Covariate-informed latent interaction models: Addressing geographic & taxonomic bias in predicting bird-plant interactions

Georgia Papadogeorgou[1], Carolina Bello[2], Otso Ovaskainen[3], David B. Dunson[4]

Abstract

Climate change and reductions in natural habitats necessitate that we better understand species' interactivity and how biological communities respond to environmental changes. However, ecological studies of species' interactions are limited by geographic and taxonomic bias which can lead to severe under-representation of certain species and distort our understanding of inter-species interactions. We illustrate that ignoring these biases can result in poor performance. We develop a model for predicting species' interactions that (a) accounts for errors in the recorded interaction networks, (b) addresses the geographic and taxonomic bias of existing studies, (c) is based on latent factors to increase flexibility and borrow information across species, (d) incorporates covariates in a flexible manner to inform the latent factors, and (e) uses a meta-analysis data set from 166 individual studies. We focus on interactions among 242 birds and 511 plants in the Brazilian Atlantic Forest, and identify 5% of pairs of species with an unrecorded interaction, but posterior probability of existing that is over 80%. Finally, we develop a permutation-based variable importance procedure and identify that a bird's body mass and a plant's fruit diameter are most important in driving the presence and detection of species interactions, with a multiplicative relationship.

keywords: Bayesian methods; bipartite graph; ecology; graph completion; latent factors; variable importance

1. Introduction

Animal-plant interactions have played a very important role in the generation of Earth's biodiversity [Ehrlich and Raven, 1964]. Dozens or even hundreds of species form complex networks of interdependences whose structure has important implications for the stability of ecosystems [Solé and Montoya, 2001]. However, climate change and the reduction in species' natural habitats necessitates that we urgently understand species' interactivity in order to better predict how biological communities will respond to environmental changes, and how these changes will affect species' interactions, equilibrium and co-existence.

Predicting and understanding species interactions is a long standing question in ecology. However, data on species' interactions are scarce and limited in their coverage. Some studies are

[1]Department of Statistics, University of Florida
[2]Swiss Federal Research Institute
[3]Department of Biological and Environmental Science, University of Jyväskylä, and Organismal and Evolutionary Biology Research Programme, University of Helsinki, and Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology
[4]Department of Statistical Science, Duke University

animal-centered and record interactions only for specific animal species. Similarly, there are studies that are plant-centered and record which animal species consume a given plant's fruit. Such studies are *taxonomically biased* in that they focus only on a subset of the species population. To learn general animal-plant interactions, the most useful studies are network studies that record *any* interaction that is observed. However, even these studies have severe limitations. They often focus on a small area where not all animal and plant species can be found, and are hence *geographically biased*. Even if the study area is well-defined, we do not have perfect knowledge of which species actually exist in the area, therefore we cannot know which interactions are even *possible to be observed* [Poisot et al., 2015]. This leads to complications in that interactions could be *un*recorded because (a) the species truly do not interact, (b) the species do not co-exist in the study area, (c) the species co-exist in the study area and truly interact but the interaction was not detected by the researchers, or (d) the interaction was detected but it was not recorded because it did not include the specific study's species of interest. Even though these biases and their implications are well-recognized in the ecological literature [Báldi and McCollin, 2003, Seddon et al., 2005, Pyšek et al., 2008, Trimble and van Aarde, 2012, Hale and Swearer, 2016, Jordano, 2016], most models for species interactions do not account for them [e.g. Bartomeus, 2013, Gravel et al., 2013]. Even though some advances are emerging in the literature [Cirtwill et al., 2019, Weinstein and Graham, 2017, Graham and Weinstein, 2018], the models therein do not provide a comprehensive treatment of species' traits and phylogenetic information. Our main focus is in addressing both sources of bias in order to understand whether *a given bird would eat the fruit of a given plant if given the opportunity*, and to learn which species traits are most important in forming and detecting these interactions.

From a statistical perspective, a bird-plant interaction network can be conceptualized as a *bipartite graph*, where the birds and plants form separate set of nodes, and a link connects one node from each set and represents that the bird would eat the plant's fruits if given the opportunity. If a certain animal-plant interaction has been recorded, the corresponding edge of the graph necessarily exists. However, absence of a recorded interaction does not mean that the interaction is not possible and the networks are measured *with error*.

Modeling the probability of connections on a graph measured without error has received a lot of attention in the statistics literature, and examples stretch across many applied fields such as social [Newman et al., 2002, Eckmann et al., 2004, Wu et al., 2010], biological [Han et al., 2004, Sporns et al., 2004, Chen and Yuan, 2006, Bullmore and Sporns, 2009], and ecological [Croft et al., 2004, Blonder and Dornhaus, 2011] networks. Since we do not hope to cover all literature in network modeling, we focus on statistical approaches for bipartite graphs. In an early approach, Skvoretz and Faust [1999] adapted the classic p^* network models to the bipartite setting. Within this string of literature, there have been a number of frequentist and Bayesian approaches performing community detection in bipartite graphs. Co-clustering was first introduced in Hartigan [1972]. Dhillon [2001] used the left and right eigenvalues of a scaled adjacency matrix to simultaneously cluster words and documents. Dhillon et al. [2003] developed an algorithm to cluster the nodes of a bipartite graph such that the optimal clustering maximizes mutual information between the clustered

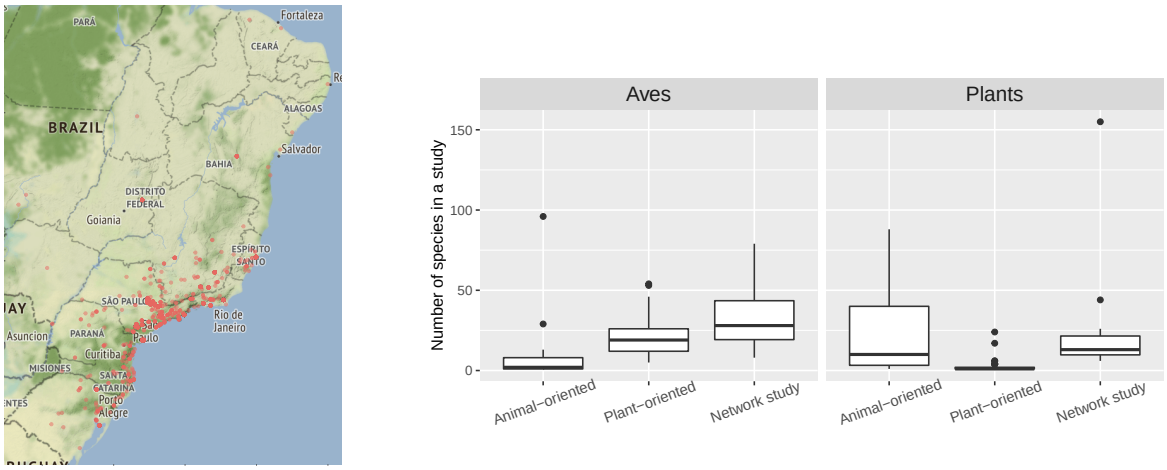
random variables. Relatedly, Banerjee et al. [2007] viewed community detection in bipartite graph as a matrix approximation problem, where the reconstruction minimizes information loss while preserving co-clustering statistics of the original graph. From a Bayesian perspective, Shan and Banerjee [2008] specified a generative model for co-clustering which allows for mixed membership of the rows and columns. Wang et al. [2011] performed simultaneous and nested clustering of the rows and columns of a bipartite graph, and Razaee et al. [2019] developed an approach to matching communities of the one set of nodes to those of the other. Co-clustering is widely used in a number of fields including genetics [Cheng and Church, 2000, Kluger et al., 2003, Madeira and Oliveira, 2004] and user-movie networks [Ungar and Foster, 1998, Hofmann and Puzicha, 1999, Yang et al., 2002].

Our approach is more closely related to network modeling using latent variables [Hoff et al., 2002, Handcock et al., 2007, Hoff, 2008] and its extension to bilinear and multilinear relationships [Hoff, 2005, 2011, 2015]. In latent variable approaches, the nodes are often embedded in a Euclidean space, and nodes' connections depend on their relative distance in this latent space, and potentially on covariates through linear terms. Since our observed networks have missing edges which we aim to learn, our approach also has ties to modeling noisy observed networks [e.g., De Choudhury et al., 2010, Jiang et al., 2011, Wang et al., 2012, Chatterjee, 2015, Priebe et al., 2015, Chang et al., 2020].

Our goal is two fold: (1) to complete the bipartite graph of bird-plant interactions given the recorded network which is measured with error, and (2) to understand which covariates are most important for driving and detecting interactions. To achieve this goal, we develop a Bayesian approach to modeling the probability of bird-plant interactions, based on a meta-analysis data set including recorded information on 166 published and unpublished studies [Bello et al., 2017] on the Brazilian Atlantic Forest. The proposed approach (a) models the probability of a link in the bipartite graph, (b) incorporates the missing data mechanism for unrecorded interactions, addressing the taxonomic and geographic bias of the individual studies, (c) uses species' trait information to inform the network model and improve precision, (d) employs a latent variable approach to link these three model components, which aids prediction of interactions of bird-plant pairs that did not co-occur, (e) quantifies our uncertainty around the estimated graph, and (f) uses posterior samples of estimated quantities and a permutation-based approach to acquire a variable importance metric for trait matching and species detectability. Even though our model employs latent factors, our approach is, to our knowledge, the first to employ latent network models for noisy networks, and to use covariates to inform the latent factors via separate models, instead of including them in the network model directly.

2. A multi-study data set of bird–plant interactions in the Atlantic Forest

We study bird-plant interactions in the Brazilian Atlantic Forest, whose area is decreasing fast, threatening its biodiversity [Ribeiro et al., 2009]. Our data include recorded bird-plant interactions from 166 published or unpublished studies. The original data set, which is thoroughly described in Bello et al. [2017], includes frugivore-plant interactions for five frugivore classes, including the



(a) Geographical bias

(b) Taxonomic bias

Figure 1: Geographical and Taxonomic Bias. Panel (a) shows the locations of recorded interactions with reported coordinate information. Panel (b) shows the number of unique ave and plant species with observed interactions within each study, by study type (animal/plant-oriented, network study).

class of Aves (birds). Since we focus on bird-plant interactions (hence excluding mammals or other classes), we incorporate the 85 studies that include at least one such interaction. Across the 85 included studies, there is a total of 6,024 recorded interactions among 242 bird and 458 plant species, and a total of 511 plant species. One of the key characteristics of our data is that unobserved interactions are not necessarily impossible. For an interaction to be registered in our data (1) both species must occur at the study site, (2) they must interact, and (3) the interaction has to be detected and recorded. Therefore, a bird-plant interaction might be unrecorded because studies have not focused on the species' territory, or do not record the given species' interactions.

Different studies took place across potentially overlapping or non-overlapping regions. Figure 1a shows the study locations of recorded bird-plant interactions with non-missing coordinate information, amounting to 68% of all recorded interactions. Most of them are located in a small geographic area along the coast and in the area near São Paulo in the southeast part of Brazil. Therefore, interactions among species that do not co-occur in this area would be less likely to be detected, implying that the data are geographically biased. On the other hand, out of the 85 studies, 19 were animal-oriented, 45 were plant-oriented, and 19 were network studies (the remaining 2 were a combination). Animal-oriented and plant-oriented studies record only a subset of the interactions that are detected: an animal-oriented study focuses on a *given* animal's diet whereas a plant-oriented focuses on learning which animals eat the fruits of a *given* plant. In contrast, the purpose of a network study is to record all detected interactions. The difference in studies' focus leads to taxonomic bias of recorded interactions which might over-represent the interactions of certain bird or plant species. Figure 1b illustrates the taxonomic bias in our data by showing the number of unique species observed in each study by study type. Animal-oriented studies have recorded interactions on a much smaller number of bird species than plant-oriented studies, and the reverse is true for the number of unique plant species.

In addition to the recorded bird-plant interactions, available data also include key bird and plant traits, measured with varying amounts of missingness. In ecology, it is believed that physical traits may influence the success of a frugivory interaction, and researchers are interested into understanding this relationship [Dehling et al., 2016, Bello et al., 2017, Descombes et al., 2019]. For plants, available data include the diameter, length and color of the plant's fruit and seed, its lipid score, and the plant's form (e.g., tree). For birds, covariates include its body mass, gape size, migration status, and frugivory score.

3. Learning species interactions addressing geographic and taxonomic bias

We use $i = 1, 2, \ldots, n_B$, and $j = 1, 2, \ldots, n_P$ to represent birds and plants respectively. For every bird i, let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip_B})'$ represent a collection of p_B physical traits, and similarly $\boldsymbol{W}_j = (W_{j1}, W_{j2}, \ldots, W_{jp_P})'$ for plant j. Recorded interactions are collapsed across studies into one $n_B \times n_P$ interaction matrix, denoted by $\boldsymbol{A}$, where $A_{ij} = 1$ implies that bird i has been recorded to interact with plant j, and $A_{ij} = 0$ otherwise. We are interested in inferring the true interaction matrix, denoted by $\boldsymbol{L}$, representing whether bird i would interact with plant j if given the opportunity ($L_{ij} = 1$), or not ($L_{ij} = 0$), and is also of dimension $n_B \times n_P$. In our study, these dimensions are 242 and 511, respectively.

We assume that there was no human error in recording interactions, and a recorded interaction was truly observed and therefore possible (hence if $A_{ij} = 1$, then $L_{ij} = 1$ necessarily). But how can we infer the probability that $L_{ij} = 1$ for pairs (i, j) for which an interaction has not been recorded? Available sources of information to estimate $\boldsymbol{L}$ are the following: (a) the recorded interactions allow us to draw inferences for other combinations of the same species, (b) species with similar traits might be involved in similar true interactions, and (c) the recorded interactions within each study provide us with information on which interactions the specific study *could* have recorded. The model presented below uses all the information in (a), (b) and (c) to infer the true interaction matrix.

3.1 The covariate-informed latent interaction model

For bird i and plant j let $\boldsymbol{U}_i = (U_{i1}, U_{i2}, \ldots, U_{iH})^T$ and $\boldsymbol{V}_j = (V_{j1}, V_{j2}, \ldots, V_{jH})^T$ denote their latent factors, respectively. The model presented below uses these factors for (a) the true interactions, (b) the species' physical traits, and (c) the probability of detecting a possible interaction. Figure 2 shows a graphical representation of our model.

First, the probability of a true interaction is modeled through the *interaction submodel*:

$$\text{logit} P(L_{ij} = 1) = \lambda_0 + \sum_{h=1}^{H} \lambda_h U_{ih} V_{jh}, \tag{1}$$

with $\lambda_h \in \mathbb{R}, h = 1, 2, \ldots, H$. In (1), the latent factors are used as in classic bipartite network models [e.g. Hoff, 2011], and they represent the species' locations in the latent space, where birds and plants in nearby locations are more likely to be connected. The interaction submodel is shown

through the arrows originating from $\boldsymbol{U}_i$ and $\boldsymbol{V}_j$ into L_{ij} in Figure 2. We link the species' traits to the probability of interaction by assuming that the latent factors $\boldsymbol{U}, \boldsymbol{V}$ are also the driving force of birds' and plants' physical traits. For appropriately chosen link functions f_m and g_l, we assume the *trait submodel*:

$$\begin{aligned} f_m^{-1}(E(X_{im} \mid \boldsymbol{U}_i)) &= \beta_{m0} + \boldsymbol{U}_i' \boldsymbol{\beta}_m, \ m = 1, 2, \ldots, p_B, \text{ and} \\ g_l^{-1}(E(W_{jl} \mid \boldsymbol{V}_j)) &= \gamma_{l0} + \boldsymbol{V}_j' \boldsymbol{\gamma}_l, \ l = 1, 2, \ldots, p_P \end{aligned} \tag{2}$$

for $\beta_{m0}, \gamma_{l0} \in \mathbb{R}$, and $\boldsymbol{\beta}_m, \boldsymbol{\gamma}_l \in \mathbb{R}^H$, shown in Figure 2 through the arrows originating from $\boldsymbol{U}_i$ to $\boldsymbol{X}_i$, and from $\boldsymbol{V}_j$ to $\boldsymbol{W}_j$. The submodel (2) implies that the latent factors can be conceived as low-dimensional summaries of the species' traits. We adopt logistic link functions for binary traits. If the trait submodel is not fully defined by its mean, additional parameters can be incorporated. For continuous traits, the link function is set to the identity function, and we incorporate a parameter for the residual variance.

The submodels (1)-(2) are interpretable from an ecological standpoint. Species' traits are believed to play an important role in whether they interact. The classic approach in network modeling includes covariates directly into the linear predictor of the interaction model in (1). However, the role of traits in an ecological network is believed to be interactive [referred to as trait matching; Fenster et al., 2015, Bender et al., 2018], and flexible approaches have performed better than including the traits directly into the model [Pichler et al., 2020]. Hence, the submodels (1)-(2) are in-line with current ecological knowledge as they represent a flexible representation of species' interactions that is driven by interactions among a low dimensional representation of the species' physical traits.

If the recorded interaction matrix $\boldsymbol{A}$ was not geographically or taxonomically biased, then the model (1)–(2) could be used to infer the true interaction matrix $\boldsymbol{L}$. However, since species' representation is biased, the model component connecting the truly possible interactions, $\boldsymbol{L}$, to the recorded interactions, $\boldsymbol{A}$, has to be incorporated. We model this component accounting for the fact that there is differential observational effort for different species. If bird i and plant j

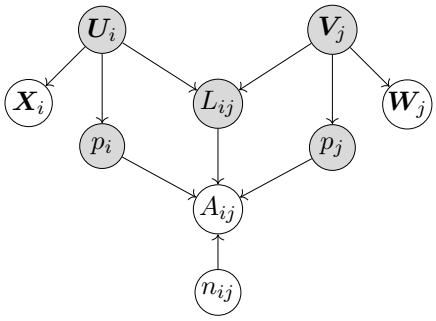


Figure 2: Graphical representation of the model. Shaded nodes represent latent variables. Bird and plant covariates are denoted by $\boldsymbol{X}_i, \boldsymbol{W}_j$, and corresponding latent factors by $\boldsymbol{U}_i, \boldsymbol{V}_j$, respectively. A recorded interaction is denoted by $A_{ij} = 1$ and possible interactions by $L_{ij} = 1$. The parameters p_i, p_j represent the probability that a species is detected, and n_{ij} denotes the number of studies that have recorded interactions of both species.

were followed and recorded across *many* studies, but no study has recorded an interaction between them ($A_{ij} = 0$), then it is more likely that the (i, j) interaction is not possible ($L_{ij} = 0$) than it being possible and not detected. On the other hand, we are more uncertain of whether the (i, j) interaction is possible if this interaction was not recorded and the species i and j have recorded interactions in a single study. To formalize this, let n_{ij} be the number of studies that have recorded at least one interaction for both i and j and consider species-specific detection probabilities which, with some abuse of notation, we denote as p_i for birds and p_j for plants. We specify the probability of observing an interaction (i, j) given whether it is possible as

$$P(A_{ij} = 1 \mid L_{ij} = l) = \begin{cases} 0, & \text{if } l = 0, \quad \text{and} \\ 1 - (1 - p_i p_j)^{n_{ij}}, & \text{if } l = 1, \end{cases} \tag{3}$$

which expresses that a study which followed both i and j (and contributes to n_{ij}) would detect a possible (i, j) interaction with probability $p_i p_j$, independently from other studies. This independence assumption has been previously employed within a related context [Weinstein and Graham, 2017]. Submodel (3) also expresses that it is impossible for a study to record a possible (i, j) interaction if the species were not detected in the study.

The missingness mechanism in (3) is key in accounting for geographic and taxonomic biases: If a species does not exist in the area under study or the species was not the focus of the study, then the study will not contribute to the count n_{ij}. Therefore, this specification treats both sources of bias simultaneously. Alternative specifications of this submodel could target the two parts of the missingness mechanism separately, by (a) using the information on the study goal (animal/plant-oriented, network study), and (b) knowledge on species' spatial distribution and co-occurrence. However, doing so would come with its own complications since available data are missing coordinate information on 32% of the recorded interactions, and knowledge of species spatial distributions during the different parts of the year is limited. This would require extensive modeling in itself, and would likely lead to increased uncertainty. We discuss model adjustments towards this direction in Section 6.

The detection of species is believed to depend on species traits such as size and behavior [Garrard et al., 2013, Troscianko et al., 2017]. In our study, a bird's body mass, whether they are solitary or gregarious, and a plant's height might affect whether their interactions are easily detected or not. For that reason, we specify the *detection submodel* to depend on the species' latent factors (which act as a low-dimensional summary of the covariates) as:

$$E[\text{logit}(p_i) \mid \boldsymbol{U}_i] = \delta_0 + \boldsymbol{U}_i^T \boldsymbol{\delta}, \quad \text{and} \quad E[\text{logit}(p_j) \mid \boldsymbol{V}_j] = \zeta_0 + \boldsymbol{V}_j^T \boldsymbol{\zeta}, \tag{4}$$

for $\delta_0, \zeta_0 \in \mathbb{R}$, and $\boldsymbol{\delta}, \boldsymbol{\zeta} \in \mathbb{R}^H$. We assume that $\text{logit}(p_i)$ and $\text{logit}(p_j)$ have conditional normal distributions with mean as specified in (4) and residual variance $\sigma^2_{p,B}$ and $\sigma^2_{p,P}$, respectively. As specified, (3)–(4) form a key model component for estimating the latent factors, since for recorded interactions it is known that $L_{ij} = 1$, which can drive estimation of the corresponding p_i, p_j ($L_{ij} \leftarrow$

$\boldsymbol{U}_i \to p_i$ in Figure 2, and similarly for the plants).

Even though the same latent factors are assumed to be the drivers of true interactions in (1), traits in (2), and the detection probability in (4), different latent factors might contribute to only a subset of the models if their corresponding coefficients are small. For example, some latent factors might inform only the probability of forming interactions, only the probability of detection, both or none. We discuss this further in Section 3.2.

3.2 Bayesian inference

The model described above is placed within the Bayesian paradigm which allows for immediate uncertainty quantification on the probability of truly possible interactions. We assume independent Gaussian prior distributions for the latent factors, but include a covariance structure across species (separately within birds, and within plants). Then, if $\boldsymbol{U}_{.h} = (U_{1h}, \ldots, U_{n_B h})^T$ and $\boldsymbol{V}_{.h} = (V_{1h}, \ldots, V_{n_P h})^T$, we specify

$$\boldsymbol{U}_{.h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_U), \quad \text{and} \quad \boldsymbol{V}_{.h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_V), \tag{5}$$

independently across h, $\boldsymbol{\Sigma}_U = \rho_U \boldsymbol{C}_U + (1-\rho_U)\boldsymbol{I}$, and $\boldsymbol{\Sigma}_V = \rho_V \boldsymbol{C}_V + (1-\rho_V)\boldsymbol{I}$, where $\boldsymbol{I}$ is the diagonal matrix, and $\boldsymbol{C}_U, \boldsymbol{C}_V$ are correlation matrices specified using subject-matter expertise (see Section 5.1 for their specification in our study which uses phylogenetic information). We specify $\rho_U, \rho_V \sim \text{Beta}(a_\rho, b_\rho)$.

Prior distributions need to be adopted for the remaining parameters which include the intercept, variance terms, and the coefficients of the latent factors in the various models: $\boldsymbol{\lambda}_{H\times 1} = (\lambda_1, \lambda_2, \ldots, \lambda_H)^T$ in (1), $\boldsymbol{B}_{H\times p_B} = (\boldsymbol{\beta}_1\ \boldsymbol{\beta}_2\ \ldots\ \boldsymbol{\beta}_{p_B})$ and $\boldsymbol{\Gamma}_{H\times p_P} = (\boldsymbol{\gamma}_1\ \boldsymbol{\gamma}_2\ \ldots\ \boldsymbol{\gamma}_{p_P})$ in (2), and $\boldsymbol{\delta}_{H\times 1}$ and $\boldsymbol{\zeta}_{H\times 1}$ in (4). Due to the complete model's high dimensionality, efficient estimation of model parameters requires either a *small* pre-specified value for H, or a moderate value of H with sufficient shrinkage of model parameters for increasing h. We follow the latter option by specifying an increasing shrinkage prior on model parameters such that the prior distribution assigns an increasing amount of weight to values close to zero as the index h increases. Specifically, we specify

$$\begin{aligned} \beta_{mh}|\tau^\beta_{mh}, \theta_h \sim N(0, \tau^\beta_{mh}\theta_h), \qquad \gamma_{lh}|\tau^\gamma_{lh}, \theta_h \sim N(0, \tau^\gamma_{lh}\theta_h) \\ \lambda_h|\tau^\lambda_h, \theta_h \sim N(0, \tau^\lambda_h\theta_h), \qquad \delta_h|\tau^\delta_h, \theta_h \sim N(0, \tau^\delta_h\theta_h), \qquad \zeta_h|\tau^\zeta_h, \theta_h \sim N(0, \tau^\zeta_h\theta_h) \end{aligned} \tag{6}$$

where $\tau^\beta_{mh}, \tau^\gamma_{lh}, \tau^\lambda_h, \tau^\delta_h, \tau^\zeta_h \sim IG(\nu/2, \nu/2)$, and

$$\begin{aligned} \theta_h \mid \pi_h \sim (1-\pi_h)P_0 + \pi_h \delta_{\theta_\infty}, \qquad \pi_h = \sum_{l=1}^{h} \omega_l, \qquad \omega_l = v_l \prod_{t=1}^{l-1}(1-v_t) \\ v_t \sim \text{Beta}(1, \alpha),\ t < H \qquad \text{and} \qquad v_H = 1. \end{aligned} \tag{7}$$

In (6), the prior variance of model coefficients is specified using parameter-specific variance terms τ and overall variance terms θ. Equation (7) specifies the truncated increasing shrinkage prior of Legramanti et al. [2019] for the overall variance terms. This prior distribution uses a stick-

breaking specification to define the mixing probabilities of a spike-and-slab prior distribution on θ_h, where P_0 is a slab distribution, and δ_{θ_∞} represents a point-mass at θ_∞. We set P_0 to be an inverse gamma distribution with parameters $(\alpha_\theta, \beta_\theta)$ and θ_∞ close to zero. For larger values of h, the prior on θ_h assigns larger weight to the point-mass rather than the slab distribution, resulting in prior distributions for the parameters in (6) that are a priori concentrated closer to zero for increasing values of h. At the same time, the parameter-specific variance terms (τ) are centered at 1 and provide flexibility to the h^{th} coefficient from each model to deviate from a $N(0, \theta_h)$ prior if, for example, θ_h is too small, and would lead to over-shrinkage of the corresponding coefficient. In that sense, the scaling parameters τ adjust the prior variance θ_h to allow for different latent variables h to be most important across the submodels (1), (2), and (4). In Supplement C.2 we see that including the parameter-specific variance terms τ only improves model performance. Inverse-gamma prior distributions are also assumed for the remaining variance components, including the residual variances of the trait models and the probability of detecting a given species.

3.3 Posterior computation

Since the posterior distribution of model parameters does not have a closed form, we approximate it using Markov Chain Monte Carlo (MCMC). Here we describe the algorithm at a high-level, but all details are included in Supplement A. All code for simulations and analyses will be made publicly available at the first author's Github page.

At each MCMC step, the entries of the true interaction matrix corresponding to recorded interactions are set to 1. Entries corresponding to interactions that were not recorded are set to 1 or 0 with weights resembling the current values of (1) while reflecting that an unrecorded interaction among species that co-existed in multiple studies is more likely to be impossible. The parameters of the interaction model in (1) are updated using the Pólya-Gamma data augmentation scheme of Polson et al. [2013] under which Pólya-Gamma random variables are drawn for all $n_B \times n_P$ pairs, conditional on which posterior distributions of logistic model parameters are normally distributed. All variance parameters are updated from inverse gamma distributions, and the parameters of continuous traits and the probability of detection have normal conditional posterior distributions. To update the parameters of the models for binary traits, we again employ that Pólya-Gamma data augmentation, draw values from Pólya-Gamma distributions for each unit and each binary covariate, and sample the parameters from their normal conditional posterior distributions. The latent factors from each set of species inform the probability of interaction, the species' covariates, and their probability of being detected, and are therefore included in a number of continuous and binary models. Despite their complicated form, conditional on all other parameters and the Pólya-Gamma draws for the interactions and the binary covariates, the latent factors have normal conditional posterior distributions. Updates for the parameters in the increasing shrinkage prior are adapted to our setting from Legramanti et al. [2019]. To update the probability of detecting a given species we perform Metropolis-Hastings steps where the proposal distribution is a Beta distribution centered at the current value. Despite the large number of parameters to be updated in this fashion $(n_B + n_P)$, we have found that these updates require minimal tuning. We update

the parameters ρ_U, ρ_V in a similar manner. Missing covariate values are common in our data set, but their imputation is based directly on (2).

3.4 Variable importance in latent interaction models

We propose a permutation-based approach to measure a covariate's importance in latent factor network models. In our study, this procedure will inform us of the relative importance of species traits for forming and detecting interactions. We briefly discuss the approach here, though further details are included in Supplement A.5. We are interested in studying the importance of the k^{th} bird trait. We use $\boldsymbol{X}_{.k}$ to denote the vector of the k^{th} covariate across all bird species. For each (i, j) pair of species, let $l_{ij}^{(r)}$ be the logit of the r^{th} posterior sample for the probability of interaction in (1), and $\boldsymbol{l}_{.j}^{(r)}$ be the vector of these probabilities across i, $(l_{1j}^{(r)}, l_{2j}^{(r)}, \ldots, l_{n_B j}^{(r)})^T$. For each posterior sample r and plant species j, we calculate the squared correlation between the predicted interaction probabilities $\boldsymbol{l}_{.j}^{(r)}$ and the covariate $\boldsymbol{X}_{.k}$. We average these values over all plant species j and posterior samples. For a large number of permutations B, we reorder the entries in $\boldsymbol{X}_{.k}$ and repeat this process. We use the number of standard deviations away from the mean of the permuted test statistics that the observed test statistic falls as a measure of variable importance. In all the steps, we only consider bird species with observed values of the covariate. The reason is that imputed covariate values are based on the latent factors which also drive the interaction model, and using them in a variable importance measure could lead to misleading conclusions on the covariate importance measure. We discuss this and other considerations in more detail in Supplement A.5. A similar approach is followed for the plant species $\boldsymbol{W}$.

3.5 Prediction for out-of-sample species

Our main focus is in predicting whether a species i^* with covariates $\boldsymbol{X}_{i^*}$ would interact with a species j^* with covariates $\boldsymbol{W}_{j^*}$, if given the opportunity, where the covariates might be available with missingness. If both species are included in the original $n_B \times n_P$ data, predictions are automatic in our MCMC scheme via updating the corresponding entry of $\boldsymbol{L}$. However, pairs of species might be partially or completely unobserved, if one or both species have no recorded interactions. Inference on the probability of interaction for pairs of two out-of-sample species could proceed by extending the observed interaction matrix to $(n_B + 1) \times (n_P + 1)$ with $n_{i^* j} = 0$ and $n_{ij^*} = 0$ for all i, j species in the original data set, and re-fitting the MCMC. However, such approach is computationally cumbersome since it would require re-fitting the MCMC for every new species. Here, we present a computationally efficient algorithm for making predictions for out-of-sample species, which combines the MCMC fit to the original data and importance sampling weighting. Mathematical and implementation details, along with predictions for pairs of species of which only one is out-of-sample, are shown in Supplement A.6.

For out-of-sample species i^*, j^*, let $\boldsymbol{\theta}^*$ denote all model parameters, $\widetilde{\boldsymbol{D}}$ observed data for all in-sample species, and $\boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*}$ denote the latent factors corresponding to i^*, j^* respectively. If

$L_{i^*j^*} \in \{0,1\}$ represents whether the species interact or not, we have that

$$P(L_{i^*j^*}=1 \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}) \propto \int \overbrace{P(L_{i^*j^*}=1 \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*})}^{(A)} \overbrace{p(\boldsymbol{X}_{i^*} \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*})\, p(\boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*, \boldsymbol{V}_{j^*})}^{(B)} \underbrace{p(\boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \mid \boldsymbol{\theta}^*)}_{(C)} \underbrace{p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})}_{(D)} \, \mathrm{d}(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*}),$$

which is the basis of our algorithm. First, we use draws from the posterior distribution of model parameters based on the original data (D) to draw latent factors for the new species i^*, j^* (C) where the correlation matrices $\boldsymbol{C}_U, \boldsymbol{C}_V$ are updated to include the new species. Based on the latent factors and the model parameters, we draw values for the indicator of whether the species interact (A). Since these draws do not account for the observed covariates, we denote the draw based on the r^{th} posterior sample by $\widetilde{L}^{(r)}_{i^*j^*}$ and up(down)-weigh the samples which use parameters and latent factors that have higher (lower) values of (B). Specifically, we let $w^{(r)}_{i^*} = w^{(r)}_{i^*1} w^{(r)}_{i^*2} \ldots, w^{(r)}_{i^*p_B}$, where $w^{(r)}_{i^*m}$ is 1 if X_{i^*m} is missing, and according to $p(X_{i^*m} \mid \boldsymbol{\theta}^{*(r)}, \boldsymbol{U}_{i^*})^{(r)}$ in (2) otherwise, and we define $w^{(r)}_{j^*}$ similarly. Then, we set the posterior probability that species i^*, j^* interact equal to $\sum_{r=1}^R w^{(r)}_{i^*j^*} \widetilde{L}^{(r)}_{i^*j^*} \Big/ \sum_{r=1}^R w^{(r)}_{i^*j^*}$, where $w^{(r)}_{i^*j^*} = w^{(r)}_{i^*} w^{(r)}_{j^*}$.

The algorithm presented above discusses how we can perform prediction for out-of-sample species, when new species i^*, j^* with covariate information $\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}$ become available. However, availability of covariate information of new species requires that we update our predictions for in-sample species since the posterior of the model parameters is now

$$p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}) \propto p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}}) \int \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^*_{i^*,j^*})}{p(\boldsymbol{\theta}^*)} p(\boldsymbol{\theta}^*_{i^*,j^*} \mid \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}) \, \mathrm{d}\boldsymbol{\theta}^*_{i^*,j^*} \tag{8}$$

for $\boldsymbol{\theta}^*_{i^*,j^*}$ denoting all model parameters for species i^*, j^* (this expression is derived in Supplement A.6). Intuitively, the covariates $\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}$ drive estimation of the latent factors for i^*, j^* (last term in (8)), and correlation of the latent factors across species would imply that the latent factors for i^*, j^* affect the latent factors for in-sample species and $p(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^*_{i^*,j^*}) \,/\, p(\boldsymbol{\theta}^*) \neq 1$. Therefore, the integral in (8) will not be equal to 1, and the posterior distribution that incorporates the new covariate values, $p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})$, will not be the same as the original posterior distribution, $p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})$.

We evaluate the performance of the computationally efficient algorithm for out-of-sample prediction and the impact that out-of-sample correlated species have in the interaction predictions of in-sample species in simulation studies.

4. Simulations

We perform simulations to evaluate the following critical aspects of our model: How does our model perform compared to 1) approaches that ignore the taxonomic and geographic biases? 2) an interaction model that depends directly on covariates? Further, we evaluated 3) how our predictive

performance varies with the amount of information that is available for a given species. Additional simulations including simulations on variable importance and trait matching are included in the Supplement, and discussed briefly below.

We considered alternative approaches that either use latent factors to link the various model components or use covariates directly, and either accommodate the probability of false negatives or not. Specifically, we considered: (Latent, observed) A model similar to the one in Section 3, but assuming that the observed interaction network is the truth, (Covariates, bias corrected) A model that allows for unobserved interactions to be possible, but for which the individual submodels depend directly on covariates instead of the latent factors, and (Covariates, observed) An interaction model that depends directly on covariates and assumes that the observed interaction network is the truth. We present the models and corresponding MCMC schemes in Supplement B, and we note here that the competing method Covariates, bias corrected is our own construction, and, to our knowledge, does not exist in the literature. We also note that the models that are based directly on the covariates do not incorporate phylogenetic information.

4.1 Data generation process imitating the observed data

Our simulations are based directly on our data on recorded bird-plant interactions in order to evaluate the models' relative performance within the context of our scientific question. Data generation followed closely several aspects of our observed data.

Sample sizes and observational effort. Across all simulations we used the observed number of bird and plant species. Further, the number of studies that observed a pair (i, j), n_{ij}, was the same as in our observed data. This ensured sufficient variability in the number of opportunities for a species to be observed in an interaction with any other. If n_{ij} is the number of studies that recorded interactions for both i and j, we define the *observational effort* for bird i as $n_i^+ = \sum_j n_{ij}$, and as $n_j^+ = \sum_i n_{ij}$ for plant j. Then, n_i^+ and n_j^+ have interquartile range 18–282 and 17–103, respectively. For each simulated data set, a random sample of 10 bird species were assumed to have $n_{ij} = 0$ for all j, and 10 plant species were assumed to have $n_{ij} = 0$ for all i to represent completely out-of sample species.

Observed interactions. We ensured that the proportion of recorded interactions in each simulated data set was approximately equal to the proportion of recorded interactions in our data, implying a quite sparse scenario (only $\approx 3.1\%$ of all pairs have a recorded interaction).

Covariates. We generated a large number of covariates $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{W}}$ from a matrix-normal distribution, with correlation across covariates equal to 0.3, and correlation across species equal to the species' phylogenetic correlation matrices $\boldsymbol{C}_B$ and $\boldsymbol{C}_P$ (discussed in Section 5.1). Some of the covariates were then transformed to binary variables using the values in $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{W}}$ as linear predictors in a Bernoulli distribution with a logistic link function. Only a subset of the generated covariates were available in the simulated data, and we consider scenarios where the important covariates are unobserved. To ensure that our generated data resembled the observed as closely as possible, we assumed that the observed covariates in each simulated scenario maintained the same structure and proportion

of missingness as in the observed data: 2 continuous and 3 binary covariates with proportion of missing values varying from $0 - 32\%$ for the first set of species, and 4 continuous and 8 binary covariates with proportion of missing values varying from $0 - 80\%$ for the second set of species. We denote the covariates that are observed in each simulated data set as $\boldsymbol{X}_i$ and $\boldsymbol{W}_j$, vectors of length five and twelve respectively, and assume that the continuous covariates are listed first, in that $X_{i1}, X_{i2}, W_{j1}, W_{j2}, W_{j3}, W_{j4}$ are continuous, and the rest are binary.

True data generative models. We generated data in three different ways:

(dgm1) All the important covariates are observed in the simulated data. Covariates X_{i1}, X_{i3}, X_{i4} ($W_{j1}, W_{j2}, W_{j5}, W_{j6}$) drive the detection of an interaction for the first (second) set of species, and the true interaction model is multiplicative in the observed covariates:

$$\text{logit}(P(L_{ij} = 1)) = \kappa_0 + \sum_{l=1}^{5} \kappa_l X_{il} W_{jl}.$$

(dgm2) Data generation in (dgm1) was based on observed covariates with a multiplicative interaction submodel. In (dgm2), the models for the detection probability are the same as in (dgm1), but the interaction model is *additive* in the observed covariates:

$$\text{logit}(P(L_{ij} = 1)) = \kappa_0 + \kappa_1 X_{i2} + \kappa_2 X_{i3} + \kappa_3 X_{i5} + \kappa_4 W_{j2} + \kappa_5 W_{j10}.$$

(dgm3) Data generation identical to (dgm1) but for which all submodels are functions of covariates that are *un*observed in the simulated data.

Importantly, in none of the above scenarios we generated data according to our model. However, the competitive model is the true data generative model in (dgm2). Therefore, simulations under (dgm2) inform us of our model's performance in this extreme scenario, and in the case where the linear predictor of our interaction submodel is misspecified. In (dgm1), we investigate the relative importance of using observed covariates instead of latent factors, in the case where all the important covariates are observed but the interaction model of the competitor model is misspecified. Then, under (dgm3), we investigate how the methods' relative performance changes when the observed covariates are correlated with the important covariates, but are themselves unobserved.

4.2 Simulation results

Methods were evaluated in terms of their predictive power in identifying true interactions and their misclassification rates. To compare the approaches in the presence of biases, predictive metrics were evaluated across the following categories: (a) Performance was evaluated separately for the observed interactions, the interactions that were not observed, and the interactions that were not observed but are truly possible. (b) Predictive metrics were also evaluated separately among pairs of species that have co-existed in studies, pairs that have not co-existed and hence it is impossible to observe their interactions, and species that were not observed in any of the studies. Doing so allows

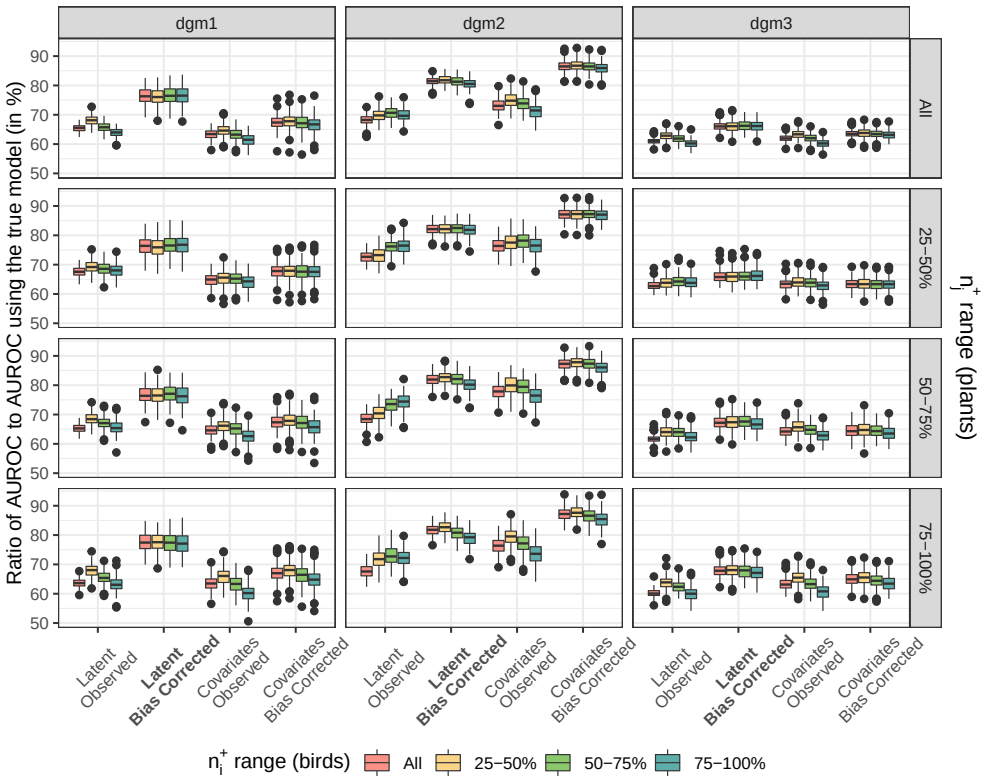


Figure 3: Methods' Predictive Performance in Simulations. Ratio of methods' AUROC to the AUROC using the true and known interaction model. Results are shown in percentages. Four methods are considered, two using latent factors and two using the observed covariates, with and without bias correction (horizontal axis). The method using latent factors with bias correction (in bold) is the proposed method. The columns represent the three data generative mechanisms: (dgm1) is a multiplicative interaction model using measured covariates, (dgm2) is an additive interaction model using measured covariates, and (dgm3) is a multiplicative model using unmeasured covariates. Results are shown by observational effort for bird and plant species, by color and row, respectively, showing the results for all species, and species at the 25-50, 50-75, and 75-100th percentile of observational effort measured using n_i^+, n_j^+.

us to evaluate model performance for in-sample prediction with different amounts of information, and out-of-sample prediction. (c) Lastly, we evaluated the methods' predictive performance based on the observational effort for each species, defined in Section 4.1.

Figure 3 shows the ratio of AUROC (area under the receiver operating characteristics curve) of each method compared to the AUROC from the true, known interaction model, when predicting the values of $\boldsymbol{L}$ among pairs with unrecorded interactions. The AUROC based on the true model (which includes unmeasured variables under (dgm3)) was approximately 83% for all three data generating processes. The results are shown by data generative mechanism, and separately by quartile of observational effort for both sets of species (based on the distribution of n_i^+, n_j^+).

First, we notice that bias correction always improves the methods' predictive performance. The performance of methods with*out* bias correction deteriorates for species with a higher observational

effort. Even though this might appear counter-intuitive at first, this is expected when considering that methods without bias correction do not accommodate that an unrecorded interaction is more likely to be impossible if there are more studies that could have observed it. This illustrates that predictions can severely suffer when the taxonomic and geographic biases are not properly accounted for. In all our simulations, we also observed that, on average, models without bias correction predicted a smaller number of possible interactions compared to their counterparts with bias correction, an observation that was also previously made by Weinstein and Graham [2017] and Graham and Weinstein [2018].

In the scenario where the competitor model is the true model (dgm2), the competitor achieves an AUROC ratio of 86.7% compared to the known model. Our method falls shortly behind with an AUROC ratio of 81.5%. This result indicates that even when the model uses all the important covariates correctly, our method performs only slightly worse for predicting missing unobserved interactions, even though it is based on a multiplicative interaction model and the true model is additive. In contrast, when the competing method uses the correct covariates but the wrong functional form for the interaction model (dgm1), the latent factor model outperforms it with AUROC ratios of 67% and 76%, respectively. Comparing the results of our method in (dgm1) and (dgm3), we see that our model effectively uses the observed covariates when those are predictive of interactions (dgm1), and it still outperforms a model based on the covariates when those are not the important ones (dgm3).

4.3 Results from additional simulation studies

We show additional simulation results in Supplement C. In Supplement C.1, we compared methods for half-in-sample and out-of-sample pairs and show that, even though the methods' predictive power is slightly lower there than compared to in-sample pairs, their relative performance remains unchanged. In Supplement C.2, we evaluated the performance of the computationally efficient out-of-sample prediction approach from Section 3.5, and we find that our importance sampling procedure performs comparably to fitting a single MCMC. We also find that model accuracy for in-sample predictions is essentially unchanged in the presence of out-of-sample correlated species, which indicates that interaction predictions for in-sample species do not have to be updated when new species become available. There, we also present simulation results under alternative specification of the prior distribution in (6), and show that including the parameter specific variance terms τ improve performance uniformly. Finally, in Supplement C.3 we show that using the procedure described in Section 3.4 leads to accurate conclusions on variable importance and trait matching, in that it accurately differentiates covariates that are important for forming or detecting interactions from those that are not. However, we notice that this procedure cannot differentiate among covariates that are important for the *presence* of interactions from those that are important for their *detection*.

5. Bird–plant interactions in the Atlantic Forest

5.1 Specifying the prior correlation matrices for the latent factors

In ecological studies, it is often assumed that the evolution of species over time follows a random walk model, and their organization in an ancestral tree leads to species that share a recent ancestor to be more similar than species that share a less recent ancestor. That means that species that are more genetically related are likely to have more similar traits and share more interactions. This ancestor and trait relationships are captured on phylogenetic trees which have been used to represent correlations across species [Ives and Helmus, 2011]. In some cases, phylogenetic information has agreed with observed correlations in species' traits or interaction profiles [Gilbert and Webb, 2007, Mariadassou et al., 2010], but not in others [Rezende et al., 2007].

A taxonomic tree is a simplification of a phylogenetic tree, in that it does not accommodate the time length of species progression, and it assumes that every branch is of the same length. We use the taxonomic tree to specify the matrix $\boldsymbol{C}$ in the latent factors' correlation structure in (5) as an approximation to the correlation matrix based on the phylogenetic tree [see Ovaskainen and Abrego, 2020, Section 6.7], which can be used if available. For the bird species, we specify

$$[C_U]_{ii'} = \begin{cases} 1, & \text{if } i = i', \\ 0.75, & \text{if they belong to the same genus (very similar),} \\ 0.5, & \text{if they belong to the same family (similar),} \\ 0.25 & \text{if they belong to the same order (somewhat similar), and} \\ 0 & \text{if they are unrelated,} \end{cases}$$

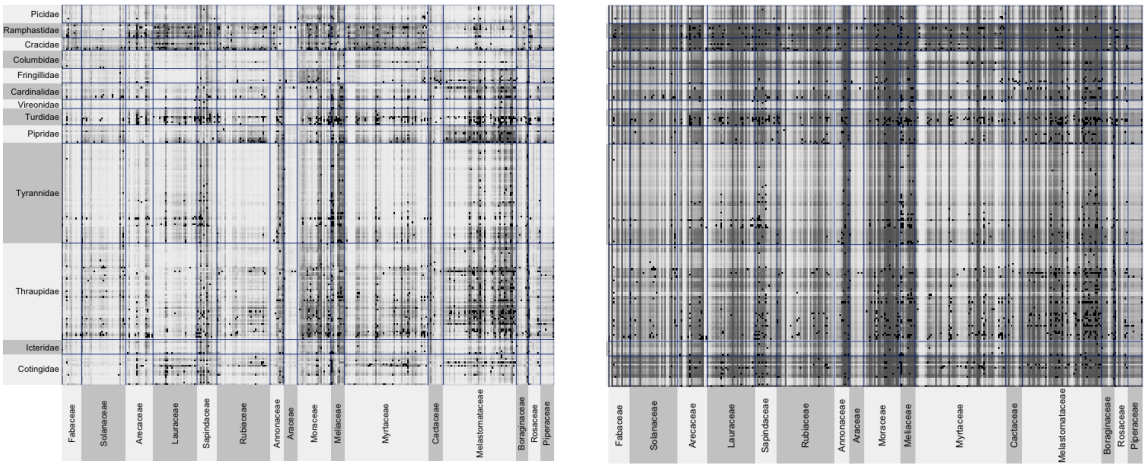
and similarly for $\boldsymbol{C}_V$, where plant species are organized in genera and genera in families. Therefore, $\boldsymbol{C}_U, \boldsymbol{C}_V$ are correlation matrices which are motivated by the species' evolutionary process. In the correlation structure of the latent factors, $\boldsymbol{\Sigma} = \rho\boldsymbol{C} + (1-\rho)\boldsymbol{I}$, the parameter ρ can be viewed as a quantification of the taxonomic importance in driving commonalities in species' traits and interaction profiles.

5.2 Estimating the graph of possible bird-plant interactions

In order to estimate $\boldsymbol{L}$ in our study, we considered the two approaches that correct for taxonomic and geographic bias discussed in Section 4: our approach, and a model for interactions which includes the covariates directly in the linear predictor. We run three chains of 80,000 iterations each, with a 40,000 burn in, and kept every 40^{th} iteration. MCMC convergence was investigated for both by studying traceplots and running means for identifiable parameters. Convergence diagnostics for our approach are shown in Appendix F. Based on similar diagnostics, we found that the MCMC of the alternative approach failed to converge based on the same number of iterations. For that reason, we excluded from this analysis the two traits of the plant species with the largest amounts of missingness (seed length and whether the species is threatened for extinction), which led to improved diagnostic plots and no detectable lack of convergence.

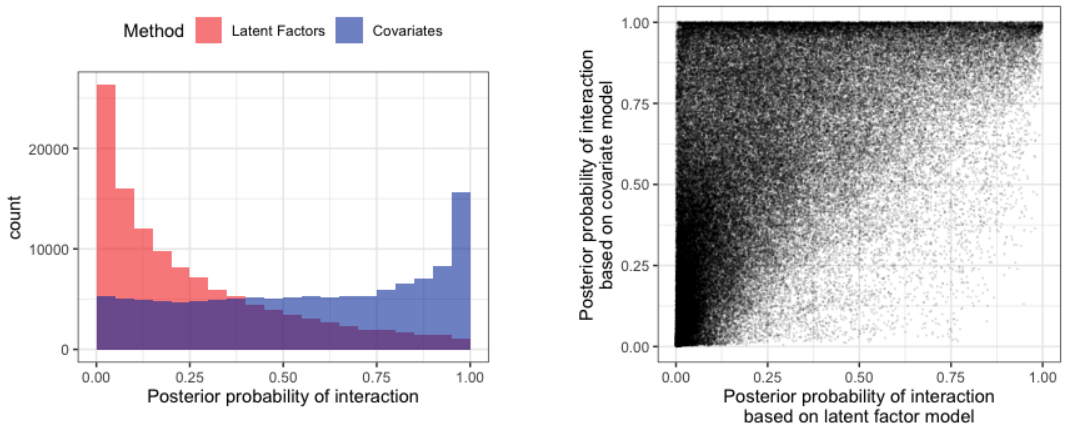
In Figure 4, we show estimates for the probability of interaction based on both models for a subset of the taxonomic families (bird/plant families with at least 5/8 species), illustrated using the vertical and horizontal blue lines. The complete results are shown in Supplement D but the conclusions remain unchanged, and the complete list of species in the order shown is included in Supplement G. The results from the two methods have important differences. According to our method (Figure 4a), species within the same family form similar interactions, as evidenced by the taxonomically-structured posterior interaction probabilities, where the blue lines separate the posterior interaction probabilities in clusters with similar values. In contrast, results from the alternative method that employs covariates directly (Figure 4b) indicate that some species interact with most other species and some species with none, as evidenced by *rows and columns* that are mostly close to one or zero. Since we do not expect this "all or none" structure in species interactions, results from the covariate model seem untrustworthy, and indicate that this model might rely on covariates too heavily. Species within the same family can belong to different genera, but we refrain from including genera information in the figure to ease visualization. However, we observed that clusters of posterior interaction probabilities from the latent factor model within a pair of species families generally corresponded to species organization by genera. The taxonomic structure in the results from our method is further supported by the posterior means (95% credible intervals) for ρ_U and ρ_V which were 0.965 $(0.929, 0.988)$ and 0.981 $(0.961, 0.994)$, respectively.

We further compare interaction results from the two models in Figure 5. There are a few important conclusions. In Figure 5a, we see that the model that employs covariates directly predicts that a large proportion of the pairs truly interact. However, it is agnostic as to whether a large



(a) Latent, Bias corrected (b) Covariates, Bias corrected

Figure 4: Posterior Probability of Possible Interactions. Posterior probability that bird species (y-axis) and plant species (x-axis) interact according to (a) the proposed method and (b) the alternative method. Species are organized in taxonomic families separated by blue vertical and horizontal lines. Only taxonomic families with at least 5 bird and 8 plant species are shown to ease visualization. Black color is used to represent interactions that are recorded in our data.



(a) (b)

Figure 5: Comparison of Model Predictions. Values close to 1 indicate that an interaction between the two species is likely to be possible. (a) Histograms of posterior interaction probabilities for all pairs of species without a recorded interaction based on the two models. (b) The predictions from the two models are plotted against each other.

portion of pairs are possible to interact or not, as evidenced by a distribution of interaction probabilities that is uniform below 0.75. This illustrates that the model that uses covariates directly cannot identify truly impossible interactions (this conclusion is further supported in Section 5.3). In contrast, our model that uses latent factors predicts that a non-negligible, but more realistic, proportion of pairs are possible to interact, and identifies a large number of pairs which are likely *impossible* to interact. When studying these predictions against each other in Figure 5b, we see that pairs that are predicted to almost surely interact based on the covariate model (values close to 1 on the y-axis) have interaction predictions based on our model that range uniformly from 0 to 1. In contrast, pairs of species that are impossible to interact based on our model (values close to 0 on the x-axis) are likely to have low probability of interactions according to the covariates model also, but these values range from 0 to 0.5. Therefore, the models seem to partially agree on which pairs of species are impossible to interact, but our model is more confident in these predictions, with posterior probabilities of interaction that are closer to zero.

5.3 Model performance in identifying truly possible interactions

We perform a variant of cross-validation to assess how well the two models fit the observed data. Since our goal is to predict which of the unrecorded interactions are truly possible, our cross-validation approach holds out a subset of the recorded interactions and studies model predictions for these pairs of species. Specifically, we randomly choose 100 recorded interactions, we set their corresponding values in the observed interaction matrix, $\boldsymbol{A}$, equal to 0, without changing their corresponding n_{ij} value, and we predict their probability of interaction. We repeat this procedure 20 times, each time holding out a different subset of recorded interactions. Then, model comparison is based on how well each model can differentiate interactions that we know are possible (the

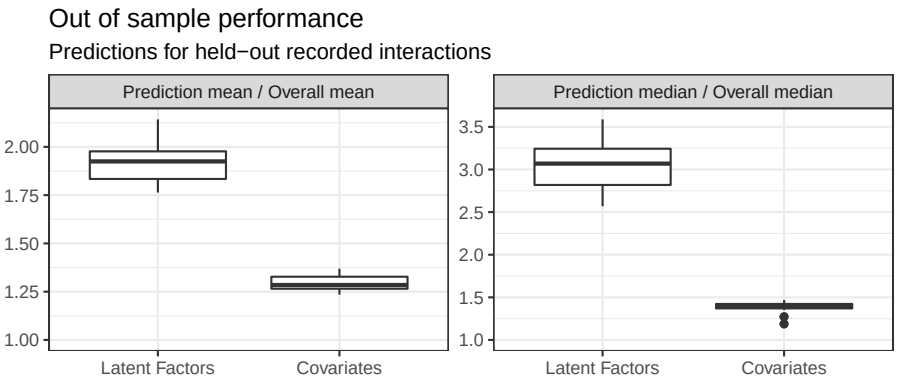


Figure 6: Cross Validation Results. For each of the 20 cross-validation iterations, we calculate the mean and median of the posterior interaction probability in the held-out, truly interacting pairs, and in the overall species population. The two panels show boxplots for the ratio of held-out to overall prediction mean (left) and median (right) by model. Higher values indicate that the true interactions were identified more clearly.

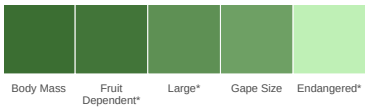
held-out, truly recorded interactions) from the predictions of interactions across all species which necessarily includes pairs that are impossible to interact. We note here that our setting forbids us from comparing model performance based on *un*recorded interactions, since those interactions are not certainly impossible.

Figure 6 shows the results. Since the two models tend to return drastically different prevalence of interactions (Figure 5a), we present results in terms of relative magnitude of average and median posterior probability of interaction in the held-out and in the overall data, separately for the two models. Figure 6 shows that the latent factor model is much more effective in differentiating the truly possible interactions from the set of all interactions compared to the model that is based directly on covariates.

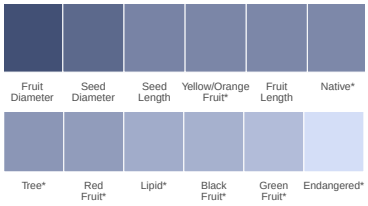
5.4 Identifying important traits for species detectability and interactions

Apart from understanding which pairs of species are possible to interact, ecologists are also interested in understanding the traits which make species interactions possible, referred to as trait matching [e.g. Fenster et al., 2015], as well as the traits that affect species' detectability [e.g. Garrard et al., 2013, Troscianko et al., 2017]. Towards that goal, Bastazini et al. [2017] studied how traits and phylogenetic information drive species interactions, and accounted for unrecorded interactions due to lack of overlap in species distributions. Pichler et al. [2020] showed that flexible models perform better than generalized linear models in both predicting interactions and identifying the important trait for these interactions.

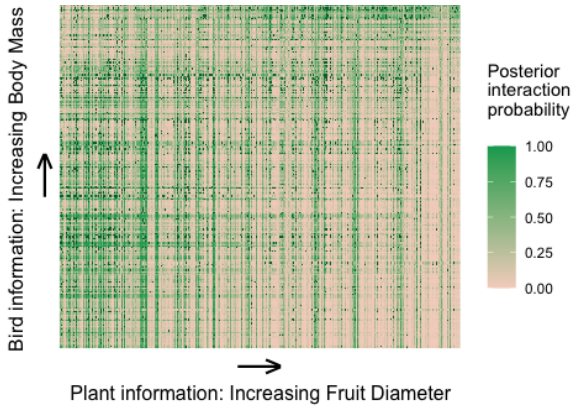
In Figure 7 we show the results of trait matching in our study. Figures 7(a-b) show the values of the variable importance measure described in Section 3.4 for bird and plant traits separately. Darker colors are used to indicate higher importance. We find that whether the species are endangered plays a minimal role in whether an interaction is possible. We identify a bird's body mass and a plant's fruit diameter as the most important traits in forming and detecting interactions among



(a) Bird Traits Importance



(b) Plant Traits Importance



(c) Interaction matrix ordered by traits

Figure 7: Trait Importance. Figures (a) and (b) show the variable importance metric of Section 3.4 for bird and plant species. Traits are ordered from most important (dark color) to least important (light color), and * is used for binary traits. Figure (c) shows the matrix of posterior probabilities of interaction where the species have been re-ordered in increasing order of body mass for birds and fruit diameter for plants, the most important traits. Orange and green are used for low and high probability of interaction, respectively, and dark green is used for recorded interactions.

species. In Figure 7(c) we plot the posterior probabilities of interaction reordering the species in increasing values of body mass for birds and fruit diameter for plants. We see that high posterior probabilities are concentrated on the upper left triangle, whereas low posterior probabilities are concentrated on the bottom right triangle. Therefore, our model estimates show that small birds are less likely to interact with plants that produce large fruits, and large fruits are most often consumed by large birds. Interestingly, our model returns estimates that indicate such interactive relationships between traits which are in line with the current ecological literature [Fenster et al., 2015, Bender et al., 2018] without having to specify this multiplicative trend parametrically.

6. Discussion

We introduced a latent factor model that uses species traits and recorded interactions in order to complete the bipartite graph of species interactivity accounting for taxonomic and geographic bias from different studies, and we proposed an approach to study variable importance in such latent network models. We found that a number of unrecorded interactions are truly possible, and identified important physical traits in forming and detecting interactions among species that are in line with current knowledge in ecology. We see a number of possible extensions. One extension could accommodate the spatial aspect of the data more directly, and treat the geographic and taxonomic biases separately. One way forward is to alter the submodel in (3) to reflect

$$P(A_{ij} = 1 \mid L_{ij} = 1) = 1 - \prod_s (1 - O_{ijs} F_{ijs} p_i p_j),$$

where F_{ijs} is equal to 1 if the focus of study s allowed for observation of both i, j species and 0 otherwise, and $O_{ijs} \sim \text{Bern}(\psi_{ijs})$ represents whether species i, j co-occur at the location of study s. A related extension could also allow for temporal variation focusing on learning whether species i and j would interact if they co-existed at location s and time t. Extensions in this direction would open the road for investigating the importance of species co-occurrence and competition in forming interactions, a topic that would require explicit geographic modeling of species. However, even if all studies provided detailed temporal and spatial information, scientists do not have perfect knowledge of species' spatial distributions, hence we could not accurately incorporate which species co-exist (and specify ψ_{ijs} above). One approach could treat ψ_{ijs} as known using published geographic maps of species distributions. Alternatively, it could be estimated incorporating geographic information such as the state, municipality or bioregion of the Atlantic forest. However, studying the co-existence of species across space is a hard problem in itself and it is the topic of joint species modeling in ecology [Ovaskainen and Abrego, 2020].

One of the key aspects of our approach is that we assume that recorded interactions are necessarily possible. Falsely recorded interactions could be accommodated by altering the model component in (3) to reflect

$$P(A_{ij} = 1 \mid L_{ij} = 0) = 1 - \prod_s (1 - p_s^{\text{error}}),$$

where p_s^{error} is a study-specific probability of mis-recording an interaction. Even though we consider this an interesting extension, we suspect that accommodating false positives could drastically affect model efficiency, especially in our very sparse scenario with only ∼3.1% of all possible pairs having a recorded interaction. An alternative approach to investigating the presence of false positives could assess recorded interactions post-hoc by examining cases where the posterior probability of interaction is low even though the interaction is recorded.

In our work, we found that using covariates to inform the latent factors performs better in (a) predicting truly impossible interactions between species, and (b) separating interactions that are possible from the rest, compared to an approach that uses the covariates directly. Even though using the covariates in the proposed manner complicates the investigation of variable importance since we cannot simply test a coefficient's statistical significance, we proposed a measure for variable importance which performed well in simulations and returned results that agree with subject-matter knowledge. We find variable importance and the identification of important covariates in latent factor models to be an interesting line of future work.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506; ERC-synergy project LIFEPLAN). Otso Ovaskainen was funded by Academy of Finland (grant no. 309581), Jane and Aatos Erkko Foundation, Research Council of Norway through its Centres of

Excellence Funding Scheme (223257). Carolina Bello acknowledges funding support from the European Research Council (ERC) under the European union's Horizon 2020 research and innovation programme (grant agreement No 787638) and the Swiss National Science Foundation (grant No. 173342), both granted to Catherine Graham.

References

András Báldi and Duncan McCollin. Island ecology and contingent theory: The role of spatial scale and taxonomic bias. *Global Ecology and Biogeography*, 12(1):1–3, jan 2003. ISSN 1466822X. doi: 10.1046/j.1466-822X.2003.00323.x.

Arindam Banerjee, Joydeep Ghosh, Srujana Merugu, and Dharmendra S Modha. A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation. Technical report, 2007.

Ignasi Bartomeus. Understanding Linkage Rules in Plant-Pollinator Networks by Using Hierarchical Models That Incorporate Pollinator Detectability and Plant Traits. *PLoS ONE*, 8(7):69200, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0069200.

Vinicius A.G. Bastazini, Pedro M.A. Ferreira, Bethânia O. Azambuja, Grasiela Casas, Vanderlei J. Debastiani, Paulo R. Guimarães, and Valério D. Pillar. Untangling the Tangled Bank: A Novel Method for Partitioning the Effects of Phylogenies and Traits on Ecological Networks. *Evolutionary Biology*, 44(3):312–324, 2017. ISSN 00713260. doi: 10.1007/s11692-017-9409-8.

Carolina Bello, Mauro Galetti, Denise Montan, Marco A Pizo, Tatiane C Mariguela, Laurence Culot, Felipe Bufalo, Fabio Labecca, Felipe Pedrosa, Rafaela Constantini, Wesley R Silva, Fernanda R da Silva, Otso Ovaskainen, and Pedro Jordano. Atlantic-frugivory: A plant-frugivore interaction dataset for the Atlantic Forest. *Ecology*, 98(1729), 2017.

Irene M.A. Bender, W. Daniel Kissling, Pedro G. Blendinger, Katrin Böhning-Gaese, Isabell Hensen, Ingolf Kühn, Marcia C. Muñoz, Eike Lena Neuschulz, Larissa Nowak, Marta Quitián, Francisco Saavedra, Vinicio Santillán, Till Töpfer, Thorsten Wiegand, D. Matthias Dehling, and Matthias Schleuning. Morphological trait matching shapes plant–frugivore networks across the Andes. *Ecography*, 41(11):1910–1919, nov 2018. ISSN 16000587. doi: 10.1111/ecog.03396.

Benjamin Blonder and Anna Dornhaus. Time-Ordered Networks Reveal Limitations to Information Flow in Ant Colonies. *PloS one*, 6(5), 2011. doi: 10.1371/journal.pone.0020298.

Ed Bullmore and Olaf Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems, mar 2009. ISSN 1471003X.

Jinyuan Chang, Eric D Kolaczyk, and Qiwei Yao. Estimation of Subgraph Densities in Noisy Networks. Technical report, 2020.

Sourav Chatterjee. Matrix estimation by Universal Singular Value Thresholding. *Annals of Statistics*, 43(1):177–214, 2015. ISSN 00905364. doi: 10.1214/14-AOS1272.

Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006. doi: 10.1093/bioinformatics/btl370.

Y Cheng and G M Church. Biclustering of expression data. In *International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 93–103, 2000. ISBN 1553-0833 (Print)\r1553-0833 (Linking). doi: 10.1007/11564126.

Alyssa R. Cirtwill, Anna Eklöf, Tomas Roslin, Kate Wootton, and Dominique Gravel. A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, 10(6):902–911, 2019. ISSN 2041210X. doi: 10.1111/2041-210X.13180.

Darren P Croft, Jens Krause, and Richard James. Social networks in the guppy (Poecilia reticulata). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271, 2004. doi: 10.1098/rsbl.2004.0206.

Munmun De Choudhury, Winter A Mason, Jake M Hofman, and Duncan J Watts. Inferring relevant social networks from interpersonal communication. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 301–310, 2010. ISBN 9781605587998. doi: 10.1145/1772690.1772722.

D Matthias Dehling, Pedro Jordano, H Martin Schaefer, Katrin Böhning-Gaese, Matthias Schleuning, and Americo Vespucio. Morphology predicts species' functional roles and their degree of specialization in plant-frugivore interactions. *Proceedings of the Royal Society B: Biological Sciences*, 283(20152444), 2016. doi: 10.1098/rspb.2015.2444.

Patrice Descombes, Alan Kergunteuil, Gaëtan Glauser, Sergio Rasmann, and Loïc Pellissier. Plant physical and chemical traits associated with herbivory in situ and under a warming treatment. *Journal of Ecology*, 108(2):733–749, 2019. ISSN 13652745. doi: 10.1111/1365-2745.13286.

Inderjit S. Dhillon. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, San Francisco, California, 2001. ACM Press. ISBN 158113391X. doi: 10.1145/502512.502550.

Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-Theoretic Coclustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98, 2003.

Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences*, 101(40):14333–14337, 2004.

Paul R. Ehrlich and Peter H. Raven. BUTTERFLIES AND PLANTS: A STUDY IN COEVOLUTION. *Evolution*, 18(4):586–608, dec 1964.

Charles B. Fenster, Richard J. Reynolds, Christopher W. Williams, Robert Makowsky, and Michele R. Dudash. Quantifying hummingbird preference for floral trait combinations: The role of selection on trait interactions in the evolution of pollination syndromes. *Evolution*, 69(5): 1113–1127, 2015. ISSN 15585646. doi: 10.1111/evo.12639.

Georgia E. Garrard, Michael A. Mccarthy, Nicholas S.G. Williams, Sarah A. Bekessy, and Brendan A. Wintle. A general model of detectability using species traits. *Methods in Ecology and Evolution*, 4(1):45–52, 2013. ISSN 2041210X. doi: 10.1111/j.2041-210x.2012.00257.x.

Gregory S Gilbert and Campbell O Webb. Phylogenetic signal in plant pathogen-host range. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):4979–4983, 2007. ISSN 00278424. doi: 10.1073/pnas.0607968104.

Catherine H. Graham and Ben G. Weinstein. Towards a predictive model of species interaction beta diversity. *Ecology Letters*, 21(9):1299–1310, sep 2018. ISSN 14610248. doi: 10.1111/ele.13084.

Dominique Gravel, Timothée Poisot, Camille Albouy, Laure Velez, and David Mouillot. Inferring food web structure from predator-prey body size relationships. *Methods in Ecology and Evolution*, 4(11):1083–1090, nov 2013. ISSN 2041210X. doi: 10.1111/2041-210X.12103.

Robin Hale and Stephen E Swearer. Ecological traps: Current evidence and future directions, 2016. ISSN 14712954.

Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, jul 2004. ISSN 0028-0836. doi: 10.1038/nature02555.

Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 170(2):301–354, mar 2007. ISSN 09641998. doi: 10.1111/j.1467-985X.2007.00471.x.

J A Hartigan. Direct Clustering of a Data Matrix. Technical Report 337, 1972.

Peter D Hoff. Bilinear Mixed Effects Models for Dyadic Data. *Journal of the American Statistical Association*, 100(469):286–295, 2005.

Peter D Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 657—-664. Curran Associates, Inc., 2008. ISBN 0378-8733. doi: 10.1016/j.socnet.2009.02.004.

Peter D Hoff. Hierarchical multilinear models for multiway data. *Computational Statistics and Data Analysis*, 55(1):530–543, 2011. doi: 10.1016/j.csda.2010.05.020.

Peter D Hoff. Multilinear Tensor Regression for Longitudinal Relational Data. *The Annals of Applied Statistics*, 9(3):1169–1193, 2015. doi: 10.1214/15-AOAS839.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, dec 2002. ISSN 0162-1459. doi: 10.1198/016214502388618906.

Thomas Hofmann and Jan Puzicha. Latent Class Models for Collaborative Filtering to appear in Proceedings of IJCAI'99. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 688–693, 1999.

Anthony R. Ives and Matthew R. Helmus. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, 81(3):511–525, 2011. ISSN 00129615. doi: 10.1890/10-1264.1.

Xiaoyu Jiang, David Gold, and Eric D. Kolaczyk. Network-based Auto-probit Modeling for Protein Function Prediction. *Biometrics*, 67(3):958–966, 2011. ISSN 0006341X. doi: 10.1111/j.1541-0420.2010.01519.x.

Pedro Jordano. Sampling networks of ecological interactions, dec 2016. ISSN 13652435.

Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research*, 13:703–716, 2003.

Sirio Legramanti, Daniele Durante, and David B. Dunson. Bayesian cumulative shrinkage for infinite factorizations. 2019.

Sara C Madeira and Arlindo L Oliveira. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45, 2004.

Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, 4(2):715–742, 2010. ISSN 19326157. doi: 10.1214/10-AOAS361.

M E J Newman, D J Watts, and S H Strogatz. Random graph models of social networks. *Proceedings of the national academy of sciences*, 99(suppl 1):2566—-2572, 2002.

Otso Ovaskainen and Nerea Abrego. *Joint Species Distribution Modelling With Applications in R*. 2020. ISBN 9781108492461.

Maximilian Pichler, Virginie Boreux, Alexandra Maria Klein, Matthias Schleuning, and Florian Hartig. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, 11(2):281–293, 2020. ISSN 2041210X. doi: 10.1111/2041-210X.13329.

Timothée Poisot, Daniel B. Stouffer, and Dominique Gravel. Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, 124(3):243–251, mar 2015. ISSN 16000706. doi: 10.1111/oik.01719.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013. ISSN 0162-1459. doi: 10.1080/01621459.2013.829001.

Carey E. Priebe, Daniel L. Sussman, Minh Tang, and Joshua T. Vogelstein. Statistical Inference on Errorfully Observed Graphs. *Journal of Computational and Graphical Statistics*, 24(4):930–953, 2015. ISSN 15372715. doi: 10.1080/10618600.2014.951049.

Petr Pyšek, David M Richardson, Jan Pergl, Vojtěch Jarošík, Zuzana Sixtová, and Ewald Weber. Geographical and taxonomic biases in invasion ecology, 2008. ISSN 01695347.

Zahra S Razaee, Arash A Amini, and Jingyi Jessica Li. Matched Bipartite Block Model with Covariates. Technical report, 2019.

Enrico L Rezende, Jessica E Lavabre, Paulo R. Guimarães, Pedro Jordano, and Jordi Bascompte. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature*, 448 (7156):925–928, 2007. ISSN 14764687. doi: 10.1038/nature05956.

Milton Cezar Ribeiro, Jean Paul Metzger, Alexandre Camargo Martensen, Flávio Jorge Ponzoni, and Márcia Makiko Hirota. The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. *Biological Conservation*, 142(6): 1141–1153, 2009. ISSN 00063207. doi: 10.1016/j.biocon.2009.02.021. URL http://dx.doi.org/10.1016/j.biocon.2009.02.021.

Philip J. Seddon, Pritpal S. Soorae, and Frédéric Launay. Taxonomic bias in reintroduction projects. *Animal Conservation*, 8(1):51–58, feb 2005. ISSN 13679430. doi: 10.1017/S1367943004001799.

Hanhuai Shan and Arindam Banerjee. Bayesian Co-clustering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 530–539, 2008.

John Skvoretz and Katherine Faust. Logit Models for Affiliation Networks. *Sociological Methodology*, 29(1):253–280, aug 1999. doi: 10.1111/0081-1750.00066.

Ricard V. Solé and J. M. Montoya. Complexity and fragility in ecological networks. *Proceedings of the Royal Society B: Biological Sciences*, 268(1480):2039–2045, oct 2001. ISSN 14712970. doi: 10.1098/rspb.2001.1767. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2001.1767.

Olaf Sporns, Dante R Chialvo, Marcus Kaiser, and Claus C Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8:418–425, 2004. doi: 10.1016/j.tics.2004.07.008.

Morgan J. Trimble and Rudi J. van Aarde. Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. *Ecosphere*, 3(12):art119, dec 2012. ISSN 2150-8925. doi: 10.1890/es12-00299.1.

Jolyon Troscianko, John Skelhorn, and Martin Stevens. Quantifying camouflage: how to predict detectability from appearance. *BMC Evolutionary Biology*, 17(1):1–13, 2017. ISSN 14712148. doi: 10.1186/s12862-016-0854-2.

Lyle H Ungar and Dean P Foster. A Formal Statistical Approach to Collaborative Filtering. *CONALD'98*, 1998.

Dan J Wang, Xiaolin Shi, Daniel A. McFarland, and Jure Leskovec. Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409, 2012. ISSN 03788733. doi: 10.1016/j.socnet.2012.01.003.

Pu Wang, Kathryn B Laskey, Carlotta Domeniconi, and Michael I Jordan. Nonparametric Bayesian Co-clustering Ensembles. In *Proceedings of the 2011 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, pages 331–342, 2011.

Ben G. Weinstein and Catherine H. Graham. On comparing traits and abundance for predicting species interactions with imperfect detection. *Food Webs*, 11(May):17–25, 2017. ISSN 23522496. doi: 10.1016/j.fooweb.2017.05.002.

Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences*, 107(44):18803–18808, 2010. doi: 10.1073/pnas.1013140107/-/DCSupplemental.

Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. Capturing Subspace Correlation in a Large Data Set. In *Proceedings 18th international conference on data engineering*, pages 517–528, 2002.

Supplementary materials for

Covariate-informed latent interaction models: Addressing geographic & taxonomic bias in predicting bird-plant interactions

by

Georgia Papadogeorgou, Carolina Bello, Otso Ovaskainen, David B. Dunson

Table of Contents

Supplement A. MCMC scheme

A.1 Some notation

1. **Outer product:** We use $\boldsymbol{A} \otimes \boldsymbol{B}$ to denote the outer product of vector $\boldsymbol{A}$ of length l_A and vector $\boldsymbol{B}$ of length l_B, where $\boldsymbol{A} \otimes \boldsymbol{B}$ is a matrix of dimension $l_A \times l_B$ with (i_1, i_2) entry equal to $A_{i_1} B_{i_2}$.

2. **Vectorization:** For a matrix $\boldsymbol{M}$ of dimension $r \times c$, denote the vectorization of $\boldsymbol{M}$ as $\text{vec}(\boldsymbol{M})$ where $\text{vec}(\boldsymbol{M})$ is a vector of length rc with entries

$$M_{11}, M_{12}, \ldots, M_{1c}, M_{21}, \ldots, M_{2c}, \ldots, M_{rc},$$

   hence unpacking first across the columns and then across the rows.

3. **Conditional distributions:** We use $p(x_1 \mid x_2, x_3)$ to denote the distribution of x_1 given x_2 and x_3, $p(x \mid \cdot)$ to denote the distribution of x given everything else, and $p(x \mid \cdot, -y)$ to denote the distribution of x given everything except y.

A.2 List of model parameters to be updated in an MCMC

Model parameters to be updated include

- the $(n_B \times n_P)$ true interaction matrix $\boldsymbol{L}$,
- the parameters of the interaction model $\boldsymbol{\lambda}$,
- the latent factors $\boldsymbol{U}, \boldsymbol{V}$ of dimension $(n_B \times H)$ and $(n_P \times H)$ respectively,
- the parameters of the trait models $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ including the residual variances σ_m^2 and σ_l^2 of continuous traits,
- the parameters of the models for the probability of observing a true interaction of a given species $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$ and the residual variances $\sigma_{p,B}^2, \sigma_{p,P}^2$,
- the probabilities themselves $\boldsymbol{p}_B = (p_1, p_2, \ldots, p_{n_B})$, and $\boldsymbol{p}_P = (p_1, p_2, \ldots, p_{n_P})$,
- the parameter in the latent factor covariance matrices ρ_U, ρ_V,
- the variance scaling parameters τ across all models,
- the parameters $\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\omega}$ and $\boldsymbol{v}$ controlling the increasing shrinkage prior, and
- covariate missing values, if applicable.

A.3 The posterior distribution

The posterior distribution of all model parameters (assuming no missing values of covariates) is

$$
\begin{aligned}
p(\text{parameters} \mid \text{Data}) \propto & \\
&\propto \prod_{m=1}^{p_B} p(\boldsymbol{X}_{\cdot m} \mid \boldsymbol{\beta}_{m\cdot}, \boldsymbol{U}, \sigma_m^2) \times \prod_{l=1}^{p_P} p(\boldsymbol{W}_{\cdot l} \mid \boldsymbol{\gamma}_{l\cdot}, \boldsymbol{V}, \sigma_l^2) \\
&\quad \times \prod_{i=1}^{n_B} \prod_{j=1}^{n_P} \Big[p(A_{ij} \mid p_i, p_j, L_{ij}) p(L_{ij} \mid \boldsymbol{\lambda}, \boldsymbol{U}_{i\cdot}, \boldsymbol{V}_{j\cdot}) \Big] \\
&\quad \times \Big\{ \prod_{i=1}^{n_B} p(\text{logit}(p_i) \mid \boldsymbol{\delta}, \boldsymbol{U}_{i\cdot}, \sigma_{p,B}^2) \Big\} \times \Big\{ \prod_{j=1}^{n_P} p(\text{logit}(p_j) \mid \boldsymbol{\zeta}, \boldsymbol{V}_{j\cdot}, \sigma_{p,P}^2) \Big\} \\
&\quad \times \prod_{h=1}^{H} \Big\{ p(\boldsymbol{U}_{\cdot h} \mid \boldsymbol{\Sigma}_B)\ p(\boldsymbol{V}_{\cdot h} \mid \boldsymbol{\Sigma}_P) \Big\} \\
&\quad \times p(\rho_U) p(\rho_V) \\
&\quad \times \prod_{m=1}^{p_B} p(\beta_{m0}) \prod_{h=1}^{H} p(\beta_{mh} \mid \tau_{mh}^{\beta}, \theta_h) \\
&\quad \times \prod_{l=1}^{p_P} p(\gamma_{l0}) \prod_{h=1}^{H} p(\gamma_{lh} \mid \tau_{lh}^{\gamma}, \theta_h) \\
&\quad \times p(\lambda_0) p(\delta_0) p(\zeta_0) \prod_{h=1}^{H} p(\lambda_h \mid \tau_h^{\lambda}, \theta_h) p(\delta_h \mid \tau_h^{\delta}, \theta_h) p(\zeta_h \mid \tau_h^{\zeta}, \theta_h) \\
&\quad \times \prod_{h=1}^{H} \left[p(\tau_h^{\delta}) p(\tau_h^{\zeta}) p(\tau_h^{\lambda}) \prod_{m=1}^{p_B} p(\tau_{mh}^{\beta}) \prod_{l=1}^{p_P} p(\tau_{lh}^{\gamma}) \right] \\
&\quad \times \prod_{h=1}^{H} p(\theta_h \mid \pi_h) p(\pi_h \mid \omega_1, \omega_2, \ldots, \omega_h) p(\omega_h \mid v_1, v_2, \ldots, v_h) p(v_h),
\end{aligned}
$$

where $p(\pi_h \mid \omega_1, \omega_2, \ldots, \omega_h)$ and $p(\omega_h \mid v_1, v_2, \ldots, v_h)$ are point mass distributions satisfying the equations in (7).

A.4 MCMC updates

Updating the true interaction matrix $\boldsymbol{L}$ We update the (i, j) entry of $\boldsymbol{L}$ in the following manner: If $A_{ij} = 1$, then L_{ij} is set to 1. If $A_{ij} = 0$, then L_{ij} is sampled using a Bernoulli distribution with

$$
p(L_{ij} = l \mid \cdot) \propto \begin{cases} 1 - p_{ij}^L, & \text{if } l = 0 \\ p_{ij}^L (1 - p_i p_j)^{n_{ij}} & \text{if } l = 1, \end{cases}
$$

where $p_{ij}^L = \text{expit}\left\{\lambda_0 + \sum_{h=1}^H \lambda_h U_{ih} V_{jh}\right\}$, and p_i, p_j are the probabilities of observing bird i and plant j in (4).

Updating the parameters $\boldsymbol{\lambda}$ of the interaction model We update these parameters using the Pólya-Gamma data-augmentation of Polson et al. [2013] in the following manner:

1. For each (i, j) pair, draw latent variables $\omega_{ij}^L \sim \text{PG}(1, \lambda_0 + \sum_h \lambda_h U_{ih} V_{jh})$. Conditional on ω_{ij}^L the contribution of L_{ij} to the likelihood is

$$p(L_{ij} \mid \omega_{ij}^L, \boldsymbol{\lambda}, \boldsymbol{U}_{i.}, \boldsymbol{V}_{j.}) \propto \exp\left\{-\frac{\omega_{ij}^L}{2}\left[\frac{L_{ij} - 1/2}{\omega_{ij}^L} - \left(\lambda_0 + \sum_{h=1}^H \lambda_h U_{ih} V_{jh}\right)\right]^2\right\},$$

   which is the kernel of a normal distribution, and can be combined with the normal prior distribution on $\boldsymbol{\lambda}$.

2. Sample $\boldsymbol{\lambda} \sim N_{H+1}(\mu_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left[\boldsymbol{D}_{UV}^T \boldsymbol{\Omega}^L \boldsymbol{D}_{UV} + (\boldsymbol{\Sigma}_0^\lambda)^{-1}\right]^{-1},$$

   and

$$\mu_{new} = \boldsymbol{\Sigma}_{new}\left[\boldsymbol{D}_{UV}^T\left(\text{vec}(\boldsymbol{L}) - 1/2\right) + (\boldsymbol{\Sigma}_0^\lambda)^{-1}\boldsymbol{\mu}_0^\lambda\right],$$

   where

   - $\boldsymbol{D}_{UV}$ is a matrix with $(n_B \times n_P)$ rows and $(H+1)$ columns, with first column equal to 1, and $(h+1)^{th}$ column equal to

$$\text{vec}(\boldsymbol{U}_{.h} \otimes \boldsymbol{V}_{.h}) = (U_{1h}V_{1h}, U_{1h}V_{2h}, \ldots, U_{1h}V_{n_P h}, U_{2h}V_{1h}, \ldots, U_{2h}V_{n_P h}, \ldots, U_{n_B h}V_{n_P h}),$$

   - $\boldsymbol{\Omega}^L$ is a matrix of dimension $(n_B n_P \times n_B n_P)$ with the entries $\text{vec}(\omega_{ij}^L)$ on the diagonal and 0 everywhere else,
   - $\boldsymbol{\Sigma}_0^\lambda$ is a diagonal matrix with entries $\sigma_0^2, \tau_1^\lambda\theta_1, \tau_2^\lambda\theta_h, \ldots, \tau_H^\lambda\theta_H$ on the diagonal (σ_0^2 is the prior variance of λ_0), and
   - $\boldsymbol{\mu}_0^\lambda$ is equal to $(\mu_0^{\lambda_0}, 0, 0, \ldots, 0)^T$, where $\mu_0^{\lambda_0}$ is the prior mean of λ_0.

Updating the variance scaling parameters τ Sample τ_{mh}^β from an inverse gamma distribution with parameters $(\nu+1)/2$ and $(\nu + \beta_{mh}^2/\theta_h)/2$. Similarly for $\tau_{lh}^\gamma, \tau_h^\delta, \tau_h^\zeta$ and τ_h^λ.

Updating the parameters of continuous traits models For a continuous trait m, the full conditional posterior distribution of $\boldsymbol{\beta}_{m\cdot} = (\beta_{m0}, \beta_{m1}, \ldots, \beta_{mH})^T$ is $N_{H+1}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left[\boldsymbol{D}_B^T \boldsymbol{D}_B / \sigma_m^2 + (\boldsymbol{\Sigma}_0^{\beta})^{-1}\right]^{-1}$$

and

$$\boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new}\left[\boldsymbol{D}_B^T \boldsymbol{X}_{\cdot m} / \sigma_m^2 + (\boldsymbol{\Sigma}_0^{\beta})^{-1}\boldsymbol{\mu}_0^{\beta}\right],$$

where

- $\boldsymbol{D}_B = (\mathbf{1} \mid \boldsymbol{U}_{\cdot 1} \mid \boldsymbol{U}_{\cdot 2} \mid \cdots \mid \boldsymbol{U}_{\cdot H})$ matrix of dimension $(n_B \times (H+1))$,
- $\boldsymbol{X}_{\cdot m}$ vector of entries for the m^{th} trait $(X_{1m}, X_{2m}, \ldots X_{n_B m})^T$,
- $\boldsymbol{\Sigma}_0^{\beta}$ diagonal matrix with entries $\sigma_0^2, \tau_{m1}^{\beta}\theta_1, \ldots, \tau_{mH}^{\beta}\theta_H$ (σ_0^2 is the prior variance of β_{m0}, and
- $\boldsymbol{\mu}_0^{\beta} = (\mu_0^{\beta_0}, 0, 0, \ldots, 0)^T$ ($\mu_0^{\beta_0}$ is the prior mean of β_{m0}).

To update the residual variance of continuous trait m, we sample σ_m^2 from an inverse gamma distribution with parameters $a_\sigma + n_B/2$ and $b_\sigma + \sum_{i=1}^{n_B}(X_{im} - (1, \boldsymbol{U}_{i\cdot}^T)^T \boldsymbol{\beta}_{m\cdot})^2/2$. Similarly we update parameters $\boldsymbol{\gamma}_{l\cdot} = (\gamma_{l0}, \gamma_{l1}, \ldots, \gamma_{lH})^T$ and σ_l^2 for continuous trait $\boldsymbol{L}$ of the other set of units.

Updating the parameters of binary traits models To update the coefficients $\boldsymbol{\beta}_{m\cdot}$ for a binary trait m we again follow the Pólya-Gamma data augmentation approach. Specifically,

1. We sample ω_{im} from $\text{PG}(1, (1, \boldsymbol{U}_{i\cdot}^T)^T \boldsymbol{\beta}_{m\cdot})$ for all $i = 1, 2, \ldots, n_B$.
2. We draw $\boldsymbol{\beta}_{m\cdot}$ from $N_{H+1}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left[\boldsymbol{D}_B^T \boldsymbol{\Omega}_m \boldsymbol{D}_B + (\boldsymbol{\Sigma}_0^{\beta})^{-1}\right]^{-1}$$

and

$$\boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new}\left[\boldsymbol{D}_B\left(\boldsymbol{X}_{\cdot m} - 1/2\right) + (\boldsymbol{\Sigma}_0^{\beta})^{-1}\boldsymbol{\mu}_0^{\beta}\right],$$

where $\boldsymbol{\Sigma}_0^{\beta}, \boldsymbol{\mu}_0^{\beta}, \boldsymbol{D}_B$ and $\boldsymbol{X}_{\cdot m}$ are as above, and Ω_m is a diagonal matrix with entries $\{\omega_{im}\}_{i=1}^{n_B}$.

Similarly we update the coefficients $\boldsymbol{\gamma}_{l\cdot} = (\gamma_{l0}, \gamma_{l1}, \ldots, \gamma_{lH})^T$ for the models of the binary traits for the other set of units.

Updating the parameters of the probability of observing an interaction The parameters $\boldsymbol{\delta}$ and $\sigma_{p,B}^2$ are updated similarly to the updates for the parameters of the continuous trait models $\boldsymbol{\beta}$ and σ_m^2,

using the same matrix $\boldsymbol{D}_B$, and setting

$$\boldsymbol{X}_{.m} = (\text{logit}(p_1), \text{logit}(p_2), \ldots, \text{logit}(p_{n_B}))^T.$$

The update of $\boldsymbol{\zeta}$ and $\sigma^2_{p,P}$ proceeds similarly.

Updating the latent factors We describe the update of the latent factors for the first set of units $\boldsymbol{U}_{.h}$ for $h = 1, 2, \ldots, H$, and updates for $\boldsymbol{V}_{.h}$ are similar. Here, we will use the Pólya-Gamma draws ω_{im} for binary traits m, and ω^L_{ij}, described above. For each $h = 1, 2, \ldots, H$, $\boldsymbol{U}_{.h}$ is drawn from $\mathcal{N}_{n_B}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left(\sum_{\substack{\text{m: } X_m \\ \text{continuous}}} \beta^2_{mh}/\sigma^2_m \boldsymbol{I}_{n_B} + \delta^2_h/\sigma^2_{p,B} \boldsymbol{I}_{n_B} + \sum_{\substack{\text{m: } X_m \\ \text{binary}}} \beta^2_{mh} \boldsymbol{\Omega}_m + \sum_{j=1}^{n_P} \lambda^2_h V^2_{jh} \boldsymbol{\Omega}^L_j + \boldsymbol{\Sigma}^{-1}_U \right)^{-1}$$

and

$$\begin{aligned} \boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new} \Bigg\{ & \sum_{\substack{\text{m: } X_m \\ \text{continuous}}} \beta_{mh}/\sigma^2_m \ \mathbf{part}(m,h) + \delta_h/\sigma^2_{p,B} \ \mathbf{part}(p,h) + \\ & \sum_{\substack{\text{m: } X_m \\ \text{binary}}} \beta_{mh} \boldsymbol{\Omega}_m \left[\left(\frac{\boldsymbol{X} - 1/2}{\boldsymbol{\omega}} \right)_m - (1 \mid \boldsymbol{U}_{.-h}) \, \boldsymbol{\beta}_{m(-h)} \right] + \\ & \sum_{j=1}^{n_P} \lambda_h V_{jh} \boldsymbol{\Omega}_j \left[\left(\frac{\boldsymbol{L} - 1/2}{\boldsymbol{\omega}} \right)_j - \left(1 \mid \boldsymbol{U}_{.-h} \boldsymbol{V}_{j(-h)}\right) \boldsymbol{\lambda}_{-h} \right] \Bigg\} \end{aligned}$$

where

- $\boldsymbol{\Omega}_j$ is used to denote, with some abuse of notation, the diagonal matrix of dimension n_B with entries representing the Pólya-Gamma draws from the interaction model involving unit j: $(\omega^L_{1j}, \omega^L_{2j}, \ldots, \omega^L_{n_B j})$,
- $\boldsymbol{\Sigma}_U$ is the covariance matrix of the latent factors in (5),
- $\mathbf{part}(m,h)$ is used to denote the residuals from the model for X_m when excluding the h^{th} latent factor, and is the following vector of length n_B:

$$\begin{aligned} \mathbf{part}(m,h) = \boldsymbol{X}_{.m} - \big(& \beta_{m0}\mathbf{1} + \beta_{m1}\boldsymbol{U}_{.1} + \cdots + \beta_{m(h-1)}\boldsymbol{U}_{.(h-1)} + \\ & \beta_{m(h+1)}\boldsymbol{U}_{.(h+1)} + \cdots + \beta_{mH}\boldsymbol{U}_{.H} \big), \end{aligned}$$

- Similarly, $\mathbf{part}(p,h)$ is used to denote a vector of length n_B including the residuals of the

model for the probability of observing when excluding the h^{th} latent factor:

$$[\text{part}(p,h)]_i = \text{logit}(p_i) - \left(\delta_0 + \delta_1 U_{i1} + \cdots + \delta_{h-1} U_{i(h-1)} + \delta_{h+1} U_{i(h+1)} + \cdots + \delta_H U_{iH}\right),$$

- $\left(\frac{\boldsymbol{X} - 1/2}{\boldsymbol{\omega}}\right)_m$ is a vector of length n_B with i^{th} element equal to $(X_{im} - 1/2)/\omega_{im}$,
- $(1 \mid \boldsymbol{U}_{.-h})$ is a matrix of dimension $n_B \times H$ representing a concatenation of a vector of 1 in the first column and the latent factors $\boldsymbol{U}$ excluding the h^{th} one,
- $\boldsymbol{\beta}_{m(-h)}$ is the vector $\boldsymbol{\beta}_m$ excluding the coefficient of the h^{th} latent factor,
- $\left(\frac{\boldsymbol{L} - 1/2}{\boldsymbol{\omega}}\right)_j$ is the diagonal matrix of dimension n_B including the transformed versions of unit j's interactions: $\left((L_{1j} - 1/2)/\omega^L_{1j}, (L_{2j} - 1/2)/\omega^L_{2j}, \ldots, (L_{n_B j} - 1/2)/\omega^L_{n_B j}\right)$,
- $\left(1 \mid \boldsymbol{U}_{.-h}\boldsymbol{V}_{j(-h)}\right)$ is the $n_B \times H$ matrix with first column equal to 1, second column equal to $V_{j1}\boldsymbol{U}_{.1} = (U_{11}V_{j1}, U_{21}V_{j1}, \ldots, U_{n_B 1}V_{j1})^T$, third column equal to $V_{j2}\boldsymbol{U}_{.2}$, up to the last column which is equal to $V_{jH}\boldsymbol{U}_{.H}$, *excluding* the h^{th} vector $V_{jh}\boldsymbol{U}_{.h}$, and *always* using the same unit j's latent factors, and
- $\boldsymbol{\lambda}_{-h}$ includes the coefficients of the interaction model excluding λ_h.

Updating the parameters of the increasing shrinkage prior In order to ease the updates of the increasing shrinkage prior parameters in (7), we introduce parameters $z_1, z_2, \ldots, z_H$ with $z_h \sim \text{Multinomial}(\omega_1, \omega_2, \ldots, \omega_H)$ and $\theta_h | z_h \sim I(z_h \leq h)\delta_{\theta_\infty} + I(z_h > h)P_0$, similarly to Legramanti et al. [2019]. Then, updates of the parameters proceeds by updating the parameters $\{v_h\}_h$ which deterministically set the values of $\{\omega_h\}_h$ and $\{\pi_h\}_h$, and updating the parameters $\{z_h\}_h$ and $\{\theta_h\}_h$.

First, updates for v_h are performed conditional on $z_1, z_2, \ldots, z_H$ by counting the number of z's with values equal or greater than h: v_h is sampled from a Beta distribution with parameters $\left(1 + \sum_{h'=1}^H I(z_{h'} = h), \alpha + \sum_{h'=1}^H I(z_{h'} > h)\right)$. Based on the sampled values for $v_1, v_2, \ldots, v_H$, the values of ω_h are updated from their deterministic relationship in (7).

Then, the variance parameters θ_h are updated using the part of the prior that is the slab P_0 or the spike δ_{θ_∞} depending on the value of the corresponding z_h:

- If $z_h \leq h$ (which happens with probability $\sum_{\omega_h} = \pi_h$) the variance component θ_h belongs to the spike part of the prior, and it is set equal to θ_∞.
- If $z_h > h$, then θ_h belongs to the P_0 part of the prior which is an inverse gamma distribution in our case, and θ_h is drawn from an inverse gamma with parameters $\alpha_\theta + (p_B + p_P + 3)/2$ and $\beta_\theta + \left(\sum_m \beta^2_{mh}/\tau^\beta_{mh} + \sum_l \gamma^2_{lh}/\tau^\gamma_{lh} + \lambda^2_h/\tau^\lambda_h + \delta^2_h/\tau^\delta_h + \zeta^2_h/\tau^\zeta_h\right)/2$.

Lastly, the parameters z_h are updated from a Multinomial distribution such that

$$p(z_h = l \mid \cdot, -\boldsymbol{\theta}) \propto \begin{cases} \omega_l \ \phi(\boldsymbol{x}; \theta_\infty \boldsymbol{\Sigma}) & \text{for } l = 1, 2, \ldots, h \\ \omega_l \ \tau(\boldsymbol{x}; 2\alpha_\theta, \beta_\theta/\alpha_\theta \boldsymbol{\Sigma}) & \text{for } l = h+1, h+2, \ldots, H, \end{cases}$$

where the vector $\boldsymbol{x}$ includes all coefficients of the h^{th} latent factors: $\boldsymbol{x} = (\boldsymbol{\beta}_{.h}^T, \boldsymbol{\gamma}_{.h}^T, \lambda_h, \delta_h, \zeta_h)^T$, $\boldsymbol{\Sigma}$ is a diagonal matrix with entries $\left((\tau_{.h}^\beta)^T, (\tau_{.h}^\gamma)^T, \tau_h^\lambda, \tau_h^\delta, \tau_h^\zeta\right)$, $\phi(\boldsymbol{x}; \theta_\infty \boldsymbol{\Sigma})$ is the density of a normal distribution centered at 0 with covariance matrix $\theta_\infty \boldsymbol{\Sigma}$ evaluated at $\boldsymbol{x}$, and $\tau(\boldsymbol{x}; 2\alpha_\theta, \beta_\theta/\alpha_\theta \boldsymbol{\Sigma})$ is the density of a multivariate t-distribution with $2\alpha_\theta$ degrees of freedom and covariance matrix $\beta_\theta/\alpha_\theta \boldsymbol{\Sigma}$ evaluated at $\boldsymbol{x}$. Note here that, even though the covariance matrix is diagonal, the density of the multivariate t-distribution is not the same as the sum of the densities from univariate t-distributions.

Updating the probability of observing an interaction Since the conditional posterior distributions of p_i, and p_j are not of known distributional form, we update them using Metropolis-Hastings. To update p_i:

- If $p_i^{(t)}$ is the value of p_i at iteration t, propose new value x from $\text{Beta}(np_i^{(t)}, n(1-p_i^{(t)}))$.
- Calculate the acceptance probability which is equal to

$$AP = \left\{ \prod_{j=1}^{n_P} \left[\frac{1-(1-xp_j)^{n_{ij}}}{1-(1-p_i^{(t)}p_j)^{n_{ij}}} \right]^{A_{ij}L_{ij}} \left[\frac{1-xp_j}{1-p_i^{(t)}p_j} \right]^{n_{ij}(1-A_{ij})L_{ij}} \right\} \times \frac{\phi(\text{logit}(x); (1 \ U_i^T)\boldsymbol{\delta}, \sigma_{p,B}^2)}{\phi(\text{logit}(p_i^{(t)}); (1 \ U_i^T)\boldsymbol{\delta}, \sigma_{p,B}^2)} \times \frac{b(p_i^{(t)}; nx, n(1-x))}{b(x; np_i^{(t)}, n(1-p_i^{(t)}))},$$

  where all other parameters are set to their most recent values, and $b(x; a, b)$ is the density of a $\text{Beta}(a, b)$ distribution evaluated at x.
- Accept x with probability AP, or stay at $p_i^{(t)}$ with probability $1 - AP$.

Similarly update the parameters p_j.

Updating the latent factor covariance parameter We update the parameters ρ_U, ρ_V using a Metropolis-Hastings step (similarly to the probability of detecting species). Specifically:

- If $\rho_U^{(t)}$ is the value of ρ_U at iteration t, propose new value x from $\text{Beta}(n\rho_U^{(t)}, n(1-\rho_U^{(t)}))$.
- Calculate the current and proposed value for the correlation matrix and denote them by $\boldsymbol{\Sigma}_U^{(t)}$ and $\boldsymbol{\Sigma}_U^x$, respectively.

- Calculate the acceptance probability which is equal to

$$AP = \frac{\prod_{h=1}^{H} \phi(\boldsymbol{U}_{.h}; \mathbf{0}, \boldsymbol{\Sigma}_U^{(t)})}{\prod_{h=1}^{H} \phi(\boldsymbol{U}_{.h}; \mathbf{0}, \boldsymbol{\Sigma}_U^{x})} \frac{b(\rho_U^{(t)}; \alpha_\rho, \beta_\rho)}{b(x; \alpha_\rho, \beta_\rho)}$$

  where all other parameters are set to their most recent values.

- Accept x with probability AP, or stay at $\rho_U^{(t)}$ with probability $1 - AP$.

Update ρ_V is a similar manner.

Update missing values of covariates If a covariate includes missing values for a subject of units, missing value imputation is straightforward and proceeds by drawing missing covariate values conditional on parameters and latent factors from (2).

A.5 Variable importance in latent factor network models

Since the latent factors are not identifiable parameters, our approach to investigate variable importance in network models is based on the posterior distribution of the probability of interaction in (1). Specifically, assume we want to investigate the importance of the k^{th} covariate for the first set of species, $\boldsymbol{X}_{.k} = (X_{1k}, X_{2k}, \ldots, X_{n_B k})^T$. Let $L_{ij}^{(r)}$ denote the logit of the fitted probability of interaction between species i and j at the r^{th} iteration of the MCMC and $\boldsymbol{l}_{.j}^{(r)}$ denote the vector of probabilities $L_{ij}^{(r)}$ for all i. In what follows we assume that species with missing information on the k^{th} covariate are excluded from both $\boldsymbol{X}_{.k}$ and $\boldsymbol{l}_{.j}^{(r)}$. Then,

1. For each posterior sample r and species j, we calculate the square correlation between $\boldsymbol{l}_{.j}^{(r)}$ and $\boldsymbol{X}_{.k}$. We denote the value by $T_{jk}^{*(r)}$.

2. We average the values of $T_{jk}^{*(r)}$ across species j and iterations r to acquire T_k^*.

3. For a large number of permutations B, do

   (a) Permute the entries in the vector $\boldsymbol{X}_{.k}$, "breaking" any relationship between the probabilities of interaction and the covariate.

   (b) Perform steps 1-2 using the permuted vector to acquire $T_k^{*(b)}$.

4. Calculate the mean and standard deviation of $T_k^{*(b)}$ across the B permutations.

5. Use $\left[T_k^* - \text{mean}(T_k^{*(b)})\right] \;/\; \text{sd}(T_k^{*(b)})$ as a measure of variable importance, separately among continuous and binary covariates.

Simulations for this measure of variable importance are shown in Supplement C.3.

We find the line of work of identifying important covariates in latent factor network models interesting, and for this reason, we include some considerations that occurred during our investigations. First, we found that the performance of the procedure described above drastically deteriorates when species with missing covariate values are *included* by using their imputed values (imputed during the MCMC). We believe that this happens because the latent factors are informed by the true interactions in $\boldsymbol{L}$ and they themselves play a role in the imputation of missing covariates. Therefore, a variable importance procedure that uses the imputed values will necessarily bias our understanding about the presence or magnitude of a link between the covariate and interactions. Second, since our approach to variable importance investigates marginal correlations between the covariate and the probability of interaction, it will perform best when the covariates are independent. This is well-known in the variable selection literature where a variable that is not important but is correlated with an important one has inflated marginal importance. We investigated the performance of using a multivariate linear regression and the absolute value of the regression coefficients as our test statistic. This alternative would alleviate some of the issues with observed correlated variables since each covariate's importance would be investigated conditional on the rest. However, we found that this approach did not perform as well as the method in Section 3.4. This might be because we included *all* species, including those with missing covariates, in the regression model, using their corresponding imputed values. Had we excluded the species with *any* missing covariate, the sample size would be dramatically lower, and it would have hurt our efficiency in identifying the important covariates. Finally, we also investigated the performance of a procedure that would find the latent factor with the largest standardized coefficient in the interaction model, and it would investigate the sum of squared residuals of that latent factor in the covariate models. This approach performed decently, but is perhaps harder to justify. An alternative variable importance procedure which we did not investigate could consider using the algorithm in Section 3.5 to make out of sample predictions for species where all but the target covariate are fixed at their observed level and the target covariate varies across all observed levels. Then, the variable's importance could be quantified by the variability in those predictions. We hope that this discussion might help interested researchers investigate this topic further.

A.6 Computationally efficient approximate algorithm for out of sample predictions

Suppose we have species i^* and species j^* with covariates $\boldsymbol{X}_{i^*}$ and $\boldsymbol{W}_{j^*}$, respectively, where the covariate vectors can include missing values. If i^*, j^* were both excluded from the original data and the MCMC fit, we refer to this pair as an *out of sample* pair. If only one of the species was excluded, the pair is referred to as a *half in sample* pair. In either case, the MCMC fit to the original data does not immediately produce estimates for $L_{i^*j^*}$, the indicator that the two species

are possible to interact. Inference on the interaction for half-in-sample or out-of-sample pairs could proceed by re-fitting the MCMC including the new species in the data. However, this approach would be computationally cumbersome, since it would require re-fitting the MCMC for every new set of species whose interactions we wish to predict, if at least one species was not recorded in the original data set.

Instead we propose an algorithm that uses samples from the posterior distribution of model parameters to predict the probability of interaction for half-in-sample and out-of sample pairs. This algorithm relies solely on the MCMC fit of the original data, and employs an importance sampling step to adjust for any differences in the two posterior distributions.

Interaction prediction for out-of sample pairs Let $\boldsymbol{\theta}^*$ denote all model parameters, and $\widetilde{\boldsymbol{D}}$ denote all observed data for all in-sample species. We denote the latent factors corresponding to units i^*, j^* as $\boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*}$ respectively. Then, we wish to predict $L_{i^*j^*}$ based on what we have learnt from the observed data and from the out-of-sample species covariates. That would amount to learning $P(L_{i^*j^*} = 1 \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})$., which we rewrite as

$$
\begin{aligned}
&P(L_{i^*j^*} = 1 \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}) = \\
&\quad = \int P(L_{i^*j^*} = 1 \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})\; p(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})\; \mathrm{d}(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*}) \\
&\quad \propto \int P(L_{i^*j^*} = 1 \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*})\; p(\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \widetilde{\boldsymbol{D}})\; p(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \mid \widetilde{\boldsymbol{D}})\; \mathrm{d}(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*}) \\
&\quad \propto \int P(L_{i^*j^*} = 1 \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*})\; p(\boldsymbol{X}_{i^*} \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*})\; p(\boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*, \boldsymbol{V}_{j^*}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad p(\boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \mid \boldsymbol{\theta}^*)\; p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})\; \mathrm{d}(\boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*})
\end{aligned}
$$

The last expression is the basis of our algorithm.

1. First, $p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})$ is the posterior distribution based on the MCMC of our original data. Hence, the posterior samples we acquired using the MCMC in Supplement A can be used to approximate this component.

2. Next, using the posterior samples of $\boldsymbol{\theta}^*$ (and the correlation matrices including the new species), we can straightforwardly sample latent factors for species i^*, j^* from $p(\boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*} \mid \boldsymbol{\theta}^*)$.

3. Based on the generated latent factors and model parameters, we can draw the interaction indicator directly from $P(L_{i^*j^*} = 1 \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}, \boldsymbol{V}_{j^*})$.

4. The previous steps did not take into consideration the components corresponding to the observed covariates. For that reason, we perform an importance sampling weight, where samples of $L_{i^*j^*}$ which use parameters and latent factors having higher (lower) values of $p(\boldsymbol{X}_{i^*} \mid \boldsymbol{\theta}^*, \boldsymbol{U}_{i^*}) p(\boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*, \boldsymbol{V}_{j^*})$ are up(down)-weighted.

We expand on how these steps are performed in what follows.

For i^*, j^* out of sample species, let $\boldsymbol{C}_U^*$, $\boldsymbol{C}_V^*$ be the extended correlation matrices. Then $\boldsymbol{C}_U^*$, $\boldsymbol{C}_V^*$ are correlation matrices of dimension $n_B + 1$ and $n_P + 1$, and the upper left $n_B \times n_B$ and $n_P \times n_P$ submatrices are $\boldsymbol{C}_U$, $\boldsymbol{C}_V$, respectively. In what follows the superscript (r) represents the r^{th} posterior sample from the MCMC out of a total of R samples.

The first step of predicting the interaction between i^* and j^* is to acquire samples for the species' latent factors based on samples from the posterior distribution of model parameters $\boldsymbol{\theta}^{*(r)}$, $r = 1, 2, \ldots, R$. We do so as follows:

- Generate latent factors for species i^*:
  For $r = 1, 2, \ldots, R$, sample U_{i^*h} from $N(\mu, \sigma^2)$ for values

$$\sigma^2 = \left[S^{(r)}\right]_{(n_B+1),(n_B+1)} - \left[\boldsymbol{S}^{(r)}\right]_{(n_B+1),(1:n_B)} \left[\left[\boldsymbol{S}^{(r)}\right]_{(1:n_B),(1:n_B)}\right]^{-1} \left[\left[\boldsymbol{S}^{(r)}\right]_{(n_B+1),(1:n_B)}\right]^T$$

  and

$$\mu = \left[\boldsymbol{S}^{(r)}\right]_{(n_B+1),(1:n_B)} \left[\left[\boldsymbol{S}^{(r)}\right]_{(1:n_B),(1:n_B)}\right]^{-1} \boldsymbol{U}^{(r)}_{1:n_B,h},$$

  where $\boldsymbol{S}^{(r)} = \rho_U^{(r)} \boldsymbol{C}_U^* + (1 - \rho_U^{(r)}) \boldsymbol{I}_{n_B+1}$, $[\boldsymbol{S}^{(r)}]_{\mathcal{A},\mathcal{B}}$ represents the submatrix of $\boldsymbol{S}^{(r)}$ with row indices in $\mathcal{A}$ and column indices in $\mathcal{B}$, and $\boldsymbol{U}^{(r)}_{1:n_B,h} = (U^{(r)}_{1h}, U^{(r)}_{2h}, \ldots, U^{(r)}_{n_B h})$.

- Generate latent factors for species j^* similarly as for species i^*, but substituting $\boldsymbol{C}_U$ for $\boldsymbol{C}_V$, ρ_U for ρ_V, $\boldsymbol{U}$ for $\boldsymbol{V}$, and n_B for n_P.

Performing these steps leads to latent factors $\boldsymbol{U}^{(r)}_{i^*} = (U_{i^*1}, U_{i^*2}, \ldots, U_{i^*H})^T$ for species i^* and $\boldsymbol{V}_{j^*} = (V_{j^*1}, V_{j^*2}, \ldots, V_{j^*H})^T$ for species j^*, for all $r = 1, 2, \ldots, R$. We use these latent factors to make an original set of predictions. For $r = 1, 2, \ldots, R$, we generate $\widetilde{L}^{(r)}_{i^*j^*}$ from a Bernoulli distribution with probability of success equal to expit $\left(\lambda_0^{(r)} + \sum_h \lambda_h^{(r)} U^{(r)}_{i^*h} V^{(r)}_{j^*h}\right)$.

However, the generated latent factors have been sampled taking only the correlation structure of the latent factors across species into consideration, and as a result the latent factors and the predictions do not use the information on the new species' covariates. To account for the covariates we perform importance weighting:

- For species i^* with generated latent factors $\boldsymbol{U}^{(r)}_{i^*}$ and covariates $\boldsymbol{X}_{i^*}$ calculate the importance sampling weight $w^{(r)}_{i^*} = w^{(r)}_{i^*1} w^{(r)}_{i^*2} \ldots, w^{(r)}_{i^*p_B}$, where $w^{(r)}_{i^*m}$ is 1 if X_{i^*m} is missing,

$$w^{(r)}_{i^*m} = \phi\left(X_{i^*m}; l^{(r)}_{i^*m}, (\sigma_m^2)^{(r)}\right)$$

if the m^{th} covariate is continuous, and

$$w_{i^*m}^{(r)} = u(X_{i^*m}; \text{expit}(l_{i^*m}^{(r)}))$$

if the m^{th} covariate is binary, where $l_{i^*m}^{(r)} = \beta_{m0}^{(r)} + (\boldsymbol{U}_{i^*}^{(r)})^T \boldsymbol{\beta}_m^{(r)}$, $\phi(\cdot; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2, and $u(\cdot; p)$ is the mass function for the Bernoulli(p) random variable.

- Similarly to the above, we acquire w_{j^*} for species j^*.
- The importance sampling weight for the pair (i^*, j^*) is then defined as $w_{i^*j^*}^{(r)} = w_{i^*}^{(r)} w_{j^*}^{(r)}$ for $r = 1, 2, \ldots, R$.

We combine these importance sampling weights with the original predicted interaction values $\widetilde{L}_{i^*j^*}^{(r)}$. Intuitively, $w_{i^*j^*}^{(r)}$ describes how in-line the generated latent factors $\boldsymbol{U}_{i^*}^{(r)}$ and $\boldsymbol{V}_{j^*}^{(r)}$ are with the species' covariate profiles, and, in a sense, how "trustworthy" the r^{th} prediction is. For this reason, we set the posterior probability for an $i^* - j^*$ interaction equal to

$$\left(\sum_{r=1}^{R} w_{i^*j^*}^{(r)} \widetilde{L}_{i^*j^*}^{(r)} \right) \Big/ \left(\sum_{r=1}^{R} w_{i^*j^*}^{(r)} \right).$$

Interaction prediction for half-in-sample pairs For half-in-sample pairs, one of the species is already included in the original data set, and samples from the posterior distribution for its latent factors are already acquired through the MCMC. Therefore, the procedure above only has to be performed for the species that are out-of-sample, and the importance sampling weights only represent one of the species. For example, if i^* is included in the original data and j^* is out-of-sample, then we use the algorithm described above to acquire $\boldsymbol{V}_{j^*}^{(r)}$, $w_{j^*}(r)$, and $\widetilde{L}_{i^*j^*}^{(r)}$ and set $w_{i^*j^*}^{(r)} = w_{j^*}^{(r)}$.

Performance of the algorithm for out-of-sample pairs The algorithm described above would theoretically return accurate predictions for $L_{i^*j^*}$ for out-of-sample units. However, importance sampling is known to have issues when weights become extremely large, and certain posterior samples dominate the weighted predictions. We evaluated the performance of the out-of-sample prediction algorithm in Supplement B.1. There, we see that prediction accuracy for out of sample pairs is comparable to prediction accuracy for out of sample pairs when they are included in the original MCMC.

Predictions for in-sample pairs The algorithms presented above discuss how we can perform prediction for out-of-sample species, and out-of-sample or half-in-sample pairs. However, the species i^*, j^* have observed covariate information, $\boldsymbol{X}_{i^*}$ and $\boldsymbol{W}_{j^*}$, and this information is not included in the

original posterior distribution $p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})$. Therefore, when covariate data on i^*, j^* become available, the posterior for $\boldsymbol{\theta}^*$ should be

$$\begin{aligned} p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}}, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}) &\propto p(\widetilde{\boldsymbol{D}} \mid \boldsymbol{\theta}^*, \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})\ p(\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*)\ p(\boldsymbol{\theta}^*) \\ &= p(\widetilde{\boldsymbol{D}} \mid \boldsymbol{\theta}^*)\ p(\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*} \mid \boldsymbol{\theta}^*)\ p(\boldsymbol{\theta}^*) \\ &= p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})\ \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})}{p(\boldsymbol{\theta}^*)} \\ &= p(\boldsymbol{\theta}^* \mid \widetilde{\boldsymbol{D}})\ \int \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^*_{i^*,j^*})}{p(\boldsymbol{\theta}^*)} p(\boldsymbol{\theta}^*_{i^*,j^*} \mid \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})\ \mathrm{d}\boldsymbol{\theta}^*_{i^*,j^*}, \end{aligned}$$

where we use $\boldsymbol{\theta}^*_{i^*,j^*}$ to denote all model parameters for species i^*, j^* (includes latent factors and detection probability).

Intuitively, since covariates are linked to latent factors and latent factors across species are correlated, the observed covariate information for out-of-sample species should affect estimation of interactions for in-sample pairs. If the species i^*, j^* were included in the MCMC with corresponding n_{i^*j} and n_{ij^*} equal to 0 for all i, j, then the covariate information $\boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*}$ for out-of-sample species would drive estimation of latent factors for the i^*, j^* units (through $p(\boldsymbol{\theta}^*_{i^*,j^*} \mid \boldsymbol{X}_{i^*}, \boldsymbol{W}_{j^*})$ in the equation above), and as a result would affect estimation of latent factors, parameters and predictions for all i, j (through the correlation of latent factors across species which implies that $p(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^*_{i^*,j^*}) \ / \ p(\boldsymbol{\theta}^*) \neq 1$ in the equation above).

This result suggests that when out of sample species with their covariate information (and non-zero correlation with in-sample species) become available, the interaction predictions for in-sample pairs should also be updated. In Supplement C.2 we evaluated the impact that out-of-sample correlated species have in the interaction predictions. There, we see that prediction of interactions for in-sample pairs when the out-of-sample species are excluded is essentially identical to the accuracy of the procedure that includes the out-of-sample species with corresponding $n-$values set to 0. This indicates that the flow of information from the out-of-sample covariates to the latent factors of in-sample species is quite weak, and ignoring it does not affect our predictions.

Supplement B. Alternative models

We consider four models that are combinations of using latent factors or covariates directly, and accommodating or not false negatives. Any overlap in the notation of coefficients among the models can be ignored. The model that uses latent factors and accommodates false negatives is the proposed one in Section 3. Here, we present the other three models. MCMC schemes are shown in Supplement E.

B.1 Model that uses covariates directly and accommodates false negatives

The first alternative model we consider includes covariates directly in model components' linear predictors and accommodates false negatives. It is specified as:

$$
\begin{aligned}
P(A_{ij} = 1 \mid L_{ij} = l) &= \begin{cases} 0, & \text{if } l = 0, \text{ and} \\ 1 - (1 - p_i p_j)^{n_{ij}}, & \text{if } l = 1 \end{cases} \\
\text{logit} P(L_{ij} = 1) &= \alpha_0 + \boldsymbol{X}_i^T \boldsymbol{\alpha}_X + \boldsymbol{W}_j^T \boldsymbol{\alpha}_W \\
\text{logit}(p_i) \mid \boldsymbol{X}_i &\sim \mathcal{N}(\delta_0 + \boldsymbol{X}_i^T \boldsymbol{\delta}, \sigma^2_{p,B}) \\
\text{logit}(p_j) \mid \boldsymbol{W}_j &\sim \mathcal{N}(\zeta_0 + \boldsymbol{W}_j^T \boldsymbol{\zeta}, \sigma^2_{p,P}).
\end{aligned}
\tag{S.1}
$$

The third and fourth lines in Supplement S.1 resemble the probability of observing model component (4) but the latent factors are substituted by the covariates. The latent factors are also substituted by covariates in the linear predictor of the interaction model (second line). The model linking the observed interaction matrix $\boldsymbol{A}$ to the true interaction matrix $\boldsymbol{L}$ (first line) is the same between this model and the one in Section 3. If the covariates include missing values, we extend Supplement S.1 to specify $E[X_{im}] = \mu_m$ and $E[W_{jl}] = \mu_l$. If the covariate is continuous, we assume it is normally distributed with variance σ^2_m and σ^2_l respectively. Doing so allows us to impute missing covariate values. We assume normal and inverse gamma prior distributions on coefficients and variance terms.

B.2 Model that uses covariates directly but does not accommodate false negatives

An alternative model we consider resembles the one in Supplement S.1 but assumes that there are no false negatives and $L_{ij} = A_{ij}$ for all i, j. Therefore, this model consists solely of the second line in Supplement S.1 with $L_{ij} = A_{ij}$.

B.3 Model that uses latent factors but does not accommodate false negatives

Lastly, we consider a version of our model that ignores the presence of false negatives, but maintains the use of latent factors to link the presence of an interaction and the model for the covariates. Therefore, since it assumes that $A_{ij} = L_{ij}$, this model specifies

$$
\begin{aligned}
&\text{logit} P(A_{ij} = 1) = \lambda_0 + \sum_{h=1}^{H} \lambda_h U_{ih} V_{jh}, \\
&f_m^{-1}(E(X_{im} \mid \boldsymbol{U}_i)) = \beta_{m0} + \boldsymbol{U}_i' \boldsymbol{\beta}_m, \ m = 1, 2, \ldots, p_B, \text{ and} \\
&g_l^{-1}(E(W_{jl} \mid \boldsymbol{V}_j)) = \gamma_{l0} + \boldsymbol{V}_j' \boldsymbol{\gamma}_l, \ l = 1, 2, \ldots, p_P
\end{aligned}
\tag{S.2}
$$

Supplement C. Additional simulation results

C.1 Simulation results for out of sample species

For out-of-sample species, we might be interested in predicting their interactions with other out-of-sample species, or with in-sample species. Pairs of species for which one is in-sample and the other is out-of-sample are referred to as "half-in-sample" and pairs for which both species are out-of-sample are referred to as "out-of-sample" pairs. Results are shown in Figure S.1.

The AUROC values for the methods are more variable here than in Figure 3 since they are based on 10 out-of-sample bird species and 10 out-of-sample plant species, but the methods' relative performance remains unchanged. Again, we see that bias correction is beneficial for learning the true interactions. However, methods performance is comparable in (dgm3) where the amount of information in the observed data is low, since all important covariates are unobserved.

C.2 Alternative uses of our method

The results shown in Section 4 and Figure S.1 for our method are based on its specification in Section 3 with out-of-sample species included in the MCMC and their interactions learnt through the MCMC updates in Supplement A. Here, we show that (a) the inclusion of the variance scaling components τ in (6) improves the method's predictive accuracy, and (b) the approximate algorithm for estimating the probability of interaction of out-of-sample species described in Section 3.5 and in more detail in Supplement A.6 performs well.

Specifically, in Figure S.2 we show the ratio of AUROC to the AUROC of the true, known model for three versions of our method: (Latent, Bias Corrected) The model as presented in Section 3, (Latent, Bias Corrected, Control Variance) The model in Section 3, setting all values of τ equal to 1 throughout, hence eliminating them, and (Latent, Bias Corrected, In-sample-only) The model as presented in Section 3 but fit over the subset of the species that were observed in at least one study: $n_i^+ > 0$ and $n_j^+ > 0$, and using the algorithm in Supplement A.6 for prediction of interactions for out-of-sample species.

Figure S.2 shows the results by type of pair (in-sample, half-in-sample, out-of-sample) and by data generating mechanism. We see that eliminating the τ-parameters can only lead to a reduction in the algorithm's predictive accuracy. This reflects that the additional flexibility offered by parameter-specific τ values is useful in improving our predictions. Further, our approximate algorithm performs reasonably in out-of-sample prediction. As expected, its performance is better when one of the species is in-sample (half-in-sample pairs) than when both species are out-of-sample (out-of-sample pairs).

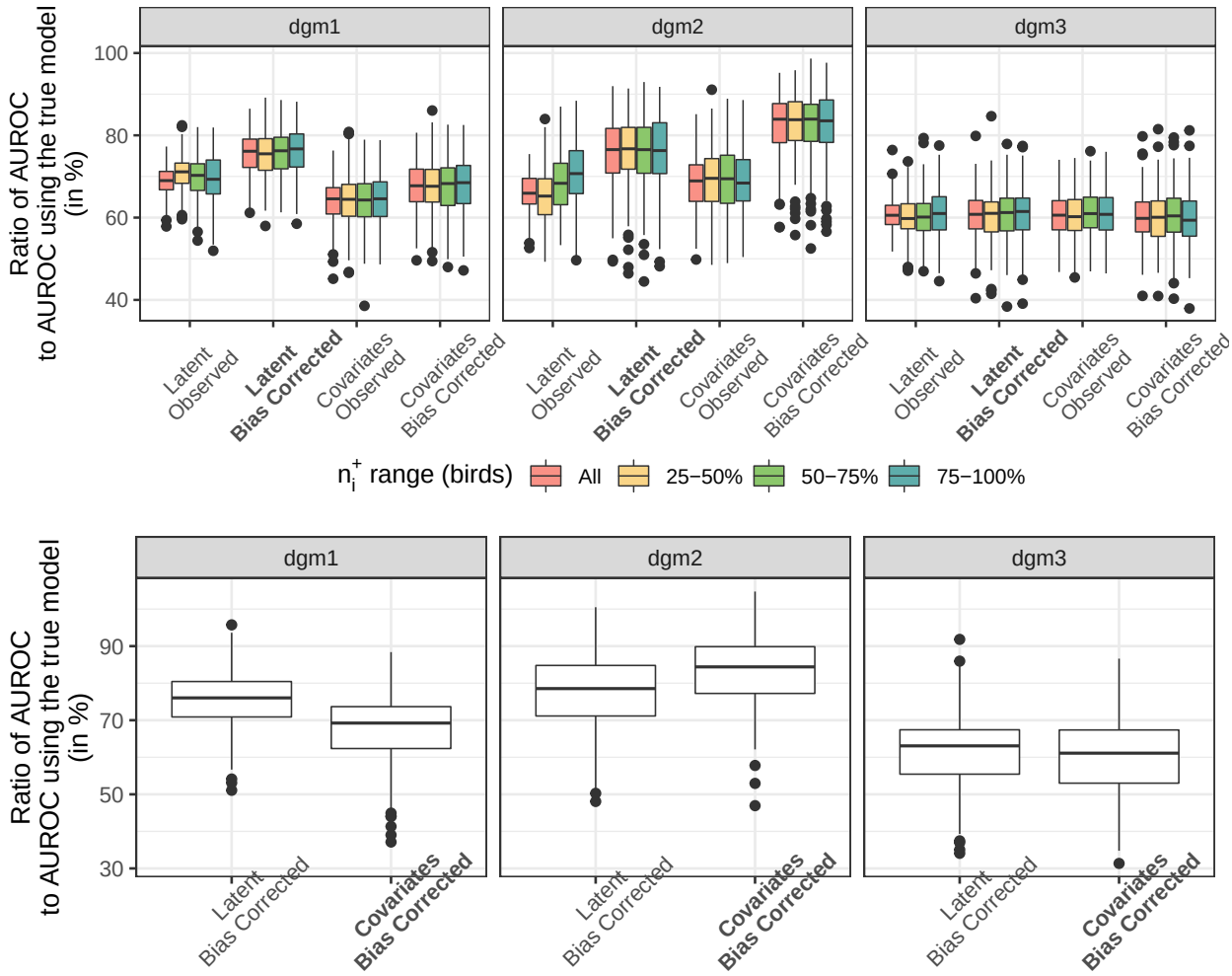


Figure S.1: Predictive Performance for Out-of-Sample Species. Ratio of methods' AUROC to AUROC using the true, known interaction model, for unrecorded interactions among in-sample bird species and out-of-sample plant species (half-in-sample pairs – top), and for out-of-sample pairs (bottom).

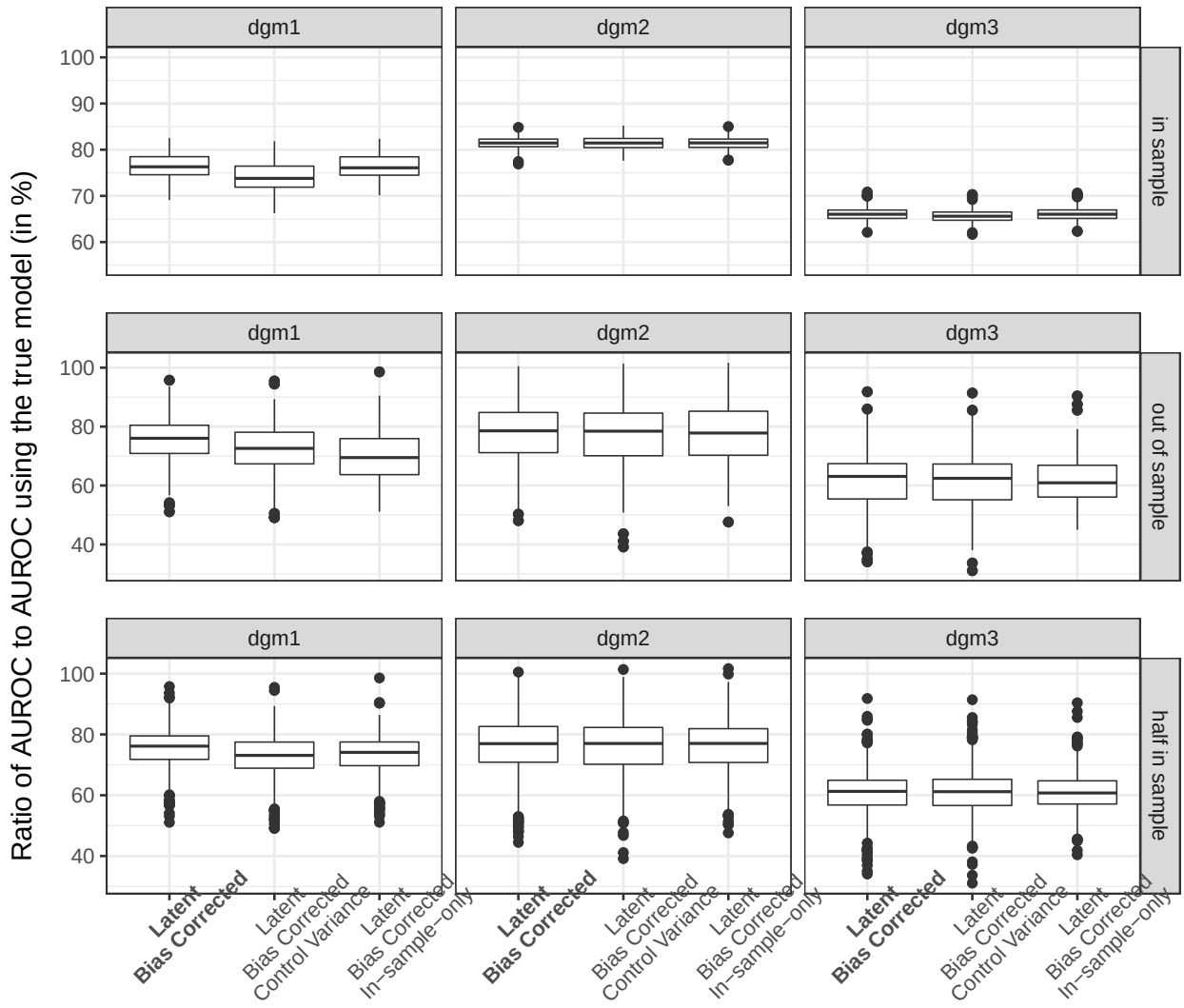


Figure S.2: Comparison of predictive power of our method under alternative specifications. The horizontal axis show the three methods considered, which represent the proposed method, the proposed method fixing all values of τ to 1, and the proposed method fit using the in-sample species only and using the algorithm in Supplement A.6 to make out-of sample predictions, respectively. The predictive power is shown by type of pair: in-sample, out-of-sample, and half-in-sample pairs, and by data generating mechanism.

C.3 Simulation results for the variable importance metric

We performed simulations to investigate the performance of the variable importance measure of Section 3.4. We considered the 3 data generative models used in the simulations Section 4. We also considered an additional scenario that is identical to (dgm1) but considers only independent covariates and uses both observed and unobserved covariates to simulate data. Which variables are used in each simulated scenario and for each model are shown in Table S.1.

Figure S.3 shows the simulation results for the variable importance measure of Section 3.4. Specifically, each panel corresponds to a different combination of covariate and data generative model and it shows the distribution across simulated data sets of the number of permutation standard deviations away from the permutation mean that the observed statistic falls. Larger values represent that the observed covariate is more informative of the probability of forming and detecting interactions between species. Dark blue color is used for covariates that are important for both forming or detecting interactions (grey shaded checkmarks in Table S.1), light blue is used for covariates that are important only for forming interactions (unshaded checkmarks in Table S.1), green is used for covariates that are important only for detecting interactions (shaded cells *without* checkmarks in Table S.1), and red is used for covariates that are important for neither (unshaded cells without checkmarks in Table S.1).

Table S.1: Variables that are used for forming and detecting interactions in the four scenarios for bird and plant species. ✓ indicates that the covariate was used in the interaction model. Shaded cells indicate that the covariate was used in the model for species detectability. The last column shows the correlation ρ for the covariates.

Bird covariates

	Continuous		Binary			Continuous	Binary	
	1	2	3	4	5	*Unobserved*		ρ
(dgm1)	✓	✓	✓	✓	✓			0.3
(dgm2)		✓	✓		✓			0.3
(dgm3)						✓	✓	0.3
(dgm4)	✓		✓		✓	✓	✓	0

Plant covariates

	Continuous				Binary				Continuous	Binary	
	1	2	3	4	5	6	10	7–9, 11-12	*Unobserved*		ρ
(dgm1)	✓	✓	✓	✓	✓						0.3
(dgm2)		✓					✓				0.3
(dgm3)									✓	✓	0.3
(dgm4)	✓		✓		✓				✓		0

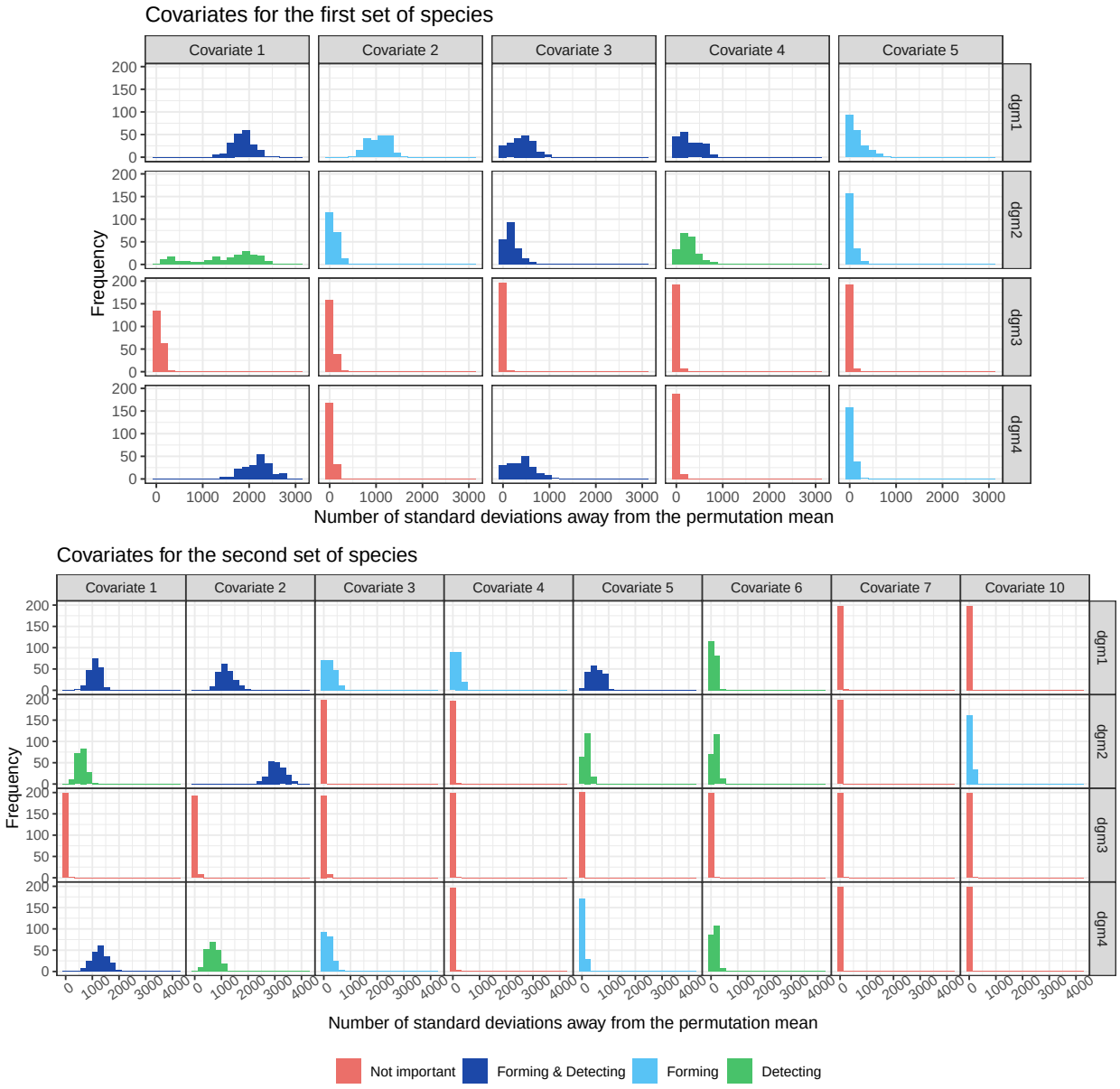


Figure S.3: Variable importance simulations. Number of standard deviations away from the permutation mean for the bird (top) and plant (bottom) traits by covariate and data generative model. Different colors are used for variables of different importance (important or not for forming and/or detecting an interaction), in agreement with Table S.1. Results for plant covariates 8, 9, 11, and 12 are similar to those of covariate 7 and are excluded.

We find that the approach correctly identifies covariates that are not important for either forming or detecting interactions, and the corresponding histograms are concentrated near 0 standard deviations, implying that the observed statistics resembles that of the permuted data sets. In contrast, the histogram for covariates that are important for both forming and detecting interactions are always well-separated from zero, and correctly identified. We find that continuous variables always have higher variable importance scores than binary variables with the same importance (for example, compare the results for covariates 1 & 3 of dgm1 of the bird species, and covariates 2 & 6 of dgm4 for the plant species). Importantly, we find that this variable importance measure (which uses the fitted values for the probability of interaction in (1)) does not only identify variables that are important for forming interactions, but also those that are only important for detecting them. This is evident from the histograms for covariates that are only important for detection which are also mostly separated from zero (green covariates). This result is not surprising when we think that the latent factors are used in both models for interactions and detection, and therefore a variable that is important for detection will likely inform the latent factors and as a result also inform the probability of interactions. We find that variables that are important for detecting interactions most often have variable importance that is higher than that of covariates that are important only for forming these interactions (comparing green to light blue histograms such as covariates 4 & 5 in dgm2 for the bird species). We think that this occurs because there is more signal in the detection model than for the interaction model, but this is just a conjecture.

Supplement D. Additional study results

Figure S.4 shows the posterior probability that an interaction between two species is possible based on our method and the alternative method. Vertical and horizontal blue lines separate different taxonomic families. A list of all the species included in our analysis in the same ordering as shown in the results is included in Appendix G. The same conclusions discussed in Section 5 also hold when showing the full set of species.

Supplement E. MCMC for alternative models

We present the MCMC schemes for the alternative models introduced in Supplement B.

E.1 Model that uses covariates directly and accommodates false negatives

We employ an MCMC scheme that resembles the one for the proposed approach in Supplement A: it uses the Pólya-Gamma data-augmentation of Polson et al. [2013] to update model parameters of the interaction model, Gibbs updates for the parameters of the probability of observing models, and

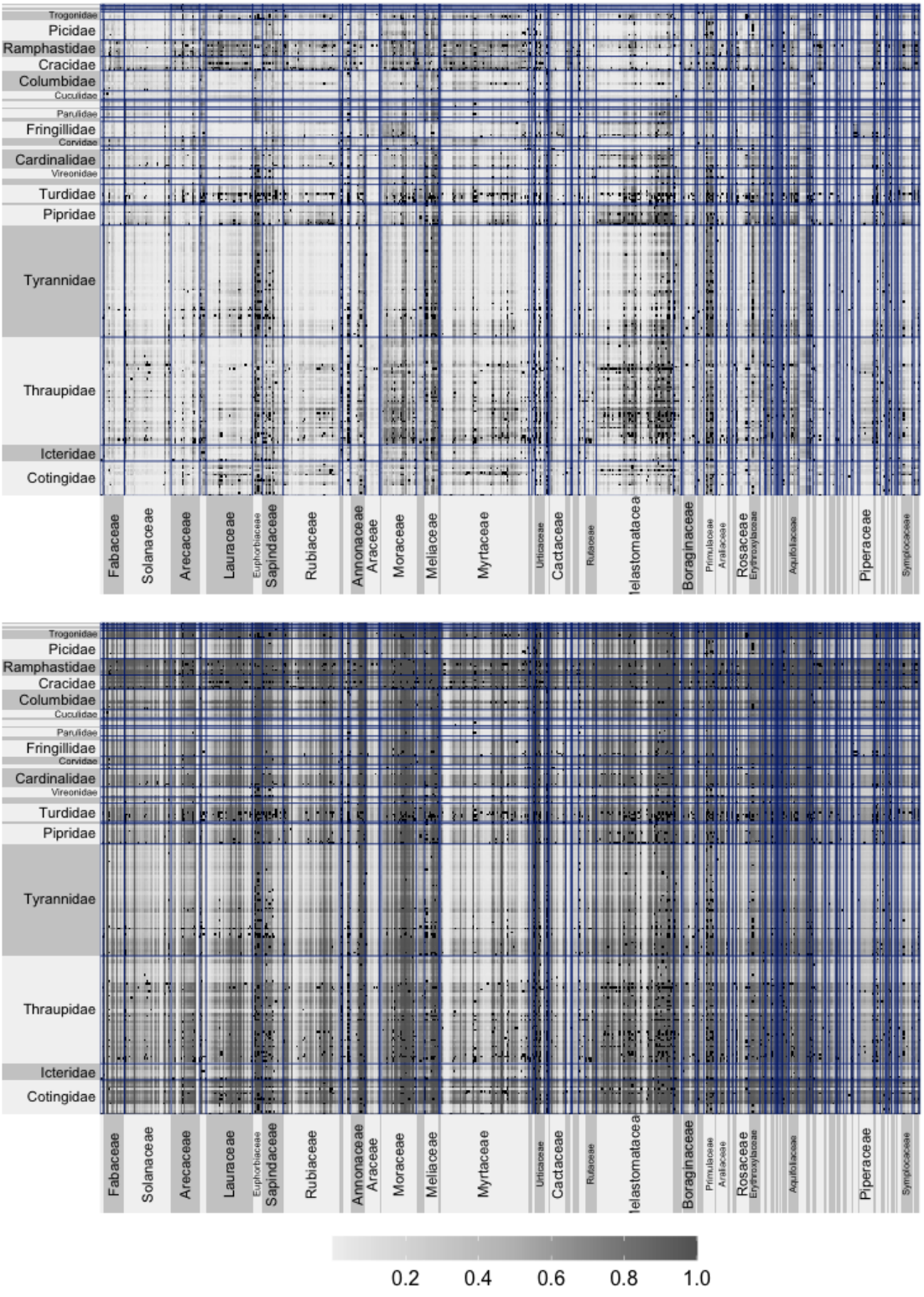


Figure S.4: Posterior Probability of Possible Interactions for all bird (y-axis) and plant (x-axis) species. Species are organized in taxonomic families separated by blue vertical and horizontal lines. Black color is used to represented interactions that are recorded in our data.

Metropolis-Hastings steps for updating the actual probabilities of observing a species. Specifically:

– The update of the true interaction matrix with entries L_{ij} proceeds exactly as in Supplement A, but for $p_{ij}^L = \text{expit}\left\{\alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha}_X + \boldsymbol{W}_j^T\boldsymbol{\alpha}_W\right\}$.

– We update the parameters of the interaction model $(\alpha_0, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_W)$ using Pólya-Gamma data-augmentation. For each (i,j) pair, we draw latent variables $\omega_{ij}^L \sim \text{PG}(1, \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha}_X + \boldsymbol{W}_j^T\boldsymbol{\alpha}_W)$. Conditional on ω_{ij}^L, we sample $(\alpha_0, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_W) \sim \mathcal{N}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left[\boldsymbol{D}^T\boldsymbol{\Omega}^L\boldsymbol{D} + \boldsymbol{\Sigma}_0^{-1}\right]^{-1}, \text{ and } \boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new}\left[\boldsymbol{D}^T\left(\text{vec}(L) - 1/2\right) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right],$$

where

1. $\boldsymbol{D}$ is a matrix with $(n_B \times n_P)$ rows and $(p_B + p_P + 1)$ columns, with first column equal to 1, each of the next p_B columns equal to $(\underbrace{X_{1m}, X_{1m}, \ldots, X_{1m}}_{n_P \text{ times}}, X_{2m}, \ldots, X_{2m}, \ldots, X_{n_Bm})$ for $m = 1, 2, \ldots, p_B$ (the i^{th} entry of the m^{th} covariate is repeated n_P number of times), and the next p_P columns are $(W_{1l}, W_{2l}, \ldots, W_{n_Pl}, W_{1l}, \ldots, W_{n_Pl}, \ldots W_{n_Pl})$ for $l = 1, 2, \ldots, p_P$ (the vector of the l^{th} covariate is repeated n_B times).
2. $\boldsymbol{\Omega}^L$ is a matrix of dimension $(n_Bn_P \times n_Bn_P)$ with the entries $\text{vec}(\omega_{ij}^L)$ on the diagonal and 0 everywhere else,
3. $\boldsymbol{\Sigma}_0 = \sigma_0^2\boldsymbol{I}_{1+p_B+p_P}$ is a diagonal matrix with prior variances, and
4. $\boldsymbol{\mu}_0$ is the vector $\boldsymbol{0}$ of length $1 + p_B + p_P$ including prior means.

– We update the parameters of the model for the probability of observing a species interaction by sampling $(\delta_0, \boldsymbol{\delta}^T)^T \sim \mathcal{N}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$, where

$$\boldsymbol{\Sigma}_{new} = \left[\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}/\sigma_{p,B}^2 + \boldsymbol{\Sigma}_0^{-1}\right]^{-1}$$

and

$$\boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new}\left[\widetilde{\boldsymbol{X}}^T[\text{logit}(p)]_{i=1}^{n_B}/\sigma_{p,B}^2 + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right],$$

where $\widetilde{\boldsymbol{X}} = (\boldsymbol{1} \mid \boldsymbol{X}_{.1} \mid \boldsymbol{X}_{.2} \mid \cdots \mid \boldsymbol{X}_{.p_B})$ is of dimension $n_B \times (p_B + 1)$, $\boldsymbol{\Sigma}_0 = \sigma_0^2\boldsymbol{I}_{p_B+1}$ is the diagonal matrix of prior variances, $\boldsymbol{\mu}_0 = \boldsymbol{0}$ is the vector of prior means, and $[\text{logit}(p)]_{i=1}^{n_B}$ is the vector of length n_B including the entries $\text{logit}(p_i)$.

To update the residual variance, we sample $\sigma_{p,B}^2$ from an inverse gamma distribution with parameters $a_0 + n_B/2$ and $b_0 + \sum_{i=1}^{n_B}(\text{logit}(p_i) - \delta_0 - \sum_{m=1}^{p_B}\delta_m X_{im})^2/2$, where a_0, b_0 are the parameters of the inverse gamma prior distribution on $\sigma_{p,B}^2$.

Similarly we update the parameters for the probability of observing an interaction for the second set of species.

– The updates for the probability of observing an interaction $\text{logit}(p_i), \text{logit}(p_j)$ proceed exactly as in Supplement A, for latent factors substituted by the observed covariates.

– Lastly, if covariates include missing values, models for the covariates are specified which include only an intercept if the covariate is binary, and an intercept and variance term if the covariate is continuous. In the presence of missing data, at each MCMC iteration we update the parameters (intercepts, residual variances for continuous covariates), and impute missing covariate values:

1. For continuous trait m with $X_{im} \sim \mathcal{N}(\mu_m, \sigma_m^2)$, and priors $\mathcal{N}(\mu_0, \sigma_0^2)$ and $IG(a_0, b_0)$ for the mean and variance parameters, we update μ_m from $\mathcal{N}(\mu_{new}, \sigma_{new}^2)$ where $\sigma_{new}^2 = [n_B/\sigma_m^2 + (\sigma_0^2)^{-1}]^{-1}$, and $\mu_{new} = \sigma_{new}^2[\sum_{i=1}^{n_B} X_{im}/\sigma_m^2 + \mu_0/\sigma_0^2]$, and update σ_m^2 from and inverse gamma distribution with parameters $a_0 + n_B/2$ and $b_0 + \sum_{i=1}^{n_B}(X_{im} - \mu_m)^2/2$. For these updates, the full vector $\boldsymbol{X}_{.m}$ is used, including the most current values of the imputed entries.

2. For the continuous trait m, we draw new values for X_{im} if this entry was missing from $\mathcal{N}(\mu_{i,new}, \sigma_{i,new}^2)$ where

$$\sigma_{i,new}^2 = \Big(\alpha_{Xm}^2 \sum_{j=1}^{n_P} \omega_{ij}^L + \delta_m^2/\sigma_{p,B}^2 + 1/\sigma_m^2\Big)^{-1},$$

$$\mu_{i,new} = \sigma_{i,new}^2 \Big[\alpha_{Xm} \sum_{j=1}^{n_P} (L_{ij} - 1/2 - \omega_{ij}^L(\alpha_0 + \boldsymbol{X}_{i(-m)}^T \boldsymbol{\alpha}_{X(-m)} + \boldsymbol{W}_j^T \boldsymbol{\alpha}_W) + \delta_m(\text{logit}(p_i) - \delta_0 - \boldsymbol{X}_{i(-m)}^T \boldsymbol{\delta}_{-m})/\sigma_{p,B}^2 + \mu_m/\sigma_m^2\Big],$$

   ω_{ij}^L are the PG draws discussed above, and the subscript $(-m)$ reflects that the m^{th} covariate (or its coefficient) is excluded.

3. For binary traits m with $X_{im} \sim \text{Bern}(\mu_m)$ we assume a normal prior on $\text{logit}(\mu_m)$ with mean μ_0 and variance σ_0^2. For the current values of μ_m, we draw n_B values from a Pólya-Gamma$(1, \text{logit}(\mu_m))$ distribution, denoted by $\omega_{im}, i = 1, 2, \ldots, n_B$. We sample a new value for $\text{logit}(\mu_m)$ from $\mathcal{N}(\mu_{new}, \sigma_{new}^2)$ where $\sigma_{new}^2 = [\sum_{i=1}^{n_B} \omega_{im} + (\sigma_0^2)^{-1}]^{-1}$ and $\mu_{new} = \sigma_{new}^2[\sum_{i=1}^{n_B}(X_{im} - 1/2) + (\sigma_0^2)^{-1}\mu_0]$.

4. For binary covariates, if X_{im} is missing, we calculate $p_{imx} = p(X_{im} = x \mid \cdot)$ for $x \in \{0, 1\}$

(up to a constant)

$$p_{imx} \propto \left[\prod_{j=1}^{n_P} (p_{ij}^L)^{L_{ij}} (1-p_{ij}^L)^{1-L_{ij}}\right] p(\text{logit}(p_i) \mid \boldsymbol{\delta}, \boldsymbol{X}_i, \sigma^2_{p.B})[\mu_m^x (1-\mu_m)^{1-x}],$$

where p_{ij}^L and the likelihood for the p_i model are calculated by setting $X_{im} = x$, and set X_{im} equal to x with probability $p_{imx}/(p_{im0} + p_{im1})$.

Updates for the missing covariates of the second set of species are identical, and the updates of all other parameters always use the most recent imputations of the missing covariate values.

E.2 Model that uses covariates directly but does not accommodate false negatives

The MCMC scheme for this model includes solely the updates of the interaction model parameters, which are the ones in Appendix E.1. In the presence of missing covariate values, models for these covariates are assumed and updates of these models' parameters are identical to the ones in Appendix E.1. However, covariate value imputation is slightly different, since here we do not assume a model for the probability of observation that depends on covariates. Therefore, covariate value imputation is like in Appendix E.1, but excluding the term from $\mu_{i,new}$ and p_{imx} corresponding to the p_i, p_j-submodel.

E.3 Model that uses latent factors but does not accommodate false negatives

The updates for the parameters in the traits models for binary or continuous traits, the parameters $\lambda_0, \boldsymbol{\lambda}$ of the interaction model, the parameters in the covariance matrix of the latent factors ρ_U, ρ_V, and the variance scaling parameters τ are the same as in Supplement A, substituting L_{ij} with A_{ij} where appropriate. Therefore, we only have to discuss updates for the latent factors and the increasing shrinkage prior:

– To update the latent factors the MCMC proceeds with an update similar to the one in Supplement A but accommodating the fact that the latent factors are no longer involved in the model for the probability of observing an interaction from a given species. We describe the update of the latent factors for the first set of units $\boldsymbol{U}_{.h}$ for $h = 1, 2, \ldots, H$, and updates for $\boldsymbol{V}_{.h}$ are similar. We use the Pólya-Gamma draws ω_{im} for binary traits m, and $\omega_{ij}^L = \omega_{ij}^A$ from the interaction model. For each $h = 1, 2, \ldots, H$, $\boldsymbol{U}_{.h}$ is drawn from $\mathcal{N}_{n_B}(\boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$ for parameters

$$\boldsymbol{\Sigma}_{new} = \left(\sum_{\substack{\text{m: } X_m \\ \text{continuous}}} \beta_{mh}^2/\sigma_m^2 \boldsymbol{I}_{n_B} + \sum_{\substack{\text{m: } X_m \\ \text{binary}}} \beta_{mh}^2 \boldsymbol{\Omega}_m + \sum_{j=1}^{n_P} \lambda_h^2 V_{jh}^2 \boldsymbol{\Omega}_j^L + \boldsymbol{\Sigma}_U^{-1}\right)^{-1}$$

and

$$\boldsymbol{\mu}_{new} = \boldsymbol{\Sigma}_{new}\Bigg\{ \sum_{\substack{\text{m: } X_m \\ \text{continuous}}} \beta_{mh}/\sigma_m^2 \; \mathbf{part}(m,h) + \sum_{\substack{\text{m: } X_m \\ \text{binary}}} \beta_{mh}\boldsymbol{\Omega}_m \left[\left(\frac{\boldsymbol{X} - 1/2}{\boldsymbol{\omega}} \right)_m - (1 \mid \boldsymbol{U}_{\cdot -h})\, \boldsymbol{\beta}_{m(-h)} \right] + \sum_{j=1}^{n_P} \lambda_h V_{jh} \boldsymbol{\Omega}_j \left[\left(\frac{\boldsymbol{L} - 1/2}{\boldsymbol{\omega}} \right)_j - \left(1 \mid \boldsymbol{U}_{\cdot -h}\boldsymbol{V}_{j(-h)}\right) \boldsymbol{\lambda}_{-h} \right] \Bigg\}$$

where $\boldsymbol{\Omega}_j, \boldsymbol{\Omega}_m, \boldsymbol{\Sigma}_U, \mathbf{part}(m,h), \left(\frac{\boldsymbol{X} - 1/2}{\boldsymbol{\omega}} \right)_m, (1 \mid \boldsymbol{U}_{\cdot -h}), \boldsymbol{\beta}_{m(-h)}, \left(\frac{\boldsymbol{L} - 1/2}{\boldsymbol{\omega}} \right)_j, \left(1 \mid \boldsymbol{U}_{\cdot -h}\boldsymbol{V}_{j(-h)}\right),$ and $\boldsymbol{\lambda}_{-h}$ are defined in Supplement A.

– The updates for the increasing shrinkage prior are as in Supplement A with two exceptions:

  · For the update of θ_h, if $z_h > h$, then θ_h is drawn from an inverse gamma with parameters $\alpha_\theta + (p_B + p_P + 1)/2$ and $\beta_\theta + \left(\sum_m \beta_{mh}^2/\tau_{mh}^\beta + \sum_l \gamma_{lh}^2/\tau_{lh}^\gamma + \lambda_h^2/\tau_h^\lambda \right)/2$.

  · For the update of z_h: z_h are updated from a Multinomial distribution such that

$$p(z_h = l \mid \cdot, -\boldsymbol{\theta}) \propto \begin{cases} \omega_l \; \phi(\boldsymbol{x}; \theta_\infty \boldsymbol{\Sigma}) & \text{for } l = 1, 2, \ldots, h \\ \omega_l \; \tau(\boldsymbol{x}; 2\alpha_\theta, \beta_\theta/\alpha_\theta \boldsymbol{\Sigma}) & \text{for } l = h+1, h+2, \ldots, H, \end{cases}$$

  where $\boldsymbol{x} = (\boldsymbol{\beta}_{\cdot h}^T, \boldsymbol{\gamma}_{\cdot h}^T, \lambda_h)^T$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with entries $\left((\tau_{\cdot h}^\beta)^T, (\tau_{\cdot h}^\gamma)^T, \tau_h^\lambda\right)$.

Supplement F. MCMC diagnostics

We evaluated convergence of MCMC scheme by studying traceplots of identifiable parameters across chains. Since latent factors and their corresponding coefficients are not identifiable, we focused our attention to: linear predictors and residual variances for the trait models, probabilities of detection and residual variances for both sets of species, and the correlation parameters ρ_U, ρ_V in the latent factors' covariance structure across species in the same set. The traceplots for a subset of parameters are shown in Figures S.5, S.6, S.7, and S.8.

We also investigated running means for the interactions indicators. If the posterior distribution is unimodal and the MCMC has converged sufficiently well, the running means converge to the same point. Running means for nine pairs of species without a recorded interaction are shown in Figure S.9.

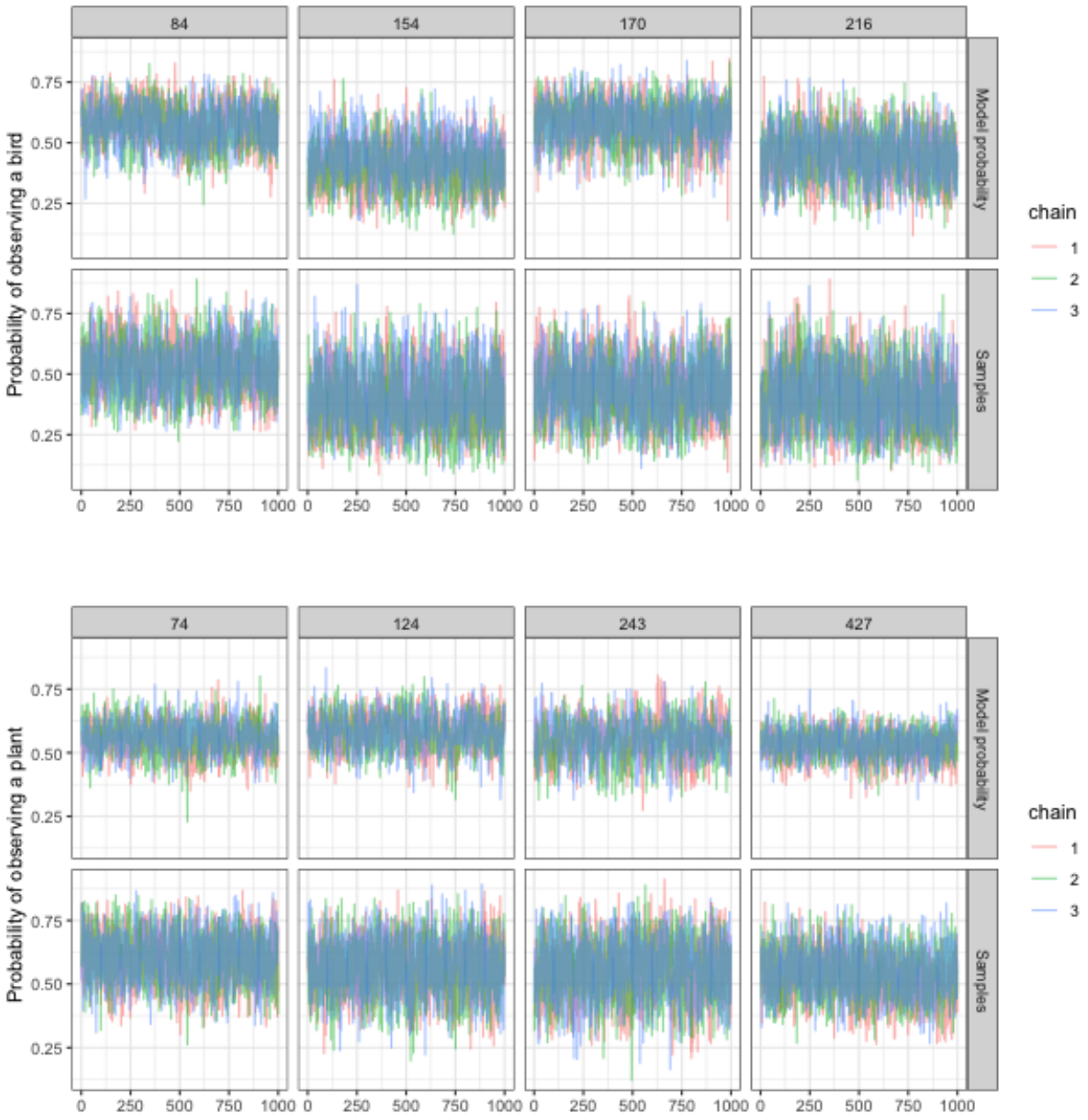


Figure S.5: Probability of detection. Traceplots for the linear predictor (Rows 1 & 3) and posterior samples (Rows 2 & 4) of the probability of detection for four randomly chosen bird (Rows 1 & 2) and four plant (Rows 3 & 4) species. Colors correspond to different MCMC chains.

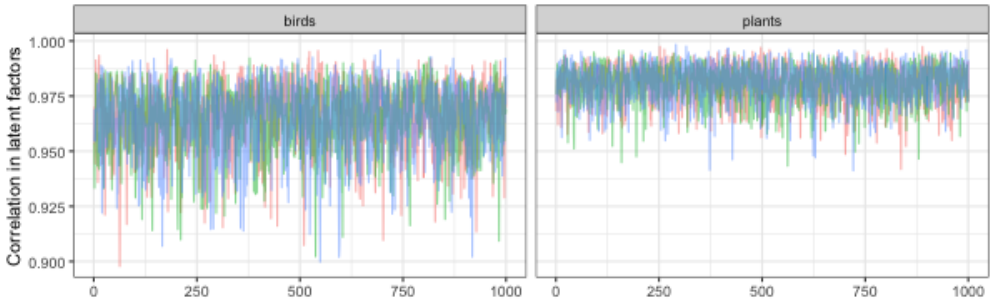


Figure S.6: Correlation of latent factors. Traceplots showing MCMC samples from the posterior distributions of ρ_U (left) and ρ_V (right). Colors correspond to different MCMC chains.

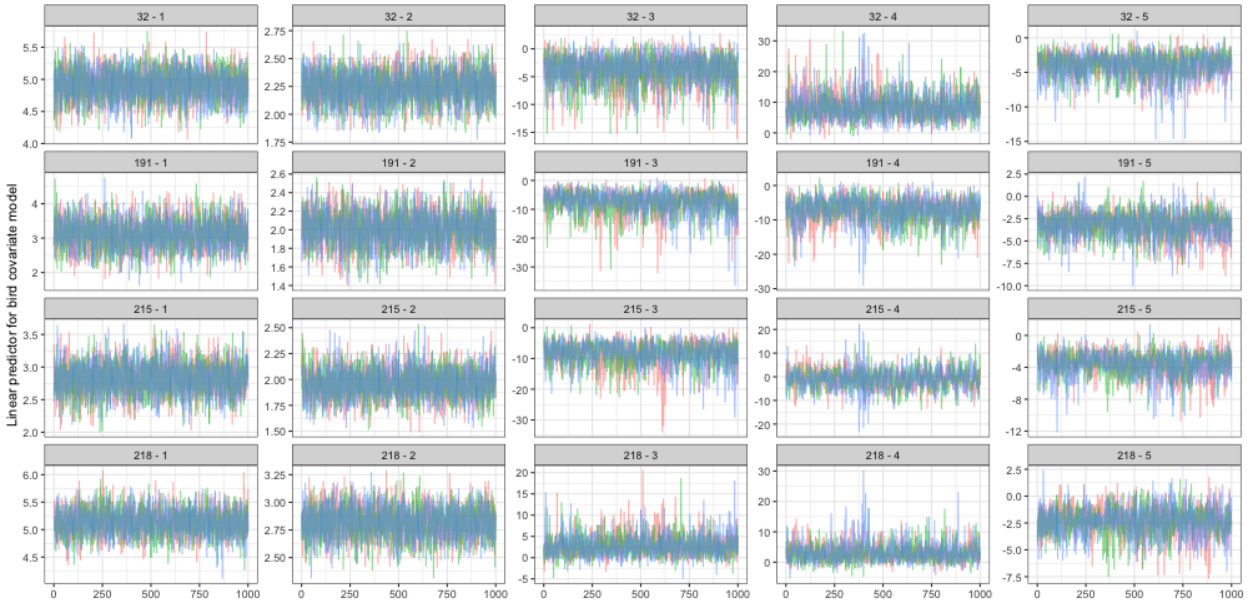


Figure S.7: Linear predictor of trait models for bird species. Traceplots for the linear predictor of all traits for four randomly chosen species of birds. The rows correspond to different bird species and the columns correspond to observed traits. The first two traits are continuous and the last three traits are binary.

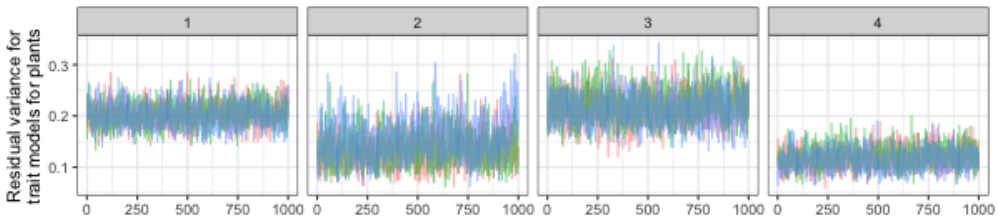


Figure S.8: Residual variances of continuous traits for plant species.

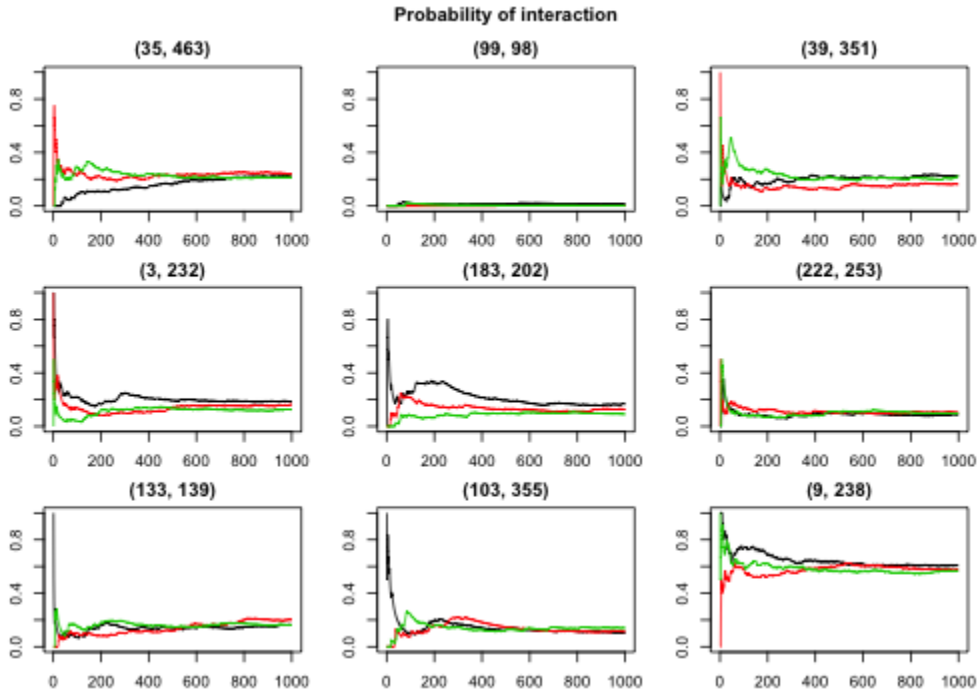


Figure S.9: Probability of species interaction. Running means of the indicator L_{ij} representing whether species i, j are possible to interact. Running means are shown for nine pairs of species without a recorded interaction.

Supplement G. List of all species included in our analysis

Tables S.2 and S.3 show the bird and plant species, respectively, that are included in our analysis along with their taxonomic information. The column "Included" denotes whether the species are shown in the results of the main text or not.

Table S.2: Bird species included in our data and their taxonomic information.

Species	Order	Family	Genus	Included
Pyroderus scutatus	Passeriformes	Cotingidae	Pyroderus	X
Pachyramphus validus	Passeriformes	Cotingidae	Pachyramphus	X
Pachyramphus castaneus	Passeriformes	Cotingidae	Pachyramphus	X
Pachyramphus viridis	Passeriformes	Cotingidae	Pachyramphus	X
Pachyramphus polychopterus	Passeriformes	Cotingidae	Pachyramphus	X
Lipaugus lanioides	Passeriformes	Cotingidae	Lipaugus	X
Lipaugus vociferans	Passeriformes	Cotingidae	Lipaugus	X
Tityra cayana	Passeriformes	Cotingidae	Tityra	X
Tityra inquisitor	Passeriformes	Cotingidae	Tityra	X
Oxyruncus cristatus	Passeriformes	Cotingidae	Oxyruncus	X
Carpornis cucullata	Passeriformes	Cotingidae	Carpornis	X
Carpornis melanocephala	Passeriformes	Cotingidae	Carpornis	X
Schiffornis virescens	Passeriformes	Cotingidae	Schiffornis	X
Procnias nudicollis	Passeriformes	Cotingidae	Procnias	X
Tijuca atra	Passeriformes	Cotingidae	Tijuca	X
Phibalura flavirostris	Passeriformes	Cotingidae	Phibalura	X
Laniisoma elegans	Passeriformes	Cotingidae	Laniisoma	X
Cacicus haemorrhous	Passeriformes	Icteridae	Cacicus	X
Cacicus chrysopterus	Passeriformes	Icteridae	Cacicus	X
Chrysomus ruficapillus	Passeriformes	Icteridae	Chrysomus	X
Molothrus bonariensis	Passeriformes	Icteridae	Molothrus	X
Icterus cayanensis	Passeriformes	Icteridae	Icterus	X
Pseudoleistes guirahuro	Passeriformes	Icteridae	Pseudoleistes	X
Psarocolius decumanus	Passeriformes	Icteridae	Psarocolius	X
Gnorimopsar chopi	Passeriformes	Icteridae	Gnorimopsar	X
Sicalis flaveola	Passeriformes	Thraupidae	Sicalis	X
Thraupis palmarum	Passeriformes	Thraupidae	Thraupis	X
Thraupis episcopus	Passeriformes	Thraupidae	Thraupis	X
Thraupis sayaca	Passeriformes	Thraupidae	Thraupis	X
Thraupis cyanoptera	Passeriformes	Thraupidae	Thraupis	X
Thraupis ornata	Passeriformes	Thraupidae	Thraupis	X
Thraupis	Passeriformes	Thraupidae	Thraupis	X
Thraupis bonariensis	Passeriformes	Thraupidae	Thraupis	X
Tachyphonus coronatus	Passeriformes	Thraupidae	Tachyphonus	X
Tachyphonus cristatus	Passeriformes	Thraupidae	Tachyphonus	X
Tachyphonus rufus	Passeriformes	Thraupidae	Tachyphonus	X
Tangara cayana	Passeriformes	Thraupidae	Tangara	X
Tangara seledon	Passeriformes	Thraupidae	Tangara	X
Tangara mexicana	Passeriformes	Thraupidae	Tangara	X
Tangara desmaresti	Passeriformes	Thraupidae	Tangara	X
Tangara cyanocephala	Passeriformes	Thraupidae	Tangara	X
Tangara cyanoptera	Passeriformes	Thraupidae	Tangara	X
Tangara preciosa	Passeriformes	Thraupidae	Tangara	X
Tangara cyanoventris	Passeriformes	Thraupidae	Tangara	X
Tangara peruviana	Passeriformes	Thraupidae	Tangara	X
Tangara	Passeriformes	Thraupidae	Tangara	X
Dacnis cayana	Passeriformes	Thraupidae	Dacnis	X

Dacnis nigripes	Passeriformes	Thraupidae	Dacnis	X
Ramphocelus carbo	Passeriformes	Thraupidae	Ramphocelus	X
Ramphocelus bresilius	Passeriformes	Thraupidae	Ramphocelus	X
Thlypopsis sordida	Passeriformes	Thraupidae	Thlypopsis	X
Conirostrum speciosum	Passeriformes	Thraupidae	Conirostrum	X
Hemithraupis guira	Passeriformes	Thraupidae	Hemithraupis	X
Hemithraupis ruficapilla	Passeriformes	Thraupidae	Hemithraupis	X
Hemithraupis flavicollis	Passeriformes	Thraupidae	Hemithraupis	X
Tersina viridis	Passeriformes	Thraupidae	Tersina	X
Chlorophanes spiza	Passeriformes	Thraupidae	Chlorophanes	X
Pipraeidea melanonota	Passeriformes	Thraupidae	Pipraeidea	X
Schistochlamys ruficapillus	Passeriformes	Thraupidae	Schistochlamys	X
Schistochlamys melanopis	Passeriformes	Thraupidae	Schistochlamys	X
Cissopis leverianus	Passeriformes	Thraupidae	Cissopis	X
Orthogonys chloricterus	Passeriformes	Thraupidae	Orthogonys	X
Trichothraupis melanops	Passeriformes	Thraupidae	Trichothraupis	X
Cyanerpes cyaneus	Passeriformes	Thraupidae	Cyanerpes	X
Stephanophorus diadematus	Passeriformes	Thraupidae	Stephanophorus	X
Nemosia pileata	Passeriformes	Thraupidae	Nemosia	X
Coryphospingus cucullatus	Passeriformes	Thraupidae	Coryphospingus	X
Coryphospingus pileatus	Passeriformes	Thraupidae	Coryphospingus	X
Volatinia jacarina	Passeriformes	Thraupidae	Volatinia	X
Sporophila caerulescens	Passeriformes	Thraupidae	Sporophila	X
Sporophila nigricollis	Passeriformes	Thraupidae	Sporophila	X
Sporophila leucoptera	Passeriformes	Thraupidae	Sporophila	X
Haplospiza unicolor	Passeriformes	Thraupidae	Haplospiza	X
Orchesticus abeillei	Passeriformes	Thraupidae	Orchesticus	X
Poospiza thoracica	Passeriformes	Thraupidae	Poospiza	X
Poospiza lateralis	Passeriformes	Thraupidae	Poospiza	X
Eucometis penicillata	Passeriformes	Thraupidae	Eucometis	X
Pyrrhocoma ruficeps	Passeriformes	Thraupidae	Pyrrhocoma	X
Elaenia flavogaster	Passeriformes	Tyrannidae	Elaenia	X
Elaenia	Passeriformes	Tyrannidae	Elaenia	X
Elaenia spectabilis	Passeriformes	Tyrannidae	Elaenia	X
Elaenia chiriquensis	Passeriformes	Tyrannidae	Elaenia	X
Elaenia mesoleuca	Passeriformes	Tyrannidae	Elaenia	X
Elaenia cristata	Passeriformes	Tyrannidae	Elaenia	X
Elaenia obscura	Passeriformes	Tyrannidae	Elaenia	X
Elaenia albiceps	Passeriformes	Tyrannidae	Elaenia	X
Elaenia parvirostris	Passeriformes	Tyrannidae	Elaenia	X
Myiodynastes maculatus	Passeriformes	Tyrannidae	Myiodynastes	X
Tyrannus melancholicus	Passeriformes	Tyrannidae	Tyrannus	X
Tyrannus savana	Passeriformes	Tyrannidae	Tyrannus	X
Tyrannus tyrannus	Passeriformes	Tyrannidae	Tyrannus	X
Pitangus sulphuratus	Passeriformes	Tyrannidae	Pitangus	X
Myiozetetes similis	Passeriformes	Tyrannidae	Myiozetetes	X
Myiozetetes cayanensis	Passeriformes	Tyrannidae	Myiozetetes	X
Myiarchus ferox	Passeriformes	Tyrannidae	Myiarchus	X
Myiarchus swainsoni	Passeriformes	Tyrannidae	Myiarchus	X
Myiarchus	Passeriformes	Tyrannidae	Myiarchus	X
Myiarchus tyrannulus	Passeriformes	Tyrannidae	Myiarchus	X
Cnemotriccus fuscatus	Passeriformes	Tyrannidae	Cnemotriccus	X
Mionectes oleagineus	Passeriformes	Tyrannidae	Mionectes	X
Mionectes rufiventris	Passeriformes	Tyrannidae	Mionectes	X
Megarynchus pitangua	Passeriformes	Tyrannidae	Megarynchus	X
Machetornis rixosa	Passeriformes	Tyrannidae	Machetornis	X
Attila rufus	Passeriformes	Tyrannidae	Attila	X
Attila phoenicurus	Passeriformes	Tyrannidae	Attila	X
Empidonomus varius	Passeriformes	Tyrannidae	Empidonomus	X

Colonia colonus	Passeriformes	Tyrannidae	Colonia	X
Phyllomyias fasciatus	Passeriformes	Tyrannidae	Phyllomyias	X
Phyllomyias griseocapilla	Passeriformes	Tyrannidae	Phyllomyias	X
Camptostoma obsoletum	Passeriformes	Tyrannidae	Camptostoma	X
Conopias trivirgatus	Passeriformes	Tyrannidae	Conopias	X
Leptopogon amaurocephalus	Passeriformes	Tyrannidae	Leptopogon	X
Tolmomyias sulphurescens	Passeriformes	Tyrannidae	Tolmomyias	X
Tolmomyias flaviventris	Passeriformes	Tyrannidae	Tolmomyias	X
Lathrotriccus euleri	Passeriformes	Tyrannidae	Lathrotriccus	X
Phylloscartes ventralis	Passeriformes	Tyrannidae	Phylloscartes	X
Phylloscartes sylviolus	Passeriformes	Tyrannidae	Phylloscartes	X
Phylloscartes oustaleti	Passeriformes	Tyrannidae	Phylloscartes	X
Knipolegus nigerrimus	Passeriformes	Tyrannidae	Knipolegus	X
Knipolegus cyanirostris	Passeriformes	Tyrannidae	Knipolegus	X
Legatus leucophaius	Passeriformes	Tyrannidae	Legatus	X
Xolmis cinereus	Passeriformes	Tyrannidae	Xolmis	X
Xolmis velatus	Passeriformes	Tyrannidae	Xolmis	X
Fluvicola nengeta	Passeriformes	Tyrannidae	Fluvicola	X
Serpophaga subcristata	Passeriformes	Tyrannidae	Serpophaga	X
Myiophobus fasciatus	Passeriformes	Tyrannidae	Myiophobus	X
Satrapa icterophrys	Passeriformes	Tyrannidae	Satrapa	X
Capsiempis flaveola	Passeriformes	Tyrannidae	Capsiempis	X
Casiornis rufus	Passeriformes	Tyrannidae	Casiornis	X
Contopus cinereus	Passeriformes	Tyrannidae	Contopus	X
Sirystes sibilator	Passeriformes	Tyrannidae	Sirystes	X
Myiopagis caniceps	Passeriformes	Tyrannidae	Myiopagis	X
Phaeomyias murina	Passeriformes	Tyrannidae	Phaeomyias	X
Chiroxiphia caudata	Passeriformes	Pipridae	Chiroxiphia	X
Chiroxiphia pareola	Passeriformes	Pipridae	Chiroxiphia	X
Manacus manacus	Passeriformes	Pipridae	Manacus	X
Pipra rubrocapilla	Passeriformes	Pipridae	Pipra	X
Pipra pipra	Passeriformes	Pipridae	Pipra	X
Ilicura militaris	Passeriformes	Pipridae	Ilicura	X
Antilophia galeata	Passeriformes	Pipridae	Antilophia	X
Neopelma aurifrons	Passeriformes	Pipridae	Neopelma	X
Neopelma pallescens	Passeriformes	Pipridae	Neopelma	X
Machaeropterus regulus	Passeriformes	Pipridae	Machaeropterus	X
Coereba flaveola	Passeriformes	Coerebidae	Coereba	
Turdus amaurochalinus	Passeriformes	Turdidae	Turdus	X
Turdus flavipes	Passeriformes	Turdidae	Turdus	X
Turdus leucomelas	Passeriformes	Turdidae	Turdus	X
Turdus rufiventris	Passeriformes	Turdidae	Turdus	X
Turdus albicollis	Passeriformes	Turdidae	Turdus	X
Turdus subalaris	Passeriformes	Turdidae	Turdus	X
Turdus fumigatus	Passeriformes	Turdidae	Turdus	X
Turdus	Passeriformes	Turdidae	Turdus	X
Catharus fuscescens	Passeriformes	Turdidae	Catharus	X
Zonotrichia capensis	Passeriformes	Emberizidae	Zonotrichia	
Arremon flavirostris	Passeriformes	Emberizidae	Arremon	
Arremon taciturnus	Passeriformes	Emberizidae	Arremon	
Vireo olivaceus	Passeriformes	Vireonidae	Vireo	X
Hylophilus amaurocephalus	Passeriformes	Vireonidae	Hylophilus	X
Hylophilus thoracicus	Passeriformes	Vireonidae	Hylophilus	X
Hylophilus poicilotis	Passeriformes	Vireonidae	Hylophilus	X
Cyclarhis gujanensis	Passeriformes	Vireonidae	Cyclarhis	X
Saltator maximus	Passeriformes	Cardinalidae	Saltator	X
Saltator similis	Passeriformes	Cardinalidae	Saltator	X
Saltator fuliginosus	Passeriformes	Cardinalidae	Saltator	X
Saltator coerulescens	Passeriformes	Cardinalidae	Saltator	X

Saltator maxillosus	Passeriformes	Cardinalidae	Saltator	X
Saltator atricollis	Passeriformes	Cardinalidae	Saltator	X
Habia rubica	Passeriformes	Cardinalidae	Habia	X
Piranga flava	Passeriformes	Cardinalidae	Piranga	X
Cyanocompsa brissonii	Passeriformes	Cardinalidae	Cyanocompsa	X
Mimus saturninus	Passeriformes	Mimidae	Mimus	
Mimus gilvus	Passeriformes	Mimidae	Mimus	
Cyanocorax cristatellus	Passeriformes	Corvidae	Cyanocorax	
Cyanocorax cyanomelas	Passeriformes	Corvidae	Cyanocorax	
Cyanocorax caeruleus	Passeriformes	Corvidae	Cyanocorax	
Cyanocorax chrysops	Passeriformes	Corvidae	Cyanocorax	
Euphonia violacea	Passeriformes	Fringillidae	Euphonia	X
Euphonia pectoralis	Passeriformes	Fringillidae	Euphonia	X
Euphonia chlorotica	Passeriformes	Fringillidae	Euphonia	X
Euphonia chalybea	Passeriformes	Fringillidae	Euphonia	X
Euphonia	Passeriformes	Fringillidae	Euphonia	X
Euphonia xanthogaster	Passeriformes	Fringillidae	Euphonia	X
Euphonia cyanocephala	Passeriformes	Fringillidae	Euphonia	X
Chlorophonia cyanea	Passeriformes	Fringillidae	Chlorophonia	X
Thamnophilus caerulescens	Passeriformes	Thamnophilidae	Thamnophilus	
Thamnophilus doliatus	Passeriformes	Thamnophilidae	Thamnophilus	
Myiothlypis flaveola	Passeriformes	Parulidae	Myiothlypis	
Setophaga pitiayumi	Passeriformes	Parulidae	Setophaga	
Basileuterus culicivorus	Passeriformes	Parulidae	Basileuterus	
Geothlypis aequinoctialis	Passeriformes	Parulidae	Geothlypis	
Estrilda astrild	Passeriformes	Estrildidae	Estrilda	
Cranioleuca pallida	Passeriformes	Furnariidae	Cranioleuca	
Synallaxis ruficapilla	Passeriformes	Furnariidae	Synallaxis	
Furnarius rufus	Passeriformes	Furnariidae	Furnarius	
Troglodytes aedon	Passeriformes	Troglodytidae	Troglodytes	
Crotophaga ani	Cuculiformes	Cuculidae	Crotophaga	
Crotophaga major	Cuculiformes	Cuculidae	Crotophaga	
Guira guira	Cuculiformes	Cuculidae	Guira	
Piaya cayana	Cuculiformes	Cuculidae	Piaya	
Patagioenas picazuro	Columbiformes	Columbidae	Patagioenas	X
Patagioenas cayennensis	Columbiformes	Columbidae	Patagioenas	X
Patagioenas	Columbiformes	Columbidae	Patagioenas	X
Patagioenas plumbea	Columbiformes	Columbidae	Patagioenas	X
Patagioenas speciosa	Columbiformes	Columbidae	Patagioenas	X
Leptotila verreauxi	Columbiformes	Columbidae	Leptotila	X
Leptotila rufaxilla	Columbiformes	Columbidae	Leptotila	X
Leptotila	Columbiformes	Columbidae	Leptotila	X
Zenaida auriculata	Columbiformes	Columbidae	Zenaida	X
Columbina talpacoti	Columbiformes	Columbidae	Columbina	X
Penelope superciliaris	Craciformes	Cracidae	Penelope	X
Penelope obscura	Craciformes	Cracidae	Penelope	X
Penelope	Craciformes	Cracidae	Penelope	X
Aburria jacutinga	Craciformes	Cracidae	Aburria	X
Ortalis guttata	Craciformes	Cracidae	Ortalis	X
Ortalis canicollis	Craciformes	Cracidae	Ortalis	X
Crax blumenbachii	Craciformes	Cracidae	Crax	X
Ramphastos dicolorus	Piciformes	Ramphastidae	Ramphastos	X
Ramphastos toco	Piciformes	Ramphastidae	Ramphastos	X
Ramphastos vitellinus	Piciformes	Ramphastidae	Ramphastos	X
Ramphastos	Piciformes	Ramphastidae	Ramphastos	X
Baillonius bailloni	Piciformes	Ramphastidae	Baillonius	X
Selenidera maculirostris	Piciformes	Ramphastidae	Selenidera	X
Pteroglossus aracari	Piciformes	Ramphastidae	Pteroglossus	X
Pteroglossus castanotis	Piciformes	Ramphastidae	Pteroglossus	X

Picumnus cirratus	Piciformes	Picidae	Picumnus	X
Picumnus nebulosus	Piciformes	Picidae	Picumnus	X
Celeus flavescens	Piciformes	Picidae	Celeus	X
Melanerpes flavifrons	Piciformes	Picidae	Melanerpes	X
Melanerpes candidus	Piciformes	Picidae	Melanerpes	X
Colaptes campestris	Piciformes	Picidae	Colaptes	X
Colaptes melanochloros	Piciformes	Picidae	Colaptes	X
Veniliornis spilogaster	Piciformes	Picidae	Veniliornis	X
Piculus aurulentus	Piciformes	Picidae	Piculus	X
Dryocopus lineatus	Piciformes	Picidae	Dryocopus	X
Trogon surrucura	Trogoniformes	Trogonidae	Trogon	
Trogon viridis	Trogoniformes	Trogonidae	Trogon	
Trogon rufus	Trogoniformes	Trogonidae	Trogon	
Trogon curucui	Trogoniformes	Trogonidae	Trogon	
Baryphthengus ruficapillus	Coraciiformes	Momotidae	Baryphthengus	
Coragyps atratus	Accipitriformes	Cathartidae	Coragyps	
Caracara plancus	Falconiformes	Falconidae	Caracara	
Aramides cajanea	Gruiformes	Rallidae	Aramides	

Table S.3: Plant species included in our data and their taxonomic information.

Species	Family	Genus	Included
Abuta selloana	Menispermaceae	Abuta	
Cissampelos andromorpha	Menispermaceae	Cissampelos	
Acacia auriculiformis	Fabaceae	Acacia	X
Andira fraxinifolia	Fabaceae	Andira	X
Cajanus cajan	Fabaceae	Cajanus	X
Copaifera langsdorffii	Fabaceae	Copaifera	X
Copaifera trapezifolia	Fabaceae	Copaifera	X
Desmodium incanum	Fabaceae	Desmodium	X
Holocalyx balansae	Fabaceae	Holocalyx	X
Hymenaea courbaril	Fabaceae	Hymenaea	X
Inga edulis	Fabaceae	Inga	X
Inga laurina	Fabaceae	Inga	X
Inga marginata	Fabaceae	Inga	X
Inga sessilis	Fabaceae	Inga	X
Samanea tubulosa	Fabaceae	Samanea	X
Acnistus arborescens	Solanaceae	Acnistus	X
Aureliana fasciculata	Solanaceae	Aureliana	X
Cestrum bracteatum	Solanaceae	Cestrum	X
Cestrum mariquitense	Solanaceae	Cestrum	X
Cestrum schlechtendalii	Solanaceae	Cestrum	X
Lycianthes pauciflora	Solanaceae	Lycianthes	X
Physalis pubescens	Solanaceae	Physalis	X
Solanum aculeatissimum	Solanaceae	Solanum	X
Solanum americanum	Solanaceae	Solanum	X
Solanum argenteum	Solanaceae	Solanum	X
Solanum bullatum	Solanaceae	Solanum	X
Solanum corymbiflorum	Solanaceae	Solanum	X
Solanum granulosoleprosum	Solanaceae	Solanum	X
Solanum inodorum	Solanaceae	Solanum	X
Solanum mauritianum	Solanaceae	Solanum	X
Solanum megalochiton	Solanaceae	Solanum	X
Solanum myrianthum	Solanaceae	Solanum	X
Solanum nigrescens	Solanaceae	Solanum	X
Solanum paranense	Solanaceae	Solanum	X
Solanum pseudoquina	Solanaceae	Solanum	X

Solanum rufescens	Solanaceae	Solanum	X
Solanum sanctae-catharinae	Solanaceae	Solanum	X
Solanum scuticum	Solanaceae	Solanum	X
Solanum subsylvestre	Solanaceae	Solanum	X
Solanum swartzianum	Solanaceae	Solanum	X
Solanum thomasiifolium	Solanaceae	Solanum	X
Solanum variabile	Solanaceae	Solanum	X
Solanum viscosissimum	Solanaceae	Solanum	X
Vassobia breviflora	Solanaceae	Vassobia	X
Acrocomia aculeata	Arecaceae	Acrocomia	X
Allagoptera arenaria	Arecaceae	Allagoptera	X
Archontophoenix cunninghamiana	Arecaceae	Archontophoenix	X
Astrocaryum aculeatissimum	Arecaceae	Astrocaryum	X
Attalea dubia	Arecaceae	Attalea	X
Bactris gasipaes	Arecaceae	Bactris	X
Elaeis guineensis	Arecaceae	Elaeis	X
Euterpe edulis	Arecaceae	Euterpe	X
Euterpe oleracea	Arecaceae	Euterpe	X
Geonoma elegans	Arecaceae	Geonoma	X
Geonoma gamiova	Arecaceae	Geonoma	X
Geonoma pauciflora	Arecaceae	Geonoma	X
Livistona chinensis	Arecaceae	Livistona	X
Livistona chinensis	Arecaceae	Livistona	X
Phoenix sylvestris	Arecaceae	Phoenix	X
Roystonea oleraceae	Arecaceae	Roystonea	X
Syagrus pseudococos	Arecaceae	Syagrus	X
Syagrus romanzoffiana	Arecaceae	Syagrus	X
Aegiphila integrifolia	Lamiaceae	Aegiphila	
Callicarpa reevesii	Lamiaceae	Callicarpa	
Vitex megapotamica	Lamiaceae	Vitex	
Vitex polygama	Lamiaceae	Vitex	
Aiouea saligna	Lauraceae	Aiouea	X
Cryptocarya aschersoniana	Lauraceae	Cryptocarya	X
Cryptocarya mandioccana	Lauraceae	Cryptocarya	X
Cryptocarya moschata	Lauraceae	Cryptocarya	X
Endlicheria paniculata	Lauraceae	Endlicheria	X
Nectandra cuspidata	Lauraceae	Nectandra	X
Nectandra grandiflora	Lauraceae	Nectandra	X
Nectandra lanceolata	Lauraceae	Nectandra	X
Nectandra megapotamica	Lauraceae	Nectandra	X
Nectandra membranacea	Lauraceae	Nectandra	X
Nectandra reticulata	Lauraceae	Nectandra	X
Ocotea aeciphila	Lauraceae	Ocotea	X
Ocotea bicolor	Lauraceae	Ocotea	X
Ocotea catharinensis	Lauraceae	Ocotea	X
Ocotea corymbosa	Lauraceae	Ocotea	X
Ocotea diospyrifolia	Lauraceae	Ocotea	X
Ocotea dispersa	Lauraceae	Ocotea	X
Ocotea macropoda	Lauraceae	Ocotea	X
Ocotea notata	Lauraceae	Ocotea	X
Ocotea odorifera	Lauraceae	Ocotea	X
Ocotea puberula	Lauraceae	Ocotea	X
Ocotea pulchella	Lauraceae	Ocotea	X
Ocotea silvestris	Lauraceae	Ocotea	X
Ocotea spixiana	Lauraceae	Ocotea	X
Ocotea teleiandra	Lauraceae	Ocotea	X
Persea alba	Lauraceae	Persea	X
Persea major	Lauraceae	Persea	X
Persea willdenovii	Lauraceae	Persea	X

Phoebe pickelli	Lauraceae	Phoebe	X
Alchornea discolor	Euphorbiaceae	Alchornea	
Alchornea glandulosa	Euphorbiaceae	Alchornea	
Alchornea sidifolia	Euphorbiaceae	Alchornea	
Alchornea triplinervia	Euphorbiaceae	Alchornea	
Sapium glandulosum	Euphorbiaceae	Sapium	
Tetrorchidium rubrivenium	Euphorbiaceae	Tetrorchidium	
Allophylus edulis	Sapindaceae	Allophylus	X
Cupania emarginata	Sapindaceae	Cupania	X
Cupania oblongifolia	Sapindaceae	Cupania	X
Cupania riodocensis	Sapindaceae	Cupania	X
Cupania vernalis	Sapindaceae	Cupania	X
Litchi chinensis	Sapindaceae	Litchi	X
Matayba elaeagnoides	Sapindaceae	Matayba	X
Matayba guianensis	Sapindaceae	Matayba	X
Paullinia carpopoda	Sapindaceae	Paullinia	X
Paullinia micrantha	Sapindaceae	Paullinia	X
Paullinia rhomboidea	Sapindaceae	Paullinia	X
Paullinia uloptera	Sapindaceae	Paullinia	X
Sapindus saponaria	Sapindaceae	Sapindus	X
Amaioua guianensis	Rubiaceae	Amaioua	X
Amaioua intermedia	Rubiaceae	Amaioua	X
Chomelia parvifolia	Rubiaceae	Chomelia	X
Coccocypselum geophiloides	Rubiaceae	Coccocypselum	X
Coccocypselum hasslerianum	Rubiaceae	Coccocypselum	X
Coffea arabica	Rubiaceae	Coffea	X
Cordiera myrciifolia	Rubiaceae	Cordiera	X
Coussarea contracta	Rubiaceae	Coussarea	X
Galium hypocarpium	Rubiaceae	Galium	X
Genipa americana	Rubiaceae	Genipa	X
Geophila macropoda	Rubiaceae	Geophila	X
Geophila repens	Rubiaceae	Geophila	X
Guettarda viburnoides	Rubiaceae	Guettarda	X
Ixora burchelliana	Rubiaceae	Ixora	X
Ixora gardneriana	Rubiaceae	Ixora	X
Ixora venulosa	Rubiaceae	Ixora	X
Margaritopsis astrellantha	Rubiaceae	Margaritopsis	X
Margaritopsis chaenotricha	Rubiaceae	Margaritopsis	X
Palicourea macrobotrys	Rubiaceae	Palicourea	X
Posoqueria latifolia	Rubiaceae	Posoqueria	X
Psychotria carthagenensis	Rubiaceae	Psychotria	X
Psychotria forsteronioides	Rubiaceae	Psychotria	X
Psychotria gracilenta	Rubiaceae	Psychotria	X
Psychotria hoffmannseggiana	Rubiaceae	Psychotria	X
Psychotria leiocarpa	Rubiaceae	Psychotria	X
Psychotria mapourioides	Rubiaceae	Psychotria	X
Psychotria nuda	Rubiaceae	Psychotria	X
Psychotria racemosa	Rubiaceae	Psychotria	X
Psychotria sessilis	Rubiaceae	Psychotria	X
Psychotria suterella	Rubiaceae	Psychotria	X
Psychotria vellosiana	Rubiaceae	Psychotria	X
Rudgea jasminoides	Rubiaceae	Rudgea	X
Rudgea recurva	Rubiaceae	Rudgea	X
Tocoyena bullata	Rubiaceae	Tocoyena	X
Tocoyena formosa	Rubiaceae	Tocoyena	X
Amaranthus hybridus	Amaranthaceae	Amaranthus	
Chamissoa altissima	Amaranthaceae	Chamissoa	
Anacardium occidentale	Anacardiaceae	Anacardium	
Lithrea molleoides	Anacardiaceae	Lithrea	

Mangifera indica	Anacardiaceae	Mangifera	
Schinus terebinthifolius	Anacardiaceae	Schinus	
Tapirira guianensis	Anacardiaceae	Tapirira	
Annona cacans	Annonaceae	Annona	X
Annona emarginata	Annonaceae	Annona	X
Annona neosericea	Annonaceae	Annona	X
Guatteria australis	Annonaceae	Guatteria	X
Guatteria sellowiana	Annonaceae	Guatteria	X
Xylopia aromatica	Annonaceae	Xylopia	X
Xylopia brasiliensis	Annonaceae	Xylopia	X
Xylopia langsdorfiana	Annonaceae	Xylopia	X
Xylopia sericea	Annonaceae	Xylopia	X
Anthurium affine	Araceae	Anthurium	X
Anthurium scandens	Araceae	Anthurium	X
Anthurium sellowianum	Araceae	Anthurium	X
Asterostigma lividum	Araceae	Asterostigma	X
Heteropsis oblongifolia	Araceae	Heteropsis	X
Heteropsis rigidifolia	Araceae	Heteropsis	X
Monstera adansonii	Araceae	Monstera	X
Philodendron appendiculatum	Araceae	Philodendron	X
Philodendron imbe	Araceae	Philodendron	X
Araucaria angustifolia	Araucariaceae	Araucaria	
Artocarpus heterophyllus	Moraceae	Artocarpus	X
Ficus carica	Moraceae	Ficus	X
Ficus benjami	Moraceae	Ficus	X
Ficus benjamina	Moraceae	Ficus	X
Ficus carica	Moraceae	Ficus	X
Ficus cestrifolia	Moraceae	Ficus	X
Ficus citrifolia	Moraceae	Ficus	X
Ficus enormis	Moraceae	Ficus	X
Ficus eximia	Moraceae	Ficus	X
Ficus guaranitica	Moraceae	Ficus	X
Ficus hirsuta	Moraceae	Ficus	X
Ficus insipida	Moraceae	Ficus	X
Ficus luschnathiana	Moraceae	Ficus	X
Ficus luschthia	Moraceae	Ficus	X
Ficus microcarpa	Moraceae	Ficus	X
Ficus organensis	Moraceae	Ficus	X
Ficus pertusa	Moraceae	Ficus	X
Ficus trigona	Moraceae	Ficus	X
Maclura tinctoria	Moraceae	Maclura	X
Morus alba	Moraceae	Morus	X
Morus nigra	Moraceae	Morus	X
Sorocea bonplandii	Moraceae	Sorocea	X
Byrsonima cydoniifolia	Malpighiaceae	Byrsonima	
Byrsonima ligustrifolia	Malpighiaceae	Byrsonima	
Byrsonima sericea	Malpighiaceae	Byrsonima	
Byrsonima variabilis	Malpighiaceae	Byrsonima	
Malpighia glabra	Malpighiaceae	Malpighia	
Cabralea canjerana	Meliaceae	Cabralea	X
Guarea guidonia	Meliaceae	Guarea	X
Guarea kunthiana	Meliaceae	Guarea	X
Guarea macrophylla	Meliaceae	Guarea	X
Melia azedarach	Meliaceae	Melia	X
Trichilia catigua	Meliaceae	Trichilia	X
Trichilia clausseni	Meliaceae	Trichilia	X
Trichilia elegans	Meliaceae	Trichilia	X
Trichilia pallida	Meliaceae	Trichilia	X
Calophyllum brasiliense	Calophyllaceae	Calophyllum	

Calyptranthes clusiifolia	Myrtaceae	Calyptranthes	X
Calyptranthes concinna	Myrtaceae	Calyptranthes	X
Campomanesia guaviroba	Myrtaceae	Campomanesia	X
Campomanesia guazumifolia	Myrtaceae	Campomanesia	X
Campomanesia neriiflora	Myrtaceae	Campomanesia	X
Campomanesia phaea	Myrtaceae	Campomanesia	X
Campomanesia xanthocarpa	Myrtaceae	Campomanesia	X
Eugenia astringens	Myrtaceae	Eugenia	X
Eugenia brasiliensis	Myrtaceae	Eugenia	X
Eugenia cerasiflora	Myrtaceae	Eugenia	X
Eugenia cuprea	Myrtaceae	Eugenia	X
Eugenia florida	Myrtaceae	Eugenia	X
Eugenia handroi	Myrtaceae	Eugenia	X
Eugenia hiemalis	Myrtaceae	Eugenia	X
Eugenia involucrata	Myrtaceae	Eugenia	X
Eugenia melanogyna	Myrtaceae	Eugenia	X
Eugenia mosenii	Myrtaceae	Eugenia	X
Eugenia neoglomerata	Myrtaceae	Eugenia	X
Eugenia oblongata	Myrtaceae	Eugenia	X
Eugenia pyriformis	Myrtaceae	Eugenia	X
Eugenia umbelliflora	Myrtaceae	Eugenia	X
Eugenia uniflora	Myrtaceae	Eugenia	X
Eugenia uruguayensis	Myrtaceae	Eugenia	X
Eugenia verticillata	Myrtaceae	Eugenia	X
Marlierea neuwiediana	Myrtaceae	Marlierea	X
Marlierea obscura	Myrtaceae	Marlierea	X
Marlierea reitzii	Myrtaceae	Marlierea	X
Marlierea suaveolens	Myrtaceae	Marlierea	X
Marlierea tomentosa	Myrtaceae	Marlierea	X
Myrceugenia myrcioides	Myrtaceae	Myrceugenia	X
Myrcia anacardiifolia	Myrtaceae	Myrcia	X
Myrcia brasiliensis	Myrtaceae	Myrcia	X
Myrcia ferruginea	Myrtaceae	Myrcia	X
Myrcia hartwegiana	Myrtaceae	Myrcia	X
Myrcia hebepetala	Myrtaceae	Myrcia	X
Myrcia ilheosensis	Myrtaceae	Myrcia	X
Myrcia oblongata	Myrtaceae	Myrcia	X
Myrcia palustris	Myrtaceae	Myrcia	X
Myrcia pubipetala	Myrtaceae	Myrcia	X
Myrcia pulchra	Myrtaceae	Myrcia	X
Myrcia spectabilis	Myrtaceae	Myrcia	X
Myrcia splendens	Myrtaceae	Myrcia	X
Myrcia tomentosa	Myrtaceae	Myrcia	X
Myrciaria glomerata	Myrtaceae	Myrciaria	X
Myrciaria cuspidata	Myrtaceae	Myrciaria	X
Myrciaria floribunda	Myrtaceae	Myrciaria	X
Myrciaria trunciflora	Myrtaceae	Myrciaria	X
Myrrhinium atropurpureum	Myrtaceae	Myrrhinium	X
Neomitranthes glomerata	Myrtaceae	Neomitranthes	X
Neomitranthes obscura	Myrtaceae	Neomitranthes	X
Plinia cauliflora	Myrtaceae	Plinia	X
Psidium cattleianum	Myrtaceae	Psidium	X
Psidium guajava	Myrtaceae	Psidium	X
Siphoneugena densiflora	Myrtaceae	Siphoneugena	X
Syzygium cumini	Myrtaceae	Syzygium	X
Carica papaya	Caricaceae	Carica	
Jacaratia spinosa	Caricaceae	Jacaratia	
Casearia decandra	Salicaceae	Casearia	
Casearia sylvestris	Salicaceae	Casearia	

Cecropia glaziovii	Urticaceae	Cecropia	
Cecropia hololeuca	Urticaceae	Cecropia	
Cecropia pachystachya	Urticaceae	Cecropia	
Coussapoa microcarpa	Urticaceae	Coussapoa	
Pourouma guianensis	Urticaceae	Pourouma	
Urera baccifera	Urticaceae	Urera	
Celtis iguanaea	Cannabaceae	Celtis	
Trema micrantha	Cannabaceae	Trema	
Cerastium glomeratum	Caryophyllaceae	Cerastium	
Cereus fernambucensis	Cactaceae	Cereus	X
Cereus hildmannianus	Cactaceae	Cereus	X
Opuntia monacantha	Cactaceae	Opuntia	X
Pereskia aculeata	Cactaceae	Pereskia	X
Pilosocereus arrabidae	Cactaceae	Pilosocereus	X
Rhipsalis campos-portoana	Cactaceae	Rhipsalis	X
Rhipsalis elliptica	Cactaceae	Rhipsalis	X
Rhipsalis paradoxa	Cactaceae	Rhipsalis	X
Rhipsalis teres	Cactaceae	Rhipsalis	X
Stephanocereus luetzelburgii	Cactaceae	Stephanocereus	X
Chrysophyllum flexuosum	Sapotaceae	Chrysophyllum	
Chrysophyllum gonocarpum	Sapotaceae	Chrysophyllum	
Chrysophyllum viride	Sapotaceae	Chrysophyllum	
Cinnamodendron dinisii	Canellaceae	Cinnamodendron	
Cissus paulliniifolia	Vitaceae	Cissus	
Cissus selloana	Vitaceae	Cissus	
Cissus striata	Vitaceae	Cissus	
Cissus verticillata	Vitaceae	Cissus	
Citharexylum myrianthum	Verbenaceae	Citharexylum	
Duranta erecta	Verbenaceae	Duranta	
Lantana camara	Verbenaceae	Lantana	
Lantana pohliana	Verbenaceae	Lantana	
Citrus reticulata	Rutaceae	Citrus	
Citrus x aurantium	Rutaceae	Citrus	
Clausena excavata	Rutaceae	Clausena	
Murraya paniculata	Rutaceae	Murraya	
Zanthoxylum hyemale	Rutaceae	Zanthoxylum	
Zanthoxylum rhoifolium	Rutaceae	Zanthoxylum	
Zanthoxylum riedelianum	Rutaceae	Zanthoxylum	
Clidemia hirta	Melastomataceae	Clidemia	X
Clidemia urceolata	Melastomataceae	Clidemia	X
Henriettea saldanhaei	Melastomataceae	Henriettea	X
Leandra acutiflora	Melastomataceae	Leandra	X
Leandra aurea	Melastomataceae	Leandra	X
Leandra australis	Melastomataceae	Leandra	X
Leandra barbinervis	Melastomataceae	Leandra	X
Leandra carassana	Melastomataceae	Leandra	X
Leandra laevigata	Melastomataceae	Leandra	X
Leandra melastomoides	Melastomataceae	Leandra	X
Leandra pilonensis	Melastomataceae	Leandra	X
Leandra refracta	Melastomataceae	Leandra	X
Leandra regnellii	Melastomataceae	Leandra	X
Leandra sabiaensis	Melastomataceae	Leandra	X
Leandra variabilis	Melastomataceae	Leandra	X
Leandra xanthocoma	Melastomataceae	Leandra	X
Miconia affinis	Melastomataceae	Miconia	X
Miconia albicans	Melastomataceae	Miconia	X
Miconia alborufescens	Melastomataceae	Miconia	X
Miconia brasiliensis	Melastomataceae	Miconia	X
Miconia budlejoides	Melastomataceae	Miconia	X

Miconia cabucu	Melastomataceae	Miconia	X
Miconia chartacea	Melastomataceae	Miconia	X
Miconia cinerascens	Melastomataceae	Miconia	X
Miconia cinnamomifolia	Melastomataceae	Miconia	X
Miconia collatata	Melastomataceae	Miconia	X
Miconia cubatanensis	Melastomataceae	Miconia	X
Miconia cuspidata	Melastomataceae	Miconia	X
Miconia discolor	Melastomataceae	Miconia	X
Miconia elegans	Melastomataceae	Miconia	X
Miconia inaequidens	Melastomataceae	Miconia	X
Miconia inconspicua	Melastomataceae	Miconia	X
Miconia latecrenata	Melastomataceae	Miconia	X
Miconia ligustroides	Melastomataceae	Miconia	X
Miconia minutiflora	Melastomataceae	Miconia	X
Miconia paniculata	Melastomataceae	Miconia	X
Miconia pepericarpa	Melastomataceae	Miconia	X
Miconia prasina	Melastomataceae	Miconia	X
Miconia pusilliflora	Melastomataceae	Miconia	X
Miconia racemifera	Melastomataceae	Miconia	X
Miconia rubiginosa	Melastomataceae	Miconia	X
Miconia sellowiana	Melastomataceae	Miconia	X
Miconia tentaculifera	Melastomataceae	Miconia	X
Miconia theizans	Melastomataceae	Miconia	X
Miconia tristis	Melastomataceae	Miconia	X
Miconia urophylla	Melastomataceae	Miconia	X
Miconia valtheri	Melastomataceae	Miconia	X
Ossaea amygdaloides	Melastomataceae	Ossaea	X
Clusia criuva	Clusiaceae	Clusia	
Clusia hilariana	Clusiaceae	Clusia	
Clusia lanceolata	Clusiaceae	Clusia	
Clusia organensis	Clusiaceae	Clusia	
Garcinia gardneriana	Clusiaceae	Garcinia	
Codonanthe cordifolia	Gesneriaceae	Codonanthe	
Cordia abyssinica	Boraginaceae	Cordia	X
Cordia axillaris	Boraginaceae	Cordia	X
Cordia corymbosa	Boraginaceae	Cordia	X
Cordia ecalyculata	Boraginaceae	Cordia	X
Cordia sellowiana	Boraginaceae	Cordia	X
Cordia silvestris	Boraginaceae	Cordia	X
Myriopus paniculatus	Boraginaceae	Myriopus	X
Varronia curassavica	Boraginaceae	Varronia	X
Costus spiralis	Costaceae	Costus	
Curatella americana	Dilleniaceae	Curatella	
Davilla elliptica	Dilleniaceae	Davilla	
Davilla rugosa	Dilleniaceae	Davilla	
Doliocarpus dentatus	Dilleniaceae	Doliocarpus	
Cybianthus peruvianus	Primulaceae	Cybianthus	
Myrsine coriacea	Primulaceae	Myrsine	
Myrsine ferruginea	Primulaceae	Myrsine	
Myrsine gardneriana	Primulaceae	Myrsine	
Myrsine lancifolia	Primulaceae	Myrsine	
Myrsine umbellata	Primulaceae	Myrsine	
Myrsine venosa	Primulaceae	Myrsine	
Daphnopsis brasiliensis	Thymelaeaceae	Daphnopsis	
Dendropanax cuneatus	Araliaceae	Dendropanax	
Hedera nepalensis	Araliaceae	Hedera	
Schefflera actinophylla	Araliaceae	Schefflera	
Schefflera angustissima	Araliaceae	Schefflera	
Schefflera arboricola	Araliaceae	Schefflera	

Schefflera macrocarpa	Araliaceae	Schefflera	
Schefflera morototoni	Araliaceae	Schefflera	
Dichorisandra thyrsiflora	Commelinaceae	Dichorisandra	
Diospyros inconstans	Ebenaceae	Diospyros	
Diospyros kaki	Ebenaceae	Diospyros	
Drimys brasiliensis	Winteraceae	Drimys	
Drimys winteri	Winteraceae	Drimys	
Eriobotrya japonica	Rosaceae	Eriobotrya	X
Prunus myrtifolia	Rosaceae	Prunus	X
Prunus persica	Rosaceae	Prunus	X
Pyracantha coccinea	Rosaceae	Pyracantha	X
Rubus brasiliensis	Rosaceae	Rubus	X
Rubus erythroclados	Rosaceae	Rubus	X
Rubus rosifolius	Rosaceae	Rubus	X
Rubus urticifolius	Rosaceae	Rubus	X
Erythroxylum ambiguum	Erythroxylaceae	Erythroxylum	
Erythroxylum argentinum	Erythroxylaceae	Erythroxylum	
Erythroxylum deciduum	Erythroxylaceae	Erythroxylum	
Erythroxylum gonocladum	Erythroxylaceae	Erythroxylum	
Erythroxylum pauferrense	Erythroxylaceae	Erythroxylum	
Erythroxylum pulchrum	Erythroxylaceae	Erythroxylum	
Erythroxylum simonis	Erythroxylaceae	Erythroxylum	
Frangula purshiana	Rhamnaceae	Frangula	
Hovenia dulcis	Rhamnaceae	Hovenia	
Scutia buxifolia	Rhamnaceae	Scutia	
Fuchsia regia	Onagraceae	Fuchsia	
Gaylussacia brasiliensis	Ericaceae	Gaylussacia	
Gaylussacia pulchra	Ericaceae	Gaylussacia	
Gaylussacia virgata	Ericaceae	Gaylussacia	
Guapira opposita	Nyctaginaceae	Guapira	
Guapira pernambucensis	Nyctaginaceae	Guapira	
Hedychium coronarium	Zingiberaceae	Hedychium	
Hedyosmum brasiliense	Chloranthaceae	Hedyosmum	
Heisteria silvianii	Olacaceae	Heisteria	
Hohenbergia ramageana	Bromeliaceae	Hohenbergia	
Humiria balsamifera	Humiriaceae	Humiria	
Hyeronima alchorneoides	Phyllanthaceae	Hyeronima	
Margaritaria nobilis	Phyllanthaceae	Margaritaria	
Richeria grandis	Phyllanthaceae	Richeria	
Hypochaeris brasiliensis	Asteraceae	Hypochaeris	
Ilex affinis	Aquifoliaceae	Ilex	
Ilex brevicuspis	Aquifoliaceae	Ilex	
Ilex microdonta	Aquifoliaceae	Ilex	
Ilex paraguariensis	Aquifoliaceae	Ilex	
Ilex pseudobuxus	Aquifoliaceae	Ilex	
Ilex theezans	Aquifoliaceae	Ilex	
Lasiacis sorghoidea	Poaceae	Lasiacis	
Megathyrsus maximus	Poaceae	Megathyrsus	
Triticum aestivum	Poaceae	Triticum	
Urochloa decumbens	Poaceae	Urochloa	
Urochloa plantaginea	Poaceae	Urochloa	
Ligustrum japonicum	Oleaceae	Ligustrum	
Ligustrum lucidum	Oleaceae	Ligustrum	
Magnolia champaca	Magnoliaceae	Magnolia	
Magnolia ovata	Magnoliaceae	Magnolia	
Marcgravia polyantha	Marcgraviaceae	Marcgravia	
Schwartzia brasiliensis	Marcgraviaceae	Schwartzia	
Maytenus aquifolia	Celastraceae	Maytenus	
Maytenus brasiliensis	Celastraceae	Maytenus	

Maytenus gonoclada	Celastraceae	Maytenus	
Maytenus littoralis	Celastraceae	Maytenus	
Schaefferia argentinensis	Celastraceae	Schaefferia	
Meliosma sellowii	Sabiaceae	Meliosma	
Melothria cucumis	Cucurbitaceae	Melothria	
Momordica charantia	Cucurbitaceae	Momordica	
Mollinedia boracensis	Monimiaceae	Mollinedia	
Mollinedia schottiana	Monimiaceae	Mollinedia	
Mollinedia triflora	Monimiaceae	Mollinedia	
Mollinedia uleana	Monimiaceae	Mollinedia	
Muntingia calabura	Muntingiaceae	Muntingia	
Musa paradisiaca	Musaceae	Musa	
Musa rosacea	Musaceae	Musa	
Ouratea polygyna	Ochnaceae	Ouratea	
Ouratea vaccinioides	Ochnaceae	Ouratea	
Passiflora actinia	Passifloraceae	Passiflora	
Passiflora edulis	Passifloraceae	Passiflora	
Peplonia organensis	Apocynaceae	Peplonia	
Peschiera catharinensis	Apocynaceae	Peschiera	
Tabernaemontana hystrix	Apocynaceae	Tabernaemontana	
Pera glabrata	Peraceae	Pera	
Phoradendron crassifolium	Santalaceae	Phoradendron	
Phoradendron piperoides	Santalaceae	Phoradendron	
Phoradendron quadrangulare	Santalaceae	Phoradendron	
Phytolacca dioica	Phytolaccaceae	Phytolacca	
Piper aduncum	Piperaceae	Piper	X
Piper amalago	Piperaceae	Piper	X
Piper corintoanum	Piperaceae	Piper	X
Piper dilatatum	Piperaceae	Piper	X
Piper gaudichaudianum	Piperaceae	Piper	X
Piper hispidinervum	Piperaceae	Piper	X
Piper miquelianum	Piperaceae	Piper	X
Piper mollicomum	Piperaceae	Piper	X
Piper tectoniifolium	Piperaceae	Piper	X
Podocarpus sellowii	Podocarpaceae	Podocarpus	
Protium heptaphyllum	Burseraceae	Protium	
Protium spruceanum	Burseraceae	Protium	
Protium widgrenii	Burseraceae	Protium	
Psittacanthus robustus	Loranthaceae	Psittacanthus	
Struthanthus concinnus	Loranthaceae	Struthanthus	
Struthanthus vulgaris	Loranthaceae	Struthanthus	
Quiina glazovii	Quiinaceae	Quiina	
Scaevola plumieri	Goodeniaceae	Scaevola	
Sloanea guianensis	Elaeocarpaceae	Sloanea	
Sloanea hirsuta	Elaeocarpaceae	Sloanea	
Smilax elastica	Smilacaceae	Smilax	
Smilax rufescens	Smilacaceae	Smilax	
Stromanthe tonckat	Marantaceae	Stromanthe	
Strychnos brasiliensis	Loganiaceae	Strychnos	
Styrax leprosus	Styracaceae	Styrax	
Styrax pohlii	Styracaceae	Styrax	
Symplocos estrellensis	Symplocaceae	Symplocos	
Symplocos glandulosomarginata	Symplocaceae	Symplocos	
Symplocos laxiflora	Symplocaceae	Symplocos	
Symplocos pubescens	Symplocaceae	Symplocos	
Symplocos revoluta	Symplocaceae	Symplocos	
Symplocos tetrandra	Symplocaceae	Symplocos	
Symplocos uniflora	Symplocaceae	Symplocos	
Turnera ulmifolia	Turneraceae	Turnera	

Virola bicuhyba	Myristicaceae	Virola
Virola gardneri	Myristicaceae	Virola
Virola sebifera	Myristicaceae	Virola
Vismia brasiliensis	Hypericaceae	Vismia