Challenges in detecting ecological interactions using sedimentary ancient DNA data

Fiona Callahan[1], Kaiyuan Li[4], and Rasmus Nielsen[1, 2, 3]

[1]Center for Computational Biology, University of California, Berkeley, California, USA
[2]Department of Integrative Biology, University of California, Berkeley, California, USA
[3]Department of Statistics, University of California, Berkeley, California, USA
[4]Biostatistics Division, School of Public Health, University of California, Berkeley, California, USA

Abstract

With increasing availability of ancient and modern environmental DNA technology, whole-community species occurrence and abundance data over time and space is becoming more available. Sedimentary ancient DNA data can be used to infer associations between species, which can generate hypotheses about biotic interactions, a key part of ecosystem function and biodiversity science. Here, we have developed a realistic simulation to evaluate five common methods from different fields for this type of inference. We find that across all methods tested, false discovery rates of inter-species associations are high under realistic simulation conditions. Additionally, we find that with sample sizes that are currently realistic for this type of data, models are typically unable to detect interactions better than random assignment of associations. We also find that at larger sample sizes, information about species abundance improves performance of these models. Different methods perform differentially well depending on the number of taxa in the dataset. Some methods (SPIEC-EASI, SparCC) assume that there are large numbers of taxa in the dataset, and we find that SPIEC-EASI is highly sensitive to this assumption while SparCC is not. We find that for small numbers of species, no method

consistently outperforms logistic and linear regression, indicating a need for further testing and methods development.

Introduction

Having a better understanding of how and why ecological communities change over time facilitates informed decisions regarding management and environmental protection [1–3]. One of the key components of this is to understand the influence of species on each other as communities assemble and during periods of environmental change [4, 5]. Inclusion of species interactions in ecological modeling changes predictions about the ecological effects of climate change [4, 6], extinction events [5], and the dynamics of whole ecosystems [2]. Species distribution data up to this point has been largely limited to spatial data, or data over short time periods relative to many of the ecological processes at play [1]. However, as sedimentary ancient DNA (sedaDNA) data becomes more available, it is becoming possible to survey populations across large spatial and temporal extents, and to simultaneously capture data across a wide range of taxa [1, 3]. This data presents the opportunity to study how different taxa have co-occurred over large spatiotemporal scales, and to make inferences about association networks among species [1, 2].

Several categories of methods have been used to infer associations between species using data from various proxies for species presence or abundance, including sed(a)DNA [7–9]. One of the most popular correlative models for spatiotemporal inference using detection/non-detection data are referred to as species distribution models (SDM) or joint species distribution models (JDSM) [10]. These are generalized linear mixed models (GLMM) with either a logit or probit link function and can include a random effect term that accounts for spatiotemporal autocorrelation [11–13]. This class of models is highly variable and can include components that account for a variety of factors, but in most cases, they are used for very few species in a single study (except see [14]). An important feature of (some of) these methods is that they account for autocorrelation between samples in space and time [11]. Failing to account for this can result in high rates of false inference [15]. However, there are still some potentially important dynamics of this data type not accounted for by these models, such as uncertainty in covariates and time

points, non-linear effects, and false detections of species.

Though SDMs have been primarily used to detect correlations between species and their abiotic environment, some have also been used to detect both biotic and abiotic interactions [11]. Wang, et al. used a SDM method in a study that investigated ecological interactions over the last fifty thousand years in the Arctic using sedaDNA data [11]. They make inferences about the ecological dynamics between humans and megafauna, plants and megafauna, and co-occurrence of different species over a large spatiotemporal scale [11]. Here, we will refer to the methods they used as SDM-INLA. Another study implemented a JSDM for spatiotemporal ordinal abundance data using a Markov Chain Monte Carlo (MCMC) algorithm [13]. They modeled the co-occurrence patterns of several frog species and of several Eucalypts, concluding that frog species had positive residual correlation not accounted for by measured environmental variables, whereas the Eucalypts had negative residual co-occurrence patterns after accounting for the measured environmental variables [13]. I will refer to this method here as JSDM-MCMC.

Network analysis methods are another set of correlative methods for detecting associations between microbes using sed(a)DNA data [16], though they are not exclusively used for microbial data [17]. They include many different approaches, but in general they use aspects of mathematical network theory in inference and interpretation of associations. These methods are often used for time series data (i.e. a single sediment core subsampled vertically, representing sampling through time), although the methods considered here do not explicitly model time or space [8, 9, 18]. SPIEC-EASI and SparCC [8, 18] use sedaDNA read abundances per species as input. EcoCopula [9], which is not explicitly designed for sedaDNA, can accommodate either count, biomass, or detection/non-detection data, and incorporates covariates and species interactions. These methods have been used to investigate microbial associations in the human gut microbiome [8], link microbial network complexity to ecosystem functionality [19], and investigate associations among a broad range of taxa in a marine ecosystem over a period of vast environmental change [17]. In the microbial context, sedimentary DNA (sedDNA) sequences are not assigned to species or taxa but rather grouped into operational taxonomic units (OTUs) based on a sequence-similarity threshold [20]. On the other hand, in studies involving plants and

megafauna, sequences are often assigned at a species or genus level [11]. Since all data are simulated in this study, we will use species, taxon, and OTU interchangeably, recognizing that there are differences between these data types in real data.

The associations inferred by these methods may arise from 1) direct interactions between species such as trophic, mutualistic, or competitive interactions, 2) indirect interactions such as similar responses to the same environment, or 3) falsely inferred associations. If species are associated through an unmeasured aspect of the environment or through an unmeasured species, associations inferred through correlative models may not be causal [21]. Therefore, we do not expect all inferred associations to be causal, but rather consider the aim of these models to be to generate hypotheses about causal interactions, which can then later be tested experimentally [21]. Additionally, false inferences will always be expected to occur at some low rate due to stochastic effects, but violations of modeling assumptions such as non-equilibrium dynamics or uncertainty in covariates can cause much higher rates of false inference [21–23].

One of the difficulties of inferring associations between species in this setting is that the number of possible pairs of taxa scales with the square of the number of taxa in the dataset [24]. In studies focusing on megafauna or plants the number of taxa is most often in the tens [11, 13], but in microbiome studies, the number of taxa can be in the hundreds [8, 19]. This means that the number of parameters that are being inferred (interactions between pairs of species) can be larger than the number of data points [8]. In these cases, additional assumptions in the methods are necessary to make inference possible [8]. Additionally, some methods consider sedaDNA as binary occurrence (detection/non-detection) data, while others consider relative read abundance data [25]. These methods assume that relative read abundance is a proxy to organismal abundance or relative biomass, but there are many potential confounders that may affect this metric [26].

Understanding the relationships between these models and the circumstances under which they succeed or fail to infer species interactions will be informative as the availability of sedaDNA increases. In particular, although we find that many of these methods do not perform well with currently available sample sizes (see Results), we are hopeful that this will be informative as data

availability increases. Additionally, one of the areas of promise presented by sedaDNA is the fact that we can potentially study the effects of taxa across kingdoms using the same samples [1], but in order to maximize this potential we will need to understand the performance of different methods across a broad range of ecological contexts.

Models to infer species interactions have been tested against each other within the category of (J)SDMs [10, 27] and within the microbial network modeling literature [8, 24], but little is known about the relative performance of methods in the two categories. Often, these methods have been tested using data simulated under simple models that do not directly incorporate ecological processes [8, 10, 27]. Due to the complexity of real ecosystems, this may severely underestimate rates of false inference made using these models because the simulated data likely meets the assumptions of inference methods better than real data. Therefore, it is important to also test methods under more challenging conditions that are based on ecological theory. To this end, we have developed a novel simulation model that uses ecological theory to simulate species abundances and simulates the sedaDNA sampling process. We compare this to a simpler simulation model that simulates sedaDNA read counts for multiple species without simulating ecological processes. We specify inter-species and species-environment interactions to create datasets for which the true interactions are known and apply different methods to the simulated data to test accuracy across a range of models, parameters, and datasets (Figure 1).

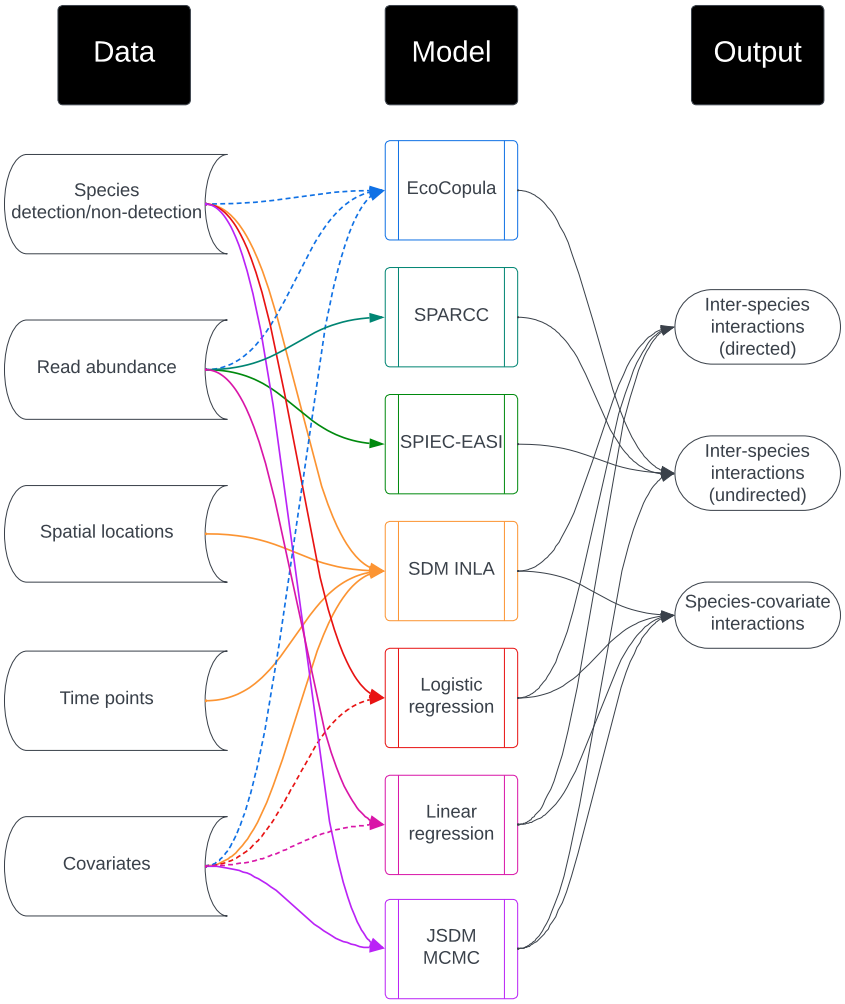

Figure 1: Inputs and outputs of inference methods tested. Dotted lined indicate optional inputs.

Materials and Methods

Simulation Models

Simulated data was generated under several different models. The first set of models, which were fairly simple simulations using a multivariate Gaussian distribution, are referred to here as covariance matrix simulations. However, because we believe that ecological data can be much more complex than this model, we also simulated data under a more complex model that incorporates ecological theory, which we refer to as the ecological simulation model. We allowed parameters in this model to vary widely, but since we recognize that this may create unrealistic scenarios, replicates with a single set of parameters that generated more realistic data were also performed.

Notation

Throughout, j indicates species (used interchangeably with taxon or OTU), s indicates a location in 2D space, and t indicates time. An omitted species index indicates the vector of values for all species. Bold symbols indicate vectors. An overview of other notation can be found in Tables 1 and 2.

Covariance matrix simulation model

In one set of simulations, species read abundances are drawn from a mean-zero multivariate Gaussian truncated at 0. We use a a fixed species interaction matrix, $\boldsymbol{\alpha}$, which is a sparse square matrix with dimension $N_{species}$ (number of species), and where the number in α_{j_1,j_2} is either 0, 1, or -1, indicating the sign of the interaction between species j_1 and species j_2. The covariance matrix for this set of simulations is defined in terms of $\boldsymbol{\alpha}$ in order to allow simulation of the same set of interactions with both the covariance matrix simulation model described in this section and the ecological simulation model described later. The correlation matrix for the Gaussian read abundances, Σ_α, is defined as a $N_{species}$ by $N_{species}$ (same dimension as $\boldsymbol{\alpha}$) matrix with ones on the diagonal, 0.1 in all matrix entries where $\alpha_{j_1,j_2} > 0$ or $\alpha_{j_2,j_1} > 0$ (positive interaction), -0.1 in all matrix entries where $\alpha_{j_1,j_2} < 0$ or $\alpha_{j_2,j_1} < 0$ (negative interaction), and 0 elsewhere. This means that the expected correlation between the reads of species with positive interactions is 0.1

and the expected correlation between species with negative interactions is -0.1. The value of 0.1 is comparable to observed correlations between species that interact in the ecological simulation model, which is an emergent property of that model rather than a parameter (See Results). Read abundances are later multiplied by 100, which changes the variance and covariance but not the correlation.

Let $a(s,t)$ be the vector of simulated read abundances for all species at location s and time t, then

$$\boldsymbol{a}^*(s,t) \sim MVNormal(\mathbf{0}, \Sigma_\alpha) \tag{1}$$

and

$$a_j(s,t) = \texttt{round}[100 \cdot max(0, a_j^*(s,t))] \tag{2}$$

Due to the mean-zero truncated Gaussian, approximately half of the read counts are 0, which may be more or less realistic depending on the dataset. In this case, this is chosen to create optimal conditions for detection/non-detection data, as these methods lose power when percent presence is very low or high (Appendix M). Additionally, counts are multiplied by 100 to put them on the same order of magnitude as several example datasets, although we also recognize that this may vary between studies and depend on the level of classification of reads (e.g. OTU vs. species vs. family level) [8, 17]. Finally, data is rounded to the nearest whole number to create more realistic count data.

Detection/non-detection data, $y_j(s,t)$, is created by setting some threshold R_{detect} for the number of reads to determine species presence, as in Wang, et al., 2021 [11].

$$y_j(s,t) = \mathbb{I}\{a_j(s,t) > R_{detect}\} \tag{3}$$

Ecological simulation model

We constructed a simulation model which takes as input 1) inter-species and species-environment interactions, 2) environment covariates, and 3) simulation hyperparameters (e.g., life history

s	point in space
t	time in unspecified units
j	species
$nSamp_j(s,t)$	number of individuals that deposited DNA
$a_j(s,t)$	sedaDNA read abundance data (species j, location s, time t)
$y_j(s,t)$	binary species detection data (species j, location s, time t)
$N_j(s,t)$	abundance (species j, location s, time t)
$K_j(s,t)$	carrying capacity of the environment (species j, location s, time t)
$M_j(s,t)$	number of migrants (species j, (to) location s, time t)
$x^*(s,t)$	measured covariates

Table 1: Variables in ecological simulation model.

traits, detection rates). The output of the simulation is sedaDNA read abundance and detection/non-detection data for each species at each point in time and space. Detection/non-detection data is created by setting a threshold and mapping the read abundance data to binary data.

This is a population-level model in which all individuals of the same species share the same dynamics and traits. It includes explicit space over which abiotic environmental covariates can vary in both space and time and an arbitrary number of species whose populations vary over space and time.

At each time step, we model migration and logistic population growth that depends on a time-varying carrying capacity. The carrying capacity at each time point and location is a function of abiotic covariates and the abundances of other species. We also model the detection process, including modeling DNA deposition, covariate measurement uncertainty, and varying numbers of sampled points (Figure 2).

Note that variables (Table 1) vary within a simulation, whereas parameters (Table 2) can be changed between simulations but are constants in a given simulation.

Mechanistic model of abundance We model abundance as logistic growth in discrete time with noise and migration from a neighborhood of locations. Simulations were initialized with species abundances at time zero of $N_j(s,0) = 10$ for $j = 1, 2, ..., N_{species}$; $s = 1, 2, ..., |L|$, where

$N_{species}$	number of species in the model
N_{cov}	number of covariates (environmental/abiotic factors)
β	$N_{species}$ x N_{cov} matrix of covariate effects on the different species
α	$N_{species}$ x $N_{species}$ matrix of species effects on each other. Note: diagonal is 0
c_i	tuning parameters
λ_j	migration rate for species j
r	intrinsic population growth rate
σ	constant controlling amount of noise in population growth
d_{mig}	migration neighborhood radius for all species
$pSample$	organism sampling rate
$lambdaReads$	read sampling rate per organism
$x(s,t)$	vector of environmental explanatory variables (covariates)
σ_{cov}	standard deviation for the covariate measurement noise
R_{detect}	detection threshold for detection/non-detection
L	set of spatial locations

Table 2: Parameters in ecological simulation model

$N_j(s,t)$ is the species abundance of species j in location s at time t, and L is the set of locations. The first 100 time points are discarded before analysis.

The algorithm then proceeds, by at every time point, $t > 0$, simulating the abundances as follows:

Migration: At location s, the number of new immigrants at time t for species j is $M_j(s,t)$, which is assumed to be Poisson distributed at a rate that depends on the number of individuals of species j that were in a neighborhood of location s at time t, and a species-specific migration rate λ_j. The migration rate varies across species, with the actual value for each species drawn from a Gamma distribution parameterized with the mean across species set to $mean_mig_rate$. The radius of the neighborhood, d_{mig}, and mean migration rate are simulation parameters (allowed to vary between simulations but constant for a given simulation). The abundance in each location is then adjusted as follows

$$N_j^*(s,t) = N_j(s,t) + M_j(s,t) \tag{4}$$

where

$$M_j(s,t) \sim \text{Poisson}(\lambda_j \underbrace{\sum_{s^* \in S} N_j(s^*, t-1)}_{\text{individuals in neighborhood}}) \tag{5}$$

and S is the set of spatial locations within radius d_{mig} of location s. Also,

$$\lambda_j \sim \text{Gamma}(\text{shape} = mean_mig_rate/0.01, \text{scale} = 0.01) \tag{6}$$

Population growth: The change in population size between time t and time $t+1$, is modeled using discrete logistic growth with Gaussian noise and with a carrying capacity that depends on other biotic and abiotic factors. The growth rate of all species, r, and the noise of the process, scaled by σ, can be adjusted for each simulation, but are constant within a simulation. We then assume that the local change in population size is

$$\Delta N_j(s,t) = \underbrace{rN_j^*(s,t)\left[1 - \frac{N_j^*(s,t)}{K_j(s,t)}\right]}_{\text{Logistic population growth}} + \underbrace{\sigma N_j^*(s,t) W_j(s,t)}_{\text{Gaussian noise}} \tag{7}$$

where $W_j(s,t) \sim Normal(0,1)$.

The variance of the Gaussian noise term in the equation above scales with the population size as in ref. [28].

We model carrying capacity as a linear function of covariates and a non-linear function of other species abundance. The vector of carrying capacities for all species, $\boldsymbol{K}(s,t)$, in the logistic growth equation depends on simulated abiotic environmental conditions, $\boldsymbol{x}(s,t)$, and the abundance of the other species at time $t-1$. The direction of the effect of each covariate and each species on other species is defined by matrices $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. $\boldsymbol{\beta}$ is a $N_{species}$ by N_{cov} (number of covariates) matrix where each entry in the matrix indicates the sign and direction of influence of a particular covariate on a particular species. $\boldsymbol{\alpha}$ is a $N_{species}$ by $N_{species}$ matrix where each entry in the matrix indicates the sign and direction of influence of a species on another species. In order to avoid inter-species effects going to infinity, the arctangent function is applied to species effects on each

other. This function has a horizontal asymptote, creating an upper bound on the effect of one species on another, representing saturation of the inter-species effect. c_1 and c_2 are constants that control the relative strength of the effects of other species versus the abiotic environment on the growth rate of each species. Carrying capacities are truncated at 0 since they should not be negative.

$$\boldsymbol{K}^*(s,t) = \underbrace{c_1 \cdot \boldsymbol{\beta} \cdot \boldsymbol{x}(s,t)}_{\text{abiotic effects}} + \underbrace{c_2 \cdot \boldsymbol{\alpha} \cdot \arctan(\boldsymbol{N}(s,t-1))}_{\text{inter-species effects}} \tag{8}$$

$$K_j(s,t) = max(K_j^*(s,t), 0) \tag{9}$$

Covariates: Covariates are simulated in two different ways for different ecological simulation sets (See Simulation Sets section of Methods).

Method 1: For the set-parameter simulations, covariates are simulated using Gaussian random walks through time. Covariates are therefore autocorrelated in time but not in space in this case. The vector of covariates for each time and location, $\boldsymbol{x}(s,t)$, is an input to the population simulation model.

$$x_i(s,t+1) = x_i(s,t) + \epsilon_t \tag{10}$$

where $\epsilon_t \sim \text{Normal}(\text{mean} = 0, \text{sd} = 0.01)$.

Method 2: In the random-parameter simulations, covariates were simulated using a spatial Gaussian random field with an exponential covariance function, where the global mean varied in time according to a deterministic sinusoidal function. The period of this function, $temporalPeriod$, and the spatial covariance parameter for the covariance function, $covar_scale_space$, are set uniformly at random within a set range for each simulation (See Table 3).

For each time point, all covariate values across space are drawn from a spatial random field with

an exponential covariance function as follows:

$$\boldsymbol{x_i}(.,t) \sim \text{MVNormal}(\mu_t, \boldsymbol{\Sigma}) \tag{11}$$

with $\mu_t = \sin\left(\frac{2\pi t}{temporalPeriod}\right)$.

The mean value of the spatial random field (global mean of the covariate), vary in time according to a sine function. $temporalPeriod$ is the period of the sine function in time, and it is drawn uniformly between 100 and 10,000 once per simulation:

$$temporalPeriod \sim \text{Unif}(\text{min} = 100, \text{max} = 10000) \tag{12}$$

$\boldsymbol{\Sigma}$ is an $|L|$ by $|L|$ covariance matrix which is calculated as follows: for spatial points s_1 and s_2, where `distance` is the Euclidean distance between the points:

$$\Sigma_{s_1,s_2} = \exp\left(\frac{-\texttt{distance}(s_1, s_2)}{covar_scale_space}\right) \tag{13}$$

With $covar_scale_space$ drawn once per simulation as:

$$covar_scale_space \sim \text{Unif}(\text{min} = 1, \text{max} = 100) \tag{14}$$

The covariance between two points decays exponentially as the distance increases, but it decays slower with increasing $covar_scale_space$.

sedaDNA read abundance and species detection model: We model sedaDNA by assuming that first, $nSamp_j(s,t)$ individuals from the population are selected to deposit DNA. $nSamp_j(s,t)$ is Binomially distributed with parameters $N_j(s,t)$ and $pSample$, the individual sampling probability. The mean of this distribution is therefore proportional to the true species abundance, $N_j(s,t)$. This is equivalent to flipping a weighted coin with weight $pSample$ for each

individual present in the location to decide whether it deposits DNA. Then, the number of sedaDNA reads, $a_j(s,t)$, is Poisson distributed with a rate proportional to $nSamp_j(s,t)$. The parameter $lambdaReads$ dictates the rate at which each individual selected to deposit DNA does so. In other words:

$$nSamp_j(s,t) \sim \text{Binom}(N_j(s,t), pSample) \tag{15}$$

$$a_j(s,t) \sim \text{Poisson}(lambdaReads \cdot nSamp_j(s,t)) \tag{16}$$

Detection/non-detection data, $y_j(s,t)$, is created by truncating the read abundances at some threshold R_{detect}.

$$y_j(s,t) = \mathbb{I}\{a_j(s,t) > R_{detect}\} \tag{17}$$

In the ecological model simulations, under many parameter settings, some species went extinct quickly, so species were filtered for at least 10% presence in the dataset. Therefore, the number of species actually analyzed is less than the number of species simulated in many cases. However, if the number of species after filtering was less than half the number originally simulated, the data set was discarded.

Covariate measurement uncertainty

Covariates are measured with Gaussian noise with variance σ_{cov}:

$$x_i^*(s,t) = x_i(s,t) + \epsilon \tag{18}$$

and $\epsilon \sim \text{Normal}(0, \sigma_{cov})$.

Inference models were tested with and without covariates included for both types of simulations, in spite of the fact that they only had an effect in the ecological simulations and not the covariance matrix simulations.

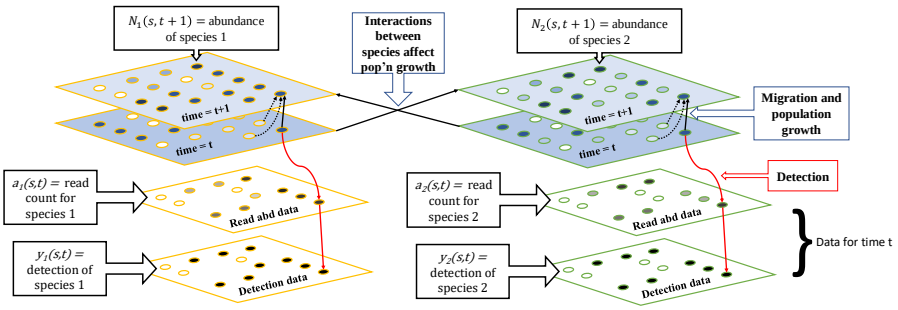

Figure 2: Ecological simulation diagram. At each time step, we model local migration and logistic population growth, which depends on the abundance of other species with specified interactions and a set of abiotic factors. We also model an sedaDNA detection process, resulting in two types of data: read abundance and detection/non-detection (detection) data. Arrows indicate the flow of information through the simulation. Dotted black arrows indicate migration, solid black arrows indicate population growth and species interactions, solid red arrows indicate data simulation.

Simulation Sets

Covariance Matrix Simulations We simulated 100 datasets with 10 species and 100 with 100 species. In each simulation, 100 spatial locations arranged in a 10 by 10 grid were simulated for 10,000 time points, but the full simulated dataset was not analyzed. Although time and space are not included in this simulation algorithm, spatial and temporal labels were included in the final dataset to allow all methods to be tested. Each simulation was subsampled to 100 samples and 10,000 samples for analysis.

Random-parameter ecological simulations We ran 100 simulations under the ecological model described above with all parameters set uniformly at random in what we assessed to be a reasonable range (ranges and distributions specified in Table 3). In each simulation, 100 spatial locations arranged in a 10 by 10 grid were simulated for 10,000 time points, but the full simulated dataset was not analyzed. We simulated 100 replicates each with 10 species and 100 species. We randomly re-sampled each dataset to 100 samples and 10,000 samples for analysis.

Another simulation set using the ecological simulation model described above was run with random parameters. This was created for analysis of the effect of different parameter settings on the success of inference. Ten species were simulated for this set, and 1000 simulations were run.

Parameter	Settings in random simulations
$N_{species}$	10 or 100 based on simulation set
N_{cov}	Uniform(3, 4, 5, ... ,$N_{species}$)
β ($N_{species}$ x N_{cov} matrix)	Uniform randomly selected $N_{species}$ entries set as 1 or -1, all others set as 0
α ($N_{species}$ x $N_{species}$ matrix)	Uniform randomly selected $N_{species}$ entries set as 1 or -1, others set as 0
c_1	200
c_2	Uniform(min = 100, max = 500)
λ_j	Gamma(shape = mean_mig_rate / 0.01, scale = 0.01) (For each species)
$mean_mig_rate$	Uniform(min = 0, max = 0.1)
r	Uniform(min = 0.01, max = 1)
σ	Uniform(min = 0, max = 0.2)
d_{mig}	Uniform(min = 12, max = 50)
$pSample$	Uniform(min = 0.001, max = 0.02)
$lambdaReads$	Uniform(min = 0.5, max = 100)
$x(s,t)$	See covariate section
σ_{cov}	Uniform(min = 0, max = 0.2)
R_{detect}	Uniform(1, 2, 3, ..., 50)

Table 3: Parameters in random simulation model. Parameters without species index were set only once per simulation for all species, times, and locations.

Set-parameter ecological simulations As these random-parameter simulations often lead to scenarios with poor performance of most methods (see Results), we also used a model with fixed parameters for which we expect somewhat better performance as correlations between species abundances are relatively large without leading to common loss of species. The final simulation set was constructed using our custom simulation model described in detail above. Specific parameter values are described in Table 4.

In this dataset, the additional noise in the simulation data was set as low as possible. For example, measurement noise in the covariates was set at 0, the detection rate was relatively high, and noise in population growth was set low. Using these parameter settings, we simulated 100 replicates each with 10 species and 100 species. We randomly subsample each dataset to 100 and 10,000 samples.

The same set of $\boldsymbol{\alpha}$ matrices (species interactions) were used in this set of simulations as for the covariance matrix simulations in order to more easily compare the two sets.

Parameter	Setting in set-parameter simulations
$N_{species}$	10 or 100 based on simulation set
N_{cov}	$N_{species}$
β ($N_{species}$ x N_{cov} matrix)	$diag(1)$
α ($N_{species}$ x $N_{species}$ matrix)	Randomly selected $N_{species}$ entries set as -1 or 1, others set as 0
c_1	200
c_2	300
λ_j	Gamma(shape = mean_mig_rate / 0.01, scale = 0.01)
$mean_mig_rate$	0.01
r	0.05
σ	0.005
d_{mig}	16 (only direct neighbors)
$pSample$	0.01
$lambdaReads$	1
$x(s,t)$	See covariate section
σ_{cov}	0
R_{detect}	5

Table 4: Parameters in set-parameters simulation. Parameters without species index were set only once per simulation for all species, times, and locations.

Testing inference models

All methods were used with default settings in the papers cited, with the exception of threshold adjustment to make receiver operating characteristic (ROC) curves. All methods that accept but do not require covariates (Logistic and linear regression, SDM-INLA, EcoCopula) were tested both with and without covariates included in the analysis (Figure 1).

Logistic Regression Logistic regression was performed using the `glm` function from the `stats` package version 4.4.0 [29]. Separate regressions were run for each species as a function of all other species (and covariates for runs with covariates). Interactions were then considered significant based on a p-value threshold. For reported false discovery rates at a single p-value threshold, the chosen threshold was 0.05 divided by the number of possible interactions being inferred (Bonferroni correction). For ROC curves, the threshold was varied from 0 to 1. Detection/non-detection data (and covariates, where applicable) were used as input to the model.

Linear Regression Linear regression was performed using the `lm` function from the `stats` package version 4.4.0 [29]. Analyses thereafter proceeded as described for logistic regression. Read

abundance data (and covariates, where applicable) were used as input to the model.

JSDM-MCMC JSDM-MCMC was performed as in Pollock, et al., 2014 [13] using JAGS version 4.3.2 [30] and package R2jags version 0.7-1 [31]. It decomposed the species co-occurrence patterns into components describing shared environmental responses and residual patterns of co-occurrence including species interactions. For each species pair, it returned a set of MCMC samples of the association parameter, and significance was determined based on whether a q-percent credible interval contained 0. To calculate false discovery rate at a single threshold, the chosen q-percent was 95%. For ROC curves, the q-percent was varied from 0 to 100%. The model was run with covariates and with detection/non-detection data as the input. Convergence of the MCMC algorithm was evaluated by examining trace plots of multiple chains and using Gelman-Rubin statistics in the R package Coda [32, 33] (Appendix C).

SDM-INLA For SDM-INLA, analysis was performed as in Wang, et al., 2021 [11] using INLA version 24.2.9 [34]. For model selection, 95% posterior credible intervals were used to determine significance and models tested were 1) all species and covariates, 2) all species without covariates, 3) only covariates without species, and 4) no species or covariates. For ROC curves, no model selection was used (models (1) and (2) form two separate curves) and the cutoff for the quantile of the posterior credible interval was varied. Detection/non-detection data, covariates (where applicable), time points, and locations were used as input to the model.

JSDM-MCMC and SDM-INLA methods were not used in the simulations with a larger numbers of species because it took a prohibitively long time for the methods to run (Appendix A).

SPIEC-EASI SPIEC-EASI was run using the R package SpiecEasi version 1.1.2 [35]. All parameters were set to default, and both the `mb` version and the `glasso` version of the model were run for comparison. The input to the model was simulated read abundances. For ROC curves, the regularization parameter, `lambda`, was adjusted to 100 different values automatically by the SpiecEasi package (`nLambda = 100`) and then results were averaged across simulations for the ordered `lambda` values. The `lambda` values were ordered but not necessarily the same values across runs of the simulation, since by default the model automatically selects the actual values.

For the model-selected results, a specific `lambda` value was selected through the default procedure.

SparCC The SpiecEasi package also implements the model SparCC, which was originally published by Friedman and Alm (2012) [18]. This model, as implemented in SpiecEasi, was also tested on simulated read counts. For ROC curves, the threshold for estimated covariance, which is used to call interactions, was adjusted. For the false discovery rate after model selection, a pseudo p-value was generated using the bootstrapping procedure described by Friedman and Alm (2012) and implemented in SpiecEasi [18, 35]. The cutoff for this pseudo p-value was set at 0.05.

EcoCopula EcoCopula was run using the R package EcoCopula version 1.0.2 [9]. The model was run with and without covariates. This method, unlike the others, can take flexible types of input data. Therefore, we tested it using the mode where it takes detection/non-detection data and where it takes read abundance data. By default, the method selects 100 values to test for the regularization parameter, `lambda`. Then it chooses one using BIC. Default selection procedure for `lambda` was used for FDR results. ROC curves were produced by selecting 10 out of the 100 `lambda` values chosen by the default model and averaging resulting true and false positive rates across simulations. The `lambda` values were ordered but not necessarily the same values across runs of the simulation. The `lambda` parameter was also set at 0 to complete the ROC curve, although this was not among the values selected by the model by default.

Counting Mistakes

The performance of these models was evaluated in several ways in order to get a more complete picture of their performance. First, we evaluated the success of the methods at detecting direct, causal interactions. Since this type of data is generally observational, and therefore can (and likely does) have significant confounding variables, we do not claim that any of these methods would detect direct, causal relationships consistently in real data. However, we believe that it is still useful to examine whether they are able to detect causal relationships when all variables are observed. For this metric, sign of interactions was ignored in the calculation of false positive and negative rates. If an interaction exists and one was inferred, this was counted as correct regardless of the interaction being positive or negative. Interactions were considered as directed, so A

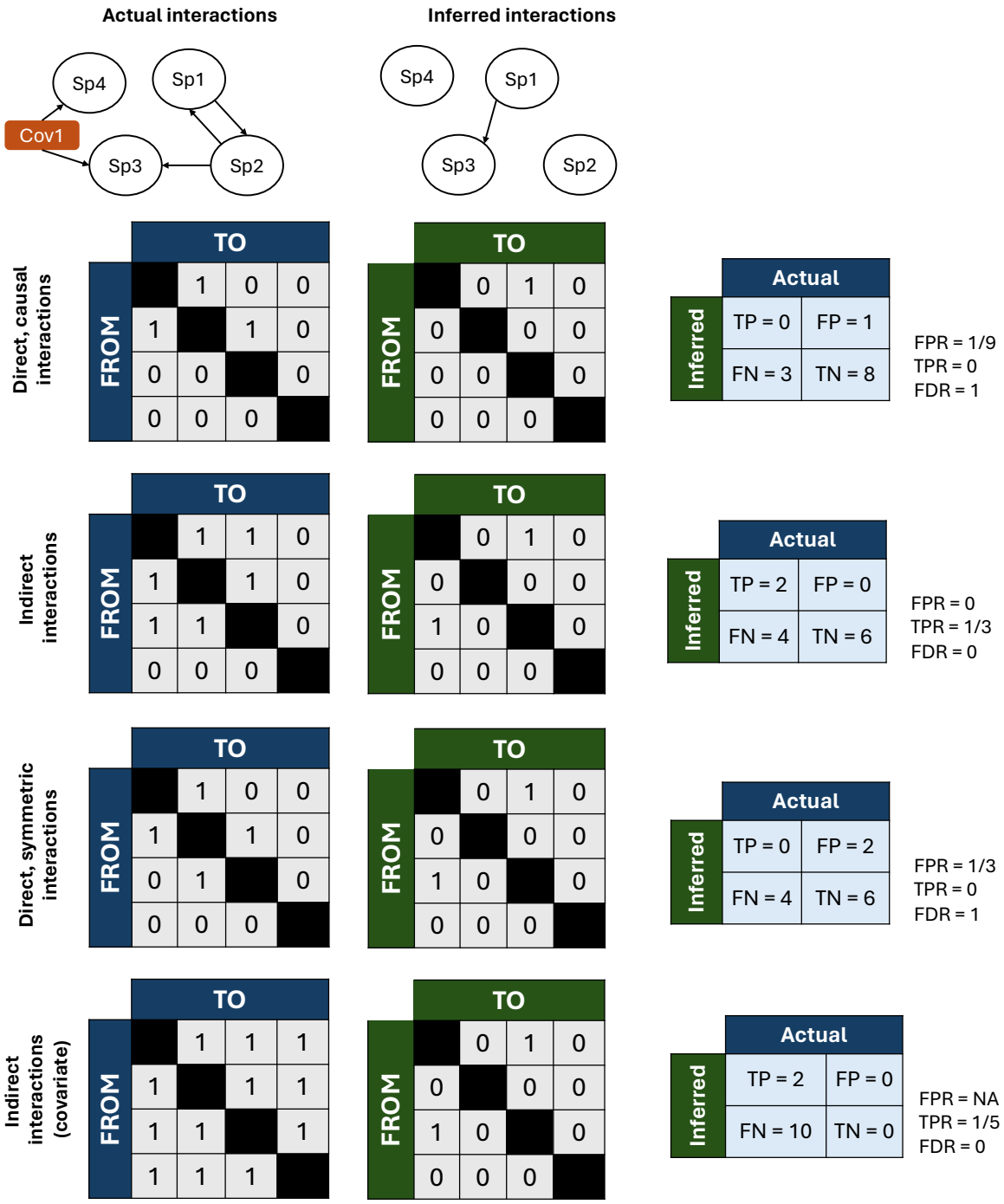

Figure 3: Schematic explaining the calculations of false positive rate (FPR), true positive rate (TPR), and false discovery rate (FDR), when interactions are considered direct (symmetric or asymmetric) versus indirect (with or without considering shared covariate interactions).

influencing B does not imply B influencing A. Therefore, for example, false positive count was calculated as the number of times where an interaction was inferred from A to B, but no interaction exists from A to B, regardless of the sign being positive or negative (Figure 3).

Second, we evaluated the success of the different methods at detecting whether there are any interactions between two species, whether they are direct or indirect through other species. For this metric, we considered an interaction to exist if the two species are connected by interactions, regardless of the direction of the interactions. We likewise consider an interaction to be inferred between two species if any inferred associations connect them (Figure 3).

Third, we evaluate the success of the models at detecting direct associations between species without direction (direct, symmetric interactions). Some methods have the theoretical potential to infer an interaction in one direction but not the other (SDM-INLA, JSDM-MCMC, logistic regression, linear regression), although this should not happen often since they are all correlative methods, while other methods always infer symmetric interactions (SPIEC-EASI, SparCC, EcoCopula). Using this metric, if an interaction exists from A to B, then we automatically assume one exists from B to A. Likewise with the inferred associations, we assume that an inferred association in one direction implies one in the other direction.

Fourth, we consider indirect interactions to include interactions through covariates. For example, if a covariate affects two species, these species are considered to interact. However, since interactions between covariates and species are not inferred, the inferred interactions were defined based on whether two species are connected by interactions between species, without regard for covariates. This metric did not cause different results than the indirect interactions metric without considering covariates (See results), but we include it here for completeness.

False discovery rate was calculated as false-positive-count/total-inferred-associations, or the probability that an inferred interaction was incorrect. This is distinct from false positive rate, which was calculated as false-positive-count/total-actual-interactions. These metrics can differ enormously, especially when there is severe class imbalance, which is the case here with many

more pairs with no actual interaction compared to the number of actual interactions (sparsity of the interaction matrix). Here, we present the false discovery rate after model selection, and the false positive rate at many thresholds as part of ROC curves.

Analysis of effect of individual simulation parameters on predictive performance

A random forest model was used to predict false discovery rate for linear and logistic regression using the simulation parameters that were set at random in the simulations. Random forest models were run in R using Ranger version 0.16.0 [36]. The model was trained on 1000 simulations, with 10 species each and with 100 or 10,000 samples per simulation. Linear and logistic regression were both tested, using no covariates for either model, and with corrected p-values (Bonferroni correction) at a 0.05 alpha level. FDR was evaluated for direct/symmetric interactions only.

The following formula was used for random forest analysis (See Table 2 for definitions and Table 3 for ranges of parameter settings):

$$FDR \sim d_{mig} + \sigma_{cov} + r + \sigma + c_2 + mean_mig_rate + pSample + lambdaReads + N_{cov} + covar_scale_space + R_{detect} \tag{19}$$

Variable importance in the random forest prediction was evaluated using permutation importance (`importance = "permutation"` in Ranger model). The random forest predictive performance was evaluated using root mean square error (RMSE) for an independent test set of 100 simulations.

Code

Simulations and SDM-INLA analyses were run in R version 4.3. All other analysis was run in R version 4.4 [29].

Code is available at https://github.com/Fiona-MC/eDNA-sims-pub.

Results

False discovery rates of direct, causal interactions are high for all inference models and simulations, but lower in some cases when interactions are considered symmetric or indirect

Several models which test for associations between different taxa (used interchangeably with species or OTUs) were tested for their effectiveness at detecting these interactions from data simulated using a mechanistic ecological model as well as a simpler covariance matrix simulation model. False discovery rate for direct, causal interactions (Figure 3) in all cases was over 50% for 100 samples (Figure 4). For 10,000 samples, the lowest false discovery rate was 48% across all simulated datasets (Figure 4).

Some of the mistakes are caused by including the direction of the inferred associations in counting mistakes. For example, overall false discovery rates are much lower for the covariance matrix simulations with 10,000 samples when the direction of the inferred interaction is ignored (interactions are considered symmetric) (Figure 4). In the covariance matrix simulations, interactions are actually symmetric, whereas in the ecological model simulations, actual interactions are not symmetric (See Methods). SDM-INLA, JSDM-MCMC, and the regression methods can in theory predict asymmetric associations, but all other methods predict associations symmetrically. These associations are nonetheless often interpreted as potentially having causal meaning. Therefore, we have included the direct, causal interactions to illustrate one of many reasons that these associations should not be interpreted as directional or causal without additional information.

There are a few outliers in the FDR results for 100 samples that may be caused by very low overall rates of inferred interactions. For example, the 0% FDR for EcoCopula in the symmetric interactions case was caused by only 4 interactions that were all inferred correctly in one dataset (out of 100) (Figure 4). In all other datasets, no interactions were inferred. Similarly, for linear regression with no covariates for the same set of simulations, only 10 interactions were inferred in total, but all ten were incorrect, resulting in an FDR of 100% (Figure 4; Appendix B). When such

low numbers of total interactions are inferred, the estimates of FDR may rely on just a few unusual cases and therefore may have high variance.

False discovery rates using the metric of indirect interactions (Figure 3) vary dramatically between methods. For some methods and simulations, all inferred interactions are correct (FDR = 0), while for others, the false discovery rate is as high as 1 (all inferred interactions incorrect) (Figure 4). EcoCopula and SPIEC-EASI claim to be able to avoid inferring indirect interactions by estimating conditional dependence of species detections (read abundances) given all other species detections (read abundances) [8, 9]. However, here we see that the false discovery rate for SPIEC-EASI is either comparable or goes down when interactions are considered indirect (Figure 4). For EcoCopula, the story is less clear between considering interactions as direct but symmetric versus indirect. There are many cases where indirect interactions are inferred much better. Notably, EcoCopula infers very few interactions in many cases for 100 samples, so these numbers should be interpreted with caution (Appendix B). For 10,000 samples in the covariance matrix simulations, there are several cases where the direct, symmetric interactions are inferred the best (Figure 4).

Whether interactions through the environment are considered correct or incorrect has almost no effect on FDR (Appendix E).

Model calibration (model selection or p-value cutoff) dramatically affects performance

False discovery rates are calculated using the default calibration for each method, but these calibration methods (model selection or choosing a p-value cutoff) vary between methods, which can cause drastic differences in the number of inferred interactions and the false discovery rate. For a comparison that does not depend on the calibration of cutoff values, refer to ROC curves. The number of total inferred interactions is as low as 0 for some (seen as NA in Figure 4), and up to over 8000 inferred interactions per simulation for others (Appendix B).

As expected for lower amounts of data, when only 100 points were sampled many methods often

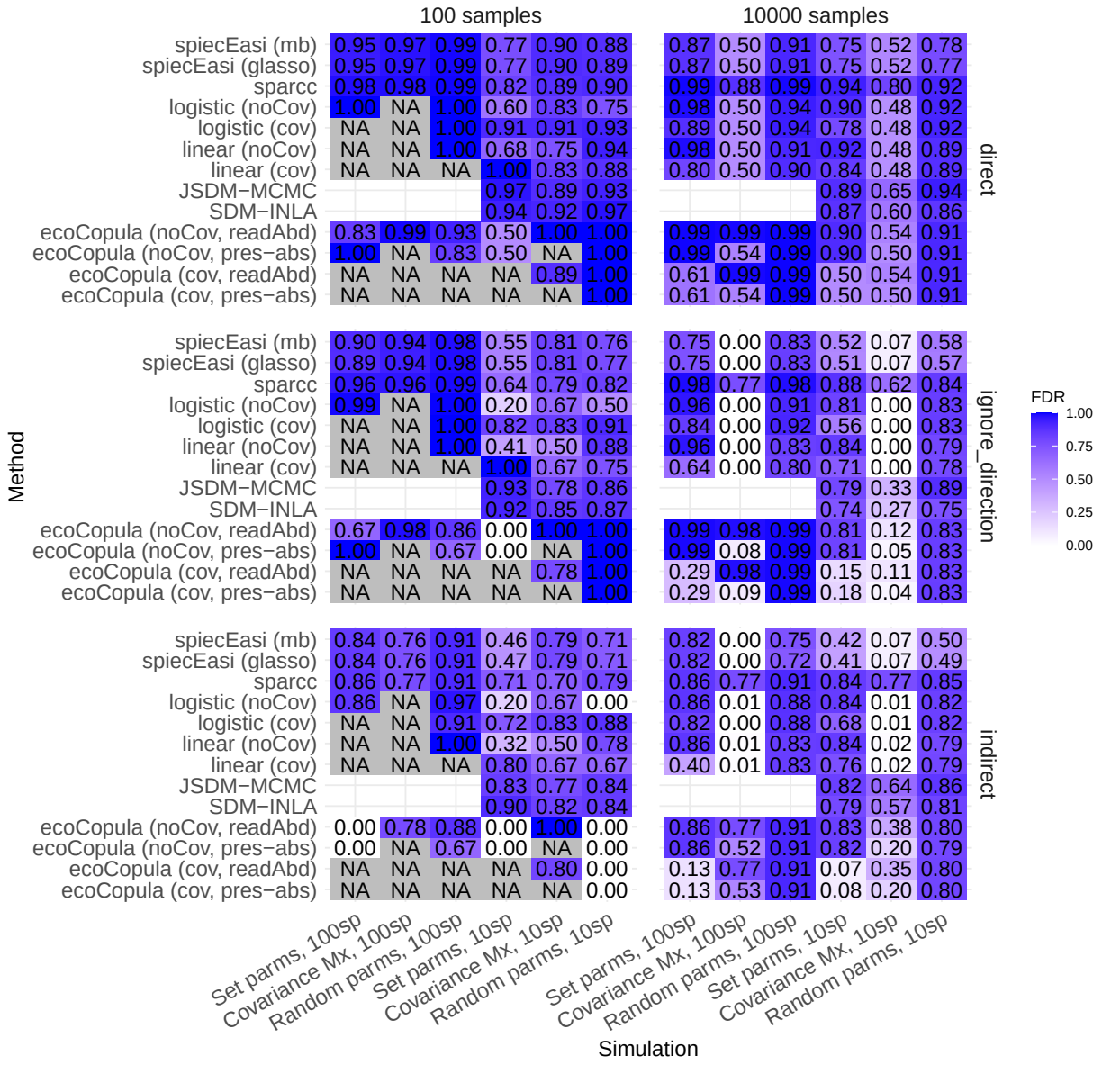

Figure 4: False discovery rates of direct interactions, direct/symmetric interactions (ignore_direction), and indirect interactions with 100 and 10,000 samples for all methods and all simulations. False discovery rate is defined as the number of false positive detections of species interactions divided by the total number of inferred interactions, with interactions defined in three ways. NA indicates that no interactions were inferred, and therefore false discovery rate is undefined. Where there is an option, `cov` refers to whether covariates were used. Similarly, where adjustable, `pres-abs` and `readAbd` refer to detection/non-detection data versus read abundance data inputs respectively.

inferred very few interactions. On the other hand, with 10,000 samples, hundreds or thousands of total interactions were inferred across the 100 simulations per dataset (Appendix B). One notable exception was SPIEC-EASI, which did not consistently infer fewer interactions when fewer points were sampled. In additional tests with as few as 10 samples, SPIEC-EASI continues to infer large numbers of interactions (Appendix D), but with so few samples it is unlikely that there is enough information in the data to infer this, indicating potential problems with calibration.

High numbers of samples are needed to infer associations between species

Increasing sample size is expected to improve the performance of statistical methods in most cases, although model misspecification can result in incorrect inference even as sample size gets very large [37]. Here, we have tested each model at a relatively low sample size, which we believe is realistic for sedaDNA studies at this time (100 samples). We have also tested all models with 10,000 samples, which is a higher number of samples than would currently be reasonable in most studies. As expected, we find that model performance is better for most methods and scenarios using a higher sample size, although FDR remains high in most cases (Figure 4).

Using only 100 samples, on average for random parameter settings in the ecological simulation model, no method performs better than a random classifier (defined as choosing to infer interactions completely at random, which is expected to follow the diagonal on a ROC curve plot) (Figure 5, 6, 7, 8). Using 100 samples, none of the methods that take detection/non-detection data (logistic regression, SDM-INLA, JSDM-MCMC, EcoCopula) perform better than a random classifier on the ecological simulation data or the covariance matrix simulation data, regardless of number of species (except SDM-INLA for the set-parameters case with 5-10 species appears to perform slightly better than random). For methods that take abundance data (SPIEC-EASI, SparCC, linear regression), there are scenarios under favorable simulation conditions (set-parameters ecological simulation) where they are able to perform better than a random classifier even with only 100 samples. Even in the covariance matrix simulation scenario, where many methods perform extremely well with 10,000 samples, 100 samples is not enough for any method to perform well (Figure 5, 6, 7, 8).

With model selection, it is often the case that very few interactions were inferred with only 100 samples, which indicates low power to detect these interactions if the method is calibrated correctly (Appendix B). When interactions are inferred, they are most often incorrect (Figure 4). Therefore, we find it to be a good sign when methods consistently infer very few interactions, since no method is able to detect the correct set of interactions. As expected, interactions are inferred more often with a higher number of samples and are more often correct (Figure 4, 5, 6). However, there is still a high chance that inferred direct interactions are incorrect under many scenarios (Figure 4). Indirect interactions are more likely to be inferred correctly, although false discovery rates are still high under many scenarios for low sample sizes and ROC curves do not improve significantly (and actually often get worse) (Figure 4, 5, 6, 7, 8). Using higher sample sizes, the false discovery rate is only consistently low in very favorable scenarios (covariance matrix simulation; note exception of set-parameters with covariates under the EcoCopula model, though this relies on very few inferred interactions) (Figure 4).

The level of correlation induced by species interactions in real data, which is expected to influence the necessary number of samples, is unknown and likely highly variable between systems. The correlation level set here in the covariance matrix simulations was 0.1. The mean empirical correlation of the samples in the ecological simulations for set-parameters simulations when there is a positive interaction is 0.168 and the mean for negative interactions is -0.170 for simulations with 10 species. For 100 species, it is 0.118 for positive and -0.105 for negative interactions. Since the level of correlation induced by strong interactions may be higher, we ran an additional test of linear and logistic regression on covariance matrix simulations where 100 species were associated in 50 independent pairs and the correlation level was set at 0.5. In these simulations, 250 samples were sufficient for linear regression without covariates to classify the samples nearly perfectly according to ROC, with logistic regression as a close second (Appendix G). Other methods were not tested on these data, although based on performance with large sample sizes in the covariance matrix simulation model, we believe that they would perform similarly well under this idealized scenario. Many factors were not modeled here so we believe this is a less realistic scenario.

Information about species abundance improves inference of associations between taxa

Some methods tested here take sedaDNA read abundance as input, which is considered a proxy for relative species abundance or relative biomass. Others take detection/non-detection data, which is derived from read abundances by setting a threshold for how many reads indicates species presence. Using detection/non-detection data is justified if read abundances are biased and therefore are not a good proxy for organism abundance. Additionally, this can be considered a way to give low-abundance taxa a higher weight in the model [38]. We consider these methods to fundamentally take the same data, but process it differently.

We find that information about species abundance improves inference of species associations, particularly with high numbers of species. In the simplest case, we compare logistic regression (detection/non-detection data) to linear regression (read abundance data). For the simulations where any methods perform better than a random classifier, linear regression performs better than logistic regression according to ROC curves (Figure 5, 6, 7, 8). According to FDR, there are exceptions to this rule due to differences in the effect of the calibration of these models (Figure 4). EcoCopula is the only other method that flexibly takes detection/non-detection or read abundance data. With detection/non-detection data, EcoCopula performs similarly to other methods (Figure 5, 6, 7, 8). With model selection, EcoCopula performs relatively well according to false discovery rate (Figure 4). There is minimal difference between the EcoCopula results using read abundance data versus detection/non-detection data in most scenarios, with the exception of covariance matrix simulations, in which the ROC curves do not differ but calibration differs significantly with 10,000 samples (Figure 4, 5, 6, 7, 8).

For larger numbers of species, linear regression performs similarly to the other methods that take the same data (SparCC, SPIEC-EASI) except in the case with 100 samples in the set-parameters simulation, which we believe is due to parameter unidentifiability (Figure 6, 8). For lower numbers of species, linear regression performs similarly to SparCC, but SPIEC-EASI is sensitive to the assumption of a large number of taxa for inference of direct interactions and therefore

performs poorly (Figure 5, 7).

For methods that take detection/non-detection data, SDM-INLA performs best when averaged across random simulation parameters at a high number of samples (Figure 5, 7). However, this result does not generalize to our set-parameters simulation, where logistic regression performs better (Figure 5, 7). The main difference between these models is that SDM-INLA models spatiotemporal autocorrelation. The effect of this may depend whether a dataset has spatiotemporal structure and whether this structure is correctly modeled by the SDM-INLA method.

In simulations with random parameters, we see less separation of the different methods in the ROC curves than when the parameters were set at a single level (Figure 5, 6). This is the effect of averaging performance across many different ecological scenarios. In some ecological contexts, some methods may perform better than others, but this may not be consistent across all contexts. As expected, the variance of the results within a method when parameters were varied was much higher than when parameters in the simulation were held constant.

Here we have not added species-level biases to simulated read abundance, which means that read abundance is correlated to species abundance, although often with significant levels of noise. We expect that more biases in read abundance would exist in real data, so we interpret this result to mean that organism abundance information is important for these models, and not necessarily that sedaDNA read abundance information improves their performance. Further investigation is needed to determine the circumstances under which sedaDNA read abundances are a good proxy for species abundances.

Inference of species associations can be worse with abiotic covariates than without

According to ROC curves, inference of species associations is often worse with environmental covariates included in the model than without them (Figure 5, 6, 7, 8). We would generally expect the inclusion of covariates to cause loss of statistical power (this would not affect ROC

curves), but improve accuracy of inference because shared environmental associations can cause correlations between species detections or abundances that are not caused by species interactions [37]. However, if many covariates are included in a model and they do not affect multiple species, correlations caused by random fluctuations in the data can cause false discovery rates to be higher (Appendix H).

Additionally, in the set of simulations with set parameters, there is a high level of correlation between some of the environment covariates and the species that they affect. In the methods that take both species read abundances or occurrences and covariates as inputs, this can cause issues with identifiability between the effect of the covariate and the species that it is correlated with (Appendix I). Therefore, it is important to check for collinearity between species and covariates if covariates are going to be included in these models.

In these simulations, the only case with shared covariates between species are the random-parameter ecological model simulations. In the set-parameter simulations, no covariates are shared between species in an attempt to assess performance of these models in the absence of confounding. In covariance matrix simulations, covariates did not affect simulated read abundances at all, but were included in the inference models to test the effect of including covariates that are, in reality, statistically independent of all species abundances (See Methods for details).

SPIEC-EASI is highly sensitive to assumptions about number of species

SPIEC-EASI explicitly assumes that the number of species is large, although the actual number of species required is not specified. There is an approximation in their model that is more accurate as the number of species gets large [8]. We find that at the 5-10 species level, SPIEC-EASI is highly sensitive to this assumption according to ROC curves when looking at direct associations (Figure 5). With 50-100 species, this sensitivity disappears and it diminishes when looking at indirect associations (Figure 6, 7). Specifically, with below 10 species, we find that SPIEC-EASI seems to have strong evidence for actual negative interactions between species and strong evidence against actual positive interactions between species (Appendix J). This effect is not observed after model

selection because the model never selects the regularization parameter in this region of the ROC curve (Figure 5, 6). Therefore, the false discovery rates after model selection are not noticeably affected (Figure 4).

Regression performs similarly to other methods with the same data

Counter to expectations, we find that in most scenarios, linear and logistic regression perform, on average, similarly to other methods with the same data input (Figure 5, 6, 7, 8). This should not be interpreted to mean that regression is the correct model for this data, but rather that the more complex methods do not improve modeling of the data structure in these scenarios. Each of the more complex methods attempts to account for different aspects of the data better than linear or logistic regression, but we find that for a variety of simulation scenarios, they do not succeed in better modeling the data generating process. It may be the case that under certain circumstances, each model does perform better, but we find that on average across many scenarios, no method recovers associations consistently better than regression. Additionally, regression methods are much faster than any of the other methods.

Many dimensions of simulation parameter space affect the success of regression models in predicting direct, symmetric interactions

In the simulations produced with random parameters, there was a great variety of resulting false discovery rates between different parameter sets. In order to better understand how the simulation parameters are affecting the success of the inference models, we created an additional 1000 data sets with random parameters (drawn from distributions specified in Table 3), and used these simulation parameters as predictors for a random forest model which predicts false discovery rates of direct, symmetric interactions. Due to the large number of simulations in this set, this was only performed for logistic and linear regression, which are much faster than other methods. We believe this is justified since no other methods outperformed these methods in tests on smaller simulation sets, though we do not know if the same simulation parameters would affect the performance of all methods.

For 10,000 samples, three simulation parameters seem to affect FDR significantly for logistic and

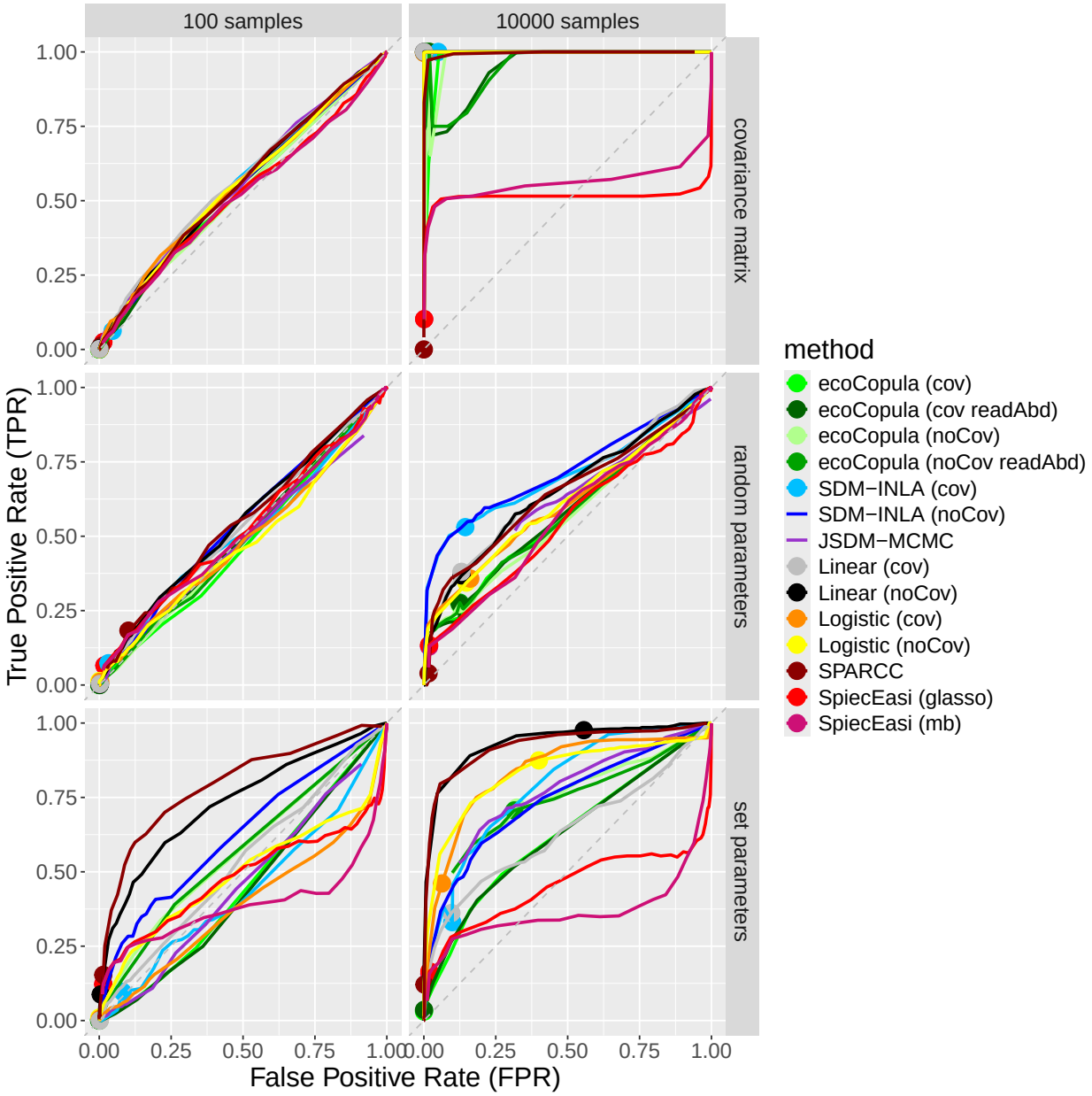

Figure 5: ROC curves for inference of direct, symmetric associations between 10 simulated species (5-10 species actually present). `cov` and `noCov` refers to whether covariates were included in the model, but in all cases, species interactions were inferred. `readAbd` is specified for EcoCopula because this method allows for different inputs (detection/non-detection data or read abundance data).

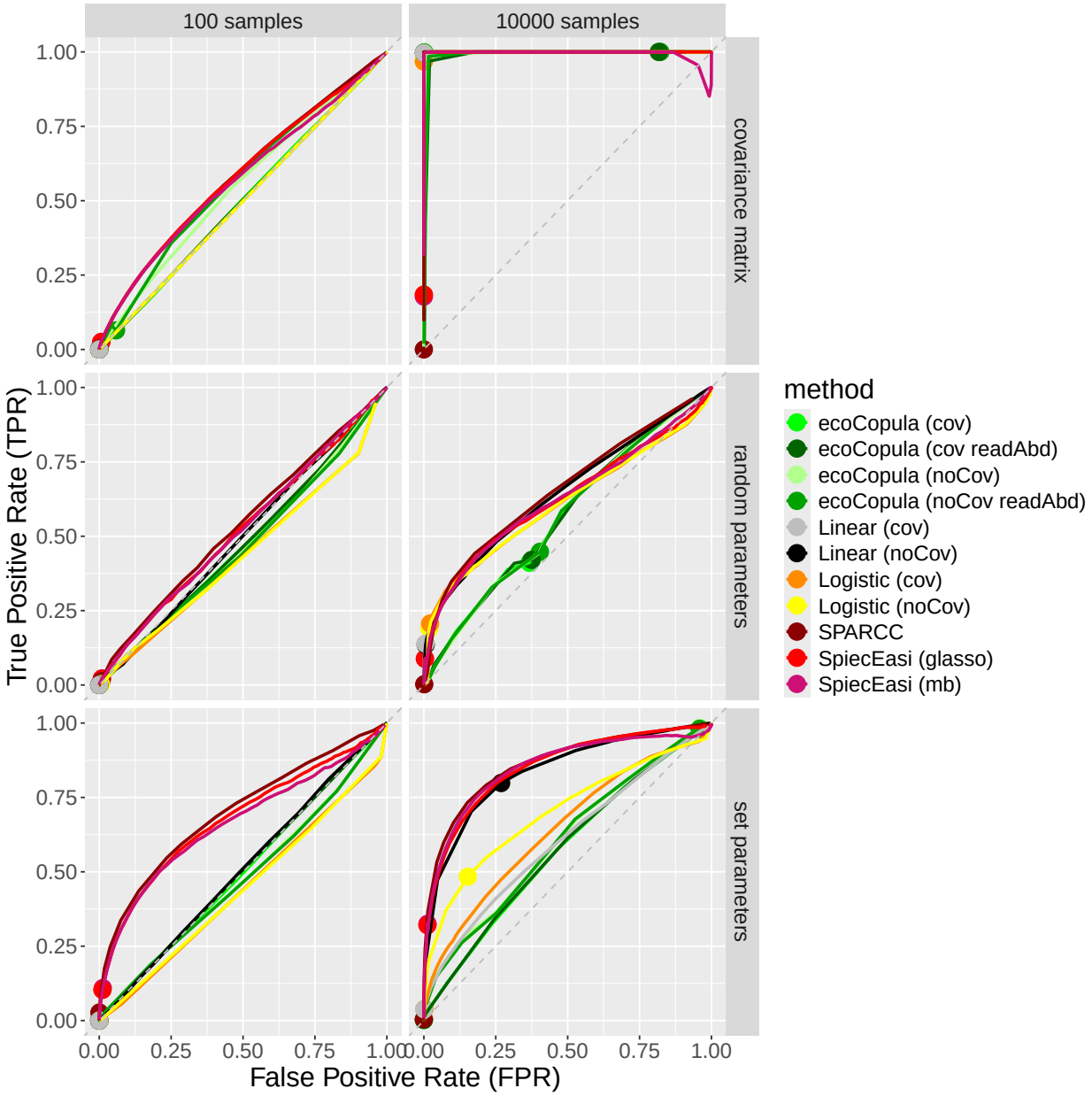

Figure 6: ROC curves for inference of direct, symmetric associations between 100 simulated species (50-100 species actually present). `cov` and `noCov` refers to whether covariates were included in the model, but in all cases, species interactions were inferred. `readAbd` is specified for EcoCopula because this method allows for different inputs (detection/non-detection data or read abundance data).

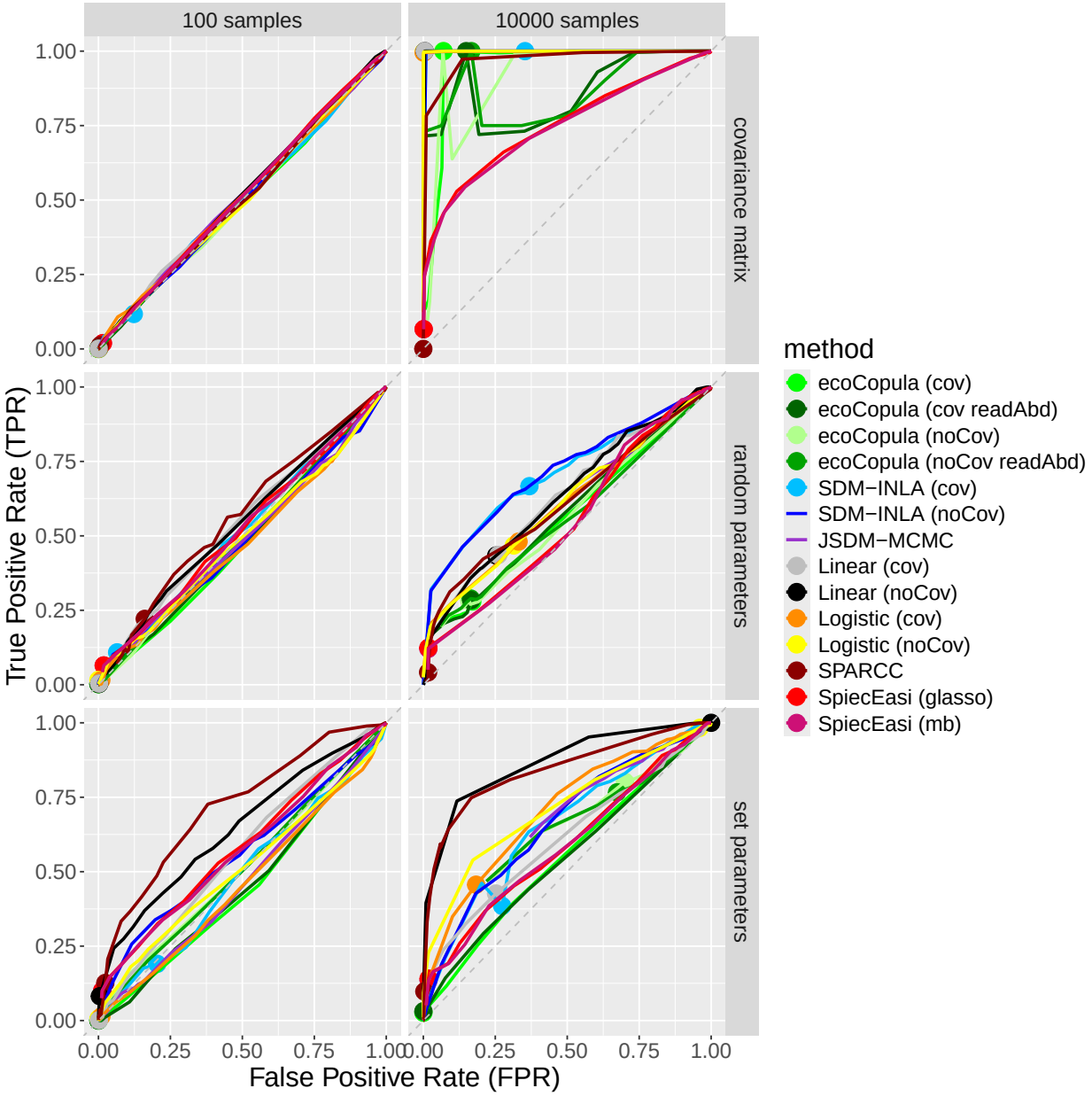

Figure 7: ROC curves for inference of indirect associations between 10 simulated species (5-10 species actually present). `cov` and `noCov` refers to whether covariates were included in the model, but in all cases, species interactions were inferred. `readAbd` is specified for EcoCopula because this method allows for different inputs (detection/non-detection data or read abundance data).

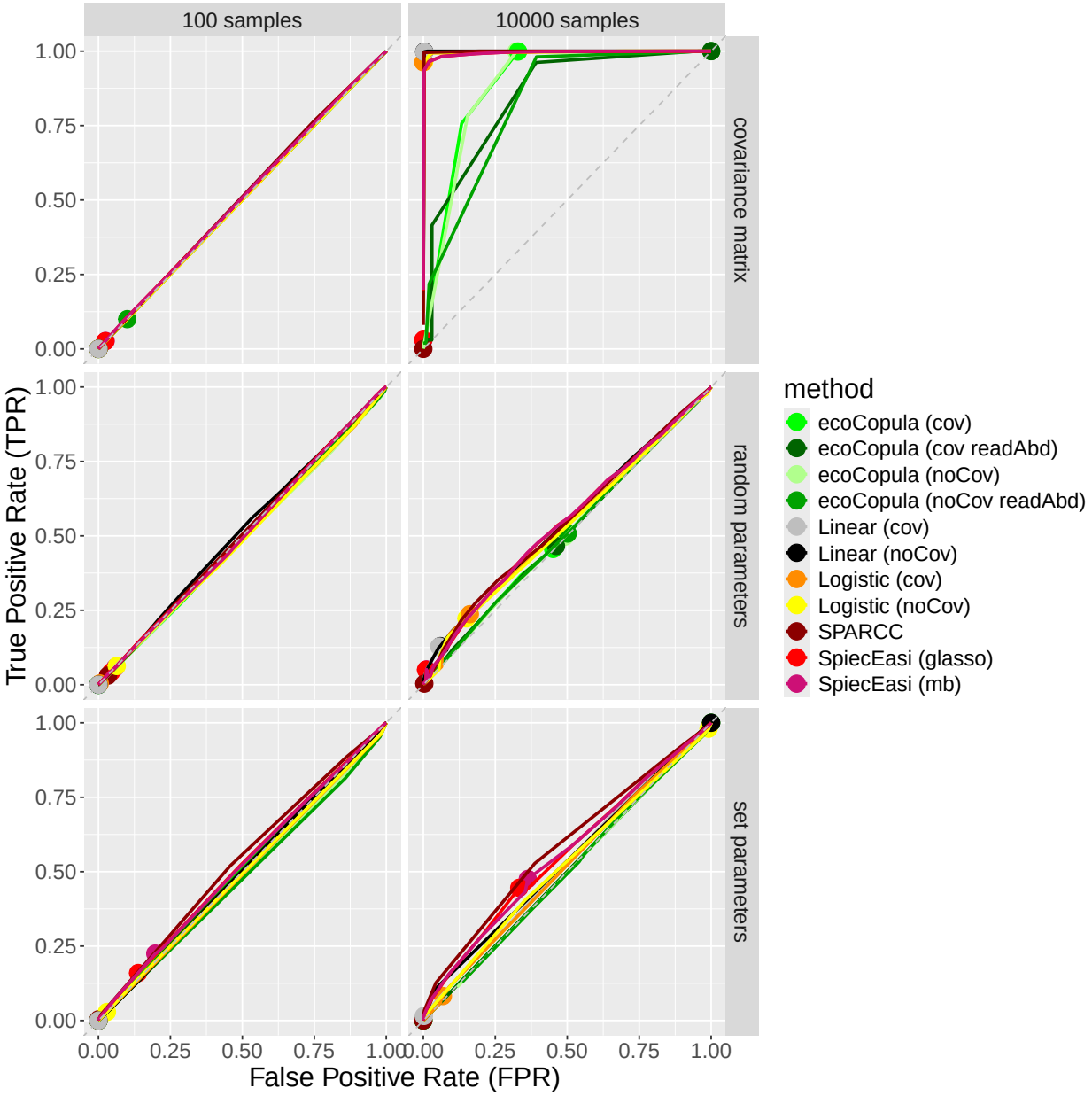

Figure 8: ROC curves for inference of indirect associations between 100 simulated species (50-100 species actually present). `cov` and `noCov` refers to whether covariates were included in the model, but in all cases, species interactions were inferred. `readAbd` is specified for EcoCopula because this method allows for different inputs (detection/non-detection data or read abundance data).

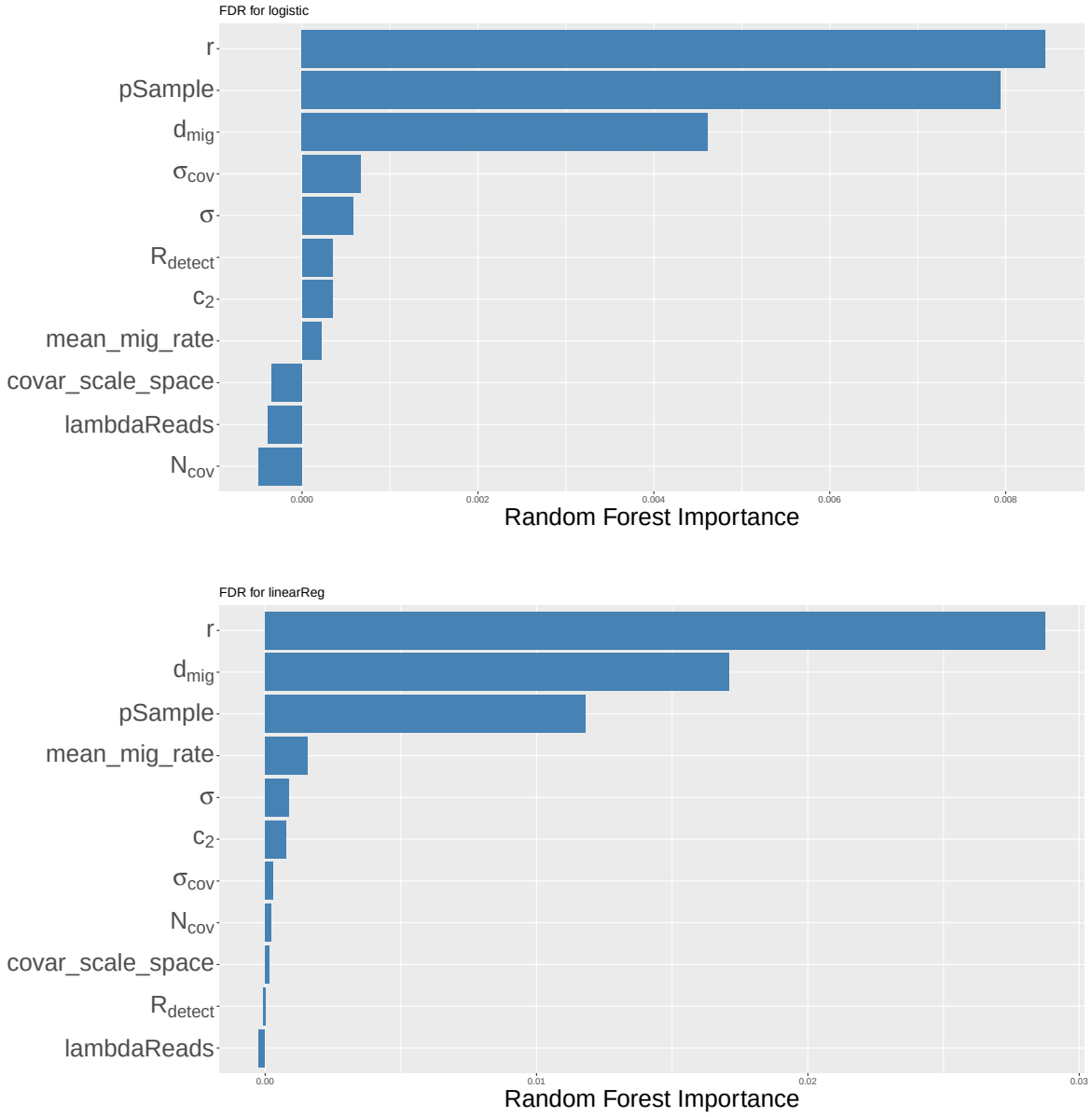

Figure 9: Importance of predictor variables in random forest predicting FDR as a function of simulation parameters. Analysis performed with 10 species in ecological simulations with random parameter settings and 10,000 samples per simulation. Random forest was trained on results from 1,000 simulations.

linear regression. For larger r (population growth rate for all species) models tend to have lower FDR (perform better) (Appendix L). Larger population growth rates will cause the populations of species to change more rapidly and to reach their carrying capacity faster, causing less lag between a change in environment and the change in the species population. Non-equilibrium dynamics caused by lag between changing conditions and population growth can cause temporal and spatial autocorrelation not accounted for by covariates. This violates the assumption of conditional independence of the samples in linear and logistic regression and can lead to a loss of information.

With a larger migration radius (d_{mig}) , the models tend to have a higher FDR (perform worse) (Appendix L). In this simulation model, migration rates are not affected by species interactions and may obscure the effect of one species on another.

The third simulation parameter that seems to affect false discovery rate in this set of simulations is the individual sampling probability ($pSample$) (Figure 9). When this parameter is higher, FDR tends to be higher (Appendix L). This effect is stronger for logistic regression than linear regression. This is a counter-intuitive result because we would expect a higher probability of sampling individuals to result in better performance of the models. However, we also observed that high observed rate of species presence can result in difficulty with inference because there is less power to infer interactions if species are present almost everywhere, which we believe to be the cause of this effect (Appendix M).

Random forest models were verified by testing for predictive accuracy on an independent test set of 100 simulations. RMSE for the test sets were evaluated against a naive predictor of FDR (median of observed FDRs). The random forest predicts FDR between 5% and 18% better than the naive predictor (Appendix K).

Discussion

Here, we have tested a range of methods which have been used to detect associations between taxa from sedaDNA data. We simulated data under a variety of models, including a simple model where all data points are independent, and a custom ecological model that leverages principles of ecological theory to create more realistic data. Explicitly simulating population dynamics through time in the presence of a changing abiotic environment will create patterns in the data that mimic those in nature. We find that in all but the most idealized scenarios, false discovery rates of species associations are high for all methods tested.

Learning how past ecosystems have changed in response to environmental change will be an essential part of understanding and mitigating present-day global climate change and making informed resource-management decisions [22]. Biotic interactions are a key piece of this puzzle [4] and sedaDNA is a new frontier in understanding ecological interactions. It has the potential to uncover interactions within and between trophic levels, model species distributions over large spatiotemporal scales, and show how biodiversity has changed over periods of massive environmental change [1–3, 11]. Robust computational methods are needed to come to accurate and reproducible conclusions, and it is important to understand the relative performance of methods under different conditions.

Many species distribution models have been developed in ecology [10], but since these methods were developed with traditional ecological survey data in mind, we find that they are often not scalable to the number of taxa that is common in sedaDNA datasets. Many of these methods account for temporal and/or spatial information in the data, and allow for information about abiotic covariates [10, 11]. However, they do not account for the compositional structure of sedDNA read abundance data. On the other hand, several methods have been developed specifically for sedDNA data, and therefore account for non-independence due to compositional sedDNA read data [8, 18]. However, many of the most commonly used methods were developed specifically for microbiome sedaDNA data and therefore assume large numbers of taxa (or OTUs) in a single data set. We find that SPIEC-EASI is very sensitive to this assumption, exhibiting

very poor performance on data with small numbers of taxa (less than 10). This is important, even in the sedaDNA field, because depending on the taxa under study and the taxonomic level of assignment, many studies may want to do similar analyses with small numbers of taxa [11, 13].

It has been previously established that inferring species interactions from spatial detection/non-detection data is a challenging statistical problem, and that there are many potential causes for false inferences from these data [37]. It has also been documented, theoretically for co-occurrence (detection/non-detection) data [37] and empirically for read abundance data [8], that high sample sizes are needed to accurately estimate associations between taxa. The challenge is in part because a large number of taxa create an even larger set of potential interactions [8], but also due to the complexity of the ecological networks involved and the abiotic factors that influence them, among other reasons [37]. From a mathematical perspective, for linear regression with all assumptions met, when the correlation is 0.1 (as in the covariance matrix simulations) using a p-value cutoff of 0.05 with Bonferroni correction, the necessary sample size to detect 80% of true interactions is nearly 2000 samples for 10 species and nearly 3000 samples for 100 species (Appendix F). If the correlation induced by species interactions is higher, the expected sample size needed would go down (Appendix F), which we also observe for simulated data (Appendix G), so lower samples sizes may be sufficient under some circumstances. However, even in these cases, false discovery rates may remain high due to model misspecification. Since very few studies using sedaDNA currently have more than a few hundred sampled points, high sample sizes are currently unrealistic, but we are hopeful that future data will rise to meet this challenge.

Sample size is not the only potential concern, since we observe high false discovery rates in our ecological simulations even with high sample sizes. The assumptions about the data generating process vary across methods. For example, none of the methods tested here account for uncertainty in covariates. Some are parametric methods that rely, for example, on residuals being independent and normal [15]. Many of them do not account for residual autocorrelation in space and/or time, which may be caused either by non-equilibrium distributions or by autocorrelation of an unmeasured abiotic factor [15, 39]. Some methods cannot detect non-linear interactions

between species and covariates or between species. These are all examples of assumptions that we have violated to varying degrees in the simulations. We have shown here that given enough samples, some methods perform moderately well with some assumptions violated, but the only case in which any method is able to reliably recover interactions with a low false discovery rate is under an extremely idealized scenario.

False discovery rates reported here depend heavily on the calibration of various methods (such as model selection or choosing a p-value cutoff). FDR reported here reflects one point on the corresponding ROC curve, which measures the performance of the model independent of calibration. Since calibration is an intrinsic part of these methods, we consider these calibrated values to be of great significance. The drastic differences in calibrated FDR values between different versions of the same method and between methods is often more a consequence of the calibration than the underlying model. For example, the ROC curves for all methods may look similar across all models in many cases, reflecting similar success of the underlying model, but the average FDR varies considerably due to poor calibration of some methods compared to others. We would also like to note that some of the extreme FDRs may be caused by low numbers of inferred interactions. With very low numbers of total inferred interactions, the variance of the average FDR may be high since simulations where no interactions are inferred are excluded from the average (FDR is undefined when 0 total interactions are inferred). The number of total inferred interactions varies dramatically between methods and simulations, so we are more certain of some values compared to others (Appendix B).

As in all simulation models, we make many assumptions about the data generating process. For example, in our ecological simulations we assume that populations in this simulation will grow according to a logistic growth curve with a changing carrying capacity through time. This may not be the optimal way to calculate growth rates for some environments and species [40]. Additionally, there has been considerable discussion in ecology about the structure of the interaction networks between species [22], and the inference methods discussed here are often used to estimate overall characteristics of the network [8]. Because actual ecological network structure is likely variable in different systems [6], we assign interactions between species randomly in our

simulations. This is likely a conservative assumption because graphs with evenly distributed neighborhoods are easier to recover than those with large hub nodes [8]. Additional assumptions of our simulation model include assuming that some species traits and data characteristics are constant across time, space, and species, which is certainly not always the case. For example, we expect that covariate uncertainty, average read abundances, and species detection rates may vary in space and time and across species [41, 42]. Our simulation model also assumes that interactions between species are constant over time and space, which may not be true over evolutionary timescales. All of these assumptions are conservative, but we recognize that we also made assumptions about the data that are less conservative. For example, strong species interactions in real data may induce higher levels of correlation than we see in our models. Additionally, other characteristics of the simulation may create unrealistic dynamics that affect the performance of the inference models. However, within our chosen simulation framework, we attempt to simulate many ecological dynamics by varying the parameters widely in our simulations. Additionally, we test the methods under a simpler simulation framework (covariance matrix simulations) that is not based on ecological theory but is expected to minimally violate assumptions of the inference methods. We note that simpler simulated data may have favored simpler methods. For example, accounting for spatial and temporal structure when it is not there or including covariates that have no effect on the data may be detrimental to some of the more complex methods.

The input to some models was read abundances, whereas other methods take only detection/non-detection data. EcoCopula was the only model that takes flexible input. We find that information about species abundances improves performance of models at detecting species interactions. However, in real data, read abundance data may be biased by taxon, which may introduce confounding factors that result in spurious correlations. For example, in real data, many factors may create biases such as species-specific DNA deposition rates [26], PCR bias [43], or bias in species assignments [42]. In our simulations, we did not include per-species bias in rates of DNA deposition, so it is not surprising that simulated read abundances have more information about species interactions than detection/non-detection data, though we do not know whether this is the case in real data. These concerns would still be applicable to some extent in detection/non-detection data, as detection rates may also be biased across taxa, space, and time,

but the influence of these biases may be lower. We also note that in microbiome studies, reads are not assigned to taxa but rather grouped into OTUs, which may have different properties with respect to read abundance biases. Further investigation will be needed to fully understand these effects and the circumstances under which it is effective to use read abundances as a proxy for species abundance.

The increasing availability of whole-community ecology data has enormous potential to uncover ecological dynamics, parameterize models, and make predictions about the future of ecosystems. However, data of this complexity must also be approached with caution, and benchmarking tools under a variety of scenarios is an essential step in understanding and interpreting these results. Further investigations into the relative success of these methods and others will be needed for a complete understanding of their performance. For example, it is possible that these methods would be more successful at recovering overall network properties as opposed to specific interactions [8]. Additionally, many other methods exist that were not tested here due to differences in data requirements and outputs, including some mechanistic models, which may perform better under certain conditions [44, 45]. An area where sedaDNA has a lot of potential is in its ability to recover species distribution data across a wide range of taxa, including microbes, plants, and megafauna, from the same samples. However, tools to analyze data on differing taxonomic scales have been developed with differing assumptions (perhaps for good reason). In order to fully take advantage of the potential of these data, we will need analysis frameworks that are able to handle many different ecological scenarios.

Appendices

A JSDM-MCMC and SDM-INLA methods time to run

B Number of Inferred Interactions

C Convergence of JSDM-MCMC

D SpiecEasi number of inferred interactions versus number of samples

E Results using different methods of counting false positives

F Sample size versus correlation for linear regression

G Covariance matrix simulations with higher correlation

H Effect of adding many unnecessary covariates

I Multicollinearity

J SPIEC-EASI performance on low number of species

K Random forest performance on test set

L Random forest ICE plots

M Actual interactions cause higher percent presence on average in ecological simulation, but logistic regression infers interactions more for species with close to 50% presence

References

1. Beng, K. C. & Corlett, R. T. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. en. *Biodiversity and Conservation* **29,** 2089–2121. ISSN: 0960-3115, 1572-9710. http://link.springer.com/10.1007/s10531-020-01980-0 (2022) (June 2020).

2. Alsos, I. G. *et al.* Using ancient sedimentary DNA to forecast ecosystem trajectories under climate change. en. *Philosophical Transactions of the Royal Society B: Biological Sciences* **379,** 20230017. ISSN: 0962-8436, 1471-2970. https://royalsocietypublishing.org/doi/10.1098/rstb.2023.0017 (2024) (May 2024).

3. Williams, J. W. *et al.* Strengthening global-change science by integrating aeDNA with paleoecoinformatics. en. *Trends in Ecology & Evolution,* S0169534723001234. ISSN: 01695347. https://linkinghub.elsevier.com/retrieve/pii/S0169534723001234 (2023) (May 2023).

4. Åkesson, A. *et al.* The importance of species interactions in eco-evolutionary community dynamics under climate change. en. *Nature Communications* **12,** 4759. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-24977-x (2024) (Aug. 2021).

5. Dussex, N. *et al.* Integrating multi-taxon palaeogenomes and sedimentary ancient DNA to study past ecosystem dynamics. en. *Proceedings of the Royal Society B: Biological Sciences* **288,** 20211252. ISSN: 0962-8452, 1471-2954. https://royalsocietypublishing.org/doi/10.1098/rspb.2021.1252 (2024) (Aug. 2021).

6. Bascompte, J., García, M. B., Ortega, R., Rezende, E. L. & Pironon, S. Mutualistic interactions reshuffle the effects of climate change on plants across the tree of life. en. *Science Advances* **5,** eaav2539. ISSN: 2375-2548. https://www.science.org/doi/10.1126/sciadv.aav2539 (2023) (May 2019).

7. Chen, W. & Ficetola, G. F. Numerical methods for sedimentary-ancient-DNA-based study on past biodiversity and ecosystem functioning. en. *Environmental DNA* **2,** 115–129. ISSN: 2637-4943, 2637-4943. https://onlinelibrary.wiley.com/doi/10.1002/edn3.79 (2024) (Apr. 2020).

8. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. en. *PLOS Computational Biology* **11** (ed Von Mering, C.) e1004226. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1004226 (2023) (May 2015).

9. Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C. & Moles, A. T. Untangling direct species associations from indirect mediator species effects with graphical models. en. *Methods in Ecology and Evolution* **10** (ed Murrell, D.) 1571–1583. ISSN: 2041-210X, 2041-210X. https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13247 (2023) (Sept. 2019).

10. Elith, J. & Graham, C. H. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. en. *Ecography* **32,** 66–77. ISSN: 09067590, 16000587. https://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2008.05505.x (2022) (Apr. 2009).

11. Wang, Y. *et al.* Late Quaternary dynamics of Arctic biota from ancient environmental genomics. en. *Nature* **600,** 86–92. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/s41586-021-04016-x (2022) (Dec. 2021).

12. Schliep, E. M. *et al.* Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. en. *Global Ecology and Biogeography* **27,** 142–155. ISSN: 1466-822X, 1466-8238. https://onlinelibrary.wiley.com/doi/10.1111/geb.12666 (2022) (Jan. 2018).

13. Pollock, L. J. *et al.* Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (¡span style="font-variant:small-caps;"¿JSDM¡/span¿). en. *Methods in Ecology and Evolution* **5** (ed McPherson, J.) 397–406. ISSN: 2041-210X, 2041-210X. https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12180 (2024) (May 2014).

14. Pichler, M. & Hartig, F. A new joint species distribution model for faster and more accurate inference of species associations from big community data. en. *Methods in Ecology and Evolution* **12,** 2159–2173. ISSN: 2041-210X, 2041-210X. https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13687 (2022) (Nov. 2021).

15. F. Dormann, C. *et al.* Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. en. *Ecography* **30,** 609–628. ISSN: 09067590. https://onlinelibrary.wiley.com/doi/10.1111/j.2007.0906-7590.05171.x (2023) (Oct. 2007).

16. Banerjee, S., Schlaeppi, K. & Van Der Heijden, M. G. A. Keystone taxa as drivers of microbiome structure and functioning. en. *Nature Reviews Microbiology* **16,** 567–576. ISSN: 1740-1526, 1740-1534. https://www.nature.com/articles/s41579-018-0024-1 (2024) (Sept. 2018).

17. Zimmermann, H. H. *et al.* Marine ecosystem shifts with deglacial sea-ice loss inferred from ancient DNA shotgun sequencing. en. *Nature Communications* **14,** 1650. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-36845-x (2024) (Mar. 2023).

18. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. en. *PLoS Computational Biology* **8** (ed Von Mering, C.) e1002687. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1002687 (2024) (Sept. 2012).

19. Wagg, C., Schlaeppi, K., Banerjee, S., Kuramae, E. E. & Van Der Heijden, M. G. A. Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. en. *Nature Communications* **10,** 4841. ISSN: 2041-1723. https://www.nature.com/articles/s41467-019-12798-y (2024) (Oct. 2019).

20. Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. en. *Briefings in Bioinformatics* **22,** 178–193. ISSN: 1477-4054. https://academic.oup.com/bib/article/22/1/178/5678919 (2024) (Jan. 2021).

21. Dormann, C. F. *et al.* Correlation and process in species distribution models: bridging a dichotomy. en. *Journal of Biogeography* **39,** 2119–2131. ISSN: 0305-0270, 1365-2699. https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2699.2011.02659.x (2024) (Dec. 2012).

22. Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C. & Dieckmann, U. Complexity and stability of ecological networks: a review of the theory. en. *Population Ecology* **60,** 319–345.

ISSN: 1438-3896, 1438-390X. https://esj-journals.onlinelibrary.wiley.com/doi/10.1007/s10144-018-0628-3 (2024) (Oct. 2018).

23. Yuan, M. M. *et al.* Climate warming enhances microbial network complexity and stability. en. *Nature Climate Change* **11,** 343–348. ISSN: 1758-678X, 1758-6798. https://www.nature.com/articles/s41558-021-00989-9 (2024) (Apr. 2021).

24. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. en. *The ISME Journal* **10,** 1669–1681. ISSN: 1751-7362, 1751-7370. https://academic.oup.com/ismej/article/10/7/1669-1681/7538221 (2024) (July 2016).

25. Wang, S. C. & Marshall, C. R. Estimating times of extinction in the fossil record. en. *Biology Letters* **12,** 20150989. ISSN: 1744-9561, 1744-957X. https://royalsocietypublishing.org/doi/10.1098/rsbl.2015.0989 (2022) (Apr. 2016).

26. Giguet-Covex, C. *et al.* New insights on lake sediment DNA from the catchment: importance of taphonomic and analytical issues on the record quality. en. *Scientific Reports* **9,** 14676. ISSN: 2045-2322. http://www.nature.com/articles/s41598-019-50339-1 (2022) (Dec. 2019).

27. Zurell, D., Pollock, L. J. & Thuiller, W. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? en. *Ecography* **41,** 1812–1819. ISSN: 09067590. https://onlinelibrary.wiley.com/doi/10.1111/ecog.03315 (2023) (Nov. 2018).

28. Dennis, B. in *Estimation and analysis of insect populations* 219–238 (Springer, 1989).

29. Team, R. C. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2024. https://www.R-project.org/.

30. Martyn Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling* in *Proceedings of the 3rd international workshop on distributed statistical computing* **124** (Vienna, Austria, 2003), 1–10.

31. Yu-Sung Su & Masanao Yajima. *R2jags: Using R to Run 'JAGS'* R package version 0.7-1. 2021. https://cran.r-project.org/package=R2jags.

32. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7.** ISSN: 0883-4237. https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Inference-from-Iterative-Simulation-Using-Multiple-Sequences/10.1214/ss/1177011136.full (2024) (Nov. 1992).

33. Martyn Plummer, Nicky Best, Kate Cowles & Karen Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6,** 7–11. https://journal.r-project.org/archive/ (2006).

34. Lindgren, F. & Rue, H. Bayesian Spatial Modelling with *R* - **INLA**. en. *Journal of Statistical Software* **63.** ISSN: 1548-7660. http://www.jstatsoft.org/v63/i19/ (2024) (2015).

35. Kurtz, Z., Mueller, C., Miraldi, E. & Bonneau, R. *SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference* R package version 1.1.2, commit d6bc1273211fef632b86f65b4b98290805b9ab5b. 2023. https://github.com/zdk123/SpiecEasi.

36. Wright, M. N. & Ziegler, A. **ranger** : A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. en. *Journal of Statistical Software* **77.** ISSN: 1548-7660. http://www.jstatsoft.org/v77/i01/ (2024) (2017).

37. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. en. *Ecology Letters* **23** (ed Jeffers, E.) 1050–1063. ISSN: 1461-023X, 1461-0248. https://onlinelibrary.wiley.com/doi/10.1111/ele.13525 (2024) (July 2020).

38. Claussen, J. C. *et al.* Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. en. *PLOS Computational Biology* **13** (ed Albert, R.) e1005361. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1005361 (2023) (June 2017).

39. Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. & Guisan, A. Measuring the relative effect of factors affecting species distribution model predictions. en. *Methods in Ecology and Evolution* **5** (ed O'Hara, R. B.) 947–955. ISSN: 2041210X. https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12203 (2023) (Sept. 2014).

40. Hatton, I. A., Mazzarisi, O., Altieri, A. & Smerlak, M. Diversity begets stability: Sublinear growth and competitive coexistence across ecosystems. en. *Science* **383,** eadg8488. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.adg8488 (2024) (Mar. 2024).

41. Capo, E. *et al.* Lake Sedimentary DNA Research on Past Terrestrial and Aquatic Biodiversity: Overview and Recommendations. en. *Quaternary* **4,** 6. ISSN: 2571-550X. https://www.mdpi.com/2571-550X/4/1/6 (2022) (Feb. 2021).

42. Dolenz, S. *et al.* Unravelling reference bias in ancient DNA datasets. en. *Bioinformatics* (ed Valencia, A.) btae436. ISSN: 1367-4811. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btae436/7705522 (2024) (July 2024).

43. Krehenwinkel, H. *et al.* Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. en. *Scientific Reports* **7,** 17668. ISSN: 2045-2322. https://www.nature.com/articles/s41598-017-17333-x (2023) (Dec. 2017).

44. Kearney, M. & Porter, W. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. en. *Ecology Letters* **12,** 334–350. ISSN: 1461023X, 14610248. https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2008.01277.x (2023) (Apr. 2009).

45. Overcast, I. *et al.* A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. en. *Molecular Ecology Resources* **21,** 2782–2800. ISSN: 1755-098X, 1755-0998. https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13514 (2022) (Nov. 2021).