OIKOS

Research article

Comparing the power of phylogenetic, trait and network structure information to predict plant–frugivore interactions

Grant Foster ✉ **and Tad A. Dallas**

Department of Biological Sciences, University of South Carolina, Columbia, SC, USA

Correspondence: Grant Foster (fostergt@email.sc.edu)

Oikos

2025: e11156

doi: 10.1002/oik.11156

Subject Editor: Will Pearse
Editor-in-Chief: Pedro Peres-Neto
Accepted 22 September 2025

Due to the constraints of limited effort and sampling error, observed species interaction networks are an imperfect representation of the 'true' underlying community. Link prediction methods allow us to construct a potentially more complete representation of a given empirical network by guiding targeted sampling of predicted links, as well as offer insight into potential interactions that may occur as species' ranges shift. Various data types can predict interactions; understanding how different kinds of information compare in their ability to predict links between different types of nodes is important. To this end, we compare random-forest regression models informed by combinations of phylogenetic structure, species traits, and latent network structure-hidden features inferred from the observed network topology – in their ability to predict interactions in a diverse network of fruiting plants and frugivorous birds in Brazil's Atlantic forest. We found that for our dataset, latent structure derived through a single-value decomposition approach was the most important determinant of model predictive performance. While incorporating trait or phylogenetic information alongside latent features had little effect on discriminatory power, they did meaningfully increase overall model accuracy. Our results highlight the potential importance of latent structural features for predicting mutualistic interactions, and encourage a clear conceptual link between prediction performance metrics and the overall goal of predicting cryptic links.

Keywords: frugivory, link prediction, mutualism, mutualistic network, seed dispersal

Introduction

Natural systems are characterized by a diverse set of interacting organisms; organizing these interaction sets into ecological networks can help us gain insight into community and ecosystem-level processes. Whether consumer–resource, host–parasite or mutualistic networks, understanding which species interact and why can give us insight into features as specific as the dynamics of one focal species all the way up to broad-scale patterns of community assembly and stability (Guimaraes 2020, Saravia et al. 2022). As such, ecologists have put a great deal of effort into observing and quantifying the interactions found in nature. However, like any data collection process, the sampling of ecological

NORDIC SOCIETY OIKOS

www.oikosjournal.org

networks is imperfect. No matter how much effort we put into characterizing a network, we may often miss observing interactions (links), or even interactors themselves (nodes) (Young et al. 2021). Moreover, the likelihood of observing a given interaction is related to factors such as the abundances of both interactors (Canard et al. 2014), with the total interaction set of rare species less likely to be represented accurately (Olesen et al. 2011a, Cirtwill et al. 2019). Errors in the construction of ecological networks, especially when biased, can lead to dramatic changes in network structure (Blüthgen et al. 2008, Young et al. 2021). Ecological networks are also temporally dynamic. Local species turnover can add or subtract links, while regional-scale extinctions, evolution or range shifts may rewire existing connections (Olesen et al. 2011b).

Increasing sampling intensity or duration could solve these problems; as sampling effort increases it is less likely for rare links to go unobserved, allowing for a more accurate representation of the true network structure. However, the resources that can be devoted to sampling are limited. While the increasing availability of passive monitoring technologies (Quintero et al. 2022), as well efforts to extract interaction data from existing large-scale occurrence datasets (Putman et al. 2021) may increase the availability on interaction data, the problem of incompletely sampling interactions will never fully disappear. A potential way to address this shortfall is to utilize predictive approaches to decide which particular species or links should be targeted for additional sampling. To this end, the goal of predicting cryptic links between nodes allows us to more productively focus limited sampling resources (Terry and Lewis 2020). Predictive approaches also allow for us to predict network organization arising from novel complements of species. Identifying likely or unlikely possible linkages between invading species and already present interactors allows us to get a more complete picture of biotic constraints on invasibility (Minoarivelo and Hui 2016), and ultimately a better understanding of how species interaction networks may be altered by anthropogenic forcing.

Mutualistic interactions are dependent on a complex set of factors including life history (Ramos-Robles et al. 2018), phenotypic traits (Rafferty and Ives 2013), co-evolutionary history (Eriksson 2016), spatial and temporal distributions (Menke et al. 2012, Fricke and Svenning 2020, Laurindo et al. 2020), as well other biotic interactions (Carreira et al. 2020). Incorporating information that either directly influences, or conveys information about these factors into link prediction frameworks should hopefully improve prediction performance (Dallas et al. 2017). When applying link prediction in practice, we often have a variety of species and interaction-level properties at our disposal, many of which may perform better or worse at predicting in certain types of networks or interactions. Understanding which types of information may be better or worse at predicting interactions in different contexts, however, is still an open avenue of research.

For plant–frugivore seed dispersal systems, information on species morphological traits (gape-width, fruit characteristics, etc), may be at the root of why many taxa do or do not interact (Moran and Catterall 2010, González-Castro et al. 2015, Bender et al. 2018). Trait matching link prediction approaches are able to use suites of continuous or discrete traits that either directly or indirectly influence interaction probability across a large number of species. These approaches perform well in situations when a small number of traits are linked to interaction probabilities across many species (e.g. flower and beak morphometric traits in flower–hummingbird pollinator systems, Pichler et al. 2020). However, while trait data may give us insight into biological underpinnings of network connections we observe, they may not always be the most appropriate for link prediction methods. Interactions may be determined by a large suite of traits that individually contribute small amounts. Alternatively, traits relevant for one taxonomic group may not be relevant for another (gape size may be predictive for birds, but not for primates). Or interaction propensities may be controlled by traits difficult to quantify, such as microhabitat usage or foraging strategies. In all of these cases, trait-matching approaches may perform relatively poorly for some or all types of nodes in a given network. In these cases, alternative information sources such as phylogenetic information or latent network structure may be useful for link prediction.

Using phylogenetic relationships between species may be useful for predicting links due to a number of factors. Firstly, traits that mediate species interactions may be phylogenetically conserved, allowing us to treat information about phylogenetic relationships between organisms as a proxy for traits. This assumption may be especially useful when traits governing interactions are numerous, hard to measure, or hard to identify. However, phylogenetic conservatism may break down at fine evolutionary scales due to local selection pressures on traits governing interactions (Pérez et al. 2007), making phylogenetic information an imperfect proxy in some contexts (Rafferty and Ives 2013). Additionally, while post hoc correlative investigation may be able to suggest potential traits of interest, this approach ultimately further abstracts from the mechanisms governing interactions even if prediction performance is quite good. Despite these shortcomings, in the use-case of targeting potential links for surveillance (in essence a prediction problem), a phylogenetic approach may still be quite useful. In addition to proxying traits, as coevolutionary history between organisms can often be an important factor governing the presence of an interaction, incorporating phylogenetic information may be vital to represent this linkage. While plant–frugivore networks may generally be less coevolutionarily linked and more asymmetric (species with few interactions tend to interact with partners with many interactions) than some other classes of ecological networks (Wheelwright and Orians 1982, Jordano 1987, Maglianesi et al. 2024), capturing nodes' shared evolutionary history may still be important for prediction. By definition phylogenetic methods require a phylogeny that includes all the species you intend to predict interactions for, causing potential limitations for predicting in systems where phylogenetic information may be incomplete. However, as the cost and time required to sequence non-model genomes has and continues to decline, the problem of data availability

continues to become a more and more surmountable barrier to prediction. To date, phylogeny-based interaction predictions have been successfully implemented in a number of both mutualistic (Braga et al. 2021) and antagonistic (Pearse and Altermatt 2013) plant–animal network systems.

Independent of the phylogenetic relationships and traits of individual nodes in a network, latent structural features of networks themselves may actually be useful for predicting missing links. In this context, the term 'latent features' refers broadly to hidden properties derived from the topology of the interaction network, which can be approximated through any number of dimensionality-reduction approaches (Poisot et al. 2021). While there are a multitude of different methods used to create latent network features for interaction prediction (single-value decomposition, random dot product graphs (Strydom et al. 2022), graph motif distributions, matrix factorization (Seo and Hutchinson 2018), etc), we focus on single-value decomposition, an eigendecomposition approach already used with some success in describing the structure of and assisting in the prediction of empirical food webs (Banville et al. 2023), host–virus associations (Poisot et al. 2021, 2023), and seed-dispersal interactions (Nunes Martinez and Mistretta Pires 2024). Once an existing network is known, calculating latent structural features is computationally tractable. While arguably even more agnostic to the mechanisms driving species interactions, this structural information may capture particular types of real linkages difficult to discover based on trait or phylogenetic information alone.

Prevailing interaction prediction approaches have often centered on a single information type, but recent work in host–parasite systems have emphasized the power of machine learning approaches to synthesize multiple data types to predict interactions (Strydom et al. 2021). Despite the use of this variety of information streams for link prediction problems, studies directly comparing the predictive capacity of phylogenetic, trait, and latent features are few, and have not been applied in plant–frugivore systems. Beyond looking at the performance of these feature classes independently however, knowing how they may be used together is another important avenue of research. By looking at the performance of models trained on combinations of different data types we are better able to understand whether different methods overlap in the types of links they predict well, and perhaps more importantly whether we can use these information streams in concert to create models that are more than the sum of their parts. To this end, we apply random forest prediction algorithms to predict plant-frugivore interactions in Atlantic forests in Brazil using combinations of species trait data, phylogenetic information, and latent structural features.

Material and methods

Trait and interaction data

The full interaction network was published by Bello et al. (2017), and includes a total of 5226 unique species interactions between 787 plant species and 342 frugivore species. This network effectively represents a regional metaweb of observed interactions, with the realized set of interactions at a given site being a subset, the exact composition of which is determined by local assembly processes. By taking a metaweb approach-predicting 'possible' interactions independent of local geographic or temporal variation in species abundance – we aim to better understand the factors associated with the potential for interaction. This general approach is shared by a number of other studies predicting potential interactions across large spatial scales. (Strydom et al. 2022, Dansereau et al. 2024, Hao et al. 2025) Due to the availability of phylogenetic and trait data, we restricted our analyses to avian–plant interactions, representing the most species-rich set of frugivores in this dataset (3856 unique interactions between 394 and 242 plant and bird species, respectively). Of our avian-frugivore species (hereafter frugivores), the number of unique plant interactions per frugivore (frugivore degree) ranged from 1 (53 species) to 120 unique plant interactions recorded for *Turdus rufiventris*; median frugivore degree was 5. The number of unique frugivore interactions per plant (plant degree) was generally lower, ranging from 1 (128 species) to 80 unique frugivore interactions recorded for *Myrsine coriacea*; median plant degree was 2. Overall this interaction network is characterized by a relatively high degree of phylogenetic generalism. Out of 187 frugivore species that interacted with more than one plant for which phylogenetic information is available, only nine interacted with a group of plants more phylogenetically related than random chance (Kembel et al. 2010, Supporting information).

The Bello et al. (2017) dataset includes a number of informative traits for both plant and frugivore nodes that could be potentially useful for prediction, many of which we used in our trait-based models. These include frugivore allometric measures such as body mass and gape size, and ecological traits such as degree of frugivory (scored 1–3). Frugivore mean body mass was unavailable for three species, while mean gape size was unavailable for 68 species. Plant traits include fruit diameter, fruit color, growth form, and fruit lipid concentration (scored 1–3). For trait-based models, species were filtered to only include those with complete sets of trait data (174/242 frugivore species). As part of our research goal was to compare the power of phylogenetic and trait data, we chose to omit these species for trait based models rather than phylogenetically impute missing trait information. Categorical variables (such as plant growth form) were transformed into a series of binary variables through one-hot encoding. Continuous data on fruit size was available for 280/394 plant species; plants without trait data were again excluded from trait-based models. Details on plant and frugivore traits used in the analysis are available in Table 1.

Phylogenetic data

For plant species, phylogenetic data was retrieved from the BIEN database, accessed using the R package 'BIEN' (Maitner et al. 2018). Out of 787 plant species in our dataset, 646 occurred in the BIEN phylogeny. 61 plant species

Table 1. Frugivore and plant traits used for interaction prediction. Non-ordinal factors such as fruit color were transformed through one-hot encoding prior to inclusion in the model.

Species	Trait	Units	Range/levels
Frugivore	Body size	g	6.60–3500
	Gape size	mm	2.80–36.29
	Degree of frugivory	–	low, medium, high
Plant	Fruit diameter	mm	1.2–325
	Fruit color	–	yellow, red, black, brown, green, other
	Fruit lipid content	–	low, medium, high
	Plant growth form	–	tree, liana, palm, scrub, other

with a recorded avian frugivore interaction did not occur in this phylogeny. For absent species with a congeneric in our phylogeny, we added those species as a polytomy at the parent node of that genera, allowing us to add an additional 46 species to our analysis. 16 species did not have congenerics present in the network, and were discarded from the analysis. Results were qualitatively the same if these polytomies are omitted from the analysis (Supporting information). Bird phylogenies were retrieved from VertLife (Jetz et al. 2012); for our analysis, we used 100 trees sampled from the Bayesian posterior distribution. Phylogenetic information was unavailable for two bird species in our network. As with plants, both species were added as polytomies at their respective parent genera. Each sampled tree was used in one of one hundred replicates for each model incorporating phylogenetic data. For both the plant and bird phylogenies we performed eigenvalue decomposition, and trained phylogenetic models on the first 4 dominant eigenvectors for plants and the first 3 dominant eigenvectors for birds. The incorporated eigenvectors explained 37.2% of variation for plants. While the exact amount of variation explained by the incorporated eigenvectors differed among sampled trees for birds due to variation in posterior samples, on average the incorporated eigenvectors explained 46.02 ± 2.29% of variation for birds.

Latent methods

We used singular value decomposition of each training interaction matrix to create continuous latent features for prediction using the *svd()* function in base R (www.r-project.org). For each training iteration, we fixed the dimensions of the training matrix to be equal to the full network, adding interactions (1's) of only the training set before creating latent features. As such, in any individual training matrix, some species with few interactions may have no recorded interactions. Latent features of training subsets were highly correlated with those derived for imputing based on the full interaction matrix (Supporting information). We used the first three axes of variation for both plants and birds as continuous predictors. For imputation of missing links across the entire network, we utilized the entire interaction matrix to construct latent features.

Model structure and comparison

We used random forest regression models informed by different feature types to make comparisons across types of information, as implemented in the R 'randomForest' package (Liaw and Wiener 2002). Random forest models are effectively an ensembled series of classification trees, each of which is applied to bootstrapped subsamples of training data and a portion of the potential predictor variables. By incorporating the information of many decision trees together, random forest techniques can often boast high predictive accuracy across a variety of ecological settings, the ability to uncover complex and nonlinear interactions between predictor variables, and internally cross-validate by repeatedly testing on 'out of bag samples' (portions of the data not included in a given bootstrap subsample) (Breiman 2001, Cutler et al. 2007). In the context of ecological networks, random forest models have been successfully applied to predict species interactions across a variety of systems (Desjardins-Proulx et al. 2017, Pichler et al. 2020, Sydenham et al. 2022). We tested a total of seven models; three using predictors from only one type of information (traits, phylogeny or latent features), three pairwise combinations of each predictor class, and one model incorporating all possible predictors. In order to validate model performance, we trained 100 iterations of each model. Like most empirically observed ecological networks, our training network was very sparse (Vázquez et al. 2009). Our total interaction matrix included 3643 observed interactions and 90 917 unobserved interactions. To address this class imbalance, for each iterations we first randomly selected 80% of our data for training, and then randomly removed unobserved interactions from the edgelist (i.e. rows of the edgelist where the interaction value is equal to 0) from the training data until we achieved a 1:3 ratio of observed:unobserved interactions (see the Supporting information for the results of alternative prevalence values). After model training, performance was evaluated on the remaining 20% test set. We analyze model performance both through area under the receiver–operator curve (AUC), which measures model discriminatory power, and root mean squared error (RMSE), which measures prediction accuracy. These two measures are commonly used to evaluate the effectiveness of predictive approaches (Norberg et al. 2019), but the choice of which metric to optimize may be different depending on the use case in question. For each model iteration we also recorded the optimal suitability threshold for classification as the value that maximized Youden's *J* statistic (the sum of specificity and sensitivity minus one, a common diagnostic statistic for dichotomous tests (Youden 1950)). For comparisons of model accuracy metrics across models, we performed post hoc two-sided pairwise *t*-tests adjusted for multiple comparisons (Holm 1979).

After validating model predictive performance using this test-train split, we then re-ran 100 iterations of each model using the full network without internal class balancing. Using the full suite of interaction information allows us to better predict potential unobserved links. We present the pairwise suitability correlations of each of these full model outputs, as well as present a list of links predicted to be highly suitable by one or more of our models but that are not observed to occur in this data set. We then used the average optimal

threshold value for each model as model-specific thresholds for classification. For each model, variable importance was quantified by the mean decrease in Gini coefficient (MDG) after permutation of each variable as implemented in the R 'randomForest' package (Liaw and Wiener 2002). A common variable importance metric for classification problems, MDG is calculated by normalizing the sum of all decreases in node purity (the ability of a given tree to correctly distinguish links from nonlinks) given permutation by the total amount of decision trees in the ensemble (Calle and Urrea 2011, Khalilia et al. 2011). A higher MDG value for a given variable indicates that variable is more important to model performance.

Results

Model performance

Using an 80–20 split of testing and training data, all models were able to predict testing interactions with reasonable accuracy and discriminatory power (Fig. 1). AUC was lowest for trait only models (0.84 ± 0.009) followed by phylogeny only models (0.84 ± 0.008). All models including latent information outperformed latent-agnostic models in terms of discriminatory power, with the greatest AUC achieved by trio (0.91 ± 0.007) and phylogeny-latent models (0.91 ± 0.006); pairwise *t*-tests indicated no significant difference between them (p = 0.14). However, prediction accuracy did not follow the same trends as discriminatory power. Latent and phylogeny-only models had the lowest overall accuracy (57.7 ± 0.40 and 56.1 ± 0.44, respectively), while the most accurate models were those combining trait information with another information type. The trait-phylogeny (RMSE: 49.3 ± 0.42) and trio (RMSE: 49.4 ± 0.44) models were the most accurate (pairwise *t*-tests indicating no difference, p = 0.8), closely followed by the traits-latent model (RMSE: 49.8 ± 0.37). The trait-only model had only middling accuracy performance however (RMSE: 52.7 ±0.54).

Variable importance on entire network

After validating model performance on novel test-data, we then retrain replicate models of each type on the full network in order to generate more robust predictions. Variable importance was quantified by mean decrease in Gini coefficient after permutation of each variable. In all models incorporating them, latent SVD features were unilaterally the most important variables for prediction (Fig. 2); all models combining latent features with other data sources ranked plant SVD axis 1 followed by frugivore SVD axis 1 as the two most important variables, with some variation as the ordering of the rest of the SVD axis importance. The latent-only model generally followed similar patterns of variable importance as composite models, though frugivore SVD axis 1 had higher importance than the corresponding plant SVD axis. For the trio model, frugivore mass and fruit diameter were the next most important traits, which was consistent with the ordering of all other trait models. While the next most important trait was frugivore gape size for the trio model, for other

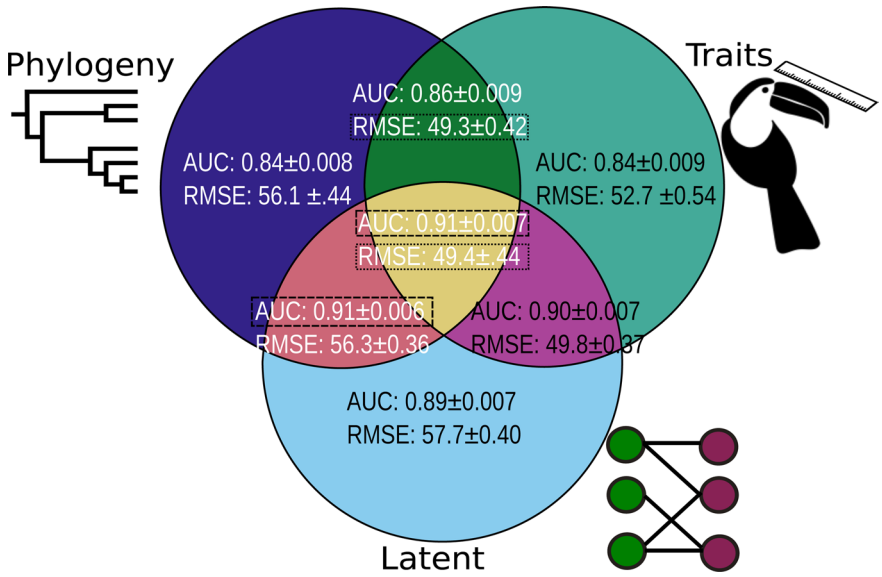

Figure 1. Summary performance metrics of all seven models, as measured by area under the receiver operating characteristic curve (AUC) and root mean square error (RMSE); highest performing models for each metric are outlined in dashed (AUC) or dotted lines (RMSE). Mean metric values are presented from 100 replicates of each model structure alongside standard deviation. Model discriminatory power between links and non-links is maximized by including latent structural features, with the inclusion of trait, phylogenetic information, or both actually slightly decreasing discriminatory power. However, inclusion of trait and phylogenetic information, while not improving AUC, does increase overall model accuracy as measured by mean root squared error. Pairwise t-test adjusted for multiple comparisons showed no significant difference between the discriminatory power of the highest performing models (Trio, PhyLatent, p = 0.13), or between the RMSE of the two highest accuracy models (Trio, PhyTraits p = 0.8).

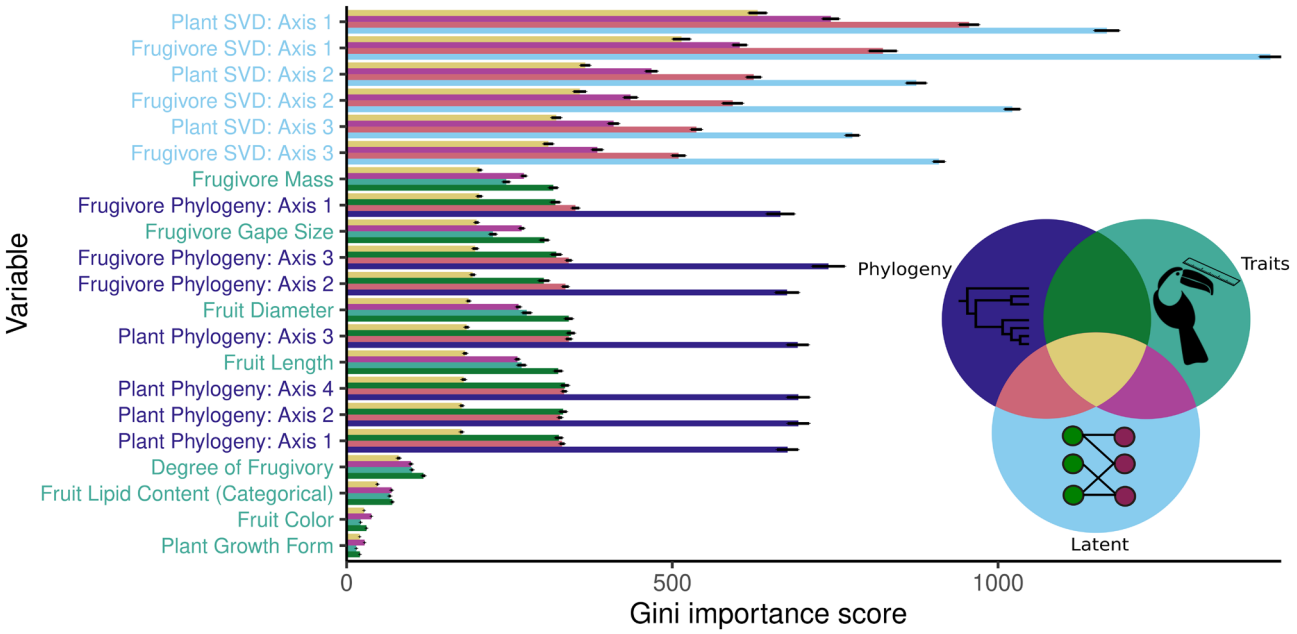

Figure 2. Variable importance across all models as measure by Gini importance score; color scheme is consistent with Fig. 1. In all models that include them, latent traits were consistently the most important variables for prediction. These were followed by continuous frugivore traits (body mass, gape size), and frugivore phylogenetic axes. Plant phylogenies and continuous trait information were generally less important for prediction than frugivore traits. Categorical plant traits (lipid content, fruit color, growth form) were the least important variables for prediction.

models incorporating trait information fruit length was the next most important trait.

In general, we found that continuous frugivore traits (body mass, gape width) were more informative to prediction than continuous plant traits. Similarly, phylogenetic information about frugivores was generally more important for prediction than phylogenetic information for plants. In the phylogenetic-only model, all three frugivore phylogenetic eigenvectors were more important than all plant eigenvectors, while for the trio model only plant eigenvalue axis 2 was more important than frugivore eigenvalue axis 2. Out of the remaining traits, the trio model was most informed by frugivory level, with the remaining categorical plant traits impacting Gini coefficient very little for all models.

Comparing models

Across the entire network, suitability values between all models were highly correlated (Fig. 3); most models were able to correctly discriminate between observed and unobserved links. While the ability to reconstitute observed links is important for model performance, focusing on model agreement on the suitability of unobserved interactions can provide insight into which models may identify different potential interactions that are yet unobserved but may occur given the appropriate ecological context. Across all potential links, Spearman's rank correlation of predicted suitability values showed low correlation between phylogenetic and trait models ($\rho = 0.643$), despite both exhibiting very similar overall performance. Phylogenetic and latent models showed the lowest agreement in suitability predictions of all pairwise combinations of models, with $\rho = 0.566$ when comparing across all links, which then dropped to $\rho = 0.032$ when only looking at unobserved links. This marked drop suggests that most of the agreement in suitability rankings between these models was in known interactions. The low agreement between the ranking of potential interactions poses a barrier to meaningfully target potential sampling based on the results of these two models used independently, highlighting the importance of composite models that synthesize multiple information types to create predictions.

Similarly, suitability predictions of composite and individual component models tended to mirror relative information type importance. For example, suitability values from the trait-latent model were more tightly correlated with those from the latent model ($\rho = 0.882$) than the trait model ($\rho = 0.773$). Table 2 presents the confusion matrix for all models after classifying points according to each models' average optimal threshold value across all training iterations. We see that at their optimal threshold value, all models exhibit markedly low false-negative rates with the exception of the trait only model, which exhibited a false-negative rate of 4.80%.

Discussion

Models incorporating latent structural features consistently outperformed models that excluded them in terms of discriminatory power, but had generally lower accuracy. In contrast, trait based and phylogenetic models displayed lower

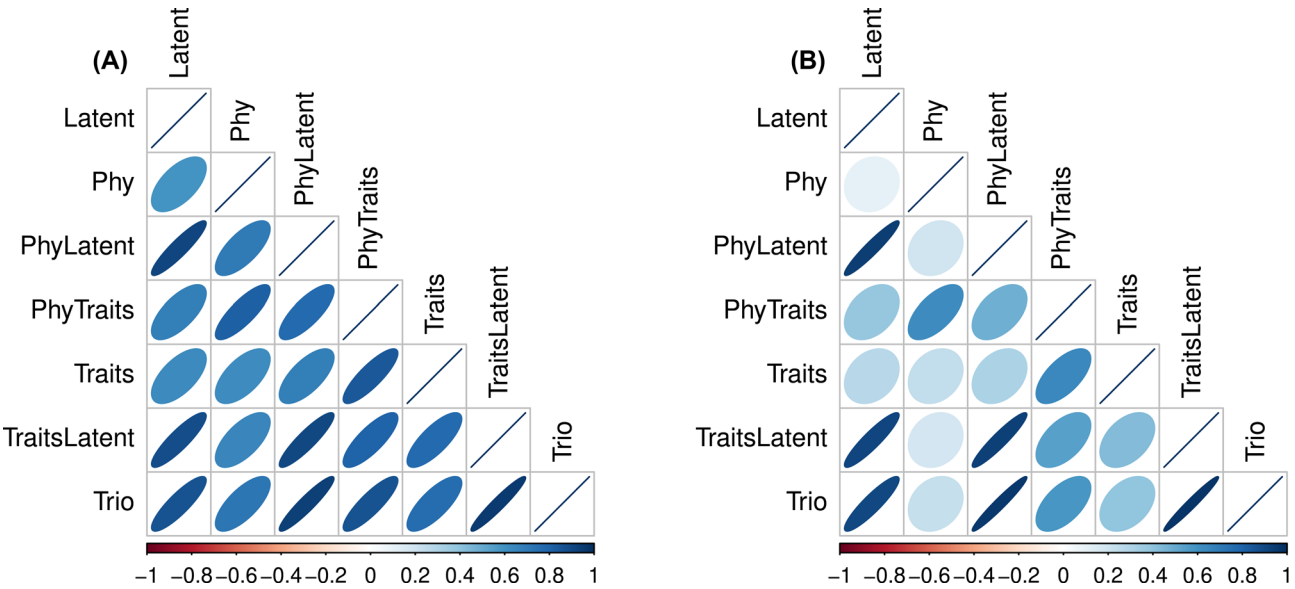

Figure 3. Pairwise Spearman's rank correlations of link suitabilities across models for all potential interaction (A), as well as only unobserved interactions (B). The latter set of links represents both true forbidden links, as well as other potential interactions not observed in our dataset.

AUC values, but a better RMSE. Overall performance of phylogenetic and trait-only models were remarkably similar in terms of both performance measures used. Incorporating latent features into each model once again yielded nearly identical discriminatory capacity, but an overall slightly higher accuracy for the trait-latent features model compared to the phylogeny-latent feature model. The discriminatory power of each individual model generally reflects the relative importance of variable information types (Fig. 2).

The strength of latent features for predicting interactions may lie in their particular suitability for relatively generalist interaction networks. Frugivorous vertebrates, for instance, often exhibit high levels of generalism in their interactions (Richardson et al. 2000), often more-so than their plant-partners (Salazar-Rivera et al. 2020). In systems characterized by low interaction specificity, latent features may be especially effective at capturing the neutral processes that underlie community assembly. Abundant, wide-ranging, or highly generalist species may drive a 'rich get richer' dynamic, where hyper-generalists that already interact with multiple species are more likely to accumulate additional partners. This pattern is reflected in our results: the 20 links with the highest predicted suitability from our trio model were spread across 16 bird species, many of which already participated in a large number of observed interactions. While the median number of interactions per bird in the dataset was 5, the median degree for these 16 species was 56. The top predicted missing link connected the azure-shouldered tanager *Thraupis cyanoptera* and *Myrsine umbellata*, two relatively common species that have indeed been documented as interacting in other datasets (Silva et al. 2002, Fricke and Svenning 2020). Host abundance, in particular, has been identified as a key driver of interaction probability in other Neotropical frugivory networks (Laurindo et al. 2020). Overall, latent features derived from network dimensionality reduction or motif completion may be especially powerful tools in the context of generalist interaction networks.

While comparing models in terms of discriminatory power highlights the importance of latent features, incorporating phylogenetic or trait information alongside latent

Table 2. True and false positive and negative rates for models trained on the entire interaction network, classified assuming the average threshold value across all training interactions that maximized Youden's *J* statistic. Percentages represent proportions in each category out of all total links each model predicted; the absolute count of each category is represented in parenthesis. Total number of predicted links differed across models according to availability of covariant data.

Model	True positives	False positives	True negatives	False negatives
Latent	3.85% (3643)	0.28% (263)	95.87% (90 654)	0.00% (0)
Phy	3.87% (3350)	0.27% (231)	95.66% (82 806)	0.29% (175)
Traits	1.86 % (778)	0.22% (92)	93.20% (39 021)	4.73% (1979)
PhyLatent	4.07% (3525)	0.124% (107)	95.80% (82 930)	0.00% (0)
PhyTraits	6.63% (2723)	0.35% (143)	93.02% (38 212)	0.01% (2)
TraitsLatent	6.59% (2757)	0.44% (185)	92.97% (38 928)	0.00% (0)
Trio	6.863% (2725)	0.15% (61)	93.22% (38 294)	0.00% (0)

features actually improves model accuracy measures (RMSE), despite slightly reducing discriminatory power. The question whether it is better to optimize one metric or the other, is largely dependent on the particular use case. As we've presented the problem of link prediction, primarily as a way to guide sampling efforts to more efficiently measure networks given limited resources, discriminatory power is likely a more apt metric. Beyond ranking different links, the relative confidence in each link is of less concern as the most likely links will be validated with additional observation data. However, if the aim was something closer to filling in a partially sampled network and then analyzing its properties, a metric like RMSE or H-measure (Hand 2009) may be more appropriate as it better allows you to directly incorporate uncertainty about particular linkages into your resultant analyses.

In models incorporating frugivore and plant phylogenetic information, the frugivore phylogeny tended to be more important to model performance. Ideally, this is likely because patterns of interactions are more phylogenetically conserved between frugivores than plants. On average, frugivores tend to interact with a more phylogenetically conserved group of plants than vice versa, with a given plant's fruit being consumed with a more phylogenetically dispersed set of avian taxa. Some of this phenomenon may be due to the frugivore phylogeny being more fully summarized by a smaller number of eigenvectors. On average, the first three eigenvectors of the decomposed frugivore phylogeny explained 46.02% of total variation across the phylogeny on average, while the first four eigenvectors only explained 37.5% of variation for plants. The difference in variation and resultant effects on prediction may be due to overall structural features of the phylogeny, or differences in the total phylogenetic divergence characterized by the tree. Frugivory is a strategy that spans the avian tree, occurring in a diverse set of taxonomic orders (Daniel Kissling et al. 2009), though this dataset reflects a relatively narrower set of frugivorous taxa. Out of our 242 avian species investigated in this study, the majority belonged to the orders *Passeriformes*, *Piciformes* and *Columbiformes*, comprising 193, 18 and 10 species, respectively. Ornithochory is also widely distributed across plant taxa in this system, a pattern reflected in other tropical systems (Kuhlmann and Ribeiro 2016, Pizo et al. 2021).

Our modeling approach represents a useful framework for predicting species interactions in a variety of systems, but care must be taken when interpreting the results in the particular context of each system.

While we ultimately find that latent features dominate prediction in our system, the potential for information leakage between test and train sets means we must be conservative in our interpretation of comparisons across models. Due to the nature of matrix decomposition approaches, there is no clear way to utilize either latent or phylogenetic features to predict out of bounds. Even when creating latent features from training subsets of interaction data, the dimensions of the training matrix must be fixed to the size of the full network in order to generate predictions for all species within the network. As such, while useful for imputing potentially missing interactions between known species, latent features are unsuitable for predicting interactions of species that are not already represented within the existing species pool. The dominance of latent features using internal cross-validation metrics support their importance in our system. However, as latent features are derived from portions of or all of the observed interaction network, their predictive capacity may be sensitive to network characteristics such as network size, connectance, and the proportion of links that are still unsampled (Supporting information). Additional work is needed to fully explore these relationships in order to explore in which contexts latent features might be more or less useful for prediction. Despite their caveats, latent features have already been successfully implemented to predict links in a variety of ecological network types, and can provide significant improvements in predictive performance in a variety of ecological (Poisot et al. 2021, 2023, Banville et al. 2023, Nunes Martinez and Mistretta Pires 2024) and non-ecological (Yeung et al. 2002, Wu et al. 2019, Zeng et al. 2020) contexts.

Future work investigating the utility of latent features to predict frugivorous interactions in a variety of empirical systems and the mechanisms through which they act represents an important avenue for future research. While there are currently no clear methods through which we can apply these methods for out-of-sample prediction, these methods are promising ways to target potentially missing links in existing networks. Similarly, our presented interpretation of the efficacy of phylogenetic information for predicting missing links assumes that phylogenetic relatedness reflects conserved traits or shared evolutionary histories that govern interactions. However, while phylogenetic relationships do impact at least some of the species traits governing interactions, there may be other confounding information (such as a phylogenetic biases in research effort) within phylogenetic information as well. While not reducing overall link prediction performance, this does mean we should once again take a more conservative interpretation of our prediction results when comparing across models, as is the case with any correlative approach.

Link prediction is a promising method to deal with the realities of incompletely sampled natural systems, and may play an increasingly important role in helping to gain insight into how interactions networks may change over time. Anthropogenic effects on climate, landcover, and biodiversity all have the potential to fundamentally alter mutualistic networks, whether through changing mutualistic assemblages themselves, or the probability of interactions within a network (Memmott et al. 2007, Tylianakis and Morris 2017, Teixido et al. 2022). Positive feedbacks within mutualistic communities may often make them more likely to exhibit alternative stable states, and changes in mutualistic communities may result in drastic changes to network structure and resulting assemblages (Lever et al. 2014, Bascompte and Scheffer 2023). Link prediction methods provide an opportunity for us to better understand which species are interacting now, and which may interact in the future. Prediction approaches can also be useful when predicting the potential

effects of invaders on mutualistic network structure (Traveset and Richardson 2014, Fricke and Svenning 2020). However, at its core, predicting which interactions are unsampled but present in a given ecological context is related to but distinct from the task of predicting how interactions may shift in response to changing biotic or abiotic conditions. If the eco-evolutionary processes governing the formation of novel interactions over short timescales are sufficiently different from those governing the maintenance and structure of long-standing interactions across a community, then different modeling approaches may vary in their effectiveness at predicting missing links versus truly novel ones. To fully understand and predict the scope of mutualistic network change, future work applying predictive approaches to temporally sampled mutualistic networks is necessary. Understanding the spatial and temporal scales at which mutualistic interactions turn over or change strength, and how we can utilize predictive approaches to forecast these changes is a vital next step for understanding the dynamics of mutualistic networks in a changing world.

Acknowledgements – We thank Dr Pedro R. Peres-Neto, who provided advice and code used to improve the cross-validation methods implemented in this manuscript.
Funding – This work has been performed with funding to TAD from the National Science Foundation (NSF-DEB-2017826) Macrosystems Biology and NEON-Enabled Science program.
Conflict of interest – The authors declare no conflict of interest.

Author contributions

Grant Foster: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (equal); Software (lead); Validation (equal); Visualization (lead); Writing – original draft (lead); Writing – review and editing (equal). **Tad A. Dallas**: Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Investigation (supporting); Methodology (equal); Project administration (lead); Resources (lead); Software (supporting); Supervision (lead); Validation (equal); Writing – review and editing (equal).

Data availability statement

All R code and data necessary to reproduce this is available on Figshare: https://figshare.com/s/544021fc48732ca5bb13 (Foster and Dallas 2025). A version-controlled history of this project is also publically available on Github at https://github.com/GTFoster/Fruglink.

Supporting information

The Supporting information associated with this article is available with the online version.

References

Banville, F., Gravel, D. and Poisot, T. 2023. What constrains food webs? a maximum entropy framework for predicting their structure with minimal biases. – PLoS Comp. Biol. 19: e1011458.
Bascompte, J. and Scheffer, M. 2023. The resilience of plant–pollinator networks. – Annu. Rev. Entomol. 68: 363–380.
Bello, C., Galetti, M., Montan, D., Pizo, M. A., Mariguela, T. C., Culot, L., Bufalo, F., Labecca, F., Pedrosa, F., Constantini, R., Emer, C., Silva, W. R., Da Silva, F. R., Ovaskainen, O. and Jordano, P. 2017. Atlantic frugivory: a plant–frugivore interaction data set for the Atlantic forest. – Ecology 98: 1729.
Bender, I. M. A., Kissling, W. D., Blendinger, P. G., Böhning-Gaese, K., Hensen, I., Kühn, I., Muñoz, M. C., Neuschulz, E. L., Nowak, L., Quitián, M., Saavedra, F., Santillán, V., Töpfer, T., Wiegand, T., Dehling, D. M. and Schleuning, M. 2018. Morphological trait matching shapes plant–frugivore networks across the Andes. – Ecography 41: 1910–1919.
Blüthgen, N., Fründ, J., Vázquez, D. P. and Menzel, F. 2008. What do interaction network metrics tell us about specialization and biological traits – Ecology 89: 3387–3399.
Braga, M. P., Janz, N., Nylin, S., Ronquist, F. and Landis, M. J. 2021. Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. – Ecol. Lett. 24: 2134–2145.
Breiman, L. 2001. Random forests. – Mach. Learn. 45: 5–32.
Calle, M. L. and Urrea, V. 2011. Stability of random forest importance measures. – Brief. Bioinform. 12: 86–89.
Canard, E., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D. and Gravel, D. 2014. Empirical evaluation of neutral interactions in host-parasite networks. – Am. Nat. 183: 468–479.
Carreira, D. C., Brodie, J. F., Mendes, C. P., Ferraz, K. M. P. and Galetti, M. 2020. A question of size and fear: competition and predation risk perception among frugivores and predators. – J. Mammal. 101: 648–657.
Cirtwill, A. R., Eklöf, A., Roslin, T., Wootton, K. and Gravel, D. 2019. A quantitative framework for investigating the reliability of empirical network construction. – Methods Ecol. Evol. 10: 902–911.
Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.
Dallas, T., Park, A. W. and Drake, J. M. 2017. Predicting cryptic links in host–parasite networks. – PLoS Comp. Biol. 13: e1005557.
Daniel Kissling, W., Böhning-Gaese, K. and Jetz, W. 2009. The global distribution of frugivory in birds. – Global Ecol. Biogeogr. 18: 150–162.
Dansereau, G., Barros, C. and Poisot, T. 2024. Spatially explicit predictions of food web structure from regional-level data. – Philos. Trans. R. Soc. B 379: 20230166.
Desjardins-Proulx, P., Laigle, I., Poisot, T. and Gravel, D. 2017. Ecological interactions and the Netflix problem. – PeerJ 5: e3644.
Eriksson, O. 2016. Evolution of angiosperm seed disperser mutualisms: the timing of origins and their consequences for coevolutionary interactions between angiosperms and frugivores. – Biol. Rev. 91: 168–186.
Foster, G. and Dallas, T. A. 2025. Data from: Comparing the power of phylogenetic, trait, and network structure information to predict plant–frugivore interactions. – Figshare , https://figshare.com/s/544021fc48732ca5bb13..
Fricke, E. C. and Svenning, J.-C. 2020. Accelerating homogenization of the global plant–frugivore meta-network. – Nature 585: 74–78.
González-Castro, A., Yang, S., Nogales, M. and Carlo, T. A. 2015. Relative importance of phenotypic trait matching and species'

abundances in determining plant–avian seed dispersal interactions in a small insular community. – AoB Plants. 7: plv017.
Guimaraes Jr, P. R. 2020. The structure of ecological networks across levels of organization. – Annu. Rev. Ecol. Evol. Syst. 51: 433–460.
Hand, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the roc curve. – Mach. Learn. 77: 103–123.
Hao, X., Holyoak, M., Zhang, Z. and Yan, C. 2025. Global projection of terrestrial vertebrate food webs under future climate and land-use changes. – Global Change Biol. 31: e70061.
Holm, S. 1979. A simple sequentially rejective multiple test procedure. – Scand. J. Stat. 6: 65–70.
Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. and Mooers, A. O. 2012. The global diversity of birds in space and time. – Nature 491: 444–448.
Jordano, P. 1987. Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries and coevolution. – Am. Nat. 129: 657–677.
Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P. and Webb, C. O. 2010. Picante: R tools for integrating phylogenies and ecology. – Bioinformatics 26: 1463–1464.
Khalilia, M., Chakraborty, S. and Popescu, M. 2011. Predicting disease risks from highly imbalanced data using random forest. – BMC Med. Inform. Decis. Mak. 11: 51.
Kuhlmann, M. and Ribeiro, J. F. 2016. Fruits and frugivores of the Brazilian cerrado: ecological and phylogenetic considerations. – Acta Bot. Bras. 30: 495–507.
Laurindo, R.d. S., Vizentin-Bugoni, J., Tavares, D. C., Mancini, M. C. S., Mello, R. and Gregorin, R. 2020. Drivers of bat roles in neotropical seed dispersal networks: abundance is more important than functional traits. – Oecologia 193: 189–198.
Lever, J. J., van Nes, E. H., Scheffer, M. and Bascompte, J. 2014. The sudden collapse of pollinator communities. – Ecol. Lett. 17: 350–359.
Liaw, A. and Wiener, M. 2002. Classification and regression by randomforest. – R News 2: 18–22.
Maglianesi, M. A., Varassin, I. G., Ávalos, G. and Jorge, L. R. 2024. A phylogenetic perspective on ecological specialisation reveals hummingbird and insect pollinators have generalist diets. – Oikos 2024: e10208.
Maitner, B. S., et al. 2018. The bien R package: a tool to access the botanical information and ecology network (bien) database. – Methods Ecol. Evol. 9: 373–379.
Memmott, J., Craze, P. G., Waser, N. M. and Price, M. V. 2007. Global warming and the disruption of plant–pollinator interactions. – Ecol. Lett. 10: 710–717.
Menke, S., Böhning-Gaese, K. and Schleuning, M. 2012. Plant–frugivore networks are less specialized and more robust at forest–farmland edges than in the interior of a tropical forest. – Oikos 121: 1553–1566.
Minoarivelo, H. O. and Hui, C. 2016. Invading a mutualistic network: to be or not to be similar. – Ecol. Evol. 6: 4981–4996.
Moran, C. and Catterall, C. P. 2010. Can functional traits predict ecological interactions? A case study using rain forest frugivores and plants in australia. – Biotropica 42: 318–326.
Norberg, A. et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. – Ecol. Monogr. 89: e01370.
Nunes Martinez, A. and Mistretta Pires, M. 2024. Estimated missing interactions change the structure and alter species roles in one of the world's largest seed-dispersal networks. – Oikos 2024: e10521.
Olesen, J. M., Bascompte, J., Dupont, Y. L., Elberling, H., Rasmussen, C. and Jordano, P. 2011a. Missing and forbidden links in mutualistic networks. – Proc. R. Soc. B 278: 725–732.
Olesen, J. M., Stefanescu, C. and Traveset, A. 2011b. Strong, long-term temporal dynamics of an ecological network. – PLoS One 6: e26455.
Pearse, I. S. and Altermatt, F. 2013. Predicting novel trophic interactions in a non-native world. – Ecol. Lett. 16: 1088–1094.
Pérez, F., Arroyo, M. T. and Medel, R. 2007. Phylogenetic analysis of floral integration in *Schizanthus* (Solanaceae): does pollination truly integrate corolla traits? – J. Evol. Biol. 20: 1730–1738.
Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. and Hartig, F. 2020. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. – Methods Ecol. Evol. 11: 281–293.
Pizo, M. A., Morales, J. M., Ovaskainen, O. and Carlo, T. A. 2021. Frugivory specialization in birds and fruit chemistry structure mutualistic networks across the neotropics. – Am. Nat. 197: 236–249.
Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N. and Carlson, C. J. 2021. Imputing the mammalian virome with linear filtering and singular value decomposition. – arXiv:2105.14973 [q-Bio].
Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Brierley, L., Albery, G. F., Gibb, R. J., Seifert, S. N. and Carlson, C. J. 2023. Network embedding unveils the hidden interactions in the mammalian virome. – Patterns 4: 100738.
Putman, B. J., Williams, R., Li, E. and Pauly, G. B. 2021. The power of community science to quantify ecological interactions in cities. – Sci. Rep. 11: 3069.
Quintero, E., Isla, J. and Jordano, P. 2022. Methodological overview and data-merging approaches in the study of plant–frugivore interactions. – Oikos 2022: e08379.
Rafferty, N. E. and Ives, A. R. 2013. Phylogenetic trait-based analyses of ecological networks. – Ecology 94: 2321–2333.
Ramos-Robles, M., Dáttilo, W., Díaz-Castelazo, C. and Andresen, E. 2018. Fruit traits and temporal abundance shape plant-frugivore interaction networks in a seasonal tropical forest. – Sci. Nat. 105: 29.
Richardson, D. M., Allsopp, N., D'Antonio, C. M., Milton, S. J. and Rejmánek, M. 2000. Plant invasions – the role of mutualisms. – Biol. Rev. 75: 65–93.
Salazar-Rivera, G. I., Dáttilo, W., Castillo-Campos, G., Flores-Estévez, N., Ramírez García, B. and Ruelas Inzunza, E. 2020. The frugivory network properties of a simplified ecosystem: birds and plants in a Neotropical periurban park. – Ecol. Evol. 10: 8579–8591.
Saravia, L. A., Marina, T. I., Kristensen, N. P., De Troch, M. and Momo, F. R. 2022. Ecological network assembly: how the regional metaweb influences local food webs. – J. Anim. Ecol. 91: 630–642.
Seo, E. and Hutchinson, R. 2018. Predicting links in plant–pollinator interaction networks using latent factor models with implicit feedback. – In: Proc. AAAI Conf. on artificial intelligence, vol. 32. No. 1, pp. 808–815.
Silva, W. R., Marco Júnior, P. D. and Hasui, É. and Gomes, V. S. 2002. Patterns of fruit–frugivore interactions in two Atlantic forest bird communities of south-eastern brazil: implications for conservation. – In: Seed dispersal and frugivory: ecology, evolution and conservation. 3rd Int. Symp.-workshop on frugivores

and seed dispersal, São Pedro, Brazil, 6–11 August 2000. CABI publishing, pp. 423–435.
Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L. and Poisot, T. 2021. A roadmap towards predicting species interaction networks (across space and time). – Philos. Trans. R. Soc. B 376: 20210063.
Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M. J., Fortin, M.-J., Hemming, V., Mercier, B., Pollock, L. J., Runghen, R., Dalla Riva, G. V. and Poisot, T. 2022. Food web reconstruction through phylogenetic transfer of low-rank network representation. – Methods Ecol. Evol. 13: 2838–2849.
Sydenham, M. A., Venter, Z. S., Reitan, T., Rasmussen, C., Skrindo, A. B., Skoog, D. I. J., Hanevik, K.-A., Hegland, S. J., Dupont, Y. L., Nielsen, A., Chipperfield, J. and Rusch, G. M. 2022. Metacomnet: a random forest-based framework for making spatial predictions of plant–pollinator interactions. – Methods Ecol. Evol. 13: 500–513.
Teixido, A. L., Fuzessy, L. F., Souza, C. S., Gomes, I. N., Kaminski, L. A., Oliveira, P. C. and Maruyama, P. K. 2022. Anthropogenic impacts on plant–animal mutualisms: a global synthesis for pollination and seed dispersal. – Biol. Conserv. 266: 109461.
Terry, J. C. D. and Lewis, O. T. 2020. Finding missing links in interaction networks. – Ecology 101: e03047.
Traveset, A. and Richardson, D. M. 2014. Mutualistic interactions and biological invasions. – Annu. Rev. Ecol. Evol. Syst. 45: 89–113.
Tylianakis, J. M. and Morris, R. J. 2017. Ecological networks across environmental gradients. – Annu. Rev. Ecol. Evol. Syst. 48: 25–48.
Vázquez, D. P., Blüthgen, N., Cagnolo, L. and Chacoff, N. P. 2009. Uniting pattern and process in plant–animal mutualistic networks: a review. – Ann. Bot. 103: 1445–1457.
Wheelwright, N. T. and Orians, G. H. 1982. Seed dispersal by animals: contrasts with pollen dispersal, problems of terminology and constraints on coevolution. – Am. Nat. 119: 402–413.
Wu, G., Liu, J. and Yue, X. 2019. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. – BMC Bioinform. 20: 134.
Yeung, M. S., Tegnér, J. and Collins, J. J. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. – Proc. Natl Acad. Sci. USA 99: 6163–6168.
Youden, W. J. 1950. Index for rating diagnostic tests. – Cancer 3: 32–35.
Young, J.-G., Valdovinos, F. S. and Newman, M. E. J. 2021. Reconstruction of plant–pollinator networks from observational data. – Nat. Commun. 12: 3911.
Zeng, M., Lu, C., Zhang, F., Li, Y., Wu, F.-X., Li, Y. and Li, M. 2020. Sdlda: lncrna-disease association prediction based on singular value decomposition and deep learning. – Methods 179: 73–80.