Journal Title Here, 2022, pp. 1–14

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Variational inference for microbiome survey data with application to global ocean data

Aditya Mishra,[1,6,*] Jesse McNichol,[2] Jed Fuhrman,[2] David Blei[1,3] and Christian L. Müller[1,4,5]

[1]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA, [2]Department of Biological Sciences, University of Southern California, LA, USA, [3]Department of Statistics and Computer Science, Columbia Universit, NY, USA, [4]Computational Health Center, Helmholtz Zentrum München, Munich, Germany, [5]Department of Statistics, LMU München, Munich, Germany and [6]Department of Statistics, University of Georgia, Athens, GA, USA
[*]Corresponding author. aditya.mishra@uga.edu

Abstract

Linking sequence-derived microbial taxa abundances to host (patho-)physiology or habitat characteristics in a reproducible and interpretable manner has remained a formidable challenge for the analysis of microbiome survey data. Here, we introduce a flexible probabilistic modeling framework, VI-MIDAS (Variational Inference for MIcrobiome survey DAta analysiS), that enables *joint* estimation of context-dependent drivers and broad patterns of associations of microbial taxon abundances from microbiome survey data. VI-MIDAS comprises mechanisms for direct coupling of taxon abundances with covariates and taxa-specific latent coupling which can incorporate spatio-temporal information *and* taxon-taxon interactions. We leverage mean-field variational inference for posterior VI-MIDAS model parameter estimation and illustrate model building and analysis using Tara Ocean Expedition survey data. Using VI-MIDAS' latent embedding model and tools from network analysis, we show that marine microbial communities can be broadly categorized into five modules, including SAR11-, Nitrosopumilus-, and Alteromondales-dominated communities, each associated with specific environmental and spatiotemporal signatures. VI-MIDAS also finds evidence for largely positive taxon-taxon associations in SAR11 or Rhodospirillales clades, and negative associations with Alteromonadales and Flavobacteriales classes. Our results indicate that VI-MIDAS provides a powerful integrative statistical analysis framework for discovering broad patterns of associations between microbial taxa and context-specific covariate data from microbiome survey data.

Key words: Microbiome; Probabilistic model; Association learning; Variational inference; Tara ocean expedition

Introduction

Microbial species are an integral part of life on earth. Ecosystems, ranging from the human gut to the global ocean, harbor trillions of bacteria, archaea, viruses, and fungi that take on essential functional roles and have developed intricate ecological relationships within their respective habitat. Over the past decades, advances in amplicon and metagenomics sequencing techniques [74, 54, 52, 70] and standardized experimental and bioinformatics workflows [63, 10, 9] have enabled the large-scale collection and dissemination of microbial survey data, including those from the seminal Human Microbiome Project [69], several gut-focused surveys [28, 64, 32, 45], the Earth Microbiome Project [25], and the Tara Ocean Expedition [67]. These surveys have reached a level of maturity and complexity that ultimately allow the estimation of statistical associations between microbial abundances, typically represented as compositional counts of Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units (OTUs), and habitat properties [67, 7], biogeochemical processes[29], and/or host health status [23, 48]. This, in turn, provides a starting point for deciphering and understanding the ecological and functional roles of different microbial clades in the ecosystem, nutrient and bio(geo)chemical dependencies, resource limitations of microbial growth, and the presence of ecological taxon-taxon interactions [18].

Here, we introduce an integrative probabilistic modeling framework that is specifically tailored to microbiome survey data and enables joint estimation of habitat-dependent drivers and broad associations patterns of microbial taxa abundances (see Figure 1). Our approach, termed VI-MIDAS (Variational Inference for MIcrobiome survey DAta analysiS), models the observed taxon abundances by *simultaneously* learning taxon-specific latent representations that leverage the effects of host or environmental factors *and* taxon-taxon associations via an item-item interaction modeling *ansatz*, originally proposed for market basket analysis [61]. As such, VI-MIDAS seamlessly extends common statistical methods for microbiome data that only focus on either statistical abundance modeling [31, 39, 13, 79, 75, 49] or microbial association estimation[37, 18, 77, 57, 26].

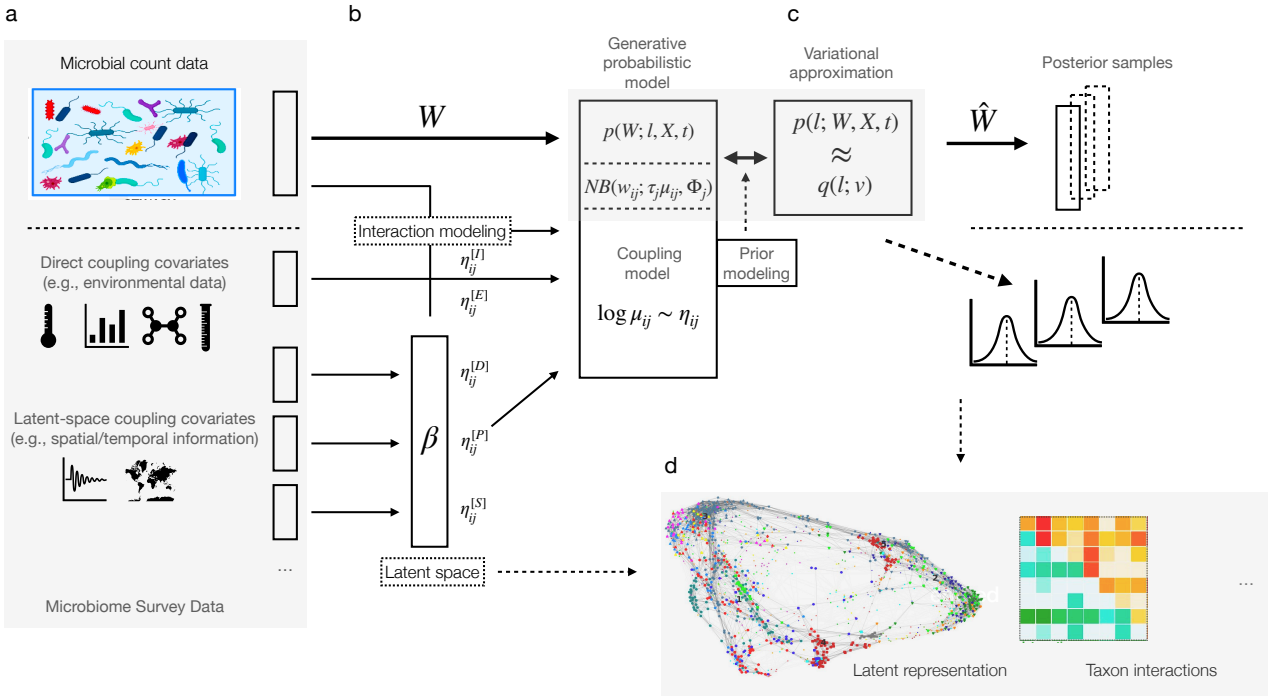

Fig. 1. Overview of the VI-MIDAS framework. **a.** VI-MIDAS integrates microbiome survey data in form of microbial abundance data W, host-associated, habitat or environmental data, and spatio-temporal information. **b.** Different data sources are coupled directly *or* indirectly through a latent space $\boldsymbol{\beta}$ to a generative model. An additional latent space taxon interaction model is included. The generative probabilistic model (e.g., Negative Binomial (NB) model) integrates covariate data via a coupling model. **c.** Variational approximation and mean-field estimation are used for Bayesian parameter estimation, resulting in posterior microbial abundance samples $\hat{W}$ and model parameter distributions. **d.** Model components, such as estimated latent representation and taxon-taxon interactions, can be used for data understanding, visualization, and downstream analysis.

VI-MIDAS uses the parametric structure of the Negative binomial distribution [46, 49] to account for the overdispersed nature of the amplicon count data and comprises two main model components: (i) a component that allows for full adjustment of taxon abundances from a user-defined subset of covariates and (ii) taxa-specific latent vectors that incorporate, e.g., spatio-temporal or environmental covariates *and* taxon-taxon interactions, thus providing a marginal characterization of each taxon. We resort to mean-field variational inference for parameter estimation of VI-MIDAS' intractable posterior distribution [8], thus complementing other recent variational approaches to microbiome data modeling, such as, e.g., Poisson principal component analysis [14], microbiome dynamics modeling [24], Dirichlet Multinomial modeling [30], multi-level modeling [42], and microbiome ordination [78].

To illustrate the complete workflow of the VI-MIDAS framework, we focus on integrative analysis of global marine microbiome survey data. The ocean microbiome is of fundamental importance for life on earth, being responsible for about half of all primary production (i.e., the production of chemical energy in organic compounds) and holds enormous potential for climate remediation [50]. Several initiatives such as the Tara Oceans Project [56] and the Simons CMAP [4] provide well-structured sequencing data, biogeochemical and environmental covariate data, and satellite-derived products that are amenable to statistical analysis. Here, we re-analyze Tara expedition data [1], originally considered in [67] to study the structure and function of the global ocean microbiome. The expedition collected ocean water samples from 68 distinct geographical locations at varying levels of depth. We will make extensive use of this dataset to motivate and describe the details of the VI-MIDAS framework as well as the learned representations and associations of the global ocean microbiome.

We start with an overview of the Tara Oceans data under study, introduce the generative model components of VI-MIDAS, and show how different data types enter the modeling framework. We then give a high-level overview of the variational parameter estimation procedure, including the selection of VI-MIDAS' hyperparameters, such as the choice of the priors and the dimensionality of the latent representation. Following model parameter inference, we illustrate how standard modularity analysis of VI-MIDAS' learned latent representation of the Tara data identifies five distinct groups of microbial consortia. We analyze the inferred modules in terms of their composition of ecologically relevant clades and discuss the derived module-specific environmental and spatiotemporal signatures. Finally, we highlight the emerging interaction pattern among ecologically relevant clades and discuss the framework in the larger context of other microbiome survey data. Further methodological details are summarized in the Supplemental Material. Code for the presented VI-MIDAS workflow is available at `http://github.com/amishra-stats/vi-midas`) and requires minimal adjustment to analyze other microbiome survey data.

[1] `http://ocean-microbiome.embl.de/companion.html`

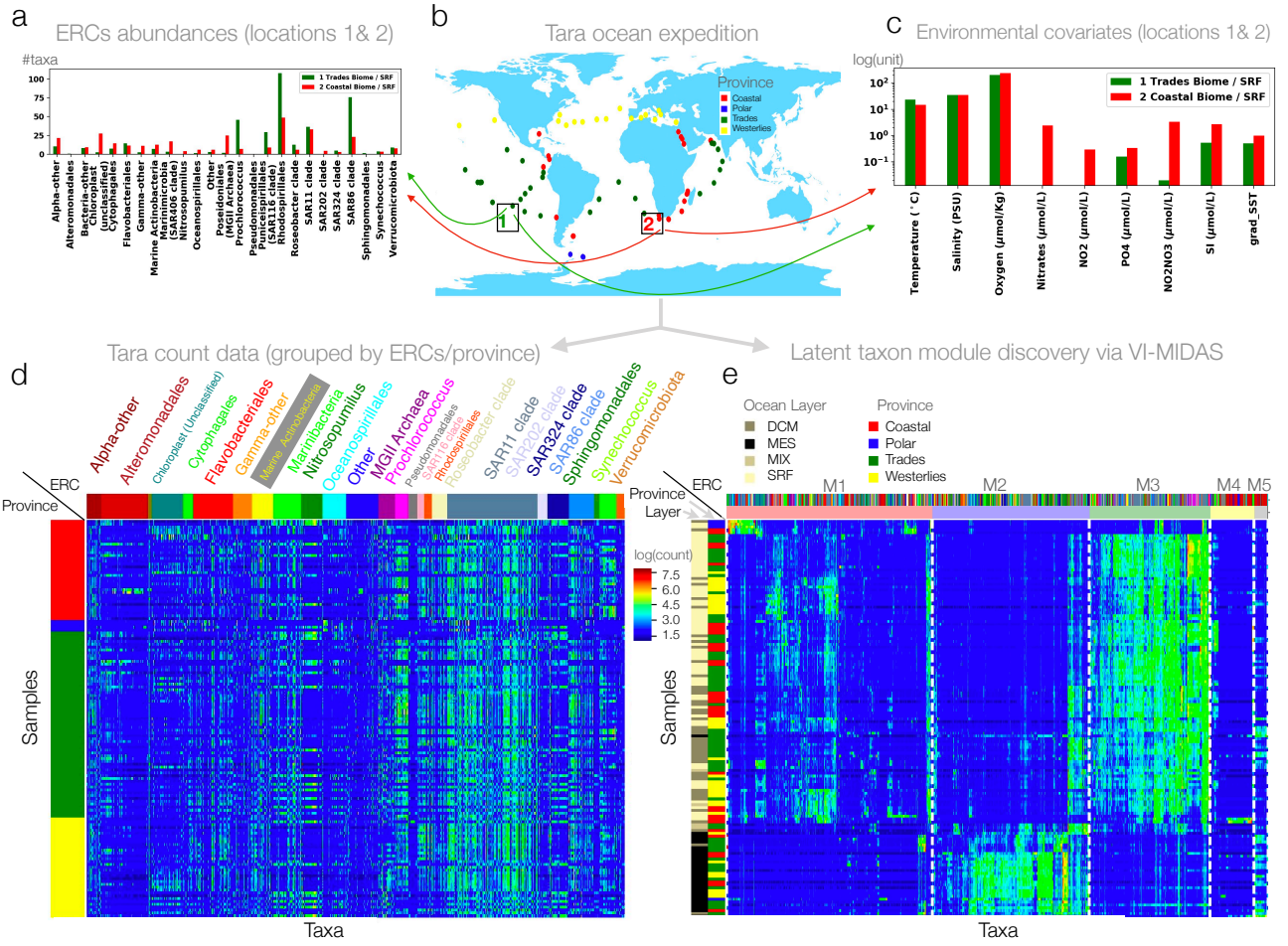

Fig. 2. Illustration of the Tara ocean data: **a.** Taxon abundance profiles, agglomerated to expert-derived ecologically relevant classes (ERCs) for two samples (red and green, marked as 1 and 2 in Figure 2b). **b.** Tara ocean sample locations. **c.** Environmental features associated with the samples marked as 1 and 2 in Figure 2 (b); **d.** Abundance profiles $\log(\mathbf{W}+1)$ of $q = 1379$ taxa at $n = 139$ distinct locations with rows highlighting province of the sample and columns grouped by ERC. **e.** Abundance profiles clustered into five modules (M1-M5) as identified by modularity analysis of the latent space $\boldsymbol{\beta}$ (see Section Modularity analysis for more details). The dashed vertical lines separate the latent modules. The five microbial modules (M1-M5) comprise 524, 400, 307, 112 and 35 taxa/OTUs, respectively. The first column shows ocean depth layer, the second column the province indicator.

Materials and Methods

Tara ocean data and ecologically relevant taxa re-classification

We consider the processed Tara expedition data, as provided at `http://ocean-microbiome.embl.de/companion.html`. The expedition collected water samples from 68 distinct geographical locations (Figure 2b) across different depths, resulting in $n = 139$ distinct samples. Across these samples, the original data comprises microbial taxa abundances profiles of more than 35,000 bacterial taxa in form of metagenomic OTUs (mOTUs) (derived using the miTAGS framework [68]).

Here, we focus on the most abundant taxa by taking the union of all mOTUs that, in each individual sample, contribute to 40% of the total library size. This filtering allows us to cover the abundance profiles of the $q = 1378$ taxa with the most significant variability and reduces the number of excess zero counts. To account for the highly variable sequencing depth across the samples, we normalize the abundance data with respect to the lowest library size via common-sum scaling [46]. Figure 2d shows the log-transformed abundance profiles $\mathbf{W} \in \mathbb{R}^{n \times q}$. Since the original taxonomic affiliations of the miTAGS are difficult to interpret, we next developed a partitioning of the selected taxa into ecologically relevant classes (ERCs). The original full taxonomy strings are too long to understand at a glance, and parsing by taxonomic level is not a good option since taxa vary widely in the depth of their annotations. For example, cyanobacteria should be annotated at the genus level or higher, but many other abundant but less described taxa do not have any taxonomic information at that level. We manually curated the data to provide a short relevant taxonomic indicator that provides a rough indicator of the ecological niche of an organism while remaining short enough to be interpreted at a glance. Some taxonomies have been altered to preserve the updated SILVA taxonomy (i.e., Betaproteobacteria is now Burkholderiales). New SILVA 138 [58] taxonomies have been used wherever possible (i.e., when the original ID was still in SILVA 138), but in cases where there was only the SILVA 108 taxonomic information, we have used our best guess. For example, if an organism had the same classification as other organisms in SILVA 108, we have often

Table 1. Environmental and spatiotemporal variables included in the VI-MIDAS model

Model component	Variables	Description
Environmental $\eta_{ij}^{[E]}$	**Environmental covariates**	Sea surface temperature (and its gradient), salinity, chlorophyll, nitrate, Nitrogen Dioxide, Phosphate, Silicon, and oxygen concentration.
Spatial (Depth) $\eta_{ij}^{[D]}$	**SRF**	Surface water layer; up to 5 m below the surface
	DCM	Deep chlorophyll maximum; approximately 17 m to 188 m below the surface; region below the surface with maximum chlorophyll concentration
	MIX	Subsurface epipelagic mixed layer; approximately 25 m to 150 m below the surface
	MES	Mesopelagic zone; approximately 250 m to 1000 m below the surface
Spatial (Longhurst Province) $\eta_{ij}^{[P]}$	**Polar biome**	Polar region in the northern and southern hemisphere characterized by low taxonomic diversity at all trophic levels.
	Westerlies biome	High-latitude region below the westerly winds
	Trades biome	Low-latitude region below the easterly trades characterized by high taxonomic diversity
	Coastal biome	Region in the upper part of the continental slope
Seasonal $\eta_{ij}^{[S]}$	**Q1, Q2, Q3, Q4**	Derived indicator of seasonal quarter when sample was taken (January to March; April-June; July-September; ctober-December)

given it the same name as its counterparts in SILVA 138. We present all our findings in terms of these 29 ERCs.

Each Tara sample also contains environmental and spatiotemporal information, including geolocation, the derived Longhurst province (biome) indicator, sampling date, ocean depth information (depth from sea surface), environmental covariates, such as, e.g., sea surface temperature (SST), and biogeochemical features such as salinity, chlorophyll, nitrate, and oxygen concentration (see Figure 2c for illustration). Table 1 summarizes the measured covariates and derived spatiotemporal indicator variables that are included in the VI-MIDAS framework and their corresponding mathematical representation.

Generative Modeling in VI-MIDAS

We seek to model the abundance profiles of q microbial taxa where we denote a single sample by the random variable $\mathbf{w} \in \mathbb{R}^q$ and the observed data from n samples by $\mathbf{W} = [w_{ij}]_{n\times q} \in \mathbb{R}^{n\times q}$. For concreteness, we illustrate model building and analysis using the Tara abundance profiles (see Figure 2(d)) of $q = 1378$ taxa but the modeling strategy is applicable to any multimodal microbiome survey.

Distributional model

VI-MIDAS posits that the overdispersed microbial count data $\mathbf{W}$ are reasonably well modeled with the Negative Binomial distribution [11, 44, 48]. While other generative statistical modeling approaches are available, including the Dirichlet Multinomial (mixture) framework [31, 71], latent Dirichlet allocation [62], and Poisson distribution models [39, 5, 75], we found the Negative Binomial model to be an excellent choice for the Tara ocean data (see Figure S1 (b) of the Supplementary Material for the over-dispersion analysis). Using the Negative Binomial distribution with mean and dispersion parameterization [11], VI-MIDAS models the jth taxa in the ith sample as:

$$p(w_{ij}; \tau_j\mu_{ij}, \phi_j) = \mathrm{NB}(w_{ij}; \tau_j\mu_{ij}, \phi_j)$$
$$= \binom{w_{ij} + \phi_j - 1}{w_{ij}} \left(\frac{\tau_j\mu_{ij}}{\tau_j\mu_{ij} + \phi_j}\right)^{w_{ij}} \left(\frac{\phi_j}{\tau_j\mu_{ij} + \phi_j}\right)^{\phi_j}. \quad (1)$$

Here, the mean parameter $\tau_j\mu_{ij}$ is the product of a taxon-specific shape parameter $\tau_j \in (0, 1)$ and the entry-specific parameter $\mu_{ij} \in \mathbb{R}^+$. The parameter $\phi_j \in \mathbb{R}^+$ is the taxon-specific dispersion parameter. Let us denote the dispersion and shape parameters for q outcomes by $\boldsymbol{\Phi} = [\phi_1, \ldots, \phi_q]$ and $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_q]$, respectively. The shape parameter $\boldsymbol{\tau}$ accounts for the disparity in abundance among microbial taxa. The generative model (1) of VI-MIDAS thus implies $\mathbb{E}(w_{ij}) = \tau_j\mu_{ij}$ and $\mathrm{Var}(w_{ij}) = \tau_j\mu_{ij} + \frac{\tau_j^2\mu_{ij}^2}{\phi_j}$. Consequently, $\mathrm{Var}(w_{ij}) > \mathbb{E}(w_{ij})$, thus making the parametric framework (1) suitable for modeling the overdispersed count data.

Modeling strategy and model components

One novelty in VI-MIDAS is the combination of ideas from generalized linear modeling [11] and compositional data analysis [2] to associate the microbial relative count data with spatiotemporal, environmental, and taxa information. Specifically, we model the log-transformed mean parameter $\boldsymbol{\mu} = [\mu_{ij}]_{n\times q}$ of the generative model (1) with two components, a consistent zero-aware geometric mean estimate t_i and a linear predictor $\boldsymbol{\eta} = [\eta_{ij}]_{n\times q} \in \mathbb{R}^{n\times q}$ as follows:

$$\log \mu_{ij} = \log t_i + \eta_{ij}, \ . \quad (2)$$

The sample-wise parameter t_i is estimated by a zero-aware geometric mean estimator, introduced in [16], which provides a principled approximation to the geometric means across all n samples in the presence of excess zeros. We detail the exact formulation of t_i and its approximation guarantees in Section 3.1 of the Supplementary Material. Including $\mathbf{O} = [\log t_1, \ldots, \log t_n]$ as an offset term in the model is necessary since we do not have access to absolute microbial abundance data, thus requiring transforming the compositional data appropriately. The second term $\boldsymbol{\eta}$ effectively models centered log-ratio (clr) transformed (rather than the original count) data and is the key component to couple habitat (or host) information to the microbial abundance profiles. VI-MIDAS introduces a novel decomposition of the component $\boldsymbol{\eta}$ that allows the incorporation of three distinct coupling mechanisms: (i) a direct coupling term for covariates, (ii) an indirect coupling term for covariates via a latent space representation, and (iii) a latent taxon-taxon interaction term.

In our ocean application, the first component, denoted by $\eta_{ij}^{[E]}$, includes all relevant environmental attributes (see first row in Table 1). All spatiotemporal features, i.e., the Longhurst **P**rovince indicator, the **D**epth information, and the **S**easonal

indicator (see second to last row in Table 1) are handled by the latent coupling term and are denoted by $\eta_{ij}^{[D]}$, $\eta_{ij}^{[P]}$, and $\eta_{ij}^{[S]}$, respectively. Lastly, statistical associations among co-occurring taxa are included via a latent interaction term $\eta_{ij}^{[I]}$, leading the following model:

$$\eta_{ij} = \eta_{ij}^{[E]} + (\eta_{ij}^{[P]} + \eta_{ij}^{[D]} + \eta_{ij}^{[S]}) + \eta_{ij}^{[I]} \,. \tag{3}$$

The following paragraphs detail the parametric form of each of the components, the nature of the underlying covariate data, and their biological relevance.

Direct coupling of environmental features

Let us denote the p covariates in the direct coupling term by $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} = [x_{ij}]_{n\times p}$. VI-MIDAS models the direct component for the jth taxa in the ith sample via

$$\eta_{ij}^{[E]} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\gamma}_{\cdot j} \,. \tag{4}$$

with $\boldsymbol{\gamma} = [\gamma_{ij}]_{p\times q} \in \mathbb{R}^{p\times q}$ denoting the matrix of all coefficients. For the Tara data, we opted to model $\eta_{ij}^{[E]}$ using following $p = 9$ covariates: sea surface temperature (SST) (and its gradient grad SST), salinity, chlorophyll, nitrate, nitrogen dioxide, phosphate, silicon, and oxygen concentration. All variables are mean-centered prior to incorporation into the model. In the original Tara analysis [67], temperature and oxygen have been identified as key drivers of taxonomic compositions. The VI-MIDAS analysis will allow a refined picture of the these general tendencies.

Latent space coupling of spatiotemporal features

VI-MIDAS offers a second mechanism for including variables of interest through latent space modeling. We denote q taxa-specific shared latent variables of size k by $\boldsymbol{\beta} = [\beta_{ij}]_{k\times q} \in \mathbb{R}^{k\times q}$. The size factor k is an application-specific hyper-parameter that controls the expressiveness of the latent space. Features are then coupled to the latent space in a multiplicative fashion.

For the Tara data, we illustrate this mechanism by coupling all available spatial and temporal indicators to the latent space component. We first consider the $r = 4$ primary provinces (or biomes): polar, Westerlies, coastal, and Trades [43]. We denote the model matrix indicating the r distinct regions of the n samples by $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times r}$ and connect it to the joint latent space via the coefficient matrix $\boldsymbol{\alpha} = [\alpha]_{r\times k} \in \mathbb{R}^{r\times k}$, leading to

$$\eta_{ij}^{[P]} = \mathbf{r}_i \boldsymbol{\alpha} \boldsymbol{\beta}_{\cdot j} \,. \tag{5}$$

Similarly, the Tara data includes samples across $b = 4$ ocean depths: surface water (SRF), deep chlorophyll maximum (DCM), the subsurface epipelagic mixed layer (MIX), and the mesopelagic zone (MES). We denote the depth indicator matrix of the n samples by $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ and connect it to the joint latent space via the coefficient matrix $\boldsymbol{\delta} = [\delta]_{b\times k} \in \mathbb{R}^{b\times k}$, leading to

$$\eta_{ij}^{[D]} = \mathbf{d}_i \boldsymbol{\delta} \boldsymbol{\beta}_{\cdot j} \,. \tag{6}$$

Finally, by parsing the sampling dates at the different Tara locations, we can associate a temporal indicator with each sample. Here, we group the samples into $m = 4$ seasons: the 1^{st} (Q1, January-March), 2^{nd} (Q2, April-June), 3^{rd} (Q3, July-September), and 4^{th} (Q4, October-December) yearly quarter, and construct the season indicator matrix $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times s}$. The coefficient matrix $\boldsymbol{\vartheta} = [\vartheta]_{m\times k} \in \mathbb{R}^{m\times k}$ couples $\mathbf{S}$ to the latent space $\boldsymbol{\beta}$, leading to

$$\eta_{ij}^{[S]} = \mathbf{s}_i \boldsymbol{\vartheta} \boldsymbol{\beta}_{\cdot j} \,. \tag{7}$$

In summary, the coupling of the described features to a shared latent space via the coefficient matrices $\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\vartheta}$ allows to quantify to what extent spatiotemporal information influences each taxon's (latent) abundance after discounting the contribution of the environmental component.

Latent modeling of taxon-taxon associations

It is well-established that the abundances of species in an ecosystem are not only driven by environmental or spatiotemporal factors but also by interactions among the species themselves [41]. While discovering detailed ecological interactions among taxa, such as, e.g., competition, mutualism, or commensalism, is beyond the reach of coarse-grained statistical models, VI-MIDAS' latent space modeling offers a principled mechanism to assess the influence of taxa *co-occurrences* on their respective abundances. We achieve this by borrowing recent ideas from market basket analysis and adopt the so-called SHOPPER utility model for interaction analysis [61]. In SHOPPER, Ruiz et al. [61] proposed a probabilistic model based on the basket data from a supermarket to learn about the latent characteristic of each item and exchangeable/complementary interactions among items. The approach uses item-specific latent variables to define an item-item interaction component. Following their setup, the "interaction", or, in the biological context, association of the jth taxa with any mth taxa is given by $\boldsymbol{\rho}_{\cdot j}^{\mathrm{T}} \boldsymbol{\beta}_{\cdot m}$ where $\boldsymbol{\rho} = [\rho]_{k\times q} \in \mathbb{R}^{k\times q}$ comprises length-k latent variables for each of the q taxa. The entries of VI-MIDAS' interaction component $\eta^{[I]}$ for the jth taxon in the ith sample are thus given by

$$\eta_{ij}^{[I]} = \begin{cases} 0, & w_{ij} = 0 \\ \frac{1}{a_i - 1} \boldsymbol{\rho}_{\cdot j}^{\mathrm{T}} \sum_{m\neq j} \mathbf{1}_{w_{im}\neq 0} \boldsymbol{\beta}_{\cdot m}, & w_{ij} \neq 0 \,, \end{cases} \,, \tag{8}$$

where $a_i = \sum_{m=1}^{q} \mathbf{1}_{w_{im}\neq 0}$ is the total number of taxa present in the ith sample. Note that the interaction term $\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\beta}$ is not symmetric. However, we can derive a symmetrized $\mathbf{I} = [I_{i,j}] \in \mathbb{R}^{q\times q}$ with each entry being computed as:

$$I_{i,j} = (\boldsymbol{\rho}_{\cdot i}^{\mathrm{T}} \boldsymbol{\beta}_{\cdot j} + \boldsymbol{\rho}_{\cdot j}^{\mathrm{T}} \boldsymbol{\beta}_{\cdot i})/2 \tag{9}$$

This allows easier downstream network analysis of potentially *positive* (mutualistic) and *negative* (competitive) associations among the taxa, or in our case, among the ecologically relevant clades.

Variational inference in VI-MIDAS

The generality and flexibility of VI-MIDAS poses a considerable challenge for fast and accurate model parameter estimation. We introduce a variational inference framework that makes estimation in VI-MIDAS feasible and illustrate its performance and parameter sensitivities using the Tara data. For ease of presentation, we summarize the key ingredients below and refer to the extensive Supplementary Information and the documented code base available at `https://github.com/amishra-stats/vi-midas`) for details.

Bayesian model and variational approximation

We begin by denoting all (latent) parameters in the VI-MIDAS framework by $\boldsymbol{\ell} = \{\boldsymbol{\alpha}, \boldsymbol{\vartheta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\tau}, \boldsymbol{\Phi}\}$ (see Table S1 of the

Supplementary Material). Given the microbial abundance data $\mathbf{W}$, the (direct) covariates $\mathbf{X}$, and the model parameters $\boldsymbol{\ell}$, we integrate the generative model (1) into a Bayesian framework where the posterior distribution reads:

$$p(\boldsymbol{\ell}; \mathbf{W}, \mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{W}; \boldsymbol{\ell}, \mathbf{X}, \mathbf{t}) p(\boldsymbol{\ell})}{p(\mathbf{W}; \mathbf{X}, \mathbf{t})}, \tag{10}$$

where $p(\mathbf{W}; \boldsymbol{\ell}, \mathbf{X}, \mathbf{t}) = \prod_{i,j} p(w_{ij}; \tau_j \mu_{ij}, \phi_j)$ denotes the likelihood of $\mathbf{W}$ and $p(\boldsymbol{\ell}) = p(\boldsymbol{\alpha})p(\boldsymbol{\delta})p(\boldsymbol{\beta})p(\boldsymbol{\gamma})p(\boldsymbol{\rho})p(\boldsymbol{\Phi})p(\boldsymbol{\tau})p(\boldsymbol{\vartheta})$ the prior distribution, respectively. To achieve good generalizabilty and interpretability of VI-MIDAS' over-parameterized model, we place sparsity-inducing Laplace priors with scale parameter λ on each of the unconstrained latent variables in the set $\{\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\vartheta}\}$. For example, the prior on $\boldsymbol{\alpha}$ reads $p(\boldsymbol{\alpha}) = \prod_{i,j} p(\alpha_{ij})$ with $p(\alpha_{ij}) = \text{Laplace}(0, \lambda)$. Furthermore, we place an inverse-Cauchy prior on the dispersion parameter $\boldsymbol{\Phi}$, i.e., $p(\phi_j) = \text{inverse-Cauchy}(0, \upsilon)$ and $p(\boldsymbol{\Phi}) = \prod_j p(\phi_j)$, and a Uniform(1,2) prior for the shape parameter $\boldsymbol{\tau}$, i.e., $\tau_j \sim \text{Beta}(1,1)$ and $p(\boldsymbol{\tau}) = \prod_j p(\tau_j)$. Choosing suitable hyperparameters for the priors will be discussed below.

In the high-dimensional setting, computing the posterior distribution is challenging because of the intractable form of the marginal distribution $p(\mathbf{W}; \mathbf{X}, \mathbf{t})$ and the non-conjugate priors on the model parameters. Markov Chain Monte Carlo (MCMC) sampling provides a helpful paradigm for obtaining samples from the posterior distribution in the Bayesian framework. However, since MCMC lacks computational efficiency in large/high-dimensional problems, we use mean-field Variational Inference (VI) [34, 72, 8] and approximate the posterior with a variational posterior distribution of the latent variable $\boldsymbol{\ell}$. Briefly, let $q(\boldsymbol{\ell}; \boldsymbol{\nu})$ be the variational posterior distribution with parameter $\boldsymbol{\nu}$. VI approximates sampling of the posterior by minimizing the Kullback-Leibler (KL) divergence,

$$\min_{\boldsymbol{\nu}} \ \text{KL}(q(\boldsymbol{\ell}; \boldsymbol{\nu}) \,||\, p(\boldsymbol{\ell}; \mathbf{W}, \mathbf{X}, \mathbf{t}))$$

such that $\text{supp}(q(\boldsymbol{\ell}; \boldsymbol{\nu})) \subseteq \text{supp}(p(\boldsymbol{\ell}; \mathbf{W}, \mathbf{X}, \mathbf{t}))$. It can be shown that the above optimization problem simplifies to maximizing the evidence lower bound (ELBO) given by

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\ell};\boldsymbol{\nu})}[\log P(\mathbf{W}, \boldsymbol{\ell}; \mathbf{X}, \mathbf{t})] - \mathbb{E}_{q(\boldsymbol{\ell};\boldsymbol{\nu})}[\log q(\boldsymbol{\ell}; \boldsymbol{\nu})], \tag{11}$$

which is a lower bound on the logarithm of the joint probability of the observations $\log P(\mathbf{W}; \mathbf{X}, \mathbf{t})$ [34]. Replacing the joint distribution $P(\mathbf{W}, \boldsymbol{\ell}; \mathbf{X}, \mathbf{t})$ with a product of likelihood and prior distribution $P(\mathbf{W}, \boldsymbol{\ell}; \mathbf{X}, \mathbf{t}) = P(\mathbf{W}; \boldsymbol{\ell}, \mathbf{X}, \mathbf{t})P(\boldsymbol{\ell})$ further simplifies the objective.

Model estimation, hyperparameter tuning, and posterior estimates

The non-convexity of the variational objective and the large number of model parameters require careful assessment of all aspects of model parameter estimation, hyperparameter tuning, and generalization capability. To estimate the parameters of the variational posterior distribution, we employ stochastic gradient descent within the automatic differentiation variational inference (ADVI) framework [36]. The key steps of ADVI are outlined in Algorithm 1 of the Supplementary Material. A prerequisite for model parameter estimation is the identification of suitable model *hyperparameters*. In VI-MIDAS, the key hyperparameters are the scale of the sparsity-inducing Laplace prior, the scale of the inverse-Cauchy prior, and the intrinsic dimensionality k of the latent space $\boldsymbol{\beta}$, respectively. VI-MIDAS tunes these parameters via random search (see Section 3.3 of the Supplementary Material for details) where the out-of-sample log-likelihood posterior predictive density (LLPD) is used for assessing optimality of the hyperparameters [22]. Due to the non-convexity of the objective and the use of stochastic optimization in VI initialization, we further evaluate the suitability of hyperparameter setting across fifty random initializations and select the hyperparameter set leading to the best averaged LLPD (see Section 3.5 of the Supplementary Material). The computational workflow is implemented in Python using the probabilistic programming language Stan [12] and is available in the GitHub repository (`https://github.com/amishra-stats/vi-midas`).

After hyperparameter tuning, we re-estimate the final model parameters on complete data. VI-MIDAS generates $m = 100$ posterior samples of each of the latent variables in the set $\boldsymbol{\ell}$ and estimates the model parameters $\boldsymbol{\ell}$ using the mean of the samples from the variational posterior distribution. The model fit is numerically evaluated using the posterior predictive check [60, 22] on the full data. The procedure requires generating m posterior samples, denoted by the random variables $\mathbf{W}^{rep} = [w_{ij}^{rep}] \in \mathbb{R}_+^{n \times q}$, and then computing the p-value of the model fit as p-value $:= p(t(\mathbf{W}^{rep}) < t(\mathbf{W}))$, where t is the test statistic. In practice, we use the test statistics $t(\mathbf{W}^{rep}) = \mathbf{E}(\log p(\mathbf{W}^{rep}|\boldsymbol{\ell}))$ and $t(\mathbf{W}) = \mathbf{E}(\log p(\mathbf{W}|\boldsymbol{\ell}))$.

Results

VI-MIDAS recapitulates broad statistical patterns of the observed species abundances

VI-MIDAS' hyperparameter tuning revealed that the setting $k = 200$, $\lambda = 0.246$, and $\nu = 0.10063$ achieved the highest average LLPD of 3.332 on the Tara data (see Figure S7 in the Supplementary Material). For this setting, a posterior predictive check on the generated samples achieved a p-value = 0.53. We thus fail to reject the null hypothesis that the posterior samples are different from the observed $\mathbf{W}$. Figure 3a and 3b the observed and estimated abundance profiles (averaged over $m = 100$ samples), respectively. Figure 3c shows the count histograms of data and model (pooled across all samples and species), and Figure 3d the Q-Q plot. We observe that, apart from the low-abundance tail of the distribution, VI-MIDAS broadly recapitulates the statistical abundances patterns across all samples and species.

VI-MIDAS identifies depth and environmental features as main drivers

We next assessed the contribution of each model component toward explaining the species abundance patterns in the Tara data. The modularity of the VI-MIDAS framework facilitates an "ablation" study (see Section 3.4 of the Supplementary Material) where each model component is excluded, followed by a re-evaluation of the out-of-sample LLPD. Table S4 (see Supplementary Materials) shows the LLPDs of the full model and the model after ablation of the environmental(E), province(P), ocean depth(D), seasonality(S), and latent interaction (I) component, respectively.

Firstly, the ablation study confirmed that all components helped improve model generalization since every ablated model has reduced out-of-sample LLPDs. While the seasonality component(S) shows comparatively little influence on explaining the abundance pattern in the current model, as previously observed for this dataset [67], the out-of-sample LLPD is reduced the most when the ocean depth(D) component is

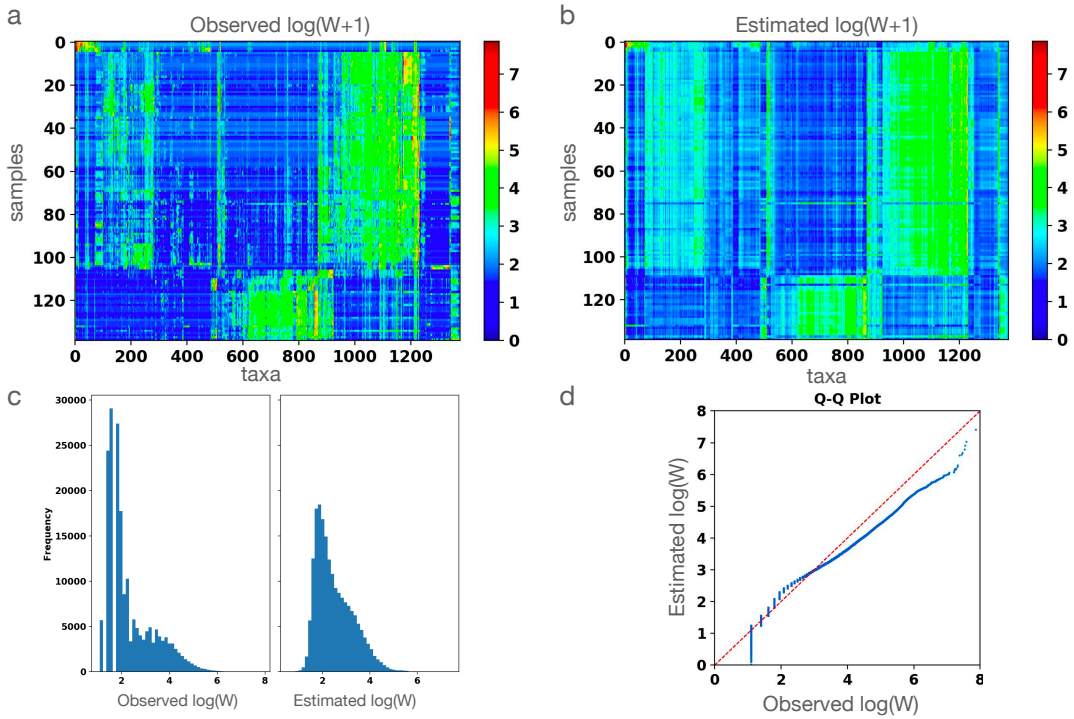

Fig. 3. Comparison of observed abundances and VI-MIDAS posterior samples: **a.** Heatmap showing the abundance profile $\log(\mathbf{W}+1)$ of 1378 species for $n = 139$ samples. **b.** Expected value of the abundance using the hyperparameter corresponding to best model fit. **c.** Histograms of observed and estimated species abundances. **d.** Q-Q plot comparing the observed and estimated abundance profile of the species.

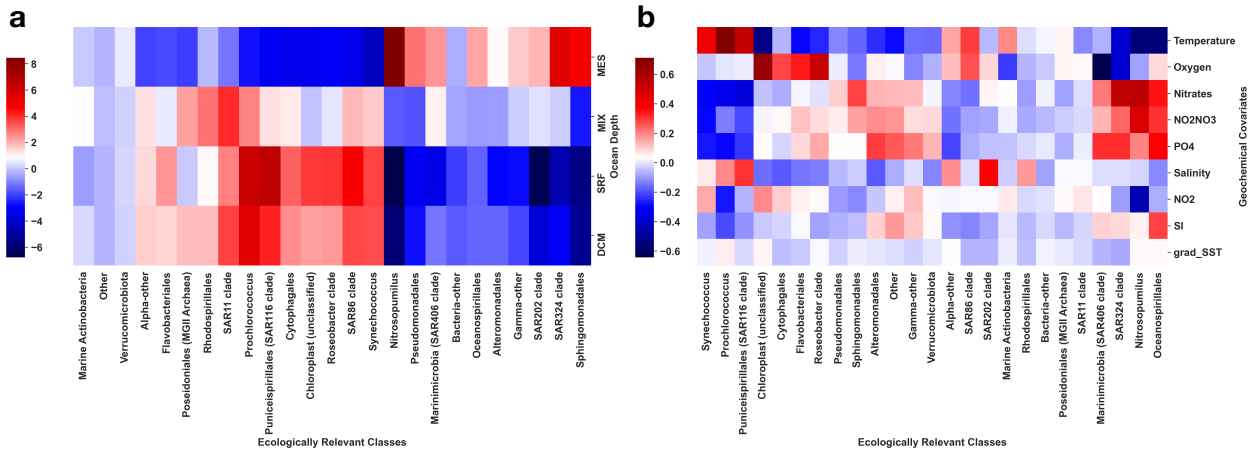

Fig. 4. Summary of the estimated average effect sizes of the influence of **a.** ocean depth (VI-MIDAS model component $\delta\boldsymbol{\beta}$) and **b.** environmental covariates (VI-MIDAS model component γ) on all ecologically relevant classes (ERCs).

ablated (LLPD=-3.3882). This reflects the well-known depth stratification of marine species between the sunlit ocean and aphotic deep ocean ecosystems. Figure S3 in the Supplementary material illustrates the learned depth stratification across all taxa, as reflected in the component $\delta\boldsymbol{\beta}$. The environmental component was identified as the second most important component with an LLPD reduction of -3.3554.

Figure 4 summarizes the estimated effects $\boldsymbol{\delta\beta}$ of the ocean depth features and the environmental effects $\boldsymbol{\gamma}$ on the abundance of species aggregated into ERCs, respectively. The ocean depth summary (Fig. 4a) reveals three distinct sets of occurrence patterns for two different groups of ERCs. One group (right most in Fig. 4a) comprises ERCs such as Nitrosopumilius, Pseudomonadales, SAR 324 clade, and Sphingomonadales which thrive in the Mesopelagic (MES) zone. A second group includes species like Prochlorococcus, SAR 116 clade, and Synechococcus, which flourish within the ecosystem of the ocean's Deep Chlorophyll Maximum (DCM) and Surface Mixed Layer (SRF) zones. The third group comprises marine Actinobacteria, Verrucomicrobiota, and others that show no dependence on depth. A summary of geochemical features highlights temperature (the top row in Fig. 4b) as the primary positive factor influencing the abundance of Synechococcus, Prochlorococcus, and

Puniceispirillales (SAR116 clade). Oxygen concentration emerges as the main positive driver of abundance for Cytophagales, Flavobacteriales, and Roseobacter clades, while Nitrates, Nitrites, and Phosphate are identified as key drivers for the SAR324 clades, Nitrosopumilus, and Oceanospirillales (four right most columns in Fig. 4b). The estimated patterns broadly recapitulate known biology about ocean microbial ecosystems.

VI-MIDAS reveals five latent microbial sub-communities

The generative model (1) of VI-MIDAS includes the taxon-specific latent variables $\boldsymbol{\beta} \in \mathbb{R}^{k \times q}$ to integrate spatiotemporal features and taxon-taxon associations. For the Tara data, VI-MIDAS' hyperparameter tuning scheme identified $k = 200$ as best latent dimension. After model estimation, the resulting $k-$dimensional latent vectors can be thought of as representing the hidden *marginal* characteristics of each of the q taxa after discounting spatiotemporal and species-species association effects, and adjusted for environmental covariates. The latent space representation thus provides an excellent opportunity to partition the different taxa into coherent sub-groups (or modules) that likely reflect functionality or niche occupation in the global ocean, independent of environmental, taxonomic or phylogenetic relatedness.

To quantify similarity between microbial taxa in the latent space, we first computed cosine distances of all pairs of the q latent vectors. This particular choice of distance allows us to bypass the non-identifiability issue of the parameter $\boldsymbol{\beta}$. We used the resulting distance matrix to construct a k-nearest neighbors graph ($k_{\text{nn}} = 10$). Figure 5 shows the latent space embedding using a force-directed layout of the k-nn graph. The latent space representation reveals several distinct microbial sub-communities, dominated by a few ERCs, including one sub-community dominated by Prochlorococcus and SAR11 clades and one dominated by Nitrosopumilus. We next performed Clauset-Newman-Moore greedy modularity analysis of the nearest neighbor graph [15] and identified five distinct modules in the latent space (see M1-M5 in Fig. 5 with top five ERCs highlighted and color-coded). Module 1 (M1) comprises Flaviobacteriales, SAR86 clades, and the Chloroplast class. SAR11 clade, SAR86 clade, and Flavobacteriales are heterotrophs with functional similarity in oxidizing carbon in the ocean [3]. Both SAR86 clade and SAR11 clade follow a similar seasonal pattern (in the Bermuda Atlantic Time Series oceanographic stations) and coexist in oligotrophic regions with less nutrient supply [73]. Module 2 (M2) includes Nitrosopumilus, Marinimicrobia, and SAR324 clades. Existing literature supports that SAR11 clade (a subgroup of a species), Marinimicrobia, and MGII Archaea are more abundant in deep sea water [76]. Module 3 (M3) comprises Prochlorococcus, SAR11, Marine Actinobacteria, and SAR86 clades, among others, all comprising dominant taxa of the sunlit ocean. The two smallest modules 4 and 5 (M4 and M5, respectively) are dominated by Alteromonadales and are separating M2 from M1 and M3. Interestingly, Module 4 also comprises Synechococcus species. This module thus hints at the known metabolic dependency of certain Alteromonadales taxa on Synechococcus (a photoautotroph) [80]. Although the latent representation does separate the majority of ERCs into distinct subgroups, we nonetheless observe that taxa of certain ERCs are spread out over the latent space, indicating different niche specialization. For instance, the SAR11 clade, one of the most abundant marine microbial taxa, is present in three different modules. Likewise, taxa in the SAR86 clade are present in both modules M1 and M3. For ease of identification, Table S3 summarizes each module in terms of the composition of the ERCs and their abundance.

Global associations between biogeography and latent microbial sub-communities

VI-MIDAS' integrative model also enables a quantitative description of the identified microbial sub-communities in terms of the direct and indirect coupling covariates. Figure 6 illustrates how the compositions of ERCs in each of the five modules are related to the most important environmental and spatial covariates.

Using the mean of the posterior sample from the VI-MIDAS model, we used the estimated $\boldsymbol{\gamma}$ as the effect sizes of the environmental features $\mathbf{X}$, $\boldsymbol{\delta\beta}$ as effect sizes of depth, and $\boldsymbol{\alpha\beta}$ as the effect sizes of the r provinces, respectively Figure 6 reports the average effect sizes of association to the four modules.

The module M1 represents taxa coexisting in the SRF and DCM zone of the ocean. The abundance of taxa in the module is associated with a higher concentration of oxygen, PO_4, and NO_2NO_3 and lower temperature and salinity. In addition to representing the taxa SAR11 clade, SAR86 clade, Chloroplast, and Flavobacteriales, the module also includes Synechococcus, Oceanospirillales, and Poseidoniales. Synechococcus is a unicellular prokaryotic autotrophic picoplankton that participates in the marine ecosystem as a primary producer via photosynthesis. Similarly, Chloroplast sequences are a signature of eukaryotic phytoplankton, though their host eukaryote is not identified in the TARA Oceans dataset. The presence of both taxa in M1 thus is consistent with environments that have higher oxygen concentrations due to photosynthesis and gas exchange with the atmosphere.

Module M2 mainly represents the species coexisting in the MES zone (200 m to 1000 m) of the ocean (see Figure 2 (e)). M2 almost exclusively represents the ERCs Nitrosopumilus and SAR324 clade. The abundance of the species in the group is associated with a lower concentration of oxygen and temperature, and higher concentrations of nitrates, PO_4, and NO_2NO_3. In the oxygen-depleted environment, Nitrosopumilus survives by oxidizing ammonia to nitrite, confirming the observed association pattern [6]. Marinomicrobia (SAR406 clade) in groups M1 and M2 allow us to distinguish subgroups of species that can survive in both deep and shallow water [76].

Module M3 comprises the highest mean abundance of all taxa is highest, primarily representing the taxa SAR11 clade, SAR86 clade, and Prochlorococcus (cyanobacteria). The abundance of the species in the group is positively associated with depth indicators {DCM, MIX, SRF} and negatively associated with MES. Among the geochemical factors, temperature, salinity, and oxygen concentration are positively associated, whereas the concentration of nitrates, PO_4, and NO_2NO_3 is negatively associated with the taxa.

Module M4 primarily represents Alteromonadales (Proteobacteria) and some Pseudomonadales (Proteobacteria) and Synechococcus. Their abundance is associated with factors such as lower salinity and higher oxygen concentration. Module M5 also primarily represents Alteromonadales. Based on its association with the ocean depth indicators and geochemical features, we conclude that these taxa can survive in a deep-sea environment characterized by lower temperatures and oxygen concentrations. Associative patterns of Alteromonadales in M4 and M5 differ significantly, suggesting distinct ERC sub-groups that populate different niches.

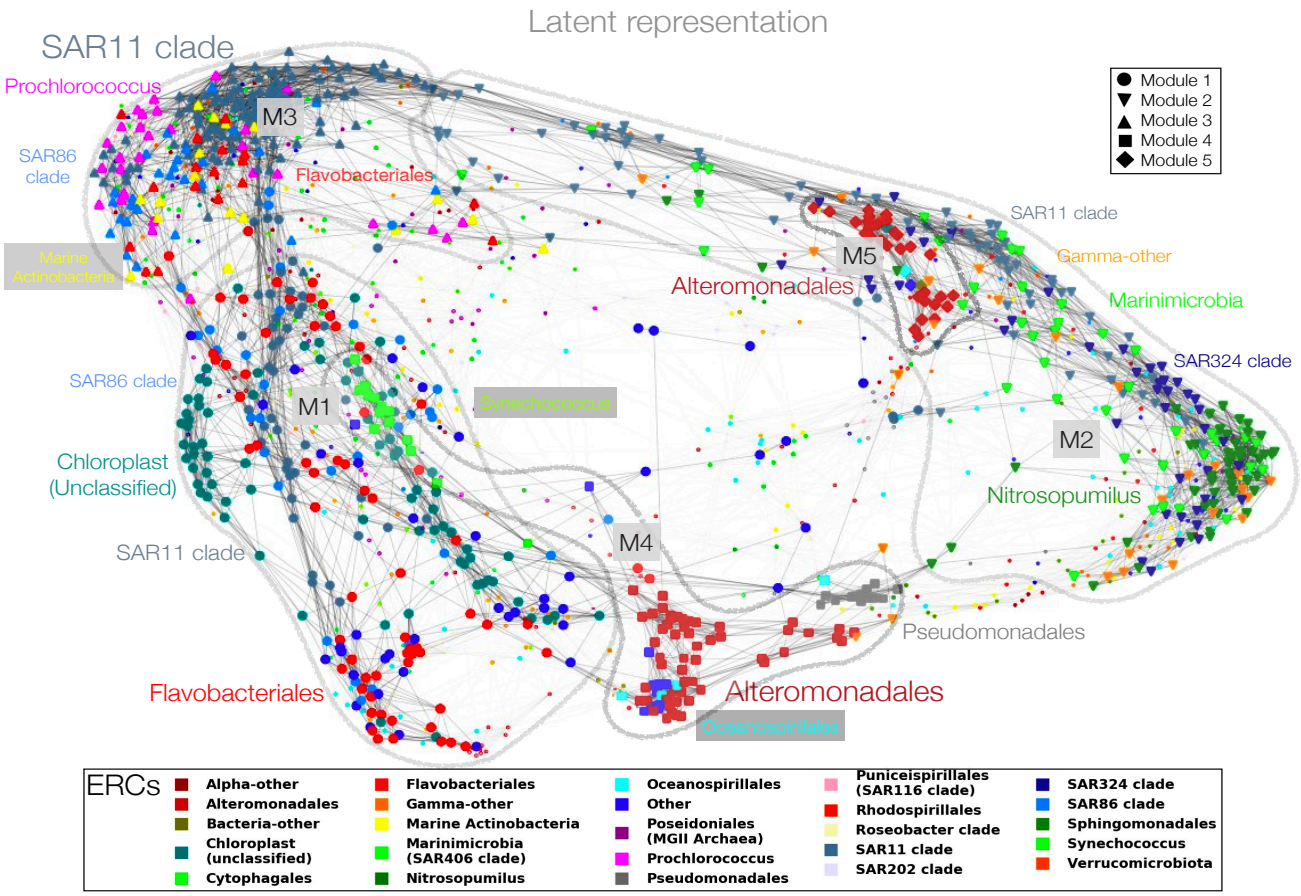

Fig. 5. Low-dimensional embedding of the latent representation $\boldsymbol{\beta}$ using a k-nearest-neighbor ($k_{\text{nn}} = 10$) graph of cosine distances. Modularity analysis reveals five distinct graph modules. We highlight 825 out of a total of 1378 taxa, comprising the top five ERCs (color-coded) in each of the five modules (see main text for further information).

Positive and Negative interactions among ERCs

VI-MIDAS includes a mechanism for learning microbial interactions adjusted for direct (here, environmental) covariates. Contrary to prominent (partial) correlation-based methods [21, 38], VI-MIDAS follows the SHOPPER utility model [61] and quantifies pairwise interactions $\mathbf{I}_{ij}$ between any two taxa i and j in terms of the latent variables $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ (see Eq. 9).

To get a high-level view of the estimated interactions, we aggregated the adjacency matrices of significant positive and negative interactions among taxa by ERCs (for a more detailed view of the most significant taxon-level interactions, we refer to Section 4 of the Supplementary Materials). Figure 7 illustrates the aggregated positive (lower triangle) and negative (upper triangle) interactions among ERCs. The diagonal entry highlights the maximum of the two types of interactions to avoid confusion (see also Section 4 of the Supplementary Materials for the matrix of ratios between positive and negative interactions). We observe that SAR11 clade and Rhodospirillales form positive interactions with almost all other ERCs. SAR11 clade and Rhodospirillales belong to the Alphaproteobacteria phylum that play a critical role in carbon and nitrogen fixation [40, 51], potentially explaining the large number of interactions. However, members of the SAR11 clade also form many negative interactions with other ERCs. Alteromonadales exhibits primarily negative interactions with other ERCs (the strongest one with SAR11).

Discussion

In recent years, multimodal and multi-omics microbiome survey data have emerged for a wide range of microbial habitats [68, 67, 47, 27, 65, 1]. These data collections hold the promise to describe and understand the functional interplay between the underlying microbial ecology and the host or the environment the microbiota resides in. Learning interactions among species and habitat characteristics from observational data remains, however, a challenging problem. To this end, we have proposed VI-MIDAS (Variational Inference for MIcrobiome survey DAta analysiS), a flexible and efficient probabilistic framework for microbiome survey data analysis.

VI-MIDAS uses the negative binomial distributional framework in combination with a principled centering transformation to model overdispersed amplicon abundance data and comprises three mechanisms to integrate concomitant covariate data into the generative model: (i) a direct coupling mechanism, (ii) an indirect latent coupling mechanism, and (iii) a latent interaction term. These terms are linearly linked to the probability distribution's mean parameter. Because of the intractable form of the marginal distribution of data, we

apply mean-field variational inference framework to learn an approximate posterior distribution of the parameters.

VI-MIDAS is available in Python and uses the probabilistic programming language Stan [12]. The implementation is available on GitHub (`https://github.com/amishra-stats/vi-midas`). The repository also includes Python scripts and Jupyter notebooks for VI-MIDAS' three-stage parameter estimation framework: hyperparameter tuning, component contribution analysis, and sensitivity analysis.

To illustrate the VI-MIDAS modeling and analysis workflow, we have used data from the global Tara expedition [67], connecting the available spatiotemporal and environmental characteristics with generative modeling of the amplicon count data. To ease interpretability, we also grouped the amplicon-derived taxa into expert-annotated ecologically relevant classes (ERCs) which may be of independent interest for the analysis of other marine sequencing data. Focusing on the $q = 1378$ most abundant taxa representing 23 ERCS, we integrated the geochemical data using the direct coupling mechanism, effectively removing influence of common environmental factors such as temperature, salinity, and elemental compositions on microbial abundances. The remaining spatiotemporal features,

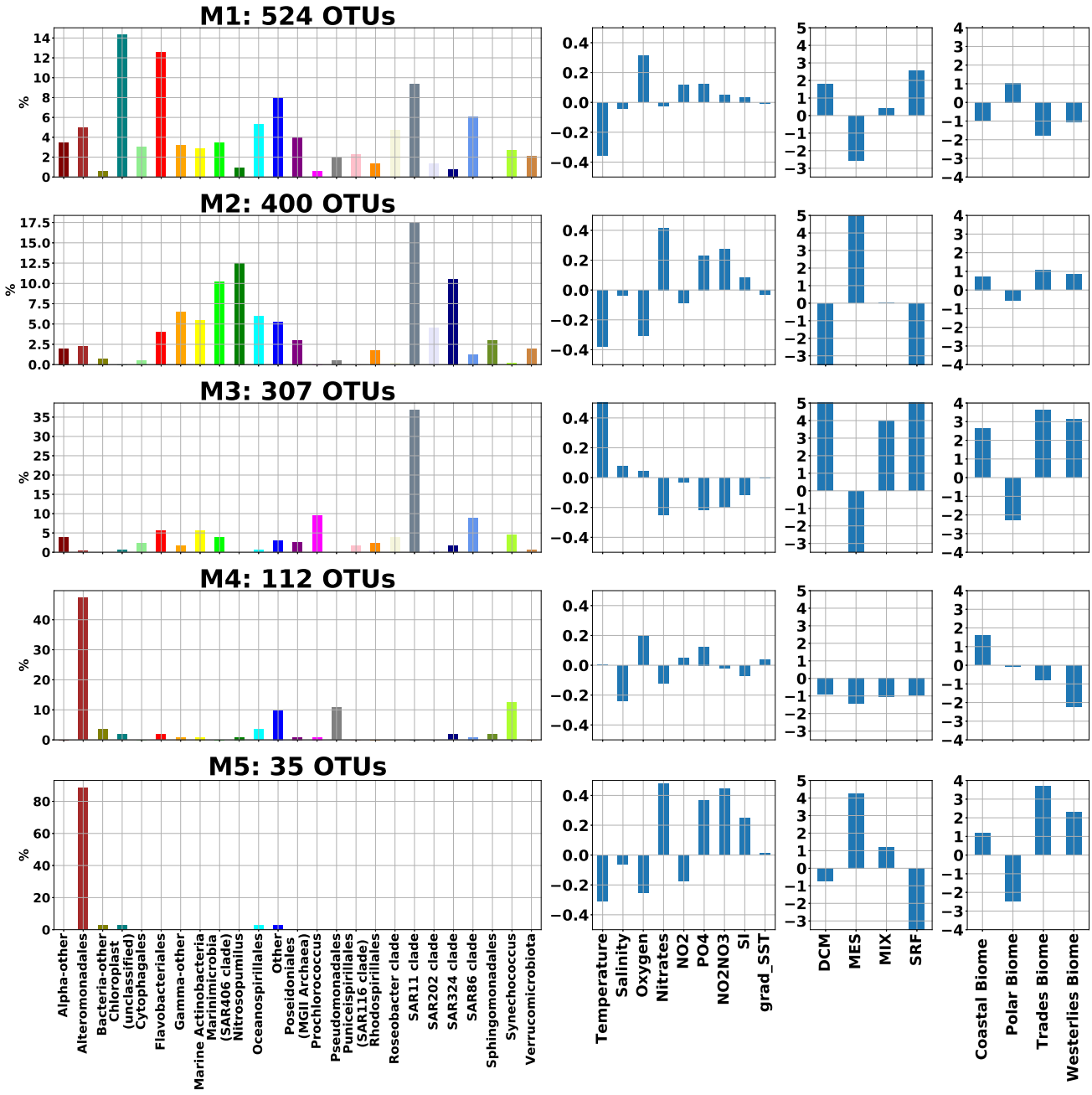

Fig. 6. Global associations between biogeography and covariates: Each row presents the average effect size of the association between the microbial abundances of taxa in a module (M1-M5) to the geochemical features, ocean depth and province/location (from left to right). A module (leftmost) is shown as the composition (in %) of the ERCs. Each module comprises different number of taxa {524, 400, 307, 112, 35}, respectively. Modules M1-M3 cover the majority of taxa, and M4-M5 two smaller Alteromonadales-dominated sub-communities.

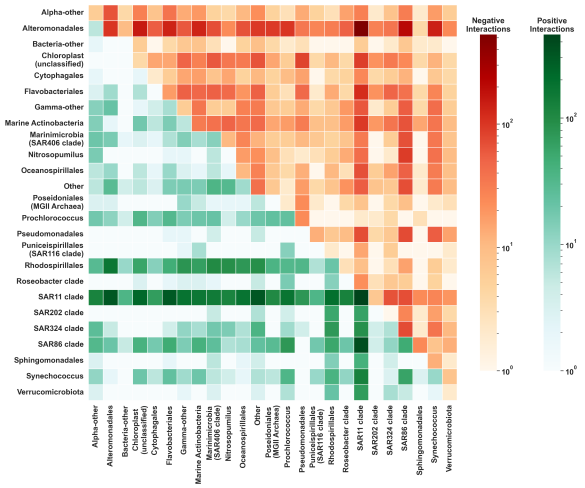

Fig. 7. Summary of taxonomic interactions: The adjacency matrices of significant positive and negative interactions among taxa are grouped and aggregated by their ERCs type. Interactions summary by the ERCs types. Lower triangle reports positive interactions, the upper triangle reports negative interactions. Diagonal entries show the maximum of either (positive or negative) self-interaction.

including season, ocean province, and depth, as well as species-species associations are integrated through the latent coupling and interaction mechanism, thus delivering a latent species representation, adjusted for the influence of all available covariates. The learned VI-MIDAS' model thus not only provides a convincing generative count model for the Tara data but also allows integrated statistical analysis of covariate feature effects and taxa abundances.

Modularity analysis of the similarity network of VI-MIDAS' latent species representation revealed that the majority of taxa (¿1200) can be categorized into three global microbial communities (M1-M3 in Figure 5), including a low-temperature/high-oxygen community (M1), dominated by Flavobacteriales and the Chloroplast ERC, a mesopelagic community (M2) dominated by SAR11, SAR324, and Nitrosopumilus, and a high-temperature community (M3) dominated by SAR11 and Prochlorococcus, the later of which is the most abundant clade in the oligotrophic subtropical and tropical oceans (see e.g., [66] and references therein). Furthermore, our analysis suggests two distinct Alteromonadales-dominated communities that show different depth and province dependencies (M4-M5) (see Figure 6 for further global associations overview). It is noteworthy that Alteromonadales also play a pivotal role in the latent interaction analysis, showing widespread negative associations with other ERCs. We posit that the potentially distinct role of Alteromonadales in the global ocean might be of interest for follow-up analysis on other data sets, including recent data on the global mesopelagic zone [59].

While our ablation study showed evidence that all VI-MIDAS components for the Tara data contribute to the quality of the generative model, the model is just one of several available alternatives. For covariate inclusion, we deliberately chose to directly adjust the microbial abundances for geochemical covariates to better carve out "hidden" relationships among the species. Nonetheless, the VI-MIDAS framework naturally enables other model constructions. For instance, one could have removed the direct coupling component and link all concomitant features to the latent space representation, or alternatively, remove the latent representation altogether and directly adjust for all covariates. We will explore such modifications in future studies. Moreover, while we chose the Negative Binomial model as base distribution for the most abundant taxa, the variational formulation lends itself to other statistical models for microbial count data, including zero-inflated or hurdle- type extensions of the Negative Binomial model [19] or the Dirichlet-Multinomial model [30, 53]. Finally, in its current state, VI-MIDAS is built on Stan [12] with tailored Python code for optimization, model selection, and analysis. The advent of extensive statistical packages in modern deep learning tools, such as Tensorflow distributions [17] or PyTorch [55], may enable efficient porting of VI-MIDAS into these general-purpose ecosystems. Paired with variational inference tools [35], would potentially allow for faster model adaptation and alternative optimization routines.

In summary, VI-MIDAS provides a novel probabilistic framework for learning environment- or host-specific feature associations, latent species characterization, and species-species interactions from microbiome survey data. With minimal adjustment, the framework is readily available for the analysis of other large-scale survey data, including gut microbiome surveys [33, 45, 20], thus representing a potentially valuable general-purpose tool for the integrated analysis of modern microbiome data collections.

Data availability

We have used microbial species abundance data from the Tara Ocean Expedition, available at (`http://ocean-microbiome.embl.de/companion.html`).

Code availability

The source code required to reproduce the results in this article is freely available at (`https://github.com/amishra-stats/vi-midas`).

Competing interests

No competing interest is declared.

References

1. The integrative human microbiome project. *Nature*, 569(7758):641–648, 2019.
2. J. (John) Aitchison. *The statistical analysis of compositional data*. Blackburn Press, Caldwell, N.J., 2003.
3. Montserrat Aldunate, Rodrigo De la Iglesia, Anthony D Bertagnolli, and Osvaldo Ulloa. Oxygen modulates bacterial community composition in the coastal upwelling waters off central chile. *Deep Sea Research Part II: Topical Studies in Oceanography*, 156:68–79, 2018.
4. Mohammad D Ashkezari, Norland R Hagen, Michael Denholtz, Andrew Neang, Tansy C Burns, Rhonda L Morales, Charlotte P Lee, Christopher N Hill, and E Virginia Armbrust. Simons collaborative marine atlas project (simons cmap): An open-source portal to share, visualize, and analyze ocean data. *Limnology and Oceanography: Methods*, 19(7):488–496, 2021.

5. Michael B. Sohn and Hongzhe Li. A glm-based latent variable ordination method for microbiome samples. *Biometrics*, 74(2):448–457, 2018.
6. Viswanathan Baskaran, Prasanna K Patil, M Leo Antony, Satheesha Avunje, Vinay T Nagaraju, Sudeep D Ghate, Suganya Nathamuni, N Dineshkumar, Shankar V Alavandi, and Kizhakedath K Vijayan. Microbial community profiling of ammonia and nitrite oxidizing bacterial enrichments from brackishwater ecosystems for mitigating nitrogen species. *Scientific reports*, 10(1):1–11, 2020.
7. Jacob Bien, Xiaohan Yan, Léo Simpson, and Christian L Müller. Tree-aggregated predictive modeling of microbiome data. *Scientific Reports*, 11(1):1–13, 2021.
8. David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
9. Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37(8):852–857, 2019.
10. Ben J Callahan, Kris Sankaran, Julia A Fukuyama, Paul J McMurdie, and Susan P Holmes. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5, 2016.
11. A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.
12. Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.
13. Jun Chen and Hongzhe Li. VARIABLE SELECTION FOR SPARSE DIRICHLET-MULTINOMIAL REGRESSION WITH AN APPLICATION TO MICROBIOME DATA ANALYSIS 1. *Ann. Appl. Stat.*, 7(1):418–442, 2013.
14. Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.
15. Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
16. Roberto de la Cruz and Jan-Ulrich Kreft. Geometric mean extension for data sets with zeros. *arXiv preprint arXiv:1806.06403*, 2018.
17. Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.
18. Karoline Faust and Jeroen Raes. Conet app: inference of biological association networks using cytoscape. *F1000Research*, 5, 2016.
19. Cindy Xin Feng. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications*, 8(1):8, 2021.
20. Sofia K Forslund, Rima Chakaroun, Maria Zimmermann-Kogadeeva, Lajos Markó, Judith Aron-Wisnewsky, Trine Nielsen, Lucas Moitinho-Silva, Thomas SB Schmidt, Gwen Falony, Sara Vieira-Silva, et al. Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature*, 600(7889):500–505, 2021.
21. Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. 2012.
22. AB Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and D Rubin. Bayesian data analysis third edition. boca raton. *FL: CRC Press.[Google Scholar]*, 2013.
23. Dirk Gevers, Subra Kugathasan, Lee a Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C Morgan, Aleksandar D Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J Xavier. The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe*, 15(3):382–92, mar 2014.
24. Travis Gibson and Georg Gerber. Robust and scalable models of microbiome dynamics. In *International Conference on Machine Learning*, pages 1763–1772. PMLR, 2018.
25. Jack A. Gilbert, Janet K. Jansson, and Rob Knight. The earth microbiome project: successes and aspirations. *BMC Biology*, 12(1):69, 2014.
26. Samantha J Gleich, Jacob A Cram, Jake L Weissman, and David A Caron. Netgam: Using generalized additive models to improve the predictive power of ecological network analyses constructed using time-series data. *ISME Communications*, 2(1):1–9, 2022.
27. Alex Gobbi, Alberto Acedo, Nabeel Imam, Rui G Santini, Rüdiger Ortiz-Álvarez, Lea Ellegaard-Jensen, Ignacio Belda, and Lars H Hansen. A global microbiome survey of vineyard soils highlights the microbial dimension of viticultural terroirs. *Communications Biology*, 5(1):241, 2022.
28. Julia K. Goodrich, Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T. Bell, Timothy D. Spector, Andrew G. Clark, and Ruth E. Ley. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, 2014.
29. Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline, Jennifer Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Tara Oceans Coordinators, Lars Stemmann, Fabrice Not, Pascal Hingamp, Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stéphane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G. Acinas, Peer Bork, Colomban de Vargas, Daniele Iudicone, Matthew B. Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):in review, 2015.
30. Joshua G Harrison, W John Calder, Vivaswat Shastry, and C Alex Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular ecology resources*, 20(2):481–497, 2020.
31. Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, 7(2):e30126, 2012.

32. M. A. Ikram, G. G. O. Brusselle, S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure, C. C. W. Klaver, T. E. C. Nijsten, R. P. Peeters, B. H. Stricker, H. Tiemeier, A. G. Uitterlinden, M. W. Vernooij, and A. Hofman. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol*, 32(9):807–850, 2017.
33. HMP Integrative. The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3):276–289, 2014.
34. Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
35. Diederik Pieter Kingma. Variational inference & deep learning: A new synthesis. 2017.
36. Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.
37. Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226, 2015.
38. Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. 11(5):e1004226, 2015.
39. Seonjoo Lee, Pauline E Chugh, Haipeng Shen, R Eberle, and Dirk P Dittmer. Poisson factor models with applications to non-normalized microrna profiling. *Bioinformatics*, 29(9):1105–1111, 2013.
40. Yufang Li, Kai Tang, Lianbao Zhang, Zihao Zhao, Xiabing Xie, Chen-Tung Arthur Chen, Deli Wang, Nianzhi Jiao, and Yao Zhang. Coupled carbon, sulfur, and nitrogen cycles mediated by microorganisms in the water column of a shallow-water hydrothermal ecosystem. *Frontiers in microbiology*, 9:2718, 2018.
41. Gipsi Lima-Mendez, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, Fabrizio Carcillo, Samuel Chaffron, J Cesar Ignacio-Espinosa, Simon Roux, Flora Vincent, et al. Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1262073, 2015.
42. Tiantian Liu, Peirong Xu, Yueyao Du, Hui Lu, Hongyu Zhao, and Tao Wang. Mzinbva: variational approximation for multilevel zero-inflated negative-binomial models for association analysis in microbiome surveys. *Briefings in Bioinformatics*, 23(1):bbab443, 2022.
43. Alan R Longhurst. *Ecological geography of the sea*. Elsevier, 2010.
44. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.
45. Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, Lindsay DeRight Goldasich, Pieter C Dorrestein, Robert R Dunn, Ashkaan K Fahimipour, James Gaffney, Jack A Gilbert, Grant Gogul, Jessica L Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A Jackson, Stefan Janssen, Dilip V Jeste, Lingjing Jiang, Scott T Kelley, Dan Knights, Tomasz Kosciolek, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V Melnik, Jessica L Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S Pollard, Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D Swafford, Varykina G Thackray, Luke R Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanac, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, and Rob Knight. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3(3), 2018.
46. Paul J. McMurdie and Susan Holmes. phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, 8(4):1–11, 2013.
47. Jacquelyn S Meisel, Geoffrey D Hannigan, Amanda S Tyldsley, Adam J SanMiguel, Brendan P Hodkinson, Qi Zheng, and Elizabeth A Grice. Skin microbiome surveys are strongly influenced by experimental design. *Journal of Investigative Dermatology*, 136(5):947–956, 2016.
48. Aditya Mishra and Christian L. Müller. Robust regression with compositional covariates. *Computational Statistics and Data Analysis*, 165:107315, 2022.
49. Aditya K Mishra and Christian L Müller. Negative binomial factor regression with application to microbiome data analysis. *Statistics in Medicine*, 41(15):2786–2803, 2022.
50. Mary Ann Moran. The global ocean microbiome. *Science*, 350(6266):aac8455, 2015.
51. Molly A Moynihan, Nathalie F Goodkin, Kyle M Morgan, Phyllis YY Kho, Adriana Lopes dos Santos, Federico M Lauro, David M Baker, and Patrick Martin. Coral-associated nitrogen fixation rates and diazotrophic diversity on a nutrient-replete equatorial reef. *The ISME journal*, 16(1):233–246, 2022.
52. GJ Olsen, DJ Lane, SJ Giovannoni, NR Pace, and DA Stahl. Microbial ecology and evolution: a ribosomal rna approach. *Annual review of microbiology*, 40:337–365, 1986.
53. Johannes Ostner, Salomé Carcy, and Christian L Müller. tasccoda: Bayesian tree-aggregated analysis of compositional amplicon and single-cell data. *Frontiers in genetics*, 12:766405, 2021.
54. N.R Pace, D.A. Stahl, D.J. Lane, and G.J. Olsen. The analysis of natural microbial populations by ribosomal rna sequences. In Marshall K.C., editor, *Advances in Microbial Ecology*, volume 9, pages 1–55. Springer, Boston, MA, 1986.
55. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
56. Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Troublé, et al. Open science resources for the discovery and analysis of tara oceans data. *Scientific data*, 2(1):1–16, 2015.
57. Stefanie Peschel, Christian L Müller, Erika von Mutius, Anne-Laure Boulesteix, and Martin Depner. NetCoMi: network construction and comparison for microbiome data in R. *Briefings in Bioinformatics*, 2020.

58. Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.
59. Janaina Rigonato, Marko Budinich, Alejandro A Murillo, Manoela C Brandão, Juan J Pierella Karlusich, Yawouvi Dodji Soviadan, Ann C Gregory, Hisashi Endo, Florian Kokoszka, Dean Vik, et al. Ocean-wide comparisons of mesopelagic planktonic community structures. *ISME communications*, 3(1):83, 2023.
60. Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.
61. Francisco JR Ruiz, Susan Athey, and David M Blei. Shopper: A probabilistic model of consumer choice with substitutes and complements. *arXiv preprint arXiv:1711.03560*, 2017.
62. Kris Sankaran and Susan P Holmes. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, 2019.
63. Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.
64. Salome Scholtens, Nynke Smidt, Morris A Swertz, Stephan JL Bakker, Aafje Dotinga, Judith M Vonk, Freerk van Dijk, Sander KR van Zon, Cisca Wijmenga, Bruce HR Wolffenbuttel, and Ronald P Stolk. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology*, 44(4):1172–1180, 2015.
65. Justin P Shaffer, Louis-Félix Nothias, Luke R Thompson, Jon G Sanders, Rodolfo A Salido, Sneha P Couvillion, Asker D Brejnrod, Franck Lejzerowicz, Niina Haiminen, Shi Huang, et al. Standardized multi-omics of earth's microbiomes reveals microbial and metabolite diversity. *Nature microbiology*, 7(12):2128–2150, 2022.
66. Alaina N Smith, Gwenn MM Hennon, Erik R Zinser, Benjamin C Calfee, Jeremy W Chandler, and Andrew D Barton. Comparing prochlorococcus temperature niches in the lab and across ocean basins. *Limnology and Oceanography*, 66(7):2632–2647, 2021.
67. Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015.
68. Shinichi Sunagawa, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10(12):1196–1199, 2013.
69. Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.
70. Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.
71. W Duncan Wadsworth, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci. An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1):1–12, 2017.
72. Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
73. Nyree J West, Cécile Lepère, Carmem-Lara de O Manes, Philippe Catala, David J Scanlan, and Philippe Lebaron. Distinct spatial patterns of sar11, sar86, and actinobacteria diversity along a transect in the ultra-oligotrophic south pacific ocean. *Frontiers in microbiology*, 7:234, 2016.
74. Fox G Woese C. Phylogenetic structure of the prokaryotic domain. *Pnas*, 74(11):5088–5090, 1977.
75. Tianchen Xu, Ryan T Demmer, and Gen Li. Zero-inflated poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101, 2021.
76. Pelin Yilmaz, Pablo Yarza, Josephine Z Rapp, and Frank O Glöckner. Expanding the world of marine bacterial and archaeal clades. *Frontiers in microbiology*, 6:1524, 2016.
77. Grace Yoon, Irina Gaynanova, and Christian L Müller. Microbial networks in spring-semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in genetics*, 10:516, 2019.
78. Yanyan Zeng, Hongyu Zhao, and Tao Wang. Model-based microbiome data ordination: A variational approximation approach. *Journal of Computational and Graphical Statistics*, pages 1–13, 2021.
79. Xinyan Zhang, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi. Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics*, 18(1):4, 2017.
80. Qiang Zheng, Yu Wang, Rui Xie, Andrew S Lang, Yanting Liu, Jiayao Lu, Xiaodong Zhang, Jun Sun, Curtis A Suttle, and Nianzhi Jiao. Dynamics of heterotrophic bacterial assemblages within synechococcus cultures. *Applied and environmental microbiology*, 84(3):e01517–17, 2018.