisms". In: *Cell Host & Microbe* 31.4 (2023), pp. 485–499.

[3] R. M. Braga, M. N. Dourado, and W. L. Araújo. "Microbial interactions: ecology in a molecular perspective". In: *brazilian journal of microbiology* 47 (2016), pp. 86–98.

[4] W. Z. Lidicker Jr. "A clarification of interactions in ecological systems". In: *Bioscience* 29.8 (1979), pp. 475–477.

[5] K. Faust and J. Raes. "Microbial interactions: from networks to models". In: *Nature Reviews Microbiology* 10.8 (2012), pp. 538–550.

[6] N. Weiland-Bräuer. "Friends or foes—microbial interactions in nature". In: *Biology* 10.6 (2021), p. 496.

[7] J. Friedman and E. J. Alm. "Inferring correlation networks from genomic survey data". In: (2012).

[8] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. "Sparse and compositionally robust inference of microbial ecological networks". In: *PLoS computational biology* 11.5 (2015), e1004226.

[9] J. Tackmann, J. F. M. Rodrigues, and C. von Mering. "Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data". In: *Cell systems* 9.3 (2019), pp. 286–296.

[10] G. Yoon, I. Gaynanova, and C. L. Müller. "Microbial networks in SPRING-Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data". In: *Frontiers in genetics* 10 (2019), p. 516.

[11] S. Peschel, C. L. Müller, E. Von Mutius, A.-L. Boulesteix, and M. Depner. "NetCoMi: network construction and comparison for microbiome data in R". In: *Briefings in bioinformatics* 22.4 (2021), bbaa290.

[12] F. G. Blanchet, K. Cazelles, and D. Gravel. "Co-occurrence is not evidence of ecological interactions". In: *Ecology Letters* 23.7 (2020), pp. 1050–1063.

[13] B. Ma, Y. Wang, S. Ye, S. Liu, E. Stirling, J. A. Gilbert, K. Faust, R. Knight, J. K. Jansson, C. Cardona, et al. "Earth microbial co-occurrence network reveals interconnection pattern across microbiomes". In: *Microbiome* 8.1 (2020), pp. 1–12.

[14] T. Zamkovaya, J. S. Foster, V. de Crécy-Lagard, and A. Conesa. "A network approach to elucidate and prioritize microbial dark matter in microbial communities". In: *The ISME journal* 15.1 (2021), pp. 228–244.

[15] A. Skwara, K. Gowda, M. Yousef, J. Diaz-Colunga, A. S. Raman, A. Sanchez, M. Tikhonov, and S. Kuehn. "Statistically learning the functional landscape of microbial communities". In: *Nature Ecology & Evolution* 7.11 (2023), pp. 1823–1833.

[16] J. Bien, X. Yan, L. Simpson, and C. L. Müller. "Tree-aggregated predictive modeling of microbiome data". In: *Scientific Reports* 11.1 (2021), p. 14505.

[17] A. S. Weiss, A. G. Burrichter, A. C. Durai Raj, A. von Strempel, C. Meng, K. Kleigrewe, P. C. Münch, L. Rössler, C. Huber, W. Eisenreich, et al. "In vitro interaction network of a synthetic gut bacterial community". In: *The ISME journal* 16.4 (2022), pp. 1095–1109.

[18] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. "Microbiome datasets are compositional: and this is not optional". In: *Frontiers in microbiology* 8 (2017), p. 294209.

[19] B. J. Callahan, P. J. McMurdie, and S. P. Holmes. "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis". In: *The ISME journal* 11.12 (2017), pp. 2639–2643.

[20] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, et al. "Quantitative microbiome profiling links gut community variation to microbial load". In: *Nature* 551.7681 (2017), pp. 507–511.

[21] A. Tkacz, M. Hortala, and P. S. Poole. "Absolute quantitation of microbiota abundance in environmental samples". In: *Microbiome* 6 (2018), pp. 1–13.

[22] C. Jian, P. Luukkonen, H. Yki-Järvinen, A. Salonen, and K. Korpela. "Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling". In: *PloS one* 15.1 (2020), e0227285.

[23] G. Galazzo, N. Van Best, B. J. Benedikter, K. Janssen, L. Bervoets, C. Driessen, M. Oomen, M. Lucchesi, P. H. van Eijck, H. E. Becker, et al. "How to count our microbes? The effect of different quantitative microbiome profiling approaches". In: *Frontiers in cellular and infection microbiology* 10 (2020), p. 403.

[24] R. L. Clark, B. M. Connors, D. M. Stevenson, S. E. Hromada, J. J. Hamilton, D. Amador-Noguez, and O. S. Venturelli. "Design of synthetic human gut microbiome assembly and butyrate production". In: *Nature communications* 12.1 (2021), p. 3254.

[25] A. L. Gould, V. Zhang, L. Lamberti, E. W. Jones, B. Obadia, N. Korasidis, A. Gavryushkin, J. M. Carlson, N. Beerenwinkel, and W. B. Ludington. "Microbiome interactions shape host fitness". In: *Proceedings of the National Academy of Sciences* 115.51 (2018), E11951–E11960.

[26] J. Aitchison and J. Bacon-Shone. "Log contrast models for experiments with mixtures". In: *Biometrika* 71.2 (1984), pp. 323–330.

[27] W. Lin, P. Shi, R. Feng, and H. Li. "Variable selection in regression with compositional covariates". In: *Biometrika* 101.4 (2014), pp. 785–797.

[28] P. Shi, A. Zhang, and H. Li. "Regression analysis for microbiome compositional data". In: (2016).

[29] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. "Balances: a new perspective for microbiome analysis". In: *MSystems* 3.4 (2018), pp. 10–1128.

[30] S. Bates and R. Tibshirani. "Log-ratio lasso: scalable, sparse estimation for log-ratio models". In: *Biometrics* 75.2 (2019), pp. 613–624.

[31] P. L. Combettes and C. L. Müller. "Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications". In: *Statistics in Biosciences* 13.2 (2021), pp. 217–242.

[32] B. Yu. "Stability". In: *Bernoulli* 19.4 (2013), pp. 1484–1500.

[33] M. Stadler, S. Lukauskas, T. Bartke, and C. L. Mueller. "asteRIa enables robust interaction modeling between chromatin modifications and epigenetic readers". In: *bioRxiv* (2024), pp. 2024–03.

[34] J. Bien, J. Taylor, and R. Tibshirani. "A lasso for hierarchical interactions". In: *The Annals of Statistics* 41.3 (June 2013).

[35] M. Hamada and C. J. Wu. "Analysis of designed experiments with complex aliasing". In: *Journal of quality technology* 24.3 (1992), pp. 130–137.

[36] J. L. Peixoto. "Hierarchical variable selection in polynomial regression models". In: *The American Statistician* 41.4 (1987), pp. 311–313.

[37] N. Meinshausen and P. Bühlmann. "Stability Selection". In: *Journal of the Royal Statistical Society, Series B* 72 (2010), pp. 417–473.

[38] R. D. Shah and R. J. Samworth. "Variable selection with error control: Another look at stability selection". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75.1 (2013), pp. 55–80.

[39] S. K. Forslund, R. Chakaroun, M. Zimmermann-Kogadeeva, L. Markó, J. Aron-Wisnewsky, T. Nielsen, L. Moitinho-Silva, T. S. Schmidt, G. Falony, S. Vieira-Silva, et al. "Combinatorial, additive and dose-dependent drug–microbiome associations". In: *Nature* 600.7889 (2021), pp. 500–505.

[40] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, et al. "Structure and function of the global ocean microbiome". In: *Science* 348.6237 (2015), p. 1261359.

[41] F. J. Poelwijk, V. Krishna, and R. Ranganathan. "The context-dependence of mutations: a linkage of formalisms". In: *PLoS computational biology* 12.6 (2016), e1004771.

[42] E. D. Weinberger. "Fourier and Taylor series on fitness landscapes". In: *Biological cybernetics* 65.5 (1991), pp. 321–330.

[43] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[44] J. Bien and R. Tibshirani. *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.9. 2020.

[45] J. Lederer and C. Müller. "Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015.

[46] Y. Wu and L. Wang. "A survey of tuning parameter selection for high-dimensional regression". In: *Annual review of statistics and its application* 7 (2020), pp. 209–226.

[47] N. Hao, Y. Feng, and H. H. Zhang. "Model selection for high-dimensional quadratic regression via regularization". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 615–625.

[48] H. Liu, K. Roeder, and L. Wasserman. "Stability approach to regularization selection (stars) for high dimensional graphical models". In: *Advances in neural information processing systems* 23 (2010).

[49] B. Bodinier, S. Filippi, T. H. Nøst, J. Chiquet, and M. Chadeau-Hyam. "Automated calibration for stability selection in penalised regression and graphical models". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* (2023), qlad058.

[50] S. Maddu, B. L. Cheeseman, I. F. Sbalzarini, and C. L. Müller. "Stability selection enables robust learning of differential equations from limited noisy data". In: *Proceedings of the Royal Society A* 478.2262 (2022), p. 20210916.

[51] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton. "Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control". In: *Proceedings of the Royal Society A* 478.2260 (2022), p. 20210904.

[52] B. Hofner and T. Hothorn. *stabs: Stability Selection with Error Control*. R package version 0.6-4. 2021.

[53] K. Lee, S. Raguideau, K. Sirén, F. Asnicar, F. Cumbo, F. Hildebrand, N. Segata, C.-J. Cha, and C. Quince. "Population-level impacts of antibiotic usage on the human gut microbiome". In: *Nature Communications* 14.1 (2023), p. 1191.

[54] V. Singh, G. Lee, H. Koh, T. Unno, and J.-H. Shin. "Butyrate producers,"The Sentinel of Gut": Their intestinal significance with and beyond butyrate, and prospective use as microbial therapeutics". In: *Frontiers in microbiology* 13 (2023), p. 1103836.

[55] A. Schwiertz, G. L. Hold, S. H. Duncan, B. Gruhl, M. D. Collins, P. A. Lawson, H. J. Flint, and M. Blaut. "Anaerostipes caccae gen. nov., sp. nov., a new saccharolytic, acetate-utilising, butyrate-producing bacterium from human faeces". In: *Systematic and applied microbiology* 25.1 (2002), pp. 46–51.

[56] P. Marquet, S. H. Duncan, C. Chassard, A. Bernalier-Donadille, and H. J. Flint. "Lactate has the potential to promote hydrogen sulphide formation in the human colon". In: *FEMS Microbiology Letters* 299.2 (2009), pp. 128–134.

[57] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, et al. "American gut: an open platform for citizen science microbiome research". In: *Msystems* 3.3 (2018), pp. 10–1128.

[58] H. S. Tee, D. Waite, G. Lear, and K. M. Handley. "Microbial river-to-sea continuum: gradients in benthic and planktonic diversity, osmoregulation and nutrient cycling". In: *Microbiome* 9 (2021), pp. 1–18.

[59] L. Wang, C. Lian, W. Wan, Z. Qiu, X. Luo, Q. Huang, Y. Deng, T. Zhang, and K. Yu. "Salinity-triggered homogeneous selection constrains the microbial function and stability in lakes". In: *Applied Microbiology and Biotechnology* 107.21 (2023), pp. 6591–6605.

[60] S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, et al. "Tara Oceans: towards global ocean ecosystems biology". In: *Nature Reviews Microbiology* 18.8 (2020), pp. 428–445.

[61] R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, et al. "Metagenomic 16S rDNA I llumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities". In: *Environmental microbiology* 16.9 (2014), pp. 2659–2671.

[62] D. P. Herlemann, J. Woelk, M. Labrenz, and K. Jürgens. "Diversity and abundance of "Pelagibacterales"(SAR11) in the Baltic Sea salinity gradient". In: *Systematic and applied microbiology* 37.8 (2014), pp. 601–604.

[63] B. J. Campbell, S. J. Lim, and D. L. Kirchman. "Controls of SAR11 subclade abundance, diversity, and growth in two Mid-Atlantic estuaries". In: *bioRxiv* (2022), pp. 2022–05.

[64] Y. Tada, R. Makabe, N. Kasamatsu-Takazawa, A. Taniguchi, and K. Hamasaki. "Growth and distribution patterns of Roseobacter/Rhodobacter, SAR11, and Bacteroidetes lineages in the Southern Ocean". In: *Polar biology* 36 (2013), pp. 691–704.

[65] S. Kraemer, A. Ramachandran, D. Colatriano, C. Lovejoy, and D. A. Walsh. "Diversity and biogeography of SAR11 bacteria from the Arctic Ocean". In: *The ISME Journal* 14.1 (2020), pp. 79–90.

[66] N. O. Eltai, A. A. Al Thani, S. H. Al Hadidi, K. Al Ansari, and H. M. Yassine. "Antibiotic resistance and virulence patterns of pathogenic Escherichia coli strains associated with acute gastroenteritis among children in Qatar". In: *BMC microbiology* 20 (2020), pp. 1–12.

[67] S. Valdezate, F. Cobo, S. Monzón, M. J. Medina-Pascual, Á. Zaballos, I. Cuesta, S. Pino-Rosa, and P. Villalón. "Genomic background and phylogeny of cfi A-positive Bacteroides fragilis strains resistant to meropenem-EDTA". In: *Antibiotics* 10.3 (2021), p. 304.

[68] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, et al. "Enterotypes of the human gut microbiome". In: *nature* 473.7346 (2011), pp. 174–180.

[69] S. Vieira-Silva, G. Falony, E. Belda, T. Nielsen, J. Aron-Wisnewsky, R. Chakaroun, S. K. Forslund, K. Assmann, M. Valles-Colomer, T. T. D. Nguyen, et al. "Statin therapy is associated with lower prevalence of gut microbiota dysbiosis". In: *Nature* 581.7808 (2020), pp. 310–315.

Supplementary information

Binary encoding of covariates in regression models

Whether to encode the input data in a regression model as $B \in \{0,1\}^{n\times p}$ or as $B \in \{-1,1\}^{n\times p}$ has an impact on the interpretations of the model coefficients. In the main effects model case with $p = 1$, for $B \in \{0,1\}^{n\times p}$, the outcome Y would be given by

$$Y = \begin{cases} \beta_0 & \text{if } B_1 = 0 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1. \end{cases}$$

The interpretation of β_0 in this case is the effect of 'absence' and the interpretation of β_1 is the difference between the effect of 'presence' and the effect of 'absence'. The interaction model, illustrated for $p = 2$, is given by

$$Y = \begin{cases} \beta_0 & \text{if } B_1 = 0 \text{ and } B_2 = 0 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1 \text{ and } B_2 = 0 \\ \beta_0 + \beta_2 & \text{if } B_1 = 0 \text{ and } B_2 = 1 \\ \beta_0 + \beta_1 + \beta_2 + \Theta_{12} & \text{if } B_1 = 1 \text{ and } B_2 = 1. \end{cases}$$

Here, β_0 is the effect of co-absence, and β_j is the effect of the difference of co-absence and presence of B_j, for $j = 1, 2$. The interaction term Θ_{12} is the additional effect when both features are 1.
If $B \in \{-1,1\}^{n\times p}$, the outcome Y in the main effects model is given by

$$Y = \begin{cases} \beta_0 - \beta_1 & \text{if } B_1 = -1 \\ \beta_0 + \beta_1 & \text{if } B_1 = 1. \end{cases}$$

The interpretation here is that β_0 is the mean effect of the two group means, and $2\beta_1$ is the difference of the two conditions in mean.
In the interaction model Y is given by

$$Y = \begin{cases} \beta_0 - \beta_1 - \beta_2 + \Theta_{12} & \text{if } B_1 = -1 \text{ and } B_2 = -1 \\ \beta_0 + \beta_1 - \beta_2 - \Theta_{12} & \text{if } B_1 = 1 \text{ and } B_2 = -1 \\ \beta_0 - \beta_1 + \beta_2 - \Theta_{12} & \text{if } B_1 = -1 \text{ and } B_2 = 1 \\ \beta_0 + \beta_1 + \beta_2 + \Theta_{12} & \text{if } B_j = 1 \text{ and } B_k = 1. \end{cases}$$

Now, β_0 represents the mean of the four group means (if the design is completely balanced this is the overall mean). The main effect coefficients β_j, $j \in \{1, 2\}$ are the average difference effects between the two conditions the respective feature can take. The interaction effect

$2\Theta_{jk}$ explains the difference between the two conditions when either both features are present or absent and when only one of the two features is present.
Moreover, there exists a linear transformation between the coefficients of both encodings. We denote all coefficients in the 0 and 1 encoding as $\tilde{\beta}$ and the coefficients in the -1 and 1 encoding as β. The transformation between both encodings in the quadratic interaction model for $p = 2$ is given by

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 - \beta_1 - \beta_2 + \Theta_{12} \\
\tilde{\beta}_1 &= 2(\beta_1 - \Theta_{12}) \\
\tilde{\beta}_2 &= 2(\beta_2 - \Theta_{12}) \\
\tilde{\Theta}_{12} &= 4\Theta_{12}.
\end{aligned}$$

For $p \geq 2$ this can be translated to a general form as by

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 - \sum_{j=1}^{p} \beta_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \Theta_{jk} \\
\tilde{\beta}_j &= 2\beta_j - 2 \sum_{\substack{k=1 \\ k \neq j}}^{p} \Theta_{jk}, \text{ for } j = 1, ..., p \\
\tilde{\Theta}_{jk} &= 4\Theta_{jk}, \text{ for } j = 1, ..., p-1, \; k = j+1, ..., p.
\end{aligned}$$

The transformation from one encoding to the other can be derived by replacing the input matrix in the model according to this equation: $B^{\{-1,1\}} = 2B^{\{0,1\}} - 1$.

Alr transformed model versus constrained log contrast model

Main effects only:

$$ALR = \sum_{j=1}^{p-1} \beta_j \log \frac{X_j}{X_p} = \sum_{j=1}^{p-1} \beta_j \log X_j - \left(\sum_{j=1}^{p-1} \beta_j \right) \log X_p$$

Thus, we can define $\beta_p := -\sum_{j=1}^{p-1} \beta_j$ and then write this as $\sum_{j=1}^{p} \beta_j \log X_j$ s.t. $\sum_{j=1}^{p} \beta_j = 0$, which corresponds to the log contrast model.

Model with interactions:

$$
\begin{aligned}
ALR &= \sum_{j=1}^{p-1} \beta_j \log \frac{X_j}{X_p} + \sum_{1 \le j,k \le p} \Theta_{jk} \log \frac{X_j}{X_p} \log \frac{X_k}{X_p} \\
&= \sum_{j=1}^{p-1} \beta_j \log X_j + \left(- \sum_{j=1}^{p-1} \beta_j \right) \log X_p \\
&+ \sum_{1 \le j < k < p} \Theta_{jk} \left(\log X_j \log X_k - \log X_j \log X_p - \log X_k \log X_p + \log^2 X_p \right) \\
&= \sum_{j=1}^{p} \beta_j \log X_j \quad (\text{taking } \sum_{j=1}^{p} \beta_j := 0) \\
&+ \sum_{1 \le j,k \le p} \Theta_{jk} \log X_j \log X_k - 2 \log X_p \sum_{j=1}^{p-1} \left(\sum_{k=1}^{p-1} \Theta_{jk} \right) \log X_j + \left(\sum_{j,k} \Theta_{jk} \right) \log^2 X_p
\end{aligned}
$$

For $j < p, \Theta_{jp} = \Theta_{pj} := -\sum_{k=1}^{p-1} \Theta_{jk}$

For $\Theta_{pp} := \sum_{1 \le j,k \le p} \theta_{jk}$.

Note: $\sum_{j=1}^{p} \Theta_{jp} = \sum_{j=1}^{p-1} \left(- \sum_{k=1}^{p-1} \Theta_{jk} \right) + \left(\sum_{1 \le j,k < p} \Theta_{jk} \right) = 0$

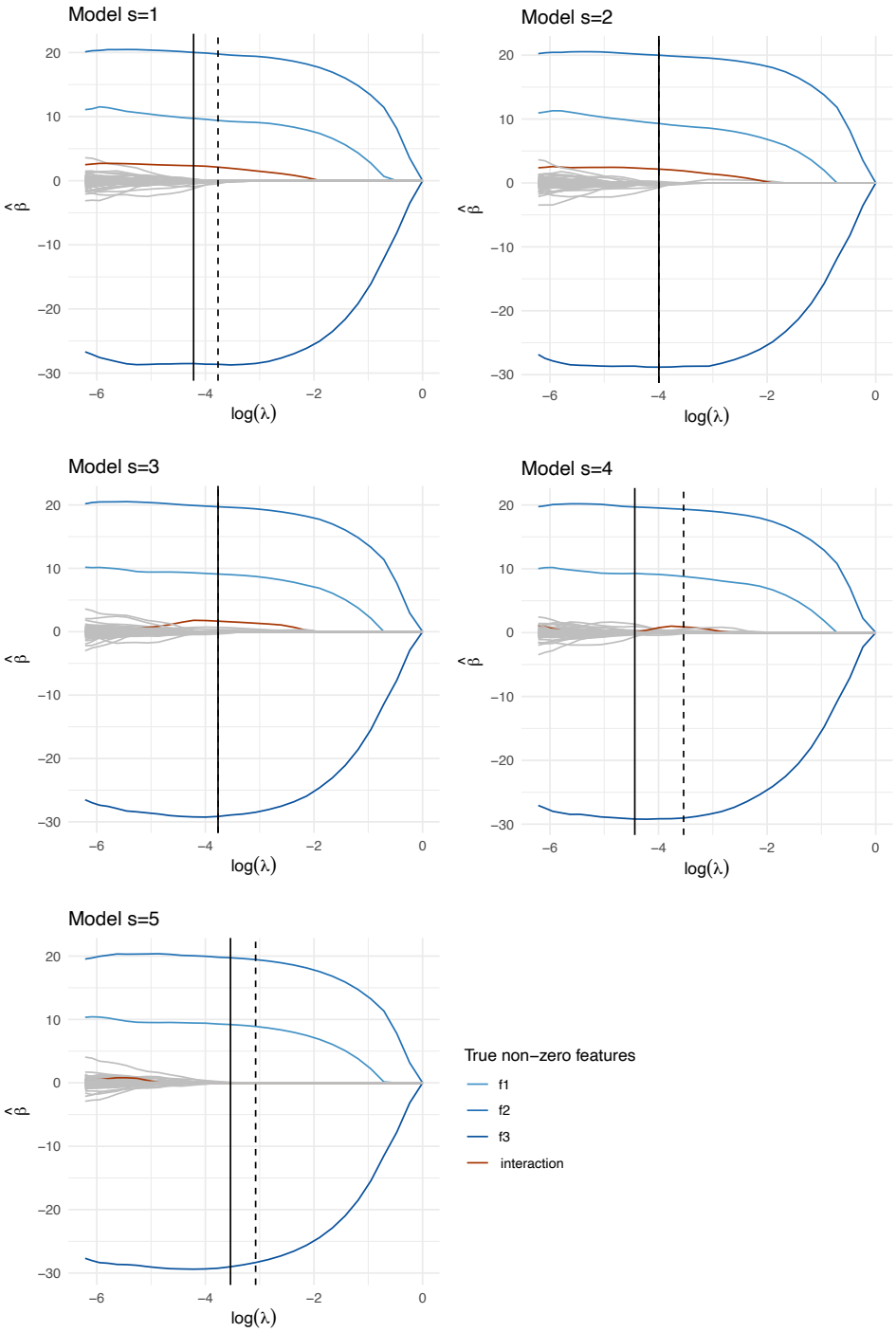

Fig S1. Solution path of the interaction model (sparse qlc) for the $S = 5$ semi-synthetic simulation setups for varying feature sparsity levels.