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

ANONYMIZED

February 28, 2025

1 Abstract

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 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 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, mutualistic network, link prediction, seed dispersal, mutualism

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 (1; 2). 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 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) (3). Moreover, the likelihood of observing a given interaction is related to factors such as the abundances of both interactors (4), with the total interaction set of rare species less likely to be represented accurately (5; 6). Errors in the construction of ecological networks, especially when biased, can lead to dramatic changes in network structure (7; 3). 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 (8).

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(9), as well efforts to extract interaction data from existing large-scale occurrence datasets (10) 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 (11). 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 (12), and ultimately a better understanding of how species interaction networks may be altered by anthropogenic forcing.

Symbiotic interactions are dependent on a complex set of factors including life history (13), phenotypic traits (14), co-evolutionary history (15), spatial and temporal distributions (16; 17; 18), as well other biotic interactions (19). Incorporating information that either directly influences, or conveys information about

these factors into link prediction frameworks should hopefully improve prediction performance (20). 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 node morphological and life history traits (gape-width, fruit characteristics, geographic or phenological or overlap), may be at the root of why many species do or do not interact (21; 22; 23). 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 (eg. flower and beak morphometric traits in flower-hummingbird pollinator systems (24). 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, making phylogenetic information an imperfect proxy for traits governing interactions (14). 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 net-

works may in generally be less coevolutionarily linked and more asymmetric some other classes of ecological networks(25; 26; 27), 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 mutualistic network systems, including plant-pollinator interactions (28).

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 properties derived from the structure of the interaction network through any number of dimensionality-reduction approaches (29). While there are a multitude of different methods used to create latent network features for interaction prediction (single-value decomposition, random dot product graphs (30), graph motif distributions, matrix factorization (31), 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 foodwebs (32) and host-virus associations (29; 33), and seed-dispersal interactions (34). 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 (35). Despite the use of this variety of information streams for link prediction problems already, 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 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 species

¹¹² trait data, phylogenetic information, and latent structural features.

Methods

¹¹⁴ **Trait and Interaction Data** The full interaction network is published by Bello et al. (36), and includes
¹¹⁵ a total of $5,226$ unique species interactions between 787 and 342 plant and frugivore species, respectively.
¹¹⁶ Due to the availability of phylogenetic and trait data, we restricted our analyses to avian-plant interactions,
¹¹⁷ representing the most species-rich set of frugivores in this dataset ($3,856$ unique interactions between 394 and
¹¹⁸ 242 plant and bird species, respectively). Of our avian-frugivore species (hereafter frugivores), numbers of
¹¹⁹ unique plant interactions per frugivore ranged from 1 (53 species) to 120 unique plant interactions recorded
¹²⁰ for *Turdus rufiventris*; median frugivore degree was 5. The number of unique frugivore interactions per plant
¹²¹ was generally lower, ranging from 1 (128 species) to 80 unique frugivore interactions recorded for *Myrsine*
¹²² *coriacea*; median plant degree was 2. Overall this interaction network is characterized by a relatively high
¹²³ degree of phylogenetic generalism. Out of 187 frugivore species that interacted with more than one plant for
¹²⁴ which phylogenetic information is available, only 9 interacted with a group of plants more phylogenetically
¹²⁵ related than random chance ((37), see supplement**??**).

¹²⁶   The Bello et al. dataset includes a number of informative traits for both plant a nd frugivore nodes.
¹²⁷ Frugivore traits include allometric measures such as body mass and gape size, life history traits such as
¹²⁸ degree of frugivory and migratory status, and population traits such as IUCN listing and population trends.
¹²⁹ Frugivore mean body mass was unavailable for 3 species, while mean gape size was unavailable for 68 species.
¹³⁰ Plant traits include fruit mass, seed diameter fruit color, growth form, fruit lipid concentration (scored 1-3),
¹³¹ IUCN listing, and whether it is a native or nonnative species. For trait-based models, species were filtered
¹³² to only include those with complete sets of trait data (174/242 frugivore species). As part of our research
¹³³ goal was to compare the power of phylogenetic and trait data, we chose to omit these species for trait based
¹³⁴ models rather than phylogenetically impute missing trait information. Categorical variables (such as plant
¹³⁵ growth form) were transformed into a series of binary variables through one-hot encoding. Continuous data
¹³⁶ on fruit size was available for 280/394 plant species; plants without trait data were again excluded from
¹³⁷ trait-based models. Details on plant and frugivore traits used in the analysis are available in Table 1).

¹³⁸ **Phylogenetic Data** For plant species, phylogenetic data was retrieved from the BIEN database, accessed
¹³⁹ using the r package 'BIEN' (38). Out of 787 plant species in our dataset, 646 occurred in the BIEN
¹⁴⁰ phylogeny. 61 plant species 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 (see supplement). Bird phylogenies were retrieved from VertLife (39); 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 respect 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 $46.02 \pm 2.29\%$ of variation for birds.

Latent Methods We used single-value decomposition (SVD) of the full interaction matrix to create continuous latent features to be used for prediction using the 'SVD()' function in base R (40). We used the first three axis of variation for both plants and birds as continuous predictors. Due to the incompatibility of latent network features with out of sample prediction, during model validation we derived node-level latent properties form the full interaction network, though each training set utilized only a portion of these nodes.

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 (41). We tested a total of 7 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(42). Our total potential interactions matrix included 3,643 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. 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 discriminator power, and mean root squared error (RMSE), which measures prediction accuracy. These

6

two measures are commonly used to evaluate the effectiveness of predictive approaches (43), but the choice of which metric to optimize may be different depending on the use case in question (see discussion). For each model iteration we also recorded the optimal suitability threshold for classification as the value that maximized Youden's J (44).

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 mean decrease in Gini coefficient (MDG) after permutation of each variable as implemented in the R `randomForest` package (41). A common variable importance metric for classification problems, MDG is calculated by normalizing the sum of all decreases node impurity by the total amount of decision trees in the ensemble (45; 46). 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 (Figure 1). AUC was lowest for trait only models (0.84 ±.009) followed by phylogeny only models (0.85 ±.009). All models including latent information outperformed latent-agnostic models in terms of discriminatory power, with the greatest AUC achieved by Latent (0.94 ±.004) and phylogeny-latent models (0.94 ±.004), both of which even outperformed the full trio model (0.93 ±.006). However, prediction accuracy did not follow the same trends as discriminatory power. Latent and trait-only models had the lowest overall accuracy (56.0, ±0.32 amd ±0.40, respectively), while the most accurate models were those combining trait information with another information type. The trio model was the most accurate (RMSE: 48.6 ±.37), closely followed by the trait-phylogeny (RMSE: 49.2 ±.36) and trait-latent models (RMSE: 49.0 ±.37). The trait-only model had only middling accuracy performance however (RMSE: 52.8 ±.57).

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 (Figure 2); all models latent models ranked plant SVD axis 1 as the most important followed by frugivore SVD axis 1, with some variation as the ordering of the rest of the SVD axis importance. For the trio model, frugivore mass and fruit diameter were the next to 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 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 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 (Figure 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 suitability of unobserved interactions can provide insight into which models may identify different potential interactions. 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 suggest that suggests an that most agreement between suitability ranking between these models was in known interactions. The low agreement between the ranking of potential interactions pose a barrier to meaningfully target potential sampling based on the results of these two models used independently.

8

Similarly, suitabilty predictions of composite and individual component models tended to mirror relative information type importance. For example, suitability values from 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 AUC values, but better a better RMSE. Overall performance of phylogenetic and trait-only models were remarkable 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 2.

The strength of latent features for prediction may stem from the fact that latent features are particularly suited to predicting interactions in relatively generalist interaction networks. Frugivorous vertebrates often tend to exhibit a high average degree of generalism (47). In systems characterized by low interaction specificity, latent features may be more efficient at capturing neutral processes that may be underlying community assembly processes. Especially abundant, wide-ranging, or generalist hosts may contribute to a "rich get richer" effect, where hyper-generalists that already interact with many species likely interact with even more partners. Host abundance in particular has been shown to be an important determinant of interaction probability in other neotropical frugivory networks (18). Latent features, such as those created through network dimensionality reduction or motif completion, may perform particularly well in generalist interaction networks.

While comparing models in terms of discriminatory power highlights the importance of latent features, incorporating phylogenetic or trait information alongside latent features actually improves model accuracy measures (RMSE), despite slightly reducing discriminatory power. The question whether it is better to

9

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 on limited resource, 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 your task was something closer to filling in a partially sampled network and then analyzing its properties, a metric like RMSE or H-measure(48) may be more appropriate as it better allows you to directly incorporate uncertainty about particular into your resultant analyses.

In models incorporating frugivore and plant phylogenitic 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. This may be due to the frugivore phylogeny being more full 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.

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, neither latent features nor phylogenetic information can be used to predict out of bounds. The decompositions must be performed on the entire network before performing the test-train split, potentially leading to information leakage. However, this phenomenon is only a potential problem when comparing performances on subdivided training data, and should not affect our predictions using the entire network. The dominance of latent features using internal cross-validation metrics support their importance in our system. 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 contexts.(32; 29; 33; 34) 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. Similarly, our presented interpretations of the effect of phylogenetic information on prediction to their ability assume they at least partially capture phylogenetically conserved traits or aspects of shared evolutionary

10

history. 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 of studiedness) 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 (49; 50; 51). 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 (52; 53). 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 (16; 54). 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 is a vital next step for understanding the dynamics of mutualistic networks in a changing world.

References

[1] Guimaraes Jr PR. 2020 The structure of ecological networks across levels of organization. *Annual Review of Ecology, Evolution, and Systematics* **51**, 433–460. (http://dx.doi.org/https://doi.org/10.1146/annurev-ecolsys-012220-120819https://doi.org/10.1146/annurev-ecolsys-012220-120819)

[2] Saravia LA, Marina TI, Kristensen NP, De Troch M, Momo FR. 2022 Ecological network assembly: how the regional metaweb influences local food webs. *Journal of Animal Ecology* **91**, 630–642.

[3] Young JG, Valdovinos FS, M EJN. 2021 Reconstruction of plant–pollinator networks from observational data. *Nature Communications* **12**, 3911. (http://dx.doi.org/https://doi.org/10.1038/s41467-021-24149-xhttps://doi.org/10.1038/s41467-021-24149-x)

[4] Canard E, Mouquet N, Mouillot D, Stanko M, Miklisova D, Gravel D. 2014 Empirical evaluation of neutral interactions in host-parasite networks. *The American Naturalist* **183**, 468–479.

[5] Cirtwill AR, Eklöf A, Roslin T, Wootton K, Gravel D. 2019 A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution* **10**, 902–911.

[6] Olesen JM, Bascompte J, Dupont YL, Elberling H, Rasmussen C, Jordano P. 2011 Missing and forbidden links in mutualistic networks. *Proceedings of the Royal Society B: Biological Sciences* **278**, 725–732.

[7] Blüthgen N, Fründ J, Vázquez DP, Menzel F. 2008 What do interaction network metrics tell us about specialization and biological traits. *Ecology* **89**, 3387–3399. (http://dx.doi.org/https://doi.org/10.1890/07-2121.1https://doi.org/10.1890/07-2121.1)

[8] Olesen JM, Stefanescu C, Traveset A. 2011 Strong, long-term temporal dynamics of an ecological network. *PloS one* **6**, e26455. (http://dx.doi.org/https://doi.org/10.1371/journal.pone.0026455https://doi.org/10.1371/journal.pone.0026455)

[9] Quintero E, Isla J, Jordano P. 2022 Methodological overview and data-merging approaches in the study of plant–frugivore interactions. *Oikos* **2022**.

[10] Putman BJ, Williams R, Li E, Pauly GB. 2021 The power of community science to quantify ecological interactions in cities. *Scientific reports* **11**, 3069.

[11] Terry JCD, Lewis OT. 2020 Finding missing links in interaction networks. *Ecology* **101**, e03047.

[12] Minoarivelo HO, Hui C. 2016 Invading a mutualistic network: to be or not to be similar. *Ecology and evolution* **6**, 4981–4996. (http://dx.doi.org/https://doi.org/10.1002/ece3.2263https://doi.org/10.1002/ece3.2263)

[13] Ramos-Robles M, Dáttilo W, Díaz-Castelazo C, Andresen E. 2018 Fruit traits and temporal abundance shape plant-frugivore interaction networks in a seasonal tropical forest. *The Science of Nature* **105**, 1–11. (http://dx.doi.org/https://doi.org/10.1007/s00114-018-1556-yhttps://doi.org/10.1007/s00114-018-1556-y)

[14] Rafferty NE, Ives AR. 2013 Phylogenetic trait-based analyses of ecological networks. *Ecology* **94**, 2321–2333. (http://dx.doi.org/https://doi.org/10.1890/12-1948.1https://doi.org/10.1890/12-1948.1)

[15] Eriksson O. 2016 Evolution of angiosperm seed disperser mutualisms: the timing of origins and their consequences for coevolutionary interactions between angiosperms and frugivores. *Biological Reviews* **91**, 168–186.

[16] Fricke EC, Svenning JC. 2020 Accelerating homogenization of the global plant–frugivore meta-network. *Nature* **585**, 74–78. (http://dx.doi.org/https://doi.org/10.1038/s41586-020-2640-yhttps://doi.org/10.1038/s41586-020-2640-y)

[17] Menke S, Böhning-Gaese K, 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.

[18] Laurindo RdS, Vizentin-Bugoni J, Tavares DC, Mancini MCS, Mello RdM, Gregorin R. 2020 Drivers of bat roles in Neotropical seed dispersal networks: abundance is more important than functional traits. *Oecologia* **193**, 189–198. (http://dx.doi.org/https://doi.org/10.1007/s00442-020-04662-4https://doi.org/10.1007/s00442-020-04662-4)

[19] Carreira DC, Brodie JF, Mendes CP, Ferraz KMP, Galetti M. 2020 A question of size and fear: competition and predation risk perception among frugivores and predators. *Journal of Mammalogy* **101**, 648–657.

[20] Dallas T, Park AW, Drake JM. 2017 Predicting cryptic links in host-parasite networks. *PLoS computational biology* **13**, e1005557.

[21] Bender IM, Kissling WD, Blendinger PG, Böhning-Gaese K, Hensen I, Kühn I, Muñoz MC, Neuschulz EL, Nowak L, Quitián M et al.. 2018 Morphological trait matching shapes plant–frugivore networks across the Andes. *Ecography* **41**, 1910–1919. (http://dx.doi.org/https://doi.org/10.1111/ecog.03396https://doi.org/10.1111/ecog.03396)

[22] González-Castro A, Yang S, Nogales M, Carlo TA. 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. (http://dx.doi.org/https://doi.org/10.1093/aobpla/plv017https://doi.org/10.1093/aobpla/plv017)

[23] Moran C, Catterall CP. 2010 Can functional traits predict ecological interactions? A case study using rain forest frugivores and plants in Australia. *Biotropica* **42**, 318–326.

[24] Pichler M, Boreux V, Klein AM, Schleuning M, Hartig F. 2020 Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution* **11**, 281–293. (http://dx.doi.org/https://doi.org/10.1111/2041-210X.13329https://doi.org/10.1111/2041-210X.13329)

[25] Maglianesi MA, Varassin IG, Ávalos G, Jorge LR. 2024 A phylogenetic perspective on ecological specialisation reveals hummingbird and insect pollinators have generalist diets. *Oikos* p. e10208.

[26] Jordano P. 1987 Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. *The American Naturalist* **129**, 657–677.

[27] Wheelwright NT, Orians GH. 1982 Seed dispersal by animals: contrasts with pollen dispersal, problems of terminology, and constraints on coevolution. *The American Naturalist* **119**, 402–413.

[28] Braga MP, Janz N, Nylin S, Ronquist F, Landis MJ. 2021 Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. *Ecology Letters* **24**, 2134–2145. (http://dx.doi.org/https://doi.org/10.1111/ele.13842https://doi.org/10.1111/ele.13842)

[29] Poisot T, Ouellet MA, Mollentze N, Farrell MJ, Becker DJ, Albery GF, Gibb RJ, Seifert SN, Carlson CJ. 2021 Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv preprint arXiv:2105.14973 [q-Bio]*.

[30] Strydom T, Bouskila S, Banville F, Barros C, Caron D, Farrell MJ, Fortin MJ, Hemming V, Mercier B, Pollock LJ et al.. 2022 Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods in Ecology and Evolution* **13**, 2838–2849.

[31] Seo E, Hutchinson R. 2018 Predicting links in plant-pollinator interaction networks using latent factor models with implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 32.

[32] Banville F, Gravel D, Poisot T. 2023 What constrains food webs? A maximum entropy framework for predicting their structure with minimal biases. *PLOS Computational Biology* **19**, e1011458. (http://dx.doi.org/https://doi.org/10.1371/journal.pcbi.1011458https://doi.org/10.1371/journal.pcbi.1011458)

[33] Poisot T, Ouellet MA, Mollentze N, Farrell MJ, Becker DJ, Brierley L, Albery GF, Gibb RJ, Seifert SN, Carlson CJ. 2023 Network embedding unveils the hidden interactions in the mammalian virome. *Patterns* **4**.

[34] Nunes Martinez A, 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* p. e10521.

[35] Strydom T, Catchen MD, Banville F, Caron D, Dansereau G, Desjardins-Proulx P, Forero-Muñoz NR, Higino G, Mercier B, Gonzalez A et al.. 2021 A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B* **376**, 20210063.

[36] Bello C, Galetti M, Montan D, Pizo MA, Mariguela TC, Culot L, Bufalo F, Labecca F, Pedrosa F, Constantini R et al.. 2017 Atlantic frugivory: a plant–frugivore interaction data set for the Atlantic Forest. *Ecology* **98**, 1729. (http://dx.doi.org/https://doi.org/10.1002/ecy.1818https://doi.org/10.1002/ecy.1818)

[37] Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010 Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464.

[38] Maitner BS, Boyle B, Casler N, Condit R, Donoghue J, Durán SM, Guaderrama D, Hinchliff CE, Jørgensen PM, Kraft NJ et al.. 2018 The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods in Ecology and Evolution* **9**, 373–379. (http://dx.doi.org/https://doi.org/10.1111/2041-210X.12861https://doi.org/10.1111/2041-210X.12861)

[39] Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012 The global diversity of birds in space and time. *Nature* **491**, 444–448. (http://dx.doi.org/https://doi.org/10.1038/nature11631https://doi.org/10.1038/nature11631)

[40] R Core Team. 2021 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[41] Liaw A, Wiener M. 2002 Classification and Regression by randomForest. *R News* **2**, 18–22.

[42] Vázquez DP, Blüthgen N, Cagnolo L, Chacoff NP. 2009 Uniting pattern and process in plant–animal mutualistic networks: a review. *Annals of botany* **103**, 1445–1457. (http://dx.doi.org/https://doi.org/10.1093/aob/mcp057https://doi.org/10.1093/aob/mcp057)

[43] Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J, Araújo MB, Dallas T, Dunson D, Elith J et al.. 2019 A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs* **89**, e01370.

[44] Youden WJ. 1950 Index for rating diagnostic tests. *Cancer* **3**, 32–35. (http://dx.doi.org/https://doi.org/10.1002/1097-0142(1950)3:1

[45] Khalilia M, Chakraborty S, Popescu M. 2011 Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making* **11**, 1–13.

[46] Calle ML, Urrea V. 2011 Stability of Random Forest importance measures. *Briefings in bioinformatics* **12**, 86–89.

[47] Richardson DM, Allsopp N, D'Antonio CM, Milton SJ, Rejmánek M. 2000 Plant invasions–the role of mutualisms. *Biological Reviews* **75**, 65–93. (http://dx.doi.org/https://doi.org/10.1017/S0006323199005435https://doi.org/10.1017/S0006323199005435)

[48] Hand DJ. 2009 Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning* **77**, 103–123. (http://dx.doi.org/https://doi.org/10.1007/s10994-009-5119-5https://doi.org/10.1007/s10994-009-5119-5)

[49] Tylianakis JM, Morris RJ. 2017 Ecological networks across environmental gradients. *Annual Review of Ecology, Evolution, and Systematics* **48**, 25–48. (http://dx.doi.org/https://doi.org/10.1146/annurev-ecolsys-110316-022821https://doi.org/10.1146/annurev-ecolsys-110316-022821)

[50] Memmott J, Craze PG, Waser NM, Price MV. 2007 Global warming and the disruption of plant–pollinator interactions. *Ecology letters* **10**, 710–717. (http://dx.doi.org/https://doi.org/10.1111/j.1461-0248.2007.01061.xhttps://doi.org/10.1111/j.1461-0248.2007.01061.x)

[51] Teixido AL, Fuzessy LF, Souza CS, Gomes IN, Kaminski LA, Oliveira PC, Maruyama PK. 2022 Anthropogenic impacts on plant-animal mutualisms: A global synthesis for pollination and seed dispersal. *Biological Conservation* **266**, 109461. (http://dx.doi.org/https://doi.org/10.1016/j.biocon.2022.109461https://doi.org/10.1016/j.biocon.2022.109461)

[52] Lever JJ, van Nes EH, Scheffer M, Bascompte J. 2014 The sudden collapse of pollinator communities. *Ecology letters* **17**, 350–359. (http://dx.doi.org/https://doi.org/10.1111/ele.12236https://doi.org/10.1111/ele.12236)

[53] Bascompte J, Scheffer M. 2022 The Resilience of Plant–Pollinator Networks. *Annual Review of Entomology* **68**. (http://dx.doi.org/https://doi.org/10.1146/annurev-ento-120120-102424https://doi.org/10.1146/annurev-ento-120120-102424)

[54] Traveset A, Richardson DM. 2014 Mutualistic interactions and biological invasions. *Annual Review of Ecology, Evolution, and Systematics* **45**, 89–113.

Tables

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	factor	low, medium, high
Plant	Fruit Diameter	mm	1.2-325
	Fruit Length	mm	1-600
	Fruit Color	mm	yellow, red, black, brown, green, other
	Fruit Lipid Content	factor	low, medium, high
	Plant Growth Form	factor	tree, liana, palm, scrub, other

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.

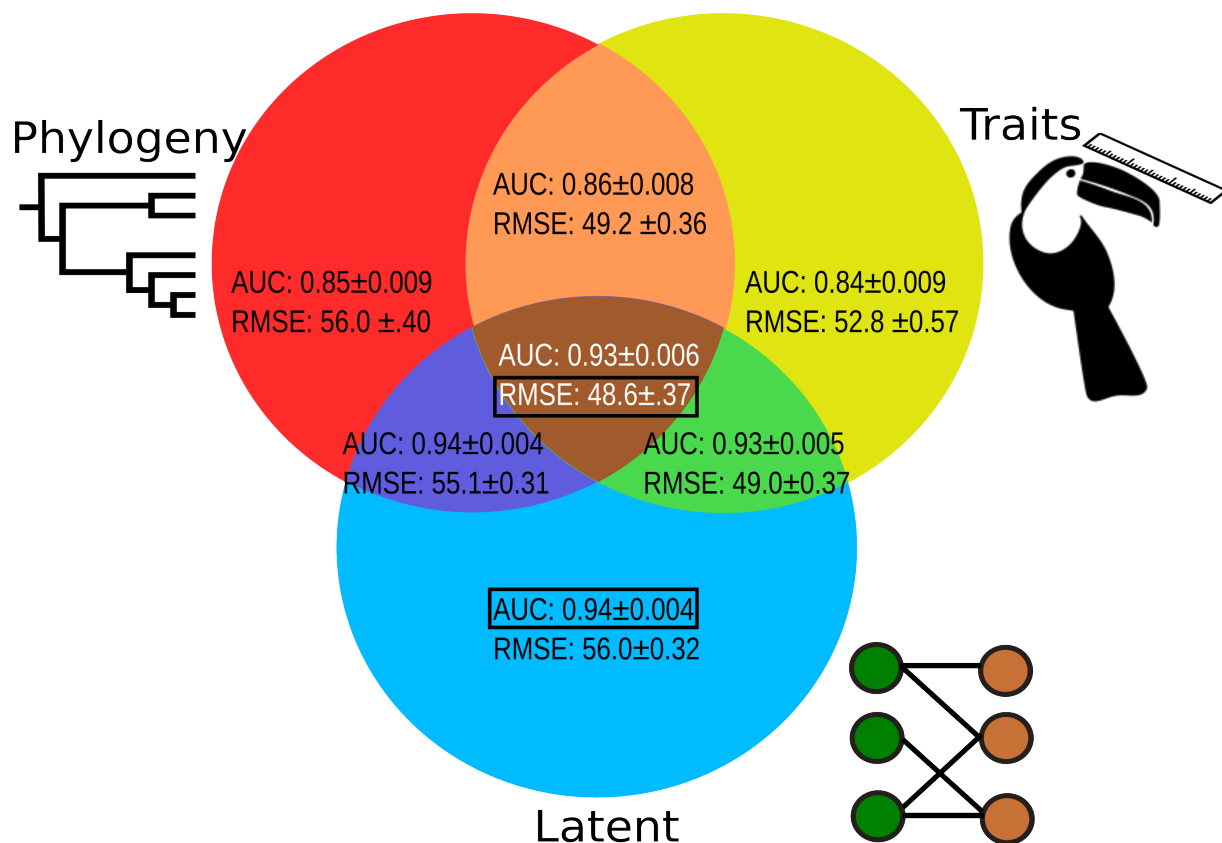| Model	True Positives	False Positives	True Negatives	False Negatives
Latent	3.85% (3,643)	0.28% (263)	95.87% (90,654)	0.00% (0)
Phy	3.87% (3,350)	0.27% (231)	95.66% (82,806)	0.29% (175)
Traits	1.86 % (778)	0.22% (92)	93.20% (39,021)	4.73% (1,979)
PhyLatent	4.07% (3,525)	0.124% (107)	95.80% (82,930)	0.00% (0)
PhyTraits	6.63% (2,723)	0.35% (143)	93.02% (38,212)	0.01% (2)
TraitsLatent	6.59% (2,757)	0.44% (185)	92.97% (38,928)	0.00% (0)
Trio	6.863% (2,725)	0.15% (61)	93.22% (38,294)	0.00% (0)

Figures

Figure 1: Summary performance metrics of all 7 models, as measured by area under the receiver operating characteristic curve (AUC) and root mean squared error (RMSE); highest performing models for each metric are outlined in black. 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.
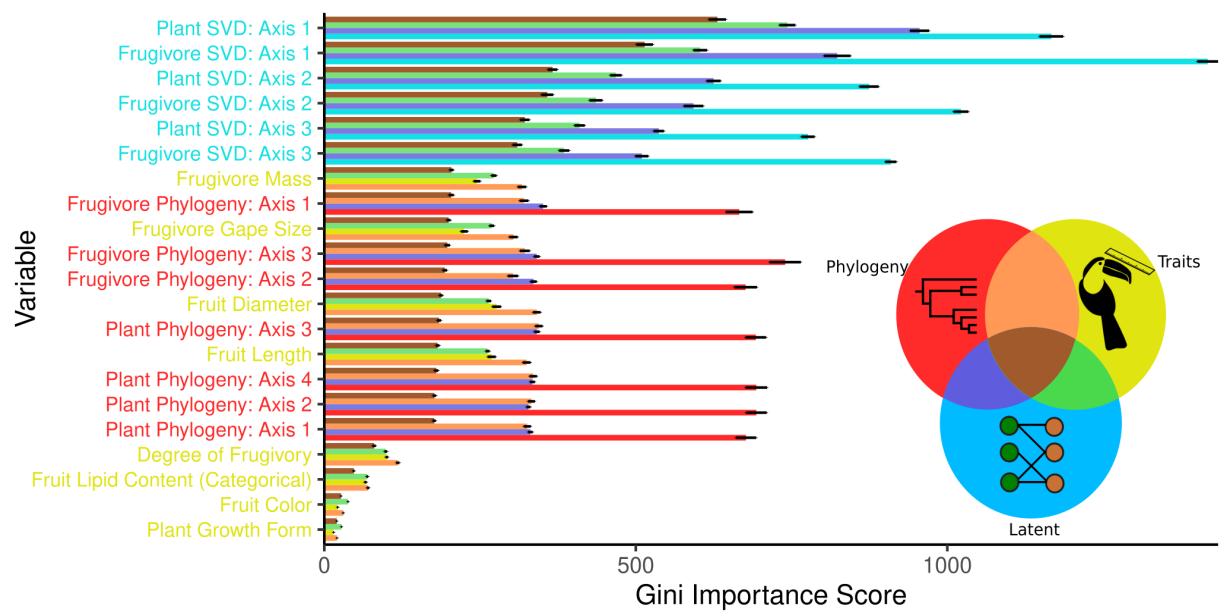
Figure 2: Variable importance across all models as measure by Gini importance score; color scheme is consistent with figure 1. In all models that include them, latent traits were consistently the most important variables for prediction. These were followed by continuous frugivore traits (body mass, gape 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.
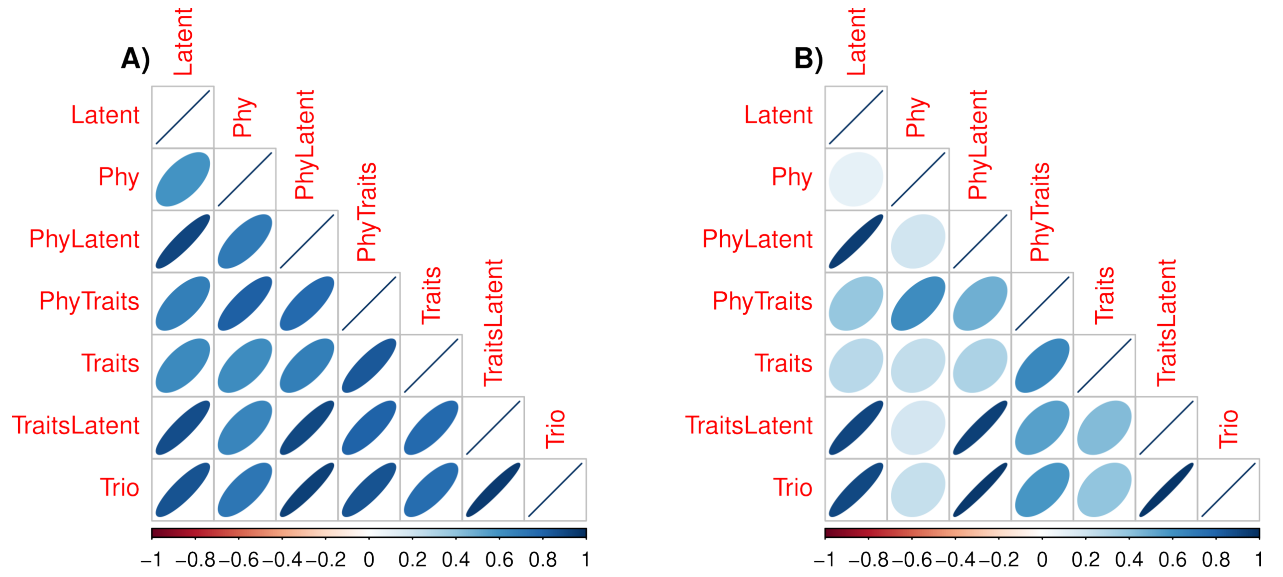
Figure 3: Pairwise Spearman's rank correlations of link suitabilities across models for all potential interaction (A), as well as only unobserved interactions (B). The latter set of links represents both true forbidden links, as well as other potential interactions not observed in our data-set.