The Annals of Applied Statistics
2020, Vol. 14, No. 1, 221–240
https://doi.org/10.1214/19-AOAS1296

A HIERARCHICAL BAYESIAN MODEL FOR PREDICTING ECOLOGICAL INTERACTIONS USING SCALED EVOLUTIONARY RELATIONSHIPS

BY MOHAMAD ELMASRI[1,*], MAXWELL J. FARRELL[2], T. JONATHAN DAVIES[3] AND DAVID A. STEPHENS[1,**]

[1]*Department of Mathematics and Statistics, McGill University,* [*]*mohamad.elmsari@mail.mcgill.ca;* [**]*david.stephens@mcgill.ca*

[2]*Department of Biology, McGill University, maxwell.farrell@mail.mcgill.ca*

[3]*Departments of Botany, and Forest and Conservation Sciences, University of British Columbia, j.davies@ubc.ca*

Identifying undocumented or potential future interactions among species is a challenge facing modern ecologists. Recent link prediction methods rely on trait data; however, large species interaction databases are typically sparse and covariates are limited to only a fraction of species. On the other hand, evolutionary relationships, encoded as phylogenetic trees, can act as proxies for underlying traits and historical patterns of parasite sharing among hosts. We show that, using a network-based conditional model, phylogenetic information provides strong predictive power in a recently published global database of host-parasite interactions. By scaling the phylogeny using an evolutionary model, our method allows for biological interpretation often missing from latent variable models. To further improve on the phylogeny-only model, we combine a hierarchical Bayesian latent score framework for bipartite graphs that accounts for the number of interactions per species with host dependence informed by phylogeny. Combining the two information sources yields significant improvement in predictive accuracy over each of the submodels alone. As many interaction networks are constructed from presence-only data, we extend the model by integrating a correction mechanism for missing interactions which proves valuable in reducing uncertainty in unobserved interactions.

1. Introduction. As we enter into a data revolution in the study of biodiversity (La Salle, Williams and Moritz (2016)), global databases of species interactions are becoming readily available (Wardeh et al. (2015), Stephens et al. (2017), Poelen, Simons and Mungall (2014)). However, most ecological networks that represent the interactions among organisms are only partially observed, and fully characterizing all interactions via systematic sampling involves substantial effort that is not feasible in most situations (Jordano (2016)). Approaches to predict highly probable, yet previously undocumented links in ecological networks will help to expand our understanding of biodiversity, and can aid in the proactive surveillance of pathogens that infect multiple host species (Farrell, Berrang-Ford and Davies (2013)).

Many potential approaches exist for link prediction in networks; a large group of them can be classified under covariate or feature models, where covariates of a pair of nodes are used to determine the likelihood of their interaction. The latent space model, introduced by Hoff, Raftery and Handcock (2002), came to augment the former approach by representing each node (i) as a point s_i in a latent low-dimensional space. The likelihood of the edge (i, j) is driven by the individual covariates of each node and a form of distance $d(s_i, s_j)$ of the corresponding pairs in the latent space. Such an approach proved valuable in link prediction for social networks for many reasons, including: (i) the abundance of covariate data in social networks, and (ii) most applications favour predictive power over interpretability.

Received January 2019; revised July 2019.

Key words and phrases. Ecological networks, composite likelihood, iterated conditional modes, presence-only networks, link prediction.

A number of recent approaches for link prediction in ecological networks rely on trait data and node-specific features, such as body size or similarity of trophic interactions (Williams and Martinez (2000), Petchey et al. (2008), Gravel et al. (2013), Bartomeus (2013), Stock et al. (2017), Dallas, Park and Drake (2017), Bastazini et al. (2017), Olival et al. (2017)). While these approaches work well for small-scale datasets, they scale poorly to large-scale ecological datasets in which traits determining species interactions are often unknown or are available only for a limited subset of species (Morales-Castilla et al. (2015)). When trait information is limited, evolutionary relationships among species may be used as a proxy to study species interactions (Webb et al. (2002)). Phylogenetic trees are a representation of the evolutionary relationships among species which provide means to quantify ecological similarity (Wiens et al. (2010)) and co-evolutionary history (Davies and Pedersen (2008)). Just as many species traits co-vary with phylogeny, species interactions are also phylogenetically structured (Gómez, Verdú and Perfectti (2010)). Incorporating phylogeny into ecological link prediction has the added benefit that it is universally applicable across all systems and offers added biological interpretability over current latent variable models.

Different approaches have been proposed to incorporate phylogeny-based similarity in link prediction (Ovaskainen et al. (2016, 2017), Chiu and Westveld (2011), Bastazini et al. (2017), Pearse and Altermatt (2013)). Despite the emerging interest in this topic, currently proposed models treat the phylogeny as fixed or linearly scaled and do not offer approaches to capture the underlying evolutionary processes that determine species differences.

Evolutionary biologists have developed methods of transforming phylogenies to represent alternative modes of evolution (Pagel (1999), Harmon et al. (2010)). Rescaling the tree using these approaches alters the dependence structure among hosts, yielding improved predictions that can also be interpreted in the context of a model of trait evolution. This allows for added flexibility in the incorporation of phylogenetic information, as the dissimilarity of potential traits underlying ecological interactions may evolve under different processes than that expected by the inferred phylogeny.

In this work we show that single-parameter (nonlinear) tree scaling based on evolutionary models improves predictive performance and allows for predictions that would otherwise be overlooked by contemporary link prediction models. Shifting away from linearity results in theoretical and computational issues. Theoretically, the conditional nature of phylogenies forces interaction probabilities to be specified conditionally on other interactions, hence, the joint distribution (if it exists) might be inaccessible. As a consequence, efficient and scalable sampling methods are required, as proposed in this work. To our knowledge, this work is the first to incorporate phylogenetic scaling in link prediction.

We develop a phylogeny-based framework for predicting undocumented links using a recent global database of host-parasite interactions (Stephens et al. (2017)). In host-parasite networks, parasite community similarity is often constrained by evolutionary distances among hosts (Gilbert and Webb (2007), Davies and Pedersen (2008), Streicker et al. (2010), Braga, Razzolini and Boeger (2015), Huang et al. (2015)). We focus on wild mammal hosts that are most closely related to domesticated ungulates and carnivores, as these species are known to harbour diseases of concern for humans and livestock (Cleaveland, Laurenson and Taylor (2001)) and include many species that are threatened with extinction due to infectious diseases (Pedersen et al. (2007)). We incorporate phylogenetic information as a weighted network, where weights quantify pairwise host similarities. This approach allows for easy expansion to different forms of dependency, if phylogenetic information is unavailable or if other dependency structures are preferred. However, we show that phylogenetic information alone can generate accurate point estimate predictions. We improve our initial point estimate by incorporating a single-parameter tree scaling model which results in posterior distributions for the probabilities of each host-parasite interaction.

We then show that this phylogeny-only model can be extended by using node-specific affinity (sociability) parameters, mimicking that of covariate-based network models, such as Hoff, Raftery and Handcock (2002), Hoff (2005), Chung and Lu (2006), Bickel and Chen (2009).

To facilitate the construction of the full joint distribution, we first augment the model using a hierarchical latent variable framework. The latent variable acts as an underlying scoring system, with higher scores attributed to more probable links. Second, we apply a method similar to the iterated conditional modes approach in auto-dependent models of Besag (1974) to deal with the conditional dependency imposed by phylogeny and include a method to account for uncertainty in unobserved interactions. Our approach allows for robust predictions for large species interaction networks with limited covariate data and can be extended to any bipartite network with a dependency structure for one of the interacting classes.

2. Data. We illustrate our framework on the Global Mammal Parasite Database version 2.0 (GMPD), described in Stephens et al. (2017). The GMPD contains over 24,000 documented associations between hosts and their parasites collected from published reports and scientific studies. The assumed interactions are based on empirical observations of associations between host-parasite pairs using a variety of evidence types (visual identification, serological tests or detection of genetic material from a parasite species in one or more host individuals). Associations are reported along with their publication reference. The GMPD gathers data on wild mammals and their parasites (including both micro- and macroparasites) which are separated into three primary databases based on host taxonomy: Primates, Carnivora, and ungulates (terrestrial hooved mammals in the orders Artiodactyla and Perissodactyla). We restricted our analyses to the ungulate and Carnivora subsets because of prior experience with these data (Farrell et al. (2015)) and tractability of the size of the resulting network.

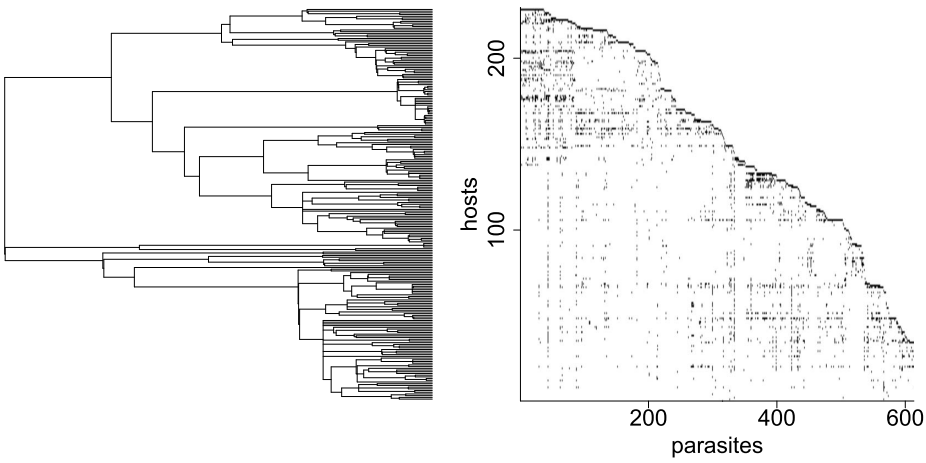
We use the ungulate and Carnivora subsets of the GMPD to construct a bipartite binary matrix, where rows represent hosts and columns parasites and documented associations (at least one piece of evidence that a parasite infects a given host species) are indicated by 1. We construct host pairwise similarities as the inverse of phylogenetic distances calculated from the mammal phylogeny of Fritz, Bininda-Emonds and Purvis (2009) which involved collapsing host subspecies to species. We excluded parasites that were not reported to species level. This resulted in a GMPD subset with 4178 pairs of interactions among 236 hosts and 1308 parasites. Out of these 1308 parasites, 695 were found to associate with a single host ($\approx$54% of parasites and $\approx$17% of total interactions).

One of the models proposed in Section 3.1 (the phylogeny-only model) can only be specified for multihost parasites. Thus, for the purpose of model comparison we remove single-host parasites, reducing the GMPD to 3483 interactions among 229 hosts and 613 parasites. In subsequent analyses we refer to the database without single-host parasites, unless otherwise specified.

Figure 1 shows the left-ordered interaction matrix $\mathbf{Z}$ of GMPD and corresponding host phylogeny. The matrix $\mathbf{Z}$ is sparse, and the degree distributions of both hosts and parasites exhibit a power-law structure (Supplementary Material, Figure S5, Elmasri et al. (2020)).

3. Bayesian hierarchical model for prediction of ecological interactions.

3.1. *Network-based latent score model.* Conditional modelling is common in many biological network models, where the class of auto-models of Besag (1974) and latent space models of Hoff, Raftery and Handcock (2002) are widely applied. One example is the use of a network-based auto-probit model in Jiang, Gold and Kolaczyk (2011), where a protein-protein association network is used as a prior to predict protein functional roles conditional



(a) host phylogeny (b) GMPD

FIG. 1. (a) *The host phylogeny and* (b) *the left-ordered interaction matrix* $\mathbf{Z}$ *of the GMPD without single-host parasites.*

on the roles of neighbouring proteins. Such network-based models rely on a pre-existing binary or weighted network with a clearly defined neighbourhood structure. Probabilities are then derived by averaging over neighbouring nodes.

Evolutionary distances among species, represented by phylogenies, translate to a fully connected weighted network. Since pairwise distances among species are measured relative to their most recent common ancestor, the same distance may be assigned to multiple host pairs. A neighbourhood structure can be constructed with weights on the fully connected network, or a threshold method can be applied, but with two main drawbacks: (i) the complexity of inferring the threshold parameter, and (ii) the interpretation of the threshold with respect to evolutionary distance.

In the case of host-parasite interactions, parasites are often found to interact with closely related hosts, but in some cases may make large jumps in the phylogeny and interact with distantly related hosts (Parrish et al. (2008), Park et al. (2018)). To account for such behaviour and to overcome the drawbacks of the threshold method, we let the probability of a host-parasite interaction be driven by the sum of evolutionary distances to the documented hosts of the parasite.

Let $\mathbf{Z}$ be an $H \times J$ host-parasite interaction matrix, where the binary variable z_{hj} denotes whether an interaction between host h and parasite j has been observed. Quantifying divergences starting from the root of the tree, let T_{hi} be a unit-free pairwise phylogenetic distances among hosts h and i, and their common ancestor k, such that $T_{hi} = T_{hk} + T_{ik} = (t_h - t_k) + (t_i - t_k)$. Phylogenetic distances are commonly measured in millions of years, so to arrive at the unit-free distance we divide all distances by the total depth of the tree.

A valid and basic conditional probability distribution of host h interacting with parasite j can be defined in terms of the pairwise phylogenetic distances from host h to all other hosts interacting with parasite j, as

$$\mathbb{P}(z_{hj} = 1 \mid \mathbf{z}_{(-h)j}) = 1 - \exp(-\delta_{hj}), \quad \delta_{hj} = \sum_{\substack{i=1 \\ i \neq h}}^{H} \frac{z_{hj}}{T_{hi}}, \tag{1}$$

where $\mathbf{z}_{(-h)j}$ is the set of interactions of the jth parasite among the H hosts ($\mathbf{z}_{\cdot j} = (z_{1j}, \ldots, z_{Hj})$), excluding that of the hth host.

The conditional distribution (1) allocates higher probabilities when closely related hosts interact with a given parasite or when many distantly related hosts also interact. The more distantly related the hosts are, the smaller the value of $1/T_{hi}$. Of course, the probability distribution in (1) is conditional on a probabilistic model for T.

The exponential choice in (1) is motivated by the power-law structure shown in Figure 1 and Supplementary Material Figure S5 (Elmasri et al. (2020)), thus we expect interaction probabilities to decay exponentially with respect to the parameters. Other probability structures are viable, though with no tractability guarantees. We later show that, under such construction, a tractable probabilistic framework is possible for a class of latent variables with tail probabilities as in (1). The next section introduces a family of single-parameter models that have meaningful biological interpretation.

3.1.1. *Evolutionary models and phylogeny transformations.* A focus of macroevolutionary research has been to develop models of trait evolution. A well-known model, and default in many ecological applications, is Brownian motion; however, transformations of the phylogenetic tree can be made to reflect alternatives in the tempo and mode of evolution. Common evolutionary models that can be defined by a single parameter transformation include the early-burst (EB), delta, kappa, lambda and the Ornstein–Uhlenbeck transformation (Pagel (1999), Harmon et al. (2010)); each scales phylogenetic distances according to a model of evolution. We term this the phylogeny-only model and define it as

$$\mathbb{P}(z_{hj} = 1 \mid \mathbf{z}_{(-h)j}) = 1 - \exp(-\delta_{hj}), \quad \delta_{hj} = \sum_{\substack{i=1 \\ i \neq h}}^{H} \frac{z_{hj}}{\phi(T_{hi}, \eta)}, \tag{2}$$

where $\phi(T_{hi}, \eta)$ is the transformed distance under a given evolutionary model controlled by a single parameter η.

With further investigation we find that the EB model stands out by displaying a nontrivial convex relationship with predictive power, as shown in Figure 2. In this figure we are evaluating the potential predictive accuracy of a simple phylogeny-only model (2) with phylogeny scaled according to the early-burst method for the database in Section 2 (for more details refer to Supplementary Material, Figure S1, Elmasri et al. (2020)).

This supports the assumption that scaled phylogenies, based on explicit models of niche or trait evolution, can result in better predictions. The EB model allows evolutionary change to

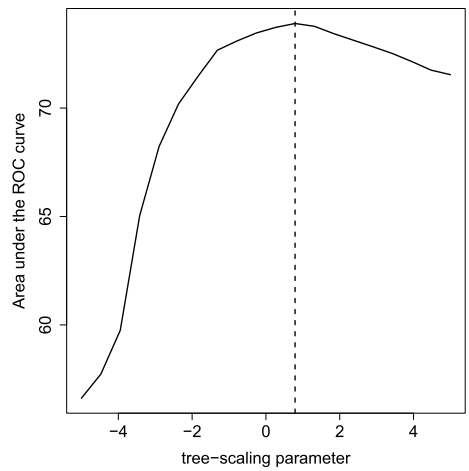
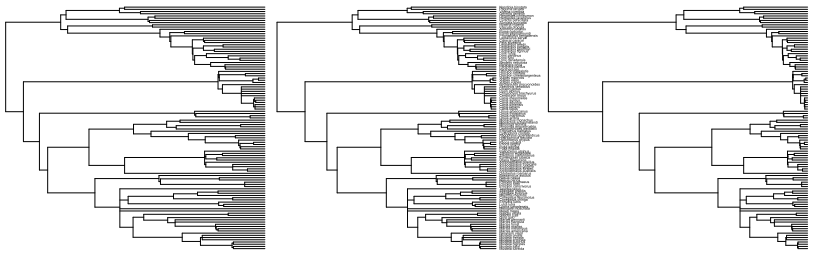


FIG. 2. *Area under the ROC curve evaluated over a fine grid under the phylogeny-only model* (2) *with early-burst tree transformational method with GMPD* (*including single-host parasites*).



(a) $EB(T, 0.02)$ (b) T (c) $EB(T, -0.02)$

FIG. 3. *Examples of the early-burst transformation in the Carnivora subset of GMPD.*

accelerate or decelerate through time, for example, evolutionary change may be fastest early in a clades history but slows through time. The rate of change in the EB model is adjusted by a single parameter $\eta \in \mathbb{R}$, with positive values of η indicating that evolution is faster earlier in history, while negative values suggests the opposite. Figure 3 illustrates the EB model for different values of η.

Under the EB model the phylogenetic distance between a pair of hosts (h, i) with a most recent common ancestor k is quantified as

$$\phi(T_{hi}, \eta) = \phi(T_{hk}, \eta) + \phi(T_{ik}, \eta) = \frac{1}{\eta}\big(e^{\eta t_h} - e^{\eta t_k}\big) + \frac{1}{\eta}\big(e^{\eta t_i} - e^{\eta t_k}\big). \tag{3}$$

Thus, for $\eta = 0$, EB reduces to the original tree distance as $\phi(T_{hi}, 0) = T_{hi}$. While this represents one form of uncertainty in the phylogeny, future work may also incorporate uncertainty in tree topology as well as distance by using posterior distributions of trees resulting from Bayesian phylogenetic inference.

3.1.2. *Full model.* Species interactions can be predicted using phylogenetic trees, though not completely, since interactions can also be driven by traits that are independent of phylogeny. In general, many network-based models assume that edge probabilities are driven by independent node affinity parameters, for example, Chung and Lu (2006), Bickel and Chen (2009) and many others. Here, we model the conditional probability of an interaction by combining both sources of information, phylogenetic distances and individual species affinities. Affinity parameters govern the general propensity for each organism to interact with members of the other class; larger affinities correlate with higher likelihood that an organism will interact. Let $\gamma_h > 0$ be the affinity parameter of host h and $\rho_j > 0$ of parasite j. The full conditional model is then

$$\mathbb{P}(z_{hj} = 1 \mid \mathbf{Z}_{-(hj)}) = 1 - \exp\big(-\gamma_h \rho_j \delta_{hj}(\eta)\big), \tag{4}$$

with $\delta_{hj}(\eta)$ as in (2) under the EB transformation, and $\mathbf{Z}_{-(hj)}$ is the interaction matrix $\mathbf{Z}$, excluding z_{hj}. The default value is $\delta_{hj}(\eta) = 1$, if no neighbouring interactions exist, reducing to the affinity-only model for this interaction. Alternative defaults are possible, such as the average pairwise distances in T.

The affinity-only model results in a workable network prediction model which has been shown in the literature on exchangeable random networks (Hoff, Raftery and Handcock (2002)). However, affinity-only models tend to generate adjacency matrices with many hyperactive columns and rows. This is because whenever a node has a sufficiently high affinity parameter, it forms edges with almost all other nodes, which is likely to be unrealistic for most ecological networks. In Section 5 we show that both models, the affinity-only and

phylogeny-only, independently result in useful predictive models that represent some variation in the data. However, each model captures different characteristics of the network and by layering them we obtain a nontrivial improvement.

Finally, we find it advantageous to use latent variables in modelling the binary variables z_{hj}. This facilitates the construction of the network joint distribution while accounting for the Markov network dependency imposed by $\delta_{hj}(\eta)$. In addition, the latent variable construction becomes essential in addressing the ambiguity associated with the case when $z_{hj} = 0$, which entails two possibilities: a yet to be observed positive interaction, or a true absence of interaction due to incompatibility (implemented in Section 4).

Thus, for each z_{hj} we define latent score $s_{hj} \in \mathbb{R}$ such that

$$z_{hj} = \begin{cases} 1 & \text{if } s_{hj} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $s_{hj} \in \mathbb{R}$ is a continuous random variable acting as a latent score determining the probability of z_{hj} being an interaction. Although unobserved, s_{hj} completely determines the binary variables z_{hj}. Therefore, the conditional model in (4) can be completely specified in terms of the latent score as

$$\mathbb{P}(z_{hj} = 1 \mid \mathbf{Z}_{-(hj)}) = \mathbf{E}[\mathbb{I}_{\{s_{hj}>0\}} \mid \mathbf{Z}_{-(hj)}] = \mathbb{P}(s_{hj} > 0 \mid \mathbf{S}_{-(hj)}), \tag{6}$$

where $\mathbf{S}_{-(hj)}$ represents the interaction matrix $\mathbf{S}$ excluding s_{hj}; it replaces $\mathbf{Z}$ as it carries the same probability events in its sign distribution.

The current formulation is flexible in the choice of distribution for $s_{hj} \mid \mathbf{S}_{-(hj}$, the only imposed requirement is absolute continuity with an exponentially decaying tail probability as in (4). One possible choice is the Gumbel with mean parameter $\log(\gamma_h \rho_j \delta_{hj}(\eta))$ and a scale of 1. Since we are only interested in positive reals, we use a zero-inflated Gumbel distribution for the latent score with the following density:

$$\mathbf{p}(s_{hj} \mid \mathbf{S}_{-(hj)}) = \tau_{hj} \exp\bigl(-s_{hj} - \tau_{hj} e^{-s_{hj}}\bigr)\mathbb{I}_{\{s_{hj}>0\}} + \exp(-\tau_{hj})\mathbb{I}_{\{s_{hj}=0\}}, \tag{7}$$

where $\tau_{hj} = \gamma_h \rho_j \delta_{hj}(\eta)$. Hence, the conditional joint distribution becomes

$$\begin{aligned} \mathbb{P}(z_{hj} = 1, s_{hj} \mid \mathbf{Z}_{-(hj)}) &= \mathbb{P}(z_{hj} = 1 \mid s_{hj})\mathbf{p}(s_{hj} \mid \mathbf{S}_{-(hj)}) \\ &= \mathbf{p}(s_{hj} \mid \mathbf{S}_{-(hj)})\mathbb{I}_{\{s_{hj}>0\}}. \end{aligned} \tag{8}$$

The choice of a zero-truncated Gumbel was made to facilitate the construction of the joint distribution, in a manner similar to the Swendsen–Wang algorithm (Swendsen and Wang (1987)), where a product of densities transforms to a sum in the exponential scale, improving the tractability of posteriors. Alternatively, the truncated exponential distribution can be used, as $s_{hj} \sim \min\{1, \mathrm{Exp}(\tau_{hj})\}$, having the tail distribution in (1) though it does not admit the direct interpretability as a latent score, as does the Gumbel distribution.

The proposed latent score model, though intricate in formulation, is no more inferentially complex than the auto-logit model of Besag (1974). The reasons we use our model are: i) it exhibits a simple joint distribution for each row of $\mathbf{Z}$ conditional on all others; ii) there are simple posterior distributions for the affinity parameters; and iii) we have the ability to correct for uncertainty using the latent score formulation (shown in Section 4). We could adopt other network-based conditional models, for example, the family of auto-models by Besag (1986). One particular example is the multivariate Gaussian latent variable with a conditional mean structure of Jiang, Gold and Kolaczyk (2011), which has a similar complexity to the phylogeny-only model in (2), modelling each column of $\mathbf{Z}$ independently without affinity parameters. Other auto-models would also require the development of an efficient sampling scheme and to make a proper choice for the conditional interaction probability, as done here.

3.2. *Prior and posterior distribution of choice parameters*. By the Hammersley–Clifford theorem (Robert and Casella (2013)), it is straightforward to verify that the joint distribution exists, as briefly shown in Supplementary Material Section S3 (Elmasri et al. (2020)). Even though the form is complicated, we do not need to access the joint density and instead may use a Gibbs sampler as in Geman and Geman (1984). An iterative algorithm can then be used to sample from conditionally independent components of the joint distribution, with the posterior sample obtained by averaging. This approach is similar in spirit to the iterated conditional modes (ICM) algorithm of Besag (1986).

In the proposed model the joint distribution of rows are conditionally independent given the rest. Let $\mathbf{Z}_{-(h.)}$ be $\mathbf{Z}$ excluding the hth row $\mathbf{z}_{h.}$. With similar notations for $\mathbf{S}$, the joint distribution of the hth row is

$$\mathbb{P}(\mathbf{z}_{h.}, \mathbf{s}_{h.} \mid \mathbf{Z}_{-(h.)}) = \gamma_h^{n_h}\left[\prod_{j=1}^{J}\left(\rho_j \delta_{hj}(\eta)\right)^{z_{hj}}\right]\exp\left(-\sum_{j=1}^{J} s_{hj} + \gamma_h \rho_j \delta_{hj}(\eta) e^{-s_{hj}}\right), \tag{9}$$

where $n_h = \sum_{j=1}^{J} z_{hj}$ such that the row-wise joint posterior distribution is

$$\mathbb{P}(\mathbf{s}_{h.}, \gamma_h, \boldsymbol{\rho}, \eta \mid \mathbf{Z}) \propto \mathbb{P}(\mathbf{z}_{h.} \mid \mathbf{s}_{h.})\mathbb{P}(\mathbf{s}_{h.} \mid \mathbf{S}_{-(h.)}, \gamma_h, \boldsymbol{\rho}, \eta)\mathbb{P}(\gamma_h)\mathbb{P}(\boldsymbol{\rho})\mathbb{P}(\eta), \tag{10}$$

where $\mathbb{P}(\mathbf{z}_{h.} \mid \mathbf{s}_{h.}) = \prod_{j=1}^{J} \mathbb{P}(z_{hj} \mid s_{hj}) = 1$ and $\boldsymbol{\rho}$ is the parasite affinity parameter set.

In a sweeping manner for $h = 1, \ldots, H$ rows of $\mathbf{Z}$, one samples γ_h from its full posterior, and $\boldsymbol{\rho}^{(h)} = (\rho_1^{(h)}, \ldots, \rho_J^{(h)})$ and $\eta^{(h)}$ from their hth row conditional posteriors. Obtaining an MCMC sample of $\boldsymbol{\rho}$ and η is done by averaging over the H samples from the row posteriors.

For prior specifications we choose a gamma distribution for both affinity parameters because of their conjugacy property. Thus, let $\gamma_h \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_\gamma, \tau_\gamma)$ and $\rho_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_\rho, \tau_\rho)$. The full posterior distributions of γ_h and the h-row partial posterior of $\rho_j^{(h)}$, respectively, are

$$\begin{aligned} \rho_j^{(h)} \mid \mathbf{z}_{h.}, \mathbf{s}_{h.} &\sim \text{Gamma}\left(\alpha_\rho + z_{hj}, \tau_\rho + \gamma_h \delta_{hj}(\eta) e^{-s_{hj}}\right), \\ \gamma_h \mid \mathbf{z}_{h.}, \mathbf{s}_{h.} &\sim \text{Gamma}\left(\alpha_\gamma + n_h, \tau_\gamma + \sum_{j=1}^{J} \rho_j \delta_{hj}(\eta) e^{-s_{hj}}\right). \end{aligned} \tag{11}$$

In the case of the scaling parameter η, we assume a constant prior for simplicity and computational stability, although this could be readily modified to any subjective prior.

The latent score is updated, given all other parameters as

$$s_{hj} \mid z_{hj}, \mathbf{S}_{-(hj)} \sim \begin{cases} \chi_0 & \text{if } z_{hj} = 0, \\ \text{tGumbel}\left(\log \gamma_h \rho_j + \log \delta_{hj}(\eta), 1, 0\right) & \text{if } z_{hj} = 1, \end{cases} \tag{12}$$

where χ_0 is an atomic measure at zero and $\text{tGumbel}(\tau, 1, 0)$ is the zero-truncated Gumbel with density

$$\frac{\exp(-(s - \tau + e^{-(s-\tau)})}{1 - \exp(-e^{\tau})} \chi_{(0,\infty)}(s).$$

The adaptive Metropolis–Hastings algorithm (Haario, Saksman and Tamminen (2001)) within Gibbs is used to update the model parameter. For additional details on the model and the MCMC method sampling algorithm, refer to Supplementary Material Section S1 (Elmasri et al. (2020)).

4. Uncertainty in unobserved interactions. In ecological networks it is unlikely that all potential links will be represented or observed. Some unobserved exist but are undocumented due to limited or biased sampling, while others may be true absences or "forbidden" links (Morales-Castilla et al. (2015)). Evidence used to support an interaction will vary depending on the nature of the system, but it is often assumed that an interaction exists if at least one piece of evidence indicates so (Jordano (2016)).

This raises concern about the uncertainty of interactions in two ways. The first is due to uncertainty in documented interactions as false positive detection errors may occur, potentially, as a result of species misidentification, sample contamination or, for parasites, unanticipated cross-reactions in serological tests (Aguirre et al. (2007)). We believe it would be useful for the scientific community to identify weakly supported interactions that may require additional supporting evidence; however, our primary motivation is identification of "novel" interactions which is complicated by uncertainty in unobserved interactions.

The second concern arises when unobserved associations are, by default, assumed to be true absences. As discussed earlier, ecological networks are often undersampled, and some fraction of unobserved interactions may occur, but are currently undocumented, or represent potential interactions that are likely to occur given sufficient opportunity. Based on this assumption, we build a measure of uncertainty in unobserved interactions by modifying our proposed model in (4). In (5) we have assumed that $\mathbb{P}(z_{hj} = 1 \mid s_{hj} > 0)$ is degenerate at 1 given s_{hj}. Thus, we have only sampled positive scores for the case when $z_{hj} = 1$, as shown in (12). As a result the posterior predictive distribution is only considered for the case when a pair has no documented associations ($z_{hj} = 0$), underlining the assumption that the data is complete and trusted. In presence-only data the objective is to model the nontrivial object $\mathbb{P}(z_{hj} = 1, \text{"a missing link"} \mid s_{hj} > 0)$. To account for such uncertainty, we attempt to approximate the proportion of interactions that are missing links in the latent space by measuring the percentage of positive scores where the input is 0 ($z_{hj} = 0$) as

$$\mathbf{p}(z_{hj} = 0 \mid s_{hj}, g) = \begin{cases} 1 & \text{if } s_{hj} = 0, \\ g & \text{if } s_{hj} > 0, \end{cases} \tag{13}$$

where g is the probability that an interaction is unobserved when the latent score indicates an interaction should exist. If g is large and close to 1, it is likely that many of the unobserved interactions could or should exist. Introducing g to the model affects all parameter estimates and the notion of $\mathbf{Z}$. Therefore, the posterior predictive distribution is now considered for both cases. For the case of a documented association, the probability of an interaction is defined in (4), and for the case of no documentation the same probability is weighted by g as shown in detail in (14).

Here, we implicitly assume that g is common to all pairs of interactions. It is possible to assign a different parameter to groups of interactions. Nonetheless, by the principle of parsimony, we favoured simplicity. This kind of construction has been used earlier by Weir and Pettitt (2000), when modelling spatial distributions to account for uncertainty in regions with unobserved statistics, and later by Jiang, Gold and Kolaczyk (2011) in modelling uncertainty in protein functions.

4.1. *Markov chain Monte Carlo algorithm.* Introducing a measure of uncertainty in the model does not alter the MCMC sampling schemes introduced in Section 3.2. The variables $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ and η are still only associated with $\mathbf{S}$; nonetheless, by introducing the measure of uncer-

tainty the conditional sampling of each individual s_{hj} is now

$$\mathbf{p}(s_{hj} \mid \mathbf{S}_{-(hj)}, \mathbf{Z}, g) = \begin{cases} \dfrac{1}{\psi(\bar{s}_{hj})}\tau_{hj}\exp\bigl(-(s_{hj}+\tau_{hj}e^{-s_{hj}})\bigr), & s_{hj}>0, z_{hj}=1, \\ 0, & s_{hj}=0, z_{hj}=1, \\ \dfrac{g}{\theta(g,\bar{s}_{hj})}\tau_{hj}\exp\bigl(-(s_{hj}+\tau_{hj}e^{-s_{hj}})\bigr), & s_{hj}>0, z_{hj}=0, \\ \dfrac{1}{\theta(g,\bar{s}_{hj})}1-\psi(\bar{s}_{hj}), & s_{hj}=0, z_{hj}=0, \end{cases} \tag{14}$$

where $\tau_{hj} = \gamma_h \rho_j \delta_{hj}^{\eta}$, $\psi(\bar{s}_{hj}) = \int_0^{\infty} \mathbf{p}(s \mid \mathbf{S}_{-(hj)})\mathbf{d}s = 1 - \exp(-\gamma_h \rho_j \delta_{hj}^{\eta})$ and $\theta(g, \bar{s}_{hj}) = g\psi(\bar{s}_{hj}) + 1 - \psi(\bar{s}_{hj})$.

Sampling the uncertainty parameter is performed using the row-wise conditional distribution as

$$\mathbb{P}(g \mid \mathbf{s}_{h\cdot}, \mathbf{z}_{h\cdot}) \propto \mathbb{P}(\mathbf{z}_{h\cdot} \mid \mathbf{s}_{h\cdot}, g)\mathbb{P}(\mathbf{s}_{h\cdot} \mid \mathbf{S}_{-(h\cdot)})\mathbb{P}(g) \propto g^{N_{-+}}(1-g)^{N_{++}}, \tag{15}$$

where $N_{-+} = \#\{(h, j) : \mathbf{z}_{hj} = 0, s_{hj} > 0\}$, $N_{++} = \#\{(h, j) : z_{hj} = 1, s_{hj} > 0\}$ and $\mathbb{P}(g)$ is the uniform distribution. Since the sampling is done by iteratively cycling through the rows of $\mathbf{Z}$, in analogy to the ICM method, a sample of g is the average of the H row samples.

5. Alternative models and comparison by cross-validation. To validate the predictive performance of the proposed latent score full model, we compare it to the two submodels of Section 3.1 (the affinity-only and the phylogeny-only models) and to the bilinear latent-distance model with two of its submodels (the bilinear and the latent-distance models) (Hoff, Raftery and Handcock (2002), Hoff (2005)). The bilinear model excludes phylogenetic information and assumes a logit formulation with an intercept and an affinity coefficient for each node; hence, it correlates with the affinity-only model in interpretation. The latent-distance mode assumes a one-dimensional latent variable for each node, and pairwise distances between nodes are the Euclidean distance between their respective latent variables, aligning it with the phylogeny-only model in interpretation. Therefore, the latent-distance model excludes explicit phylogenetic information, and distances are not informed by the association matrix, as in the case of our phylogeny-only model. Additional latent dimensions can be added to the latent-distance model which might improve prediction, though at an extra cost of interpretation. The bilinear latent-distance model combines both former components, and all three variates of this model are implemented using latentnet R-package (Krivitsky and Handcock (2008, 2017)), as `ergmm(Z~rsociality)`, `ergmm(Z~euclidean(d=1))`, and `ergmm(Z~rsociality+euclidean(d=1))`, respectively. The latentnet package readily provides alternative forms of distances, though, for this dataset, we found that the Euclidean distance has a better performance.

Finally, we also compare our proposed model to a nearest-neighbour (NN) algorithm in which we set the distances between hosts proportional to the number of parasite species they share, also known as the Jaccard distance. This form of distance does not require additional data other than $\mathbf{Z}$. Hence, for this algorithm we let the probability of a host-parasite interaction be equal to the average number of host-neighbours with documented association to the parasite, within the k-closest host-neighbours. A host can share different sets of parasites with different hosts, though at times the size of the different sets might be the same, yielding exact Jaccard pairwise distances to multiple hosts. Therefore, we let k be driven by the number of shared parasites, excluding the parasite of interest. For example, $k = 2$ would define a neighbourhood of all hosts that have at least the second highest number of shared parasites for a host of interest. In brief, k is chosen by cross-validation; the details of the optimization criterion is discussed later.

Link probabilities in many network models, as the one proposed here, are driven by the count of links of their respective nodes. Hence, in cross-validation it is natural to hold a random portion of the observed links out from the training set and validate with them. In our settings the predictive performance of each model is evaluated using the average of five-fold cross-validations, where in each fold we set a different set of the observed interactions ($z_{hj} = 1$) in $\mathbf{Z}$ to unknowns ($z_{hj} = 0$) while attempting to predict them using the remaining interactions. The same folds are used across all evaluated models. The predictive performance of each fold is assessed methodologically, using the proper scoring rules proposed by Gneiting and Raftery (2007) and Ehm et al. (2016), graphically, using the receiver operating characteristic (ROC) curves, and numerically, using the percentage of recovered interactions.

The recent work of Ehm et al. (2016) has shown that, under unimportant regularity conditions, every score (loss) function consistent for the probability of binary events admits a representation as a mixture of the form

$$L(p, y) = \int_0^1 L_\theta(p, y)\, \mathrm{d}\mathbf{H}(\theta),$$

with $\mathbf{H}$ being a nonnegative measure and

$$L_\theta(p, y) = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1 - \theta, & y = 1,\ p \le \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

for a predictive probability p of binary event y, and $\theta \in [0, 1]$. The choice of the mixing measure $\mathbf{H}$ determines the score function. For example, when $\mathbf{H}$ is twice the Lebesgue measure, L is the ubiquitous Brier score with $L(p, 0) = p^2$ and $L(p, 1) = (1 - p)^2$. For alternative score functions of dichotomous events, please refer to Table 1 in Gneiting and Raftery (2007).

In applied problems θ in (16) has an economic interpretation, for example, in binary settings θ can represent the cost of a false positive prediction, $1 - \theta$ is the cost of a false negative, while true positive has no cost. Hence, for a fixed θ, an optimal strategy is to predict positive events with probability $> \theta$ and negative events with probability $< \theta$. This has a direct implication on model comparison; if a model receives consistently a lower mean score for every θ in comparison to alternative models, then the model dominates in predictive power. The choice of a proper scoring function becomes irrelevant in this case as the model would dominate for any other proper scoring rule (Ehm et al. (2016)).

In empirical settings one can compare competing models graphically by plotting the so-called Murphy's diagrams which display, for each model considered, the mean of the elementary score function L_θ over different values of $\theta \in [0, 1]$. In our settings, for a fixed value of θ, we calculate the average of L_θ over the test set of each cross-validation fold with posterior predictive from its training set. The final score curve for each model is the average of scores over cross-validation folds.

The ROC curves is a popular graphical tool for the assessment of discrimination ability in binary prediction problems. For each model an ROC curve is obtained by thresholding the predictive probabilities of the full unknowns in each cross-validation fold, calculating the true and false positive rates on each fold and then averaging them over the five-folds. With this process the posterior predictive interaction matrix is obtained at the threshold value that maximizes the area under the ROC curve (AUC). Moreover, for each fold the k parameter of the NN algorithm is chosen as the value that maximizes the AUC over the training set of that fold.

The phylogeny-only model in (2) is ill-formulated for the case of single-host parasites, since $\delta_{hj}(\eta) = 0$. Therefore, for comparison across the models, each held-out portion is constructed to ensure that at least two interactions are kept in each column of $\mathbf{Z}$. By this restriction each held-out portion is approximately 11% and 13% of documented associations for the datasets with and without single-host parasites, respectively.

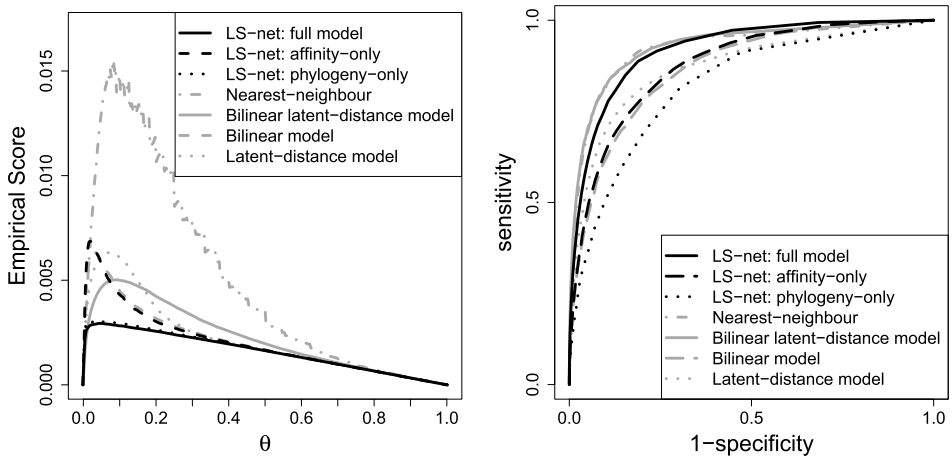
6. Results.

6.1. *Parameter estimation for the latent score full model.* For the GMPD we first fit the model proposed in Section 3.1. We run 20,000 MCMC iterations and the same for burn-in for posterior estimates. In total we have $J + H + 1$ parameters to estimate: an affinity parameter for each host and each parasite, and a tree scaling parameter for the host phylogeny.

Standard convergence diagnostics showed that all parameters had converged (For convergence and diagnostic plots refer to Supplementary Material, Section S4 (Elmasri et al. (2020))). It is worth noting that the posterior distributions of the host parameters (γ) show large variation which reflect that some hosts are more likely to interact with parasites or have been more intensively studied. The magnitude of the unit-free scaling parameter η is found to concentrate around 1.702 with 95% credible interval as $(0.391, 5.805)$, indicating accelerating evolution compared to the original tree.

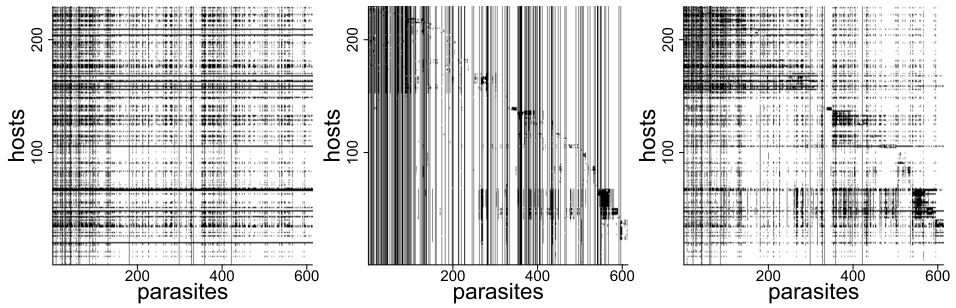
From Figure 4(a), the predictive performance of the proposed LS-net full model dominates its competitors, with the NN algorithm performing the least well. All other models, except the phylogeny-only, have mixed performance making it harder to infer predictive dominance. Nonetheless, it is worth noting that phylogeny-only model performs equivalently to LS-net full model, contrary to other neighbourhood-based conditional models. The weaker performance of the Jaccard-based neighbourhood models (NN and latent-distance) in comparison to the phylogeny-only model suggest phylogeny may provide more power over Jaccard distances in predicting host-parasite interactions. Jaccard distances, based on parasite sharing, should in principle mimic evolutionary distances for hosts and parasites that are relatively well studied and show phylogenetic structure among hosts, as in the GMPD. As a result, phylogeny-based models may be more favourable than the NN algorithm for sparser datasets. Murphy's diagrams on GMPD, including single-host-parasites, follow similar pattern and are depicted in Supplementary Material Figure S6 (Elmasri et al. (2020)).

Evident from five-fold average ROC curves in Figure 4(b), the LS-net full model outperforms its two submodels and their counterparts, which confirms the notion that each of the simpler models capture different characteristics of the data, and layering them yields better



(a) Murphy's diagrams (b) ROC curves

FIG. 4. *Murphy's diagrams and ROC curves of the latent score network* (*LS-net*) *model and two of its submodels*, *in comparison to competing models*, *the NN algorithms*, *the bilinear latent-distance models and two of its submodels* (*bilinear and latent-distance*). *Results are based on an average of five-fold cross-validations on GMPD*, *excluding single-host parasites*.



(a) affinity-only (b) phylogeny-only (c) full model

FIG. 5. *Posterior association matrix comparison for the GMPD, excluding single-host parasites: between the affinity-only (left), phylogeny-only (middle) and full model (right).*

results. The NN algorithm and the bilinear latent-distance model seem to have equivalent performance to the LS-net full model. Although the performance of the phylogeny-only model is subpar to its analogs (NN and the latent-distance model), it outperforms significantly in Murphy's diagrams (Figure 4(a)) which is a stronger indicator of predictive performance than ROC curves.

For a visual interpretation Figure 5 illustrates posterior predictive matrices for the affinity-only (Figure 5(a)), phylogeny-only (Figure 5(b)) and the full model (Figure 5(c)). To show the full effect of different models, posterior predictive probabilities for all interactions in $\mathbf{Z}$, observed and unobserved, are used to generate the matrices in Figure 5. From these figures the affinity-only model does not appear to account for any neighbourhood structure and results in hyperactive hosts, while the phylogeny-only model results in greater differences among parasites. The overall shape of the original $\mathbf{Z}$ in Figure 1 is best captured by the full model. In addition, the full model generates clear blocks of interacting hosts and parasites, reflecting interactions among particular host clades. For predictive matrices of competing models, please refer to Supplementary Material Figure S3 (Elmasri et al. (2020)).

For an analytical comparison we followed the recommendation of Demšar (2006) to use the single-sided paired Wilcoxon signed-rank test on the five-fold cross-validations rather than a fully Bayesian method (which could be implemented using a Wilcoxon-like statistic derived from posterior predictive samples). The paired test version is used since all models are tested using the same folds. For the GMPD, excluding single-host parasites, we obtain a p-value less than 0.035 when comparing the full model to all other models, except the bilinear latent-distance model and the NN algorithm. This indicates, for a 5% level of significance, that the full model outperforms its two submodels and the bilinear submodels, namely the bilinear and the latent-distance models. The combined bilinear latent-distance model and the NN algorithm are of equivalent statistical performance to the full model.

Single-host parasites comprise a nonnegligible portion of the total interactions ($\approx$17%) and including them in the calculation of host affinity parameters increases predictive performance, even though they are not included in the cross-validation set. To assess the effect of including single-host parasites on model performance, we repeated all analyses while keeping these in the original data. Table 1 shows the five-fold average AUC and true positive prediction results when the single-host parasites are kept or removed from the GMPD. The predictive strength of the full model is now more evident. The increase in AUC for the full model is directly attributed to the inclusion of single-host parasites, since both the AUC and the percent of ones recovered increased for the same held-out portion. This pattern is also more evident in the phylogeny-only model. For the other competing models the AUC increase is coupled with a weaker improvement in the recovery of positive interactions, suggesting a stronger explanatory power of phylogenetic distances over Jaccard-based distances.

TABLE 1
Area under the curve and prediction values for tested models

Model	No single-host parasites		With single-host parasites	
	AUC	% 1's recovered	AUC	% 1's recovered
LS-net: full model	0.921	87.46	0.959	92.56
LS-net: affinity-only	0.876	80.99	0.933	88.79
LS-net: phylogeny-only	0.823	79.75	0.917	90.34
Nearest-neighbour	0.926	88.61	0.948	91.10
Bilinear latent-distance model	0.929	86.35	0.936	86.40
Bilinear model	0.868	78.55	0.914	84.05
Latent-distance model	0.872	77.32	0.890	78.69

Since the single-host parasites are not part of the hold-out set, we infer that the improved AUC is due to the increased proportion of zeros in the larger database, as the held-out portion is kept constant. For the GMPD, including single-host parasites, the single-sided paired Wilcoxon signed rank test results in a p-value less than 0.035 when comparing the full model to all other models. This indicates a stronger performance in comparison to the results of the GMPD, excluding single-host parasites in terms of all measures, the proper scoring rules, ROC curves, and percent of ones recovered.

Computationally, we found that our ICM method, implemented in R, runs at least as fast as the latentnet R-package and, most of the time, twice as fast. For more details refer to Supplementary Material Table S1 (Elmasri et al. (2020)).

6.2. *Uncertainty in unobserved interactions.* We improve our latent score model by accounting for uncertainty in unobserved interactions, as shown in Section 4. This addition increases the posterior predictive accuracy by estimating the proportion of missing interactions in the latent space, and reducing scores for unobserved interactions. Using the model in Section 4, we infer the uncertainty parameter g, using 20,000 MCMC iterations with half as burn-in. The posterior mean of g is found to be 0.232 (posterior histograms in Figure 6(a)). Documented associations in the GMPD are identified through systematic searches of peer-

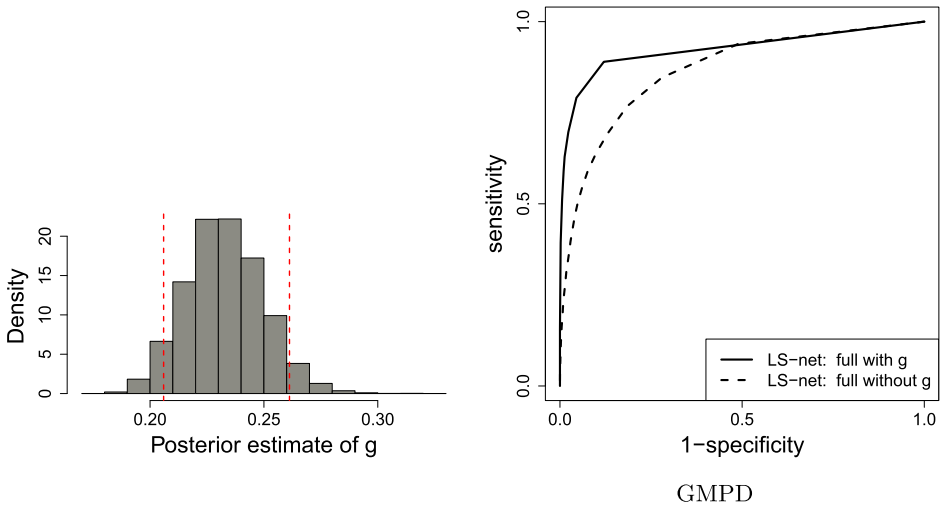


FIG. 6. *Posterior histogram for g (left) for GMPD, and comparison of ROC curves (right) for the full model with g and without g for GMPD, excluding single-host parasites.*

reviewed articles that support an interaction, therefore, we expect those associations to be of high confidence, reflecting the relatively low value of g.

Introducing g to the model affects all interactions, including known ones. Therefore, to measure the predictive accuracy, we require a different cross-validation method to the one of Section 5. We divide the GMPD into two sets, a training and a validation set. Since associations in the GMPD are sourced only from peer-reviewed articles, we were able to use information on article publication dates to create the training and test datasets. This mimics the discovery of interactions in the system rather than random hold-out of observations. Taking the earliest annotated year for each association, we set the training set as all associations documented prior to and including 2006, and the validation set as all associations up to 2010. There are 3755 pairs of documented associations in the GMPD, including single-host parasites, up to and including 2006. By 2010, the associations increased to 4178, with 236 hosts and 1308 parasites, approximately a 10% increase. For the training sets using the GMPD up to 2006, we used an average of five-fold cross-validations, constructed as in Section 5, to estimate the parameters of the model, where each fold ran for 20,000 iterations with half as burn-in. Since the full model is used, cross-validation is no longer restricted to multihost parasites as in Section 5, nonetheless, to avoid empty columns at least one interaction is kept for each parasite.

Figure 6(b) illustrates the improvement in potential predictive accuracy between the models with or without g. Essentially, incorporating uncertainty results in probability estimates for all interactions, undocumented and documented, where the former is penalized proportional to g. This reduces the overlap in posterior probability densities between interacting and noninteracting pairs, refer to Supplementary Material Figure S4 (Elmasri et al. (2020)) for the posterior histogram of both categories.

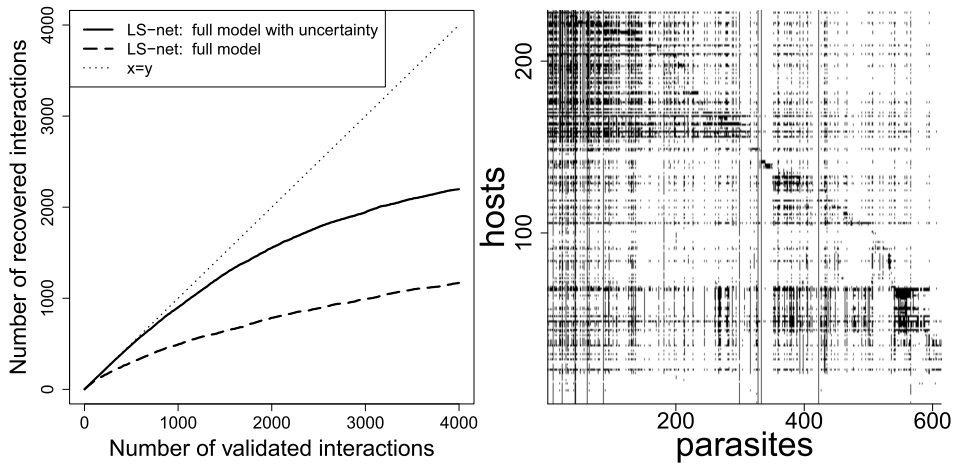
The model with g outperforms the full model on both AUC and proportion of positive interactions predicted, including and excluding the single-host parasites (Table 2). These results represent the evaluation on the whole dataset, up to 2010, not only the held-out and undocumented portions as in Section 5. The model with g is able to predict 90.90% of the documented interactions in the 2010 GMPD, approximately 3798 out of 4178 interactions, where the model without g predicts approximately 194 fewer interactions.

Another method of model comparison is through the proportion of recovered interactions from the full data. This can be quantified by sorting all pairwise interactions, based on their posterior predictive probabilities, and examining the top x pairs with the highest predictive probabilities, as they represent interactions with highest confidence. By counting the number of true interactions recovered in those x selected pairs and by scaling x from 1 to 4000, we find the model with g again outperforms the full model by recovering more than double the number of interactions (Figure 7(a)). Finally, for comparison with Figure 5, the posterior interaction matrix for the model with g excluding single-host parasites is shown in Figure 7(b).

Incorporating phylogenetic information identifies interactions that would not be considered likely under the affinity-only model. To illustrate this, we plot the number of documented interactions (node degree) for both hosts and parasites included in the 100 most probable, yet

TABLE 2
Area under the curve and prediction values for the model with(out) g

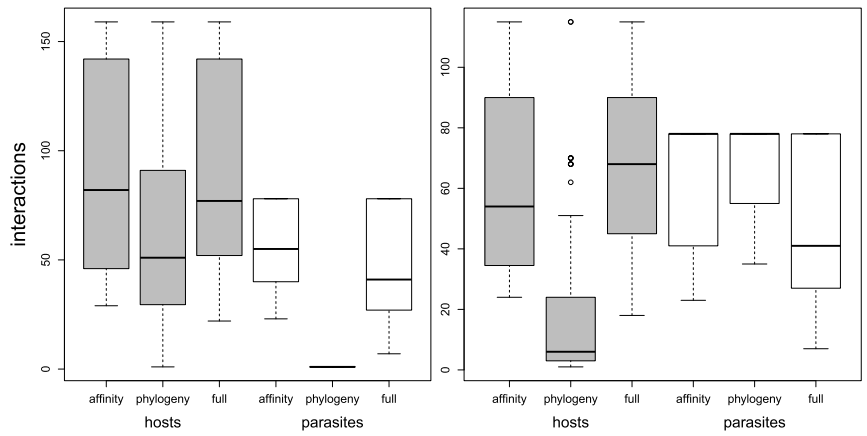
	No single-host parasites		With single-host parasites	
	AUC	% 1's recovered	AUC	% 1's recovered
With g	0.924	88.98	0.944	90.90
Without g	0.865	76.80	0.918	86.26



(a) recovery power (b) GMPD with g no single-host parasites

FIG. 7. *Number of pairwise recovered interactions from the original* 2010 *GMPD data* (*left*) *and the posterior interaction matrix for the* 2010 *GMPD*, *excluding single-host parasite*, *using the model that accounts for uncertainty with g*.

previously undocumented, interactions for each model (Figure 8). When fitting the model, including single-host parasites, we find that the top 100 predicted links for the phylogeny-only model tend to include hosts and parasites with fewer observed links in the original data. In fact, all top 100 novel predictions made by the phylogeny-only model include parasites that have one documented interaction, all of which would be given low probability by preferential attachment models (including our affinity only model). By contrast, in the top 100 predictions made by the affinity only model, the parasite with the fewest number of observed interactions has 26 known host species. This suggests that the inclusion of phylogenetic information allows the identification of highly probable interactions for rare or understudied species.



(a) with single-host (b) without single-host

FIG. 8. *Comparison of the number of documented interactions* (*node degree*) *for both hosts* (*grey*) *and parasites* (*white*) *included in the* 100 *most probable*, *yet previously undocumented*, *interactions across each of the three sub-models*. *Results are split into models*: (a) *with single-host parasites*, *and* (b) *without single-host parasites*.

7. Discussion. We introduced a latent score model for link prediction in ecological networks and illustrate it using a recently published global database of host-parasite interactions. The proposed model is a combination of two separate models, an affinity based exchangeable random network model overlaid with a Markov network dependency informed by phylogeny (2). The affinity-only model is characterized by independent affinity parameters for each species, while the phylogeny-only model is characterized by a scaled species similarity matrix. Both parts perform reasonably well alone, and by overlaying them the posterior prediction is significantly improved.

Many advantages arise from integrating host phylogenies. By utilizing known evolutionary models, phylogenies provide remarkable predictive power, comparable to state-of-the-art latent-distance models (Hoff, Raftery and Handcock (2002), Hoff (2005)), but with added biological interpretation. Such tree-scaling models could also be integrated in existing link prediction frameworks, such as that outlined by Chiu and Westveld (2011) and elaborated by Ovaskainen et al. (2017). However, computational issues might arise as the full joint distribution becomes intractable and is not guaranteed to exist.

To our knowledge, our framework is the first to attempt to incorporate this type of evolutionary information in link-prediction models. Computational issues arise from integrating this procedure, and we solve this by imposing minimal conditions on the latent variable, which produces promising results. We used the early-burst model to scale the phylogeny, but any other evolutionary model that scales the species covariance matrix could be fit.

While we incorporated phylogeny as the dependence structure, the model can easily accommodate different similarity matrices or types of dependence in an additive manner. For host-parasite networks, host traits or geographic overlap, or parasite similarity based on phylogeny, taxonomy or traits may improve prediction (Pedersen et al. (2005), Davies and Pedersen (2008), Luis et al. (2015)). Introducing different similarity measures affects the model characteristics in two ways: it changes the topology of the probability domain, and it increases the number of parameters to estimate due to introduced scaling parameters. The latter is easily integrated since the number of estimated parameters increases by one for each new scaling parameter. It is also possible to introduce different tree scaling parameters for different host-groups, allowing for a richer representation and added flexibility with minimum cost, which should improve performance. In addition, covariate data, such as species traits, can easily be integrated in the model in an additive manner. For example, set $\tau_{hj} = \gamma_h \rho_j \delta(\eta) \exp(-\beta_i x_i - \beta_j x_j)$. Alternatively, they could be included in a hierarchical manner as a function of the affinity parameter. Each case represents a different interpretation, with covariates in the former driving the interaction probability directly while the latter influence the affinity parameters.

A particular dependence structure that does not require additional data is a similarity based on the number of shared interactions, as used in the NN algorithm (Section 5). In host-parasite networks, parasite community similarity is often well predicted by evolutionary distance among hosts (Gilbert and Webb (2007), Davies and Pedersen (2008)). In this case the NN similarity is likely capturing some of the phylogenetic structure in the network and could be a reasonable approach if a reliable phylogeny is unavailable. However, as phylogeny is estimated independently from the interaction data, it will likely be more robust to incomplete sampling of the original network than NN type dependence structures.

Many ecological and other real world networks display power-law degree distributions (Albert and Barabási (2002)). This is also the case with the host-parasite database used in this paper, where both hosts and parasites exhibit power-law degree distributions. The affinity-only version of the proposed model in (4) has been shown to generate a power-law behaviour when a generalized gamma process is used (Brix (1999), Lijoi, Mena and Prünster (2007), Caron and Fox (2017)). In fact, when $\gamma_h = \gamma$ for all h, the affinity-only model behaves much

like the stable Indian buffet process of Teh and Gorur (2009) that has a power-law behaviour. Nonetheless, we find the full model to show a significant improvement in predictive accuracy over the affinity-only model, though it does not yield a degree distribution with a power-law. However, when accounting for uncertainty in the full model, the posterior predictions we regain a power-law degree distribution for hosts and parasites (Supplementary Material, Figure S5, Elmasri et al. (2020)). It would be interesting in future work to explore which other network properties are maintained using this model.

While the intent of this research is to identify undocumented interactions, this model can also account for uncertainty in observed interactions. In this case, our model may be used to identify weakly supported interactions that are false positives or sampling artefacts in the literature that may benefit from additional investigation. In the case of host-parasite interactions, our approach could form an integral component of proactive surveillance systems for emerging diseases (Farrell, Berrang-Ford and Davies (2013)). However, the framework illustrated here is not limited to host-parasite networks but is well suited to multiple ecological networks such as plant-herbivore, flower-pollinator or predator-prey interactions.

Acknowledgements. We would like to thank the McGill Statistics-Biology Exchange Group (S-BEX) and organizers Russell Steele, Zofia Taranu and Amanda Winegardner for fostering an environment that led to this collaboration. We also thank the Davies lab at McGill for critical feedback throughout model development and writing, and the Macroecology of Infectious Disease Research Coordination Network (funded by NSF DEB 1316223) for providing early versions of the GMPD. DS was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). MJF was funded by an NSERC Vanier CGS, and ME by FQRNT and NSERC PDF.

SUPPLEMENTARY MATERIAL

Supplement: A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships (DOI: 10.1214/19-AOAS1296SUPP; .pdf). Full model development, MCMC diagnostics and additional results.

REFERENCES

AGUIRRE, A. A., KEEFE, T. J., REIF, J. S., KASHINSKY, L., YOCHEM, P. K., SALIKI, J. T., STOTT, J. L., GOLDSTEIN, T., DUBEY, J. P. et al. (2007). Infectious disease monitoring of the endangered Hawaiian monk seal. *J. Wildl. Dis.* **43** 229–241.

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. MR1895096 https://doi.org/10.1103/RevModPhys.74.47

BARTOMEUS, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PLoS ONE* **8** e69200.

BASTAZINI, V. A. G., FERREIRA, P. M. A., AZAMBUJA, B. O., CASAS, G., DEBASTIANI, V. J., GUIMARÃES, P. R. and PILLAR, V. D. (2017). Untangling the tangled bank: A novel method for partitioning the effects of phylogenies and traits on ecological networks. *Evol. Biol.* **44** 312–324.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BRAGA, M. P., RAZZOLINI, E. and BOEGER, W. A. (2015). Drivers of parasite sharing among Neotropical freshwater fishes. *J. Anim. Ecol.* **84** 487–497.

BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.* **31** 929–953. MR1747450 https://doi.org/10.1239/aap/1029955251

CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

CHIU, G. S. and WESTVELD, A. H. (2011). A unifying approach for food webs, phylogeny, social networks, and statistics. *Proc. Natl. Acad. Sci. USA* **108** 15881–15886.

CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks. CBMS Regional Conference Series in Mathematics* **107**. Amer. Math. Soc., Providence, RI. MR2248695 https://doi.org/10.1090/cbms/107

CLEAVELAND, S., LAURENSON, M. K. and TAYLOR, L. H. (2001). Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **356** 991–999.

DALLAS, T., PARK, A. W. and DRAKE, J. M. (2017). Predicting cryptic links in host-parasite networks. *PLoS Comput. Biol.* **13** 1–15.

DAVIES, T. J. and PEDERSEN, A. B. (2008). Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proc. Biol. Sci.* **275** 1695–1701. https://doi.org/10.1098/rspb.2008.0284

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30. MR2274360

EHM, W., GNEITING, T., JORDAN, A. and KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505–562. MR3506792 https://doi.org/10.1111/rssb.12154

ELMASRI, M., FARRELL, M., DAVIES, T. J. and STEPHENS, D. A. (2020). Supplement to "A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships." https://doi.org/10.1214/19-AOAS1296SUPP.

FARRELL, M. J., BERRANG-FORD, L. and DAVIES, T. J. (2013). The study of parasite sharing for surveillance of zoonotic diseases. *Environ. Res. Lett.* **8** 015036.

FARRELL, M. J., STEPHENS, P. R., BERRANG-FORD, L., GITTLEMAN, J. L. and DAVIES, T. J. (2015). The path to host extinction can lead to loss of generalist parasites. *J. Anim. Ecol.* **84** 978–984.

FRITZ, S. A., BININDA-EMONDS, O. R. P. and PURVIS, A. (2009). Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecol. Lett.* **12** 538–549.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GILBERT, G. S. and WEBB, C. O. (2007). Phylogenetic signal in plant pathogen-host range. *Proc. Natl. Acad. Sci. USA* **104** 4979–4983.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GÓMEZ, J. M., VERDÚ, M. and PERFECTTI, F. (2010). Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature* **465** 918–21.

GRAVEL, D., POISOT, T., ALBOUY, C., VELEZ, L. and MOUILLOT, D. (2013). Inferring food web structure from predator-prey body size relationships. *Methods Ecol. Evol.* **4** 1083–1090.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504 https://doi.org/10.2307/3318737

HARMON, L. J., LOSOS, J. B., JONATHAN DAVIES, T., GILLESPIE, R. G., GITTLEMAN, J. L., BRYAN JENNINGS, W., KOZAK, K. H., MCPEEK, M. A., MORENO-ROARK, F. et al. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64** 2385–2396.

HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295. MR2156838 https://doi.org/10.1198/016214504000001015

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HUANG, S., DRAKE, J. M., GITTLEMAN, J. L. and ALTIZER, S. (2015). Parasite diversity declines with host evolutionary distinctiveness: A global analysis of carnivores. *Evolution* **69** 621–630.

JIANG, X., GOLD, D. and KOLACZYK, E. D. (2011). Network-based auto-probit modeling for protein function prediction. *Biometrics* **67** 958–966. MR2829270 https://doi.org/10.1111/j.1541-0420.2010.01519.x

JORDANO, P. (2016). Sampling networks of ecological interactions. *Funct. Ecol.* **30** 1883–1893.

KRIVITSKY, P. N. and HANDCOCK, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *J. Stat. Softw.* **24**. https://doi.org/10.18637/jss.v024.i05

KRIVITSKY, P. N. and HANDCOCK, M. S. (2017). latentnet: Latent position and cluster models for statistical networks. The Statnet Project. R package version 2.8.0. Available at http://www.statnet.org.

LA SALLE, J., WILLIAMS, K. J. and MORITZ, C. (2016). Biodiversity analysis in the digital era. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **371** 20150337.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 715–740. MR2370077 https://doi.org/10.1111/j.1467-9868.2007.00609.x

LUIS, A. D., O'SHEA, T. J., HAYMAN, D. T. S., WOOD, J. L. N., CUNNINGHAM, A. A., GILBERT, A. T., MILLS, J. N. and WEBB, C. T. (2015). Network analysis of host-virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecol. Lett.* **18** 1153–1162.

MORALES-CASTILLA, I., MATIAS, M. G., GRAVEL, D. and ARAÚJO, M. B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.* **30** 347–356.

OLIVAL, K. J., HOSSEINI, P. R., ZAMBRANA-TORRELIO, C., ROSS, N., BOGICH, T. L. and DASZAK, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature* **546** 646–650. https://doi.org/10.1038/nature22975

OVASKAINEN, O., ABREGO, N., HALME, P. and DUNSON, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7** 549–555.

OVASKAINEN, O., TIKHONOV, G., NORBERG, A., GUILLAUME BLANCHET, F., DUAN, L., DUNSON, D., ROSLIN, T. and ABREGO, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20** 561–576.

PAGEL, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401** 877–884.

PARK, A., FARRELL, M., SCHMIDT, J., HUANG, S., DALLAS, T., PAPPALARDO, P., DRAKE, J., STEPHENS, P., POULIN, R. et al. (2018). Characterizing the phylogenetic specialism–generalism spectrum of mammal parasites. *Proc. R. Soc. Lond., B Biol. Sci.* **285** 20172613.

PARRISH, C. R., HOLMES, E. C., MORENS, D. M., PARK, E.-C., BURKE, D. S., CALISHER, C. H., LAUGHLIN, C. A., SAIF, L. J. and DASZAK, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* **72** 457–70.

PEARSE, I. S. and ALTERMATT, F. (2013). Predicting novel trophic interactions in a non-native world. *Ecol. Lett.* **16** 1088–1094.

PEDERSEN, A. B., ALTIZER, S., POSS, M., CUNNINGHAM, A. A. and NUNN, C. L. (2005). Patterns of host specificity and transmission among parasites of wild primates. *Int. J. Parasitol.* **35** 647–657.

PEDERSEN, A. B., JONES, K. E., NUNN, C. L. and ALTIZER, S. (2007). Infectious diseases and extinction risk in wild mammals. *Conserv. Biol.* **21** 1269–1279.

PETCHEY, O. L., BECKERMAN, A. P., RIEDE, J. O. and WARREN, P. H. (2008). Size, foraging, and food web structure. *Proc. Natl. Acad. Sci. USA* **105** 4191–4196.

POELEN, J. H., SIMONS, J. D. and MUNGALL, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* **24** 148–159.

ROBERT, C. P. and CASELLA, G. (2013). *Monte Carlo Statistical Methods. Springer Texts in Statistics*. Springer, New York. MR1707311 https://doi.org/10.1007/978-1-4757-3071-5

STEPHENS, P. R., PAPPALARDO, P., HUANG, S., BYERS, J. E., FARRELL, M. J., GEHMAN, A., GHAI, R. R., HAAS, S. E., HAN, B. et al. (2017). Global mammal parasite database version 2.0. *Ecology* **98** 1476–1476.

STOCK, M., POISOT, T., WAEGEMAN, W. and DE BAETS, B. (2017). Linear filtering reveals false negatives in species interaction data. *Sci. Rep.* **7** 1–8.

STREICKER, D. G., TURMELLE, A. S., VONHOF, M. J., KUZMIN, I. V., MCCRACKEN, G. F. and RUPPRECHT, C. E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* **329** 676–679. https://doi.org/10.1126/science.1188836

SWENDSEN, R. H. and WANG, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

TEH, Y. W. and GORUR, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems* **22** 1838–1846.

WARDEH, M., RISLEY, C., MCINTYRE, M. K., SETZKORN, C. and BAYLIS, M. (2015). Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data* **2** 150049. https://doi.org/10.1038/sdata.2015.49

WEBB, C. O., ACKERLY, D. D., MCPEEK, M. A. and DONOGHUE, M. J. (2002). Phylogenies and community ecology. *Ann. Rev. Ecolog. Syst.* **33** 475–505.

WEIR, I. S. and PETTITT, A. N. (2000). Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 473–484. MR1824553 https://doi.org/10.1111/1467-9876.00206

WIENS, J. J., ACKERLY, D. D., ALLEN, A. P., ANACKER, B. L., BUCKLEY, L. B., CORNELL, H. V., DAMSCHEN, E. I., JONATHAN DAVIES, T., GRYTNES, J.-A. et al. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.* **13** 1310–1324.

WILLIAMS, R. J. and MARTINEZ, N. D. (2000). Simple rules yield complex food webs. *Nature* **404** 180–183.