nature communications

Article

https://doi.org/10.1038/s41467-026-68769-7

Predicting missing links in food webs using stacked models and species traits

Received: 27 November 2024

Accepted: 16 January 2026

Published online: 03 February 2026

Check for updates

Lucy B. Van Kleunen[1,2,3] ✉, Laura E. Dee[4], Kate L. Wootton[5], François Massol[6,7] & Aaron Clauset[1,8,9] ✉

Networks are a powerful way to represent the complexity of large ecological systems. However, most ecological networks, such as food webs, contain only partial lists of species interactions. Computational methods for inferring missing links can facilitate field work and investigations of ecological processes. Here, we describe a stacked generalization approach to predict missing links in food webs that accounts for ecological assumptions including link direction. Tests of this method on synthetic food webs show that it can learn to optimally combine structural and trait-based predictions. On a global database of 290 food webs, the method often achieves near-perfect performance, performs better when it can exploit both species traits and network structure, and is principally driven by a subset of ecologically-interpretable predictors. Furthermore, we find that link predictability varies with ecosystem and network characteristics. These results show broad applicability of stacked generalization for predicting and understanding ecological interactions.

Many complex social, technological, and biological systems can be represented as networks, defined as a set of nodes, e.g., individual species, people, genes, or even places, along with their pairwise interactions, e.g., feeding relationships between species in food webs, friendships in social networks, regulatory interactions between genes, or traffic flows among places. However, nearly all empirical networks are incomplete, because real links can be unobserved, unmeasured, hidden, or inaccessible. For example, in food webs representing species feeding relationships, both hard to observe feeding events and rare interactions may be missing. Although many methods now exist for inferring such missing links based on their correlation with a network's partially observed structure[1–4], we lack highly-accurate methods that leverage the particular characteristics of food webs to make ecologically accurate predictions of species interactions.

Broadly, link prediction methods based on network structure can be grouped into three classes[2]: those that predict missing links based on (i) the pattern of links local to where a missing link may occur, (ii) large-scale models of the entire network's structure (e.g., grouping structure), and (iii) node proximity within a learned embedding of the network. Systematic evaluations of link prediction methods using large corpora of structurally diverse empirical networks indicate that there is no universally best method for all networks[2], and the best approach depends on the particular network. Among modern link prediction methods, the meta-learning approach of stacked generalization[5], or model stacking, is a state-of-the-art technique that can learn from patterns among observed network interactions how to optimally combine many individual link predictors to produce highly accurate predictions in real-world social, biological, and technological networks[2]. Using existing stacking methods, missing links in social networks are the easiest to recover, while missing links in biological networks, including food webs, remain substantially harder to predict[2].

[1]Department of Computer Science, University of Colorado, Boulder, CO, USA. [2]Department of Public Health and Primary Care, KU Leuven Campus Kulak, Kortrijk, Belgium. [3]Itec, imec research group at KU Leuven, Kortrijk, Belgium. [4]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. [5]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand. [6]University of Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000, Lille, France. [7]Sorbonne Université, Université Paris Cité, Université Paris Est Créteil, CNRS, INRAE, IRD, Institut d'Ecologie et des Sciences de l'Environnement de Paris (UMR7618), 75005, Paris, France. [8]BioFrontiers Institute, University of Colorado, Boulder, CO, USA. [9]Santa Fe Institute, Santa Fe, NM, USA. ✉e-mail: lucyvankleunen@gmail.com; aaron.clauset@colorado.edu

Hence, tailored approaches are required in particular domains like predicting missing links in food webs. Food webs are often used as models of ecosystem structure to assess ecosystem vulnerability to disturbances in theoretical studies and applied conservation contexts[6,7]. Assembling a food web is labor-intensive, often requiring researchers to carefully identify and combine feeding interactions recorded in the literature with new field observations or experimental results[8,9], as well as collect or assemble trait data for member species. Feeding links between species can be identified via a number of methods, including expert elicitation, direct observation in the field or in the lab, and molecular analyses of gut content, feces, tissues, or museum specimens. However, because the number of possible feeding links grows quadratically with the number of species considered, while the number of true feeding links typically grows only linearly, distinguishing every true link from all true non-links in even a modest-sized food web can be prohibitive. Hence, the links present in most food-web datasets are incompletely sampled[10,11]. More accurate methods for predicting missing links in food webs would benefit both empirical and theoretical studies in community ecology by increasing the efficiency of collecting species interaction data in the field and providing more reliable insights into questions about ecosystem stability, conservation efforts, and tests of ecological theories[12].

Food webs have three distinguishing characteristics that existing stacking methods do not account for in missing link prediction. First, species attributes, or traits, like feeding mode, trophic level, body mass, and metabolic type constrain the set of ecologically feasible feeding links[13,14]. Several studies match species foraging traits with vulnerability traits that constrain interactions (known as trait matching)[13,15–20]. Second, feeding links are directional. That is, we must not just predict that an interaction exists between two species, but also the correct direction of the feeding interaction and whether it is reciprocated. Third, while food webs are similar to social networks in often exhibiting skewed degree distributions[21] and compartmentalized grouping or community structure[22], they also have structural properties that are different from social networks, in particular exhibiting fewer triangles and a globally hierarchical structure[1,23]. State-of-the-art model stacking approaches do not currently exploit node attributes or link directionality, and are not customized to expect global hierarchical structure, which limits their utility for making accurate predictions in food webs. Here, we develop a stacking model specifically designed to exploit these features to make more accurate predictions of missing feeding links.

Food webs provide an ideal setting for exploring how the stacking model approach can be adapted to a specific class of biological networks where interactions are directional. We build on substantial previous work on applying individual link prediction methods to partially observed food webs[15,24–29] and other ecological networks[1,11,14,20,29,30]. Mirroring work on networks in general[2], work on missing link prediction in food webs has found that there is no universally best predictor for such missing links[11,20,24–26,29]. Meta-learning exploits the fact that many prediction methods work well in practice but do so using complementary underlying signals. By learning to optimally combine these signals, meta-learning can substantially improve prediction accuracies. Here, we build upon past explorations in ecology that have combined prediction methods via averaging, multiplication, or summation of predictors[11,31]. Model stacking generalizes the approach of combining methods by algorithmically constructing an optimal predictive distribution from individual predictors for a particular data set[5]. We do so in this case over an ensemble of simple predictors, many of which have not been used previously for ecological link prediction. Such a meta-learning approach is an attractive strategy for missing link prediction in food webs because it allows us to be relatively agnostic about the theoretical basis of particular distinct predictors, while also using data to guide their combination into a single prediction algorithm that takes advantage of whatever structural regularities are present. In addition, the resulting model can often be interpreted to yield insights into the underlying processes shaping the network[32].

Here, we comprehensively develop a stacking model for missing link prediction that follows ecological assumptions about food webs. Given the importance of species traits in determining interactions, we add predictors based on species traits (node attributes), taking into account both assortative and disassortative structure. We extend the approach from Ghasemian et al. 2020[2] to directed networks, adapt 6 predictors based on network connectivity patterns to the specific food-web context based on their expected hierarchical structure, and add K-nearest-neighbor predictors to the ensemble. We first evaluate this method using a class of synthetic food webs with known structure, which allows us to systematically vary the degree to which links exist due to node attributes or network structure. We then apply the method to a global database of 290 food webs with species trait annotations[33].

We find that model stacking is able to optimize prediction for a given network based on whether species traits, connectivity patterns, or a combination thereof are most useful for predicting missing links. This approach is highly predictive of missing interactions in real food webs, and the best performance is generally achieved by combining both types of information. By assessing how performance varies across ecosystem types and network characteristics, we find that missing links are easiest to predict in terrestrial belowground food webs, and in food webs that are larger, have better taxonomic resolution, are more connected, and are less modular. Further, our results illustrate some of the ecological insights that can be obtained with a method that flexibly learns highly accurate prediction rules for specific food webs. Across 290 food webs, we find that ecosystem type correlates with the relative performance of attribute vs. structure-based predictors as groups, and with which individual predictors are most important for predicting missing links. At the same time, we identify a subset of predictors that are broadly important across ecosystem types, suggesting common underlying processes that structure these networks. These results demonstrate how a model stacking approach that is adapted specifically to ecological networks can produce both highly accurate missing link predictions in food webs and provide insights into food-web organization for the development and verification of ecological theory.

Results

In missing link prediction, the true network $G = (V, E)$ is defined by a set of nodes V and a set of edges or links E, but is incompletely observed. The observed network $G' = (V, E')$ has the same set of nodes V but only the observed subset of edges $E' \subset E$. Our stacked generalization model for food webs, adapted from Ref. 2 for directed and attributed networks, is described in detail in the Methods section. Briefly, this supervised learning approach uses an ensemble of missing link predictors, based on network structure and node attributes, to learn how to score all potentially missing (unobserved) links in a particular observed food web G' such that higher scoring candidates are more likely to be missing links. Hence, this approach defines a network-specific model that learns from a particular network's observed edges and unique characteristics.

In contrast to past work[2], our stacked model adapts individual link predictors to food webs by grounding them in ecological assumptions. For example, the common neighbors predictor predicts that a link is more likely to be missing if it would connect a pair of nodes who have many neighbors in common; we modify this idea so that the direction of the predicted connection aligns with ecological assumptions about trophic hierarchies (Fig. 1), and we define a set of predictors based on assortativity among node attributes (Fig. 2).

https://doi.org/10.1038/s41467-026-68769-7

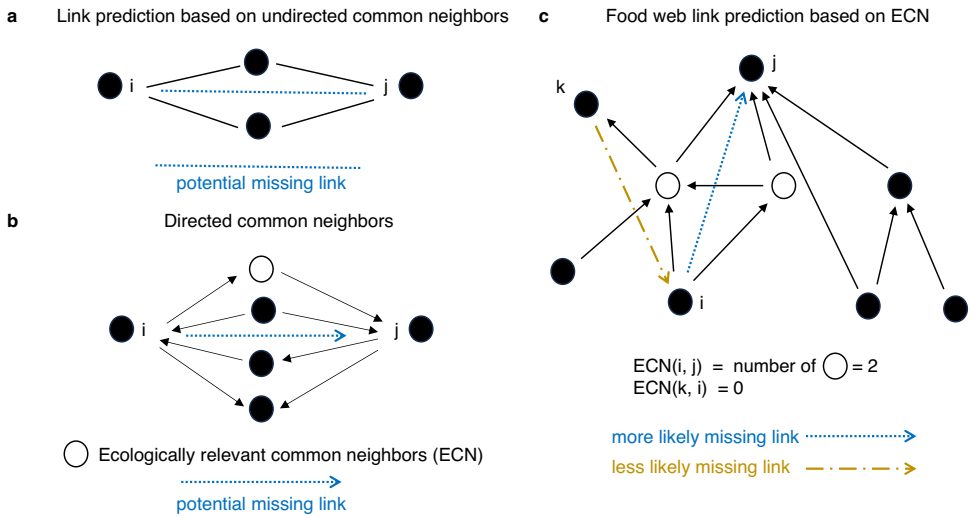

Fig. 1 | Adapting topological predictors to food webs. a In undirected networks, the common neighbors predictor assumes that the more neighbors two unconnected nodes i and j have in common, the greater the likelihood that (i, j) is a missing link. **b** For an unconnected pair i, j in a directed network, there are four distinct arrangements of directed connections with a common neighbor of i and j. If there is a missing link between i and j, the arrangement whose edge directions align with the network's trophic hierarchy is the more ecologically likely. **c** For example, the ecologically relevant common neighbors (ECN) predictor predicts a missing link (i, j) (dashed arrow) because i, j have two ecologically relevant common neighbors (open circles), and not the link (k, i) (dot-dashed arrow) which have none.

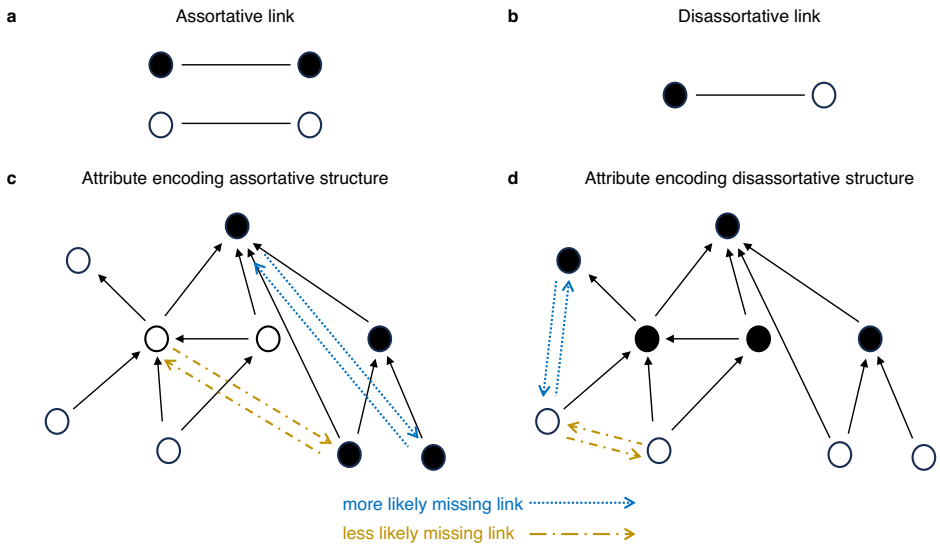

Fig. 2 | Assortative and disassortative patterns. Node attributes can encode **a** assortative patterns, e.g., common environmental conditions, and **b** disassortative patterns, e.g., a species trait that differs by trophic level, which can be exploited to predict missing links. **c** When node attributes encode assortative information, missing links tend to occur between nodes with similar attributes. **d** In contrast, when node attributes encode disassortative information, missing links will tend to occur between nodes with different attributes. The stacked model (see text) allows us to include both types of attribute predictors and to use data to learn which are most useful in a given network.

Performance on synthetically generated networks

We first evaluate the accuracy of the stacked model using synthetically generated food webs with node attributes. In this controlled setting, the true data generating processes are known, allowing us to calculate the theoretical maximum accuracy, and we can adjust the extent to which the probability of an edge depends on its nodes' attributes. These tests allowed us to evaluate the stacking model performance on a difficult task for the model, and to see whether the model could learn to predict missing links using network structure, node attributes, or a combination.

In empirical networks with node annotations, observed node attributes often correlate with the network structure, and hence they also correlate with missing links; however, other structural patterns relevant to missing link prediction do not appear to correlate with node attributes[13,24,26]. We incorporate these patterns into our synthetic networks by defining a parameter $\rho \in [0, 1]$, which tunes the probability that an edge's existence depends on node attributes. When $\rho = 0$, the network structure is completely independent of all node attributes and when $\rho = 1$, the structure is completely determined by node attributes. This range of dependency is accomplished by creating a pair of anchor networks with the same number of nodes and approximately the same number of edges, one with strong latent topological structure unrelated to node attributes and one with structure fully determined by node attributes. To generate a network

with some mixture of these patterns, ρ specifies the fraction of edges sampled from the first anchor network, with the remaining edges sampled from the second.

To generate anchor networks with network structure that is independent of node attributes but with structure that is nevertheless similar to that found in food webs, we used the stochastic block model (SBM)[34]. In the SBM, nodes are assigned to groups and the probability of a link (i, j) depends on the group assignments of the nodes i, j. To emulate role-based trophic grouping structure in food webs, e.g., between predators, herbivores, and primary producers, the direction of the generated edges were chosen to replicate expected hierarchical structure in food webs (see Materials & Methods). To generate anchor networks with network structure that is fully determined by node attributes, we used the random geometric graph model (RGG)[35] to generate networks with assortative or disassortative structure. Assortative structure is found in food webs when, e.g., node attributes related to environmental conditions correlate with interaction probability (e.g., fish swimming depth[13]). Disassortative structure is found in food webs when, for example, nodes differ in traits between trophic levels. In the RGG, the probability that a pair of nodes is connected is given by an attachment function parameterized by the Euclidean distance $d(i, j)$ between a pair of nodes' trait vectors, e.g., a decreasing function of distance in the case of assortative networks (see Materials & Methods).

Using synthetic networks to measure the performance of link prediction algorithms allows us to calculate the theoretical maximum prediction performance[2] in terms of the standard Area Under the ROC (Receiver Operating Characteristics) Curve (ROC-AUC) statistic[36], using the underlying probability of missing edges in the synthetic network model (see Supplementary Note 1, Supplementary Fig. 1), and to measure performance systematically. The ROC-AUC provides a threshold- and scale-invariant measure of an algorithm's ability to distinguish missing links (true positives) from non-edges (true negatives). We performed tests on these synthetic networks by dividing the observed links uniformly at random into 5 equal-sized groups and performing 5-fold cross validation for each algorithm's link prediction performance. That is, in each iteration, we removed (held out) a distinct 20% of the observed links from each food web (validation set), and we predicted these missing links using models trained on the other 80% of links (training set, see Materials & Methods). Under this scheme, we systematically evaluated three different versions of the stacked model for varying ρ (Fig. 3): a model with structural predictors only (structure-only model) (47 predictors, Supplementary Table 1), a model with node attribute predictors only (attribute-only model) (10 fixed predictors, raw attribute values, and ratios of numeric attributes, Supplementary Table 2), and a model with both types of predictors (full model). The structure-only model also included a subset of 4 nearest neighbor predictors based entirely on network structure and the full model was the only model containing 8 nearest neighbor predictors that combined node attributes with a node's local topology (Supplementary Table 3).

In this experiment, the performance of the structure model was highest when the synthetic network's topology was drawn from the SBM anchor network ($\rho \approx 0.0$), and nearly as high in the limiting case of drawing edges only from the RGG anchor network ($\rho \approx 1.0$), in both assortative and disassortative cases. This U shape in the performance reflects the fact that, like SBM networks, RGG networks also have specific topological patterns in network connectivity (e.g. clustering of (dis)similar nodes) that the structure model can exploit to predict missing links. However, these patterns are distinct from those induced by group-based generation of edges, and hence the model's performance was lowest when edges were drawn with nearly equal probability from the SBM anchor and the RGG anchor networks ($\rho \approx 0.6$). In the limiting case of all edges drawn from the SBM anchor network, the structure model matched the theoretical maximum accuracy, and was only slightly suboptimal in the limiting case of all edges drawn from the RGG anchor network.

The performance of the attribute model was lowest when a majority of synthetic network's topology was drawn from the SBM anchor network ($\rho < 0.5$), and was only slightly better than chance (ROC-AUC ≈ 0.56) when more than 80% of edges were drawn from the SBM anchor network. In contrast, the attribute model's performance improved steadily as the fraction of edges drawn from the RGG anchor network increased above 20%. And, in the limiting case of all edges drawn from the RGG anchor network, the attribute model matched the theoretical maximum accuracy.

The full model performed well across all values of ρ, matching or exceeding the structure model for small values of ρ and matching or exceeding the attribute model for large values of ρ. The full model also exceeded the performance of both alternative models in the more difficult middle range of the mixing parameter $\rho \approx 0.5$, demonstrating that the stacking model successfully learned how to best combine structure and attribute based predictors for missing link prediction, without knowing the underlying generative process.

Performance on empirical food webs

We then evaluated the three stacked models—structure-only, attribute-only, and full models—using a large global database of empirical food webs[33,37], which includes 290 networks across five ecosystem types (lakes, marine, streams, terrestrial aboveground, and terrestrial belowground) with a common set of species traits as attributes for each node: log body mass, movement type, and metabolic type (Table 1, see the Supplementary Note 2, Supplementary Fig. 2–8, Supplementary Table 4–6 for details on the food-web database and pre-processing). We expected these traits to constrain interactions based on prior analyses of predator-prey interactions in this database[37]. We considered nodes at the species-lifestage level given traits might differ across lifestages for the same species, though we found that aggregating traits across all lifestages of the same species in a food web dataset did not change results (Supplementary Fig. 3, Supplementary Table 7).

For each food web, we divide the observed links uniformly at random into five equal-sized groups and perform 5-fold cross validation to assess each algorithm's performance in realistic settings. We repeat this procedure for five independent iterations, thus averaging results over 25 iterations per food web. As with the synthetic networks, we measure algorithm performance using the ROC-AUC statistic. In addition, we measure the Area Under the Precision-Recall Curve (PR-AUC) statistic. The PR-AUC provides a complementary measure to the ROC-AUC by emphasizing an algorithm's ability to recover missing links (true positives) rather than simply its ability to correctly assign positives and negatives, which is useful in this context given that most food webs are sparse, and thus the missing link prediction problem is typically imbalanced[31]. Across our evaluations, we examine performance relative to random baselines for both ROC-AUC and PR-AUC metrics. The baseline for the ROC-AUC is 0.5, while the baseline for PR-AUC differs by food web and is equal to the proportion of true positives (missing links) in the test set.

All three models produce mean ROC-AUC and PR-AUC scores far above the baselines for each empirical food web. Reflecting the versatility we observed on synthetic networks, the full model gives the highest mean performance across the 290 networks on both ROC-AUC and PR-AUC metrics (Fig. 4a, b), with mean ROC-AUC $= 0.95 \pm 0.06$ and PR-AUC $= 0.68 \pm 0.20$ (mean $\pm$ stddev), versus 0.94 ± 0.06 and 0.62 ± 0.20, respectively, for the structure-only model, and 0.88 ± 0.06 and 0.35 ± 0.20 for the attribute-only model (Supplementary Table 8). At the same time, however, on about 10% of the individual food webs for ROC-AUC and PR-AUC, the attribute-only or structure-only models marginally outperformed the other models (Fig. 4c, d). Together, these results indicate that (i) both network structure and species traits are

https://doi.org/10.1038/s41467-026-68769-7

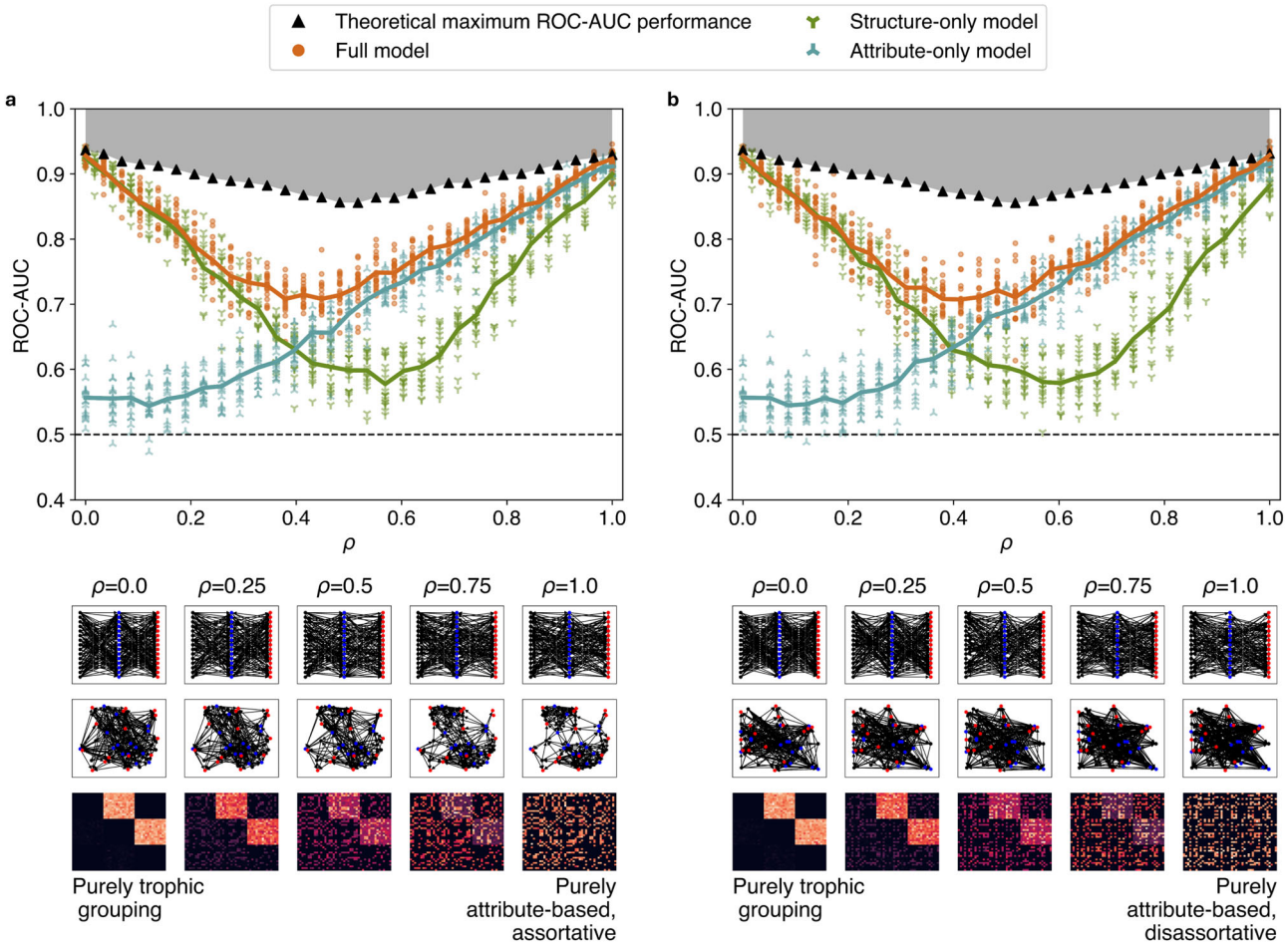

Fig. 3 | Link prediction on synthetic networks with known structure. We used three stacked models: a structure-only model (51 predictors, green Ys), an attribute-only model (20 predictors, blue inverted Ys), and a full model (orange circles) that includes both structure and attributes (79 predictors). Synthetic networks are a variable mix (see text) of purely trophic grouping structure without attributes ($\rho = 0$), and purely **a** assortative or **b** disassortative attribute-based connections without trophic groupings ($\rho = 1$). Thumbnails show example visualizations of mixture networks for specific choices of ρ, generated with only two numeric node attributes for easy visualization: in the first row, nodes are separated into three trophic groups (black, blue, and red); in the second row the nodes are still colored by group, but nodes are positioned based on the two node attributes; and the third row shows adjacency matrices. Main panels show the Area Under the ROC (Receiver Operating Characteristics) Curve (ROC-AUC) performance as a function of the mixing parameter ρ for 20 iid synthetic networks evaluated at each of 30 ρ values, in each case averaged across 5-fold cross validation (N=3000 results for each of the three models, N=600 average results across folds per model shown on the scatter plot). Solid lines connect the mean performance across the 20 networks for each model at each ρ value. The baseline ROC-AUC is shown at 0.5 (dashed line) and the theoretical maximum performance at each ρ (black triangles). For both types of attribute pattern, the full model exhibits the best ROC-AUC performance at all values of ρ, matching the accuracy of the structure-only model when $\rho = 0$ and of the attribute-only model when $\rho = 1$. Moreover, at intermediate values of ρ, the full model performs better than either alternative model.

Table 1 | Species trait data included as node attributes in all 290 empirical food web studied here

Trait name	Type	Definition
Metabolic type	Categorical (one-hot encoded)	9 possible values: invertebrate, primary producer, ectotherm vertebrate, endotherm vertebrate, detritus, heterotrophic fungi, heterotrophic bacteria, dead organic material, other
Movement type	Categorical (one-hot encoded)	7 possible values: walking, swimming, sessile, flying, other, floating, other_nonliving
Log_{10} mass	Numeric (float)	Log base 10 of mean mass in grams of the population involved in the interaction

useful for predicting missing links in food webs, and (ii) the most accurate predictions are typically achieved by combining structural and trait information, i.e., they encode marginally different and complementary information about the existence of links.

The food webs in the empirical corpus we evaluated can be divided into five ecosystem types, which allows us to compare how link prediction performance varies with ecosystem characteristics. As with the aggregate results, we find that the full model is highly accurate at distinguishing missing links (true positives) from non-edges (true negatives). Moreover, the full model performs best on average in all five ecosystem types based on both performance metrics (Fig. 5, Supplementary Fig. 9, Supplementary Table 9), ranging from ROC-AUC = 0.99 ± 0.01 on terrestrial below-ground food webs to 0.93 ± 0.06 on marine food webs and 0.93 ± 0.07 on terrestrial above-ground food webs. Precision-recall AUC ranged from PR-AUC = 0.82 ± 0.11 on terrestrial below-ground food webs to 0.48 ± 0.18 on terrestrial above-ground food webs. In each ecosystem, the structure-only model performed only marginally worse on average than the full

https://doi.org/10.1038/s41467-026-68769-7

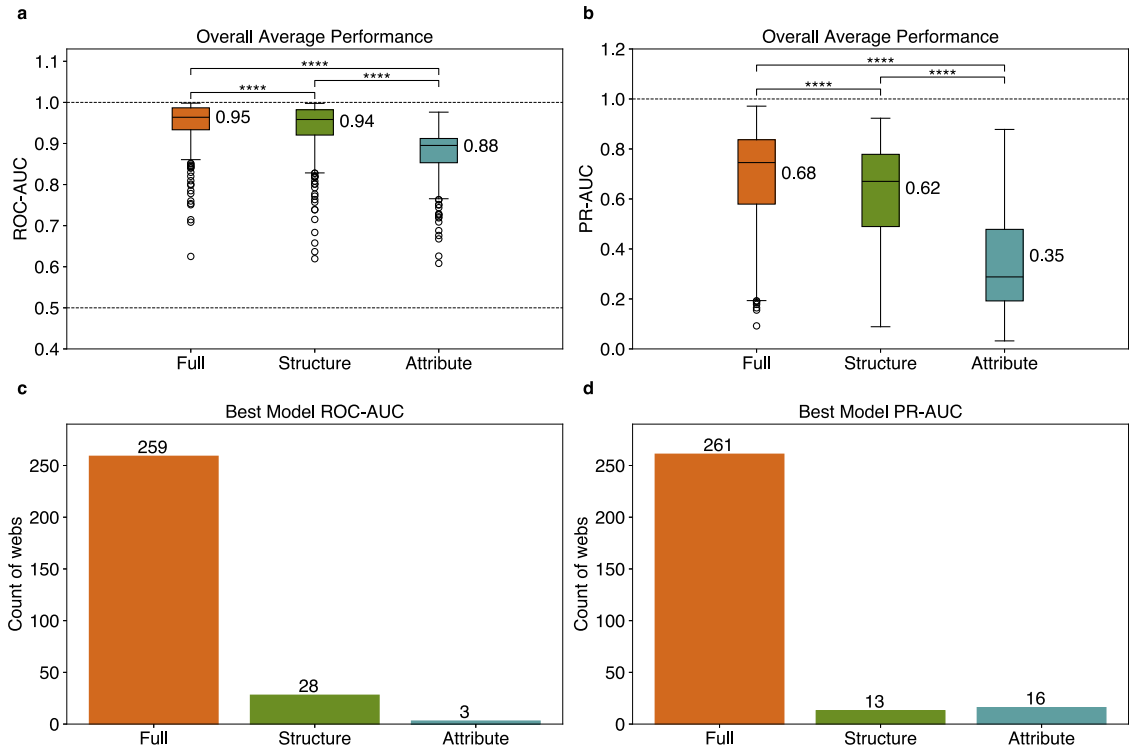

Fig. 4 | Link prediction performance on 290 food webs. For stacked models using structure-only predictors (Structure), attribute-only predictors (Attribute), and both (Full), models are evaluated via **a** ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and **b** PR-AUC (Area Under the Precision-Recall Curve) metrics. Performance for each of the three models is calculated for each food web by averaging across five independent iterations of evaluating across five unique folds (*N*=25 results per food web per model, N=21750 results total across iterations, folds, food webs, and models). Mean performance is displayed for each model across food webs. Significant differences in mean model performance based on false discovery rate (FDR) adjusted (Benjamini-Hochberg method) within-subjects pair-wise two-sided t-tests are shown, where **** indicates a p value < 0.0001, along with (**c**,**d**) the respective counts of the number of food webs for which a particular model produced the best average score for each metric. Box plots indicate median (middle line), 25, 75th percentile (box), the farthest data point lying within 1.5x the inter-quartile range (whiskers) as well as outliers (single points). Details are in Supplementary Table 8.

model, with a larger step down in mean accuracy for the attribute-only model in each case. Marine food webs yield the smallest gap in performance between the structure-only and attribute-only stacked models (ROC-AUC = 0.91 ± 0.07 vs. 0.88 ± 0.07, respectively; PR-AUC = 0.60 ± 0.20 vs. 0.48 ± 0.21, respectively), while terrestrial aboveground food webs yield the largest gap for ROC-AUC (= 0.92 ± 0.08 vs. 0.79 ± 0.05, respectively) and terrestrial below-ground yield the largest gap for PR-AUC (= 0.76 ± 0.11 vs. 0.21 ± 0.06, respectively).This modest variability in absolute prediction accuracy across the five ecosystem types suggests that ecosystem characteristics play an important but marginal role in determining the relative importance of structural and trait characteristics in whether a link exists or not.

An advantage of model stacking is that the supervised learning algorithm that sits atop the individual predictors can itself be inspected to learn which predictors the model identified as more or less useful in making accurate predictions. In our stacked models, Gini importance scores within the full models provide a quantitative measure of predictor utility (Fig. 6). In the full model, many of the structural predictors were among the most important, including two of the ecological predictors we adapted here (ecological PA, ecological Adamic/Adar index), with the EPA predictor achieving a higher relative importance. Predictors based on K-nearest neighbors (KNN), whether based entirely on structure or a combination of structure and node attributes, were also particularly helpful overall, as were predictors based on low-rank approximations of the network, and some predictors encoding the centrality of node *j* (the consumer species). The $\log_{10}$ mass ratio between nodes was the most important attribute-based predictor. The top predictors varied across ecosystem types (Supplementary Fig. 10), with a subset of predictors (ecological PA, KNN predictors, low-rank approximation predictors) appearing as important across multiple ecosystem types, including when calculated using an alternative importance metric (permutation importance, Supplementary Fig. 11). Given the importance of the KNN predictors, we also tested versions of the full and structure-only models restricted to only KNN predictors. We found that predictive performance slightly decreased compared to the models with all predictors but remained high, indicating that these predictors alone could encode much of the information necessary for predicting missing links (Supplementary Fig. 12).

Predictability depends on food web characteristics

The large size of the empirical corpus of food webs allows us to investigate the determinants of a model's predictive performance as a function of a food web's characteristics, e.g., the fraction of species with missing body mass measures, whether parasites were removed or not, the distribution of species body masses, the food web's size, and various summary statistics of the food web's network structure. To do this, we first perform univariable beta regressions that model missing link predictive performance (ROC-AUC or PR-AUC) as a function of food-web characteristics: (i) metadata and data processing (11 numeric features, Supplementary Table 10), (ii) global network topology (5 features, Supplementary Table 11), and (iii) network assortativity (7 features, Supplementary Table 12), adjusting for multiple comparisons. Empirical distributions of these features are given in the Supplementary Information (Supplementary Fig. 5–8, Supplementary Data 1). We then inspect the beta regression mean coefficient estimates

https://doi.org/10.1038/s41467-026-68769-7

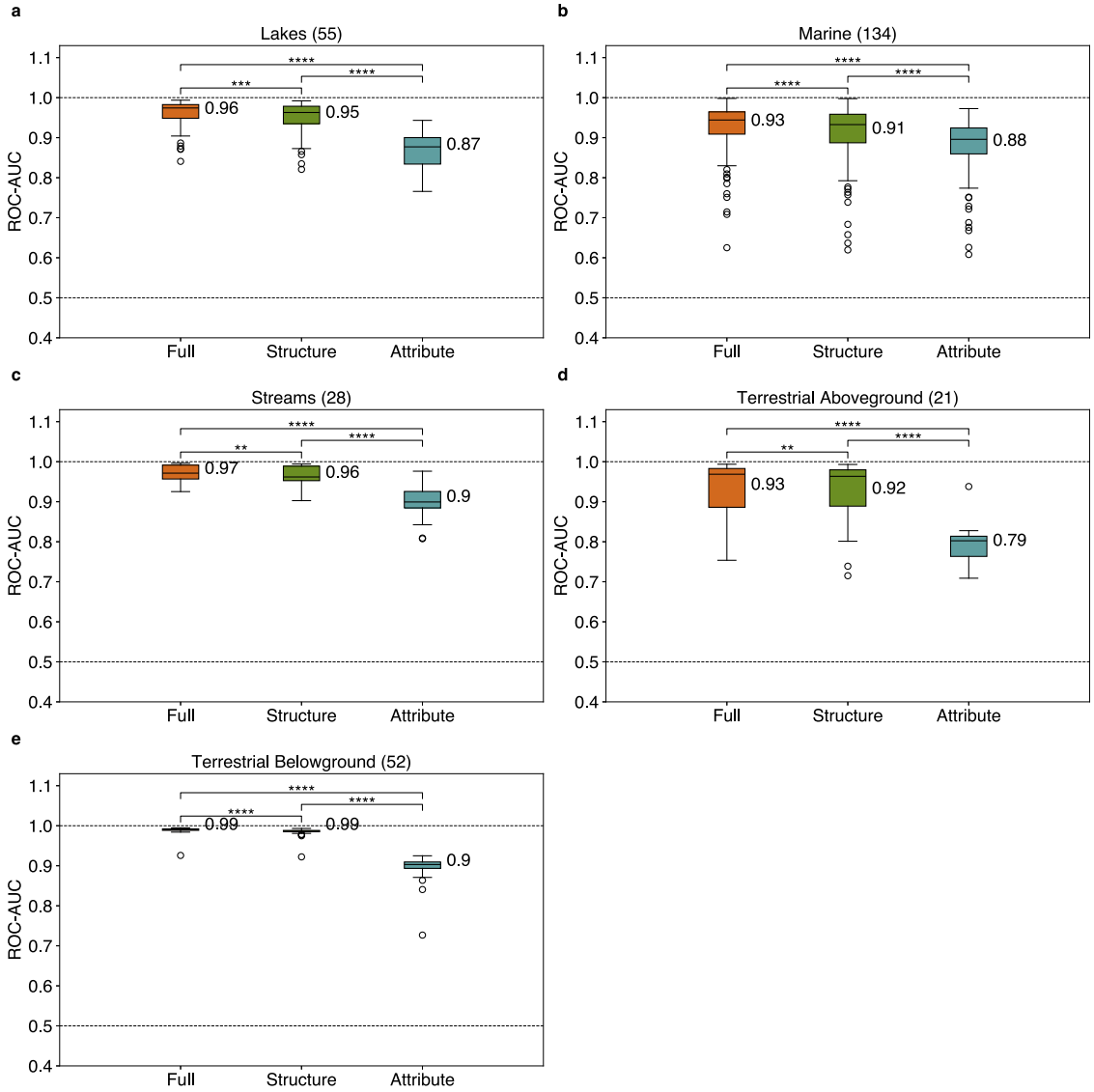

Fig. 5 | Link prediction performance by food web ecosystem type. a Lakes, **b** Marine, **c** Streams, **d** Terrestrial Aboveground, and **e** Terrestrial Belowground, for stacked models using structure-only predictors (Structure), attribute-only predictors (Attribute), and both (Full). Performance for each of the three models is calculated for each food web by averaging across five independent iterations of evaluating across five unique folds (N=25 results per food web per model, N=21750 results total across iterations, folds, food webs, and models). The number of food webs in each ecosystem type is indicated in parentheses, mean ROC-AUC (Area Under the Receiver Operating Characteristic Curve) is displayed for each model across food webs in that ecosystem type, and significant differences in mean model performance based on false discovery rate adjusted (Benjamini-Hochberg method) within-subjects pair-wise two-sided t-tests are shown, where **** indicates a p value < 0.0001, *** indicates a p value < 0.001, ** indicates a p value < 0.01. Box plots indicate median (middle line), 25th, 75th percentile (box), the farthest data point lying within 1.5x the inter-quartile range (whiskers) as well as outliers (single points). Details are in Supplementary Table 9.

of the features to assess which characteristics of a food web correlate with more or less accurate predictions of missing links.

We find significant correlations between many network features and predictive performance for both ROC-AUC and PR-AUC (Fig. 7, Supplementary Data 2, un-scaled coefficient estimates are shown in Supplementary Fig. 13). For both ROC-AUC and PR-AUC, we found better performance on food webs that had lower proportions of nodes with an unclassified taxonomic level (Supplementary Fig. 14a, Supplementary Fig. 15a). This relationship was significant for all three models for the ROC-AUC. For the PR-AUC, this relationship was significant for the structure-only and full models, and for the attribute-only model was significant when controlling for the number of nodes in the network (Supplementary Fig. 16b). We also found that the attribute-only model performed worse for networks with a higher proportion of nodes resolved at a taxonomic level higher than species for both metrics (Supplementary Fig. 14b, Supplementary Fig. 15b). Together, these results show that taxonomic resolution of the food webs was one of the factors that correlated with missing link predictive performance.

However, these significant trends related to taxonomic level were ecosystem-type dependent (Supplementary Fig. 17–21). Notably, without controlling for the number of nodes in the network, a negative trend of performance with the proportion of nodes with an unclassified taxonomic levels was observed for all three models for marine food webs and with higher taxonomic level proportion for stream and lake food webs, but results across other ecosystem types were positive or mixed.

Looking at global topological metrics per network, we found that larger food webs (log number of nodes) generally had better link prediction performance, a trend that was significant for all three

https://doi.org/10.1038/s41467-026-68769-7

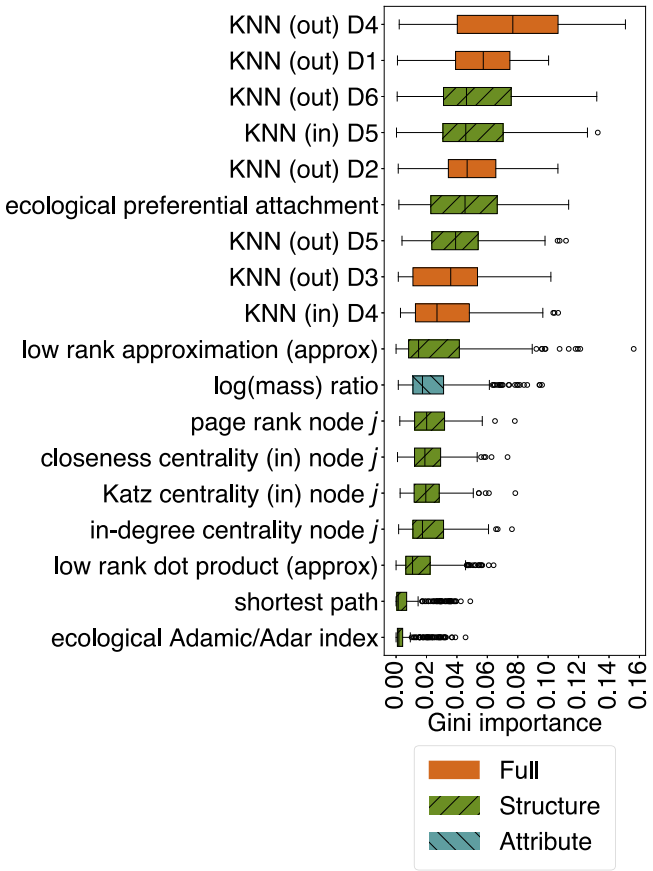

Fig. 6 | Top features by importance. To determine the top features, feature importance scores from the full model containing both structure and attribute predictors (106 predictors in total) were averaged across the 290 food webs split by ecosystem type, with five folds and five iterations per food web (N=7250 results per feature across folds, iterations, and food webs). Structure predictors (out of 51) are based only on network structure, attribute predictors are based on node attributes (out of 47), and full predictors (out of 8) combine information about structure and node attributes in a single predictor and are only included in the full model. KNN predictors are based on K-Nearest Neighbors. Node j refers to the consumer species. D1 is the Euclidean distance between the full normalized attribute vectors, D2 is the Manhattan distance between the full normalized attribute vectors, D3 is the Manhattan distance between the binary part of the attribute vectors, D4 is the Jaccard distance between the binary part of the attribute vectors, D5 is the Jaccard distance between in-neighbor sets, and D6 is the Jaccard distance between out-neighbor sets. The set of top predictors shown here (18 total) was selected by taking the union of the top 10 predictors from each ecosystem type. Box plots indicate median (middle line), 25th, 75th percentile (box), the farthest data point lying within 1.5x the inter-quartile range (whiskers) as well as outliers (single points).

models for the ROC-AUC metric overall and for the structure-only and full models for the PR-AUC metric (Supplementary Fig. 14c, Supplementary Fig. 15c). However, there was a significant negative trend for the attribute-only model performance and log number of nodes for the PR-AUC metric, and the directions of the overall trends for the PR-AUC metric partially differed when looking at only stream, only lake, or only terrestrial belowground food webs (Supplementary Fig. 17–21). We found the same overall trends for average degree (Supplementary Fig. 14d, Supplementary Fig. 15d). However, these results differed when controlling for network size (Supplementary Fig. 16), and in this case we observed that average degree significantly correlated with better performance for both metrics and all three models. After controlling for network size, we found that connectance had a significant positive correlation with better performance for both metrics (Supplementary Fig. 16). Together, these results indicate that link prediction was generally easier in cases where there was more link information available to the model both globally and locally.

We also found that link prediction performance was better for all models for food webs with lower modularity. This result was significant for all three models for PR-AUC, and for the attribute-only model for ROC-AUC without controlling for network size, and for all three models and both metrics when controlling for network size (Supplementary Fig. 14e, Supplementary Fig. 15e, Supplementary Fig. 16). Though, this correlation did not hold looking only at lake or terrestrial aboveground food webs (Supplementary Fig. 17–21). For the seven network assortativity features, we observed trends that were mixed across the features, ecosystem types, and models, particularly after controlling for network size (Supplementary Fig. 16), indicating that assortativity did not display a universal trend with predictive performance.

Discussion

Food webs provide a broadly useful representation of the ecological complexity of species interactions. However, food webs are nearly always incompletely sampled because of the large number of potential interactions and the labor required to observe them, particularly rare interactions. Hence, more accurate methods for estimating missing links in a partially observed food web with commonly available species traits would improve the accuracy of data on species interactions, the efficiency of collecting it, and the utility of food-web analyses and modeling. Here we evaluated the utility of model stacking—a state-of-the-art meta-learning technique for link prediction[2] that learns to combine multiple predictors into a single algorithm—for improving the accuracy of link prediction in food webs. Using this approach, we investigated the relative utility of species traits vs. species interactions for predicting missing links in food webs, how prediction accuracy varies with ecosystem type and network characteristics, and the relative utility of various individual link predictors.

Our broad analyses of synthetic food webs with known structure and of 290 real-world food webs indicates that species traits and observed network structure are both useful for predicting missing links—often because they are correlated—but, on average, structural predictors tend to produce more accurate predictions of missing links than do species traits. This result indicates that even when no trait data is available for nodes, structure-based methods can be used effectively to predict missing links in food webs. Of course, individual networks may be better explained by available traits alone or by structure alone, as has been found in previous work[20], and we find some evidence of this in our own results (Figs. 4, 5).

However, the most accurate predictions are generally obtained by combining traits and structural predictors within a single algorithm that can learn their relative importance in a particular food web, a result that is in alignment with prior work[15,26,28]. Moreover, across the empirical food webs studied here, we find that missing links tend to become more predictable when food webs are larger, have better taxonomic resolution, have higher connectance, and are less modular (Fig. 7, Supplementary Fig. 16). We observed better average performance on the food-web database than on the synthetic networks, which is interesting. It is possible that the empirical networks have more correlation between structure and traits than do the synthetic networks in the most difficult mid-ρ range of our mixed network tests. Understanding the origin of the higher accuracy on empirical data would be a useful direction for future work.

We also find that prediction accuracy and trait usefulness varies across ecosystem types: in the empirical food webs we studied, links in terrestrial belowground ecosystems are easiest to predict, while links in marine and terrestrial aboveground ecosystems are hardest to predict and traits are least useful for prediction in terrestrial aboveground ecosystems (Fig. 5). However, we had the fewest (21) food webs from terrestrial aboveground ecosystems (Supplementary Fig. 2a),

https://doi.org/10.1038/s41467-026-68769-7

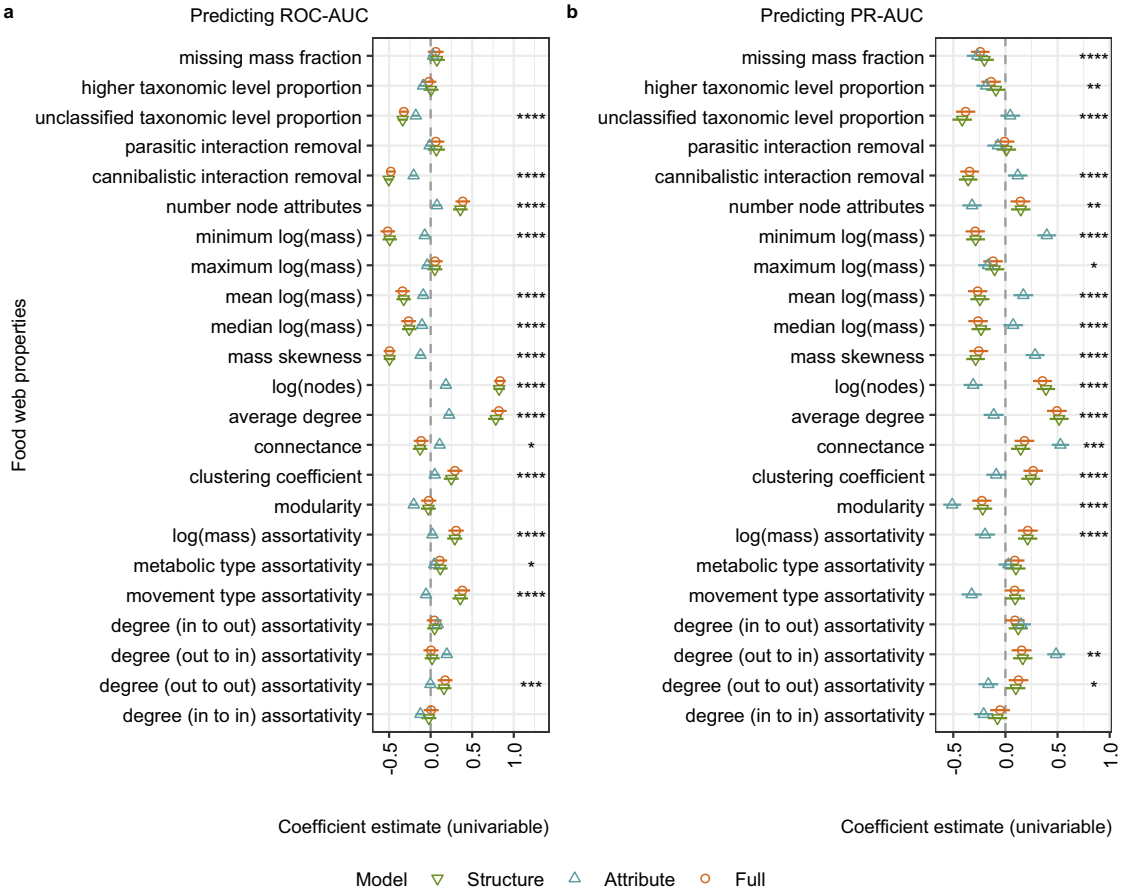

Fig. 7 | Correlates of missing link predictability. Mean coefficient estimates in univariable beta regressions between 23 food-web properties and the mean **a** ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and **b** PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction across food webs (N=290 samples for each regression, N=138 regressions across 23 properties, two performance metrics, and three models). Estimates are shown for the performance of the structure-only model (green inverted trianges), the attribute-only model (blue triangles), and the full model (orange circles). All food-web properties were first z score normalized so that the coefficients would be on comparable scales (un-scaled coefficient estimates are shown in Supplementary Fig. 13). Whiskers show 95% confidence intervals and a vertical line at 0 represents neither a positive nor negative correlation. The link function was set as logit and coefficients are shown on a log odds scale. Significance is based on the full model performance, where **** indicates a p value ≤0.0001, *** indicates a p value ≤0.001, ** indicates a p value ≤0.01, and * indicates a p value ≤0.05 (FDR adjusted, Benjamini-Hochberg method). Details are in Supplementary Data 2.

impeding generalizations. Although we do not test this possibility, the differences in methods used by different research teams to construct a food web may drive structural differences influencing predictor performance. For example, previous work has shown that there are distinct structural signatures in ecological networks based on construction methodology[38], which may influence which structural predictors are upweighted in the ensemble learned by model stacking.

Predicting missing links using species traits depends on what trait information is available for a given food web, and most empirical networks today include relatively few traits. For example, our analyses use just three traits (Table 1). Substantial prior work has indicated that body mass plays a fundamental role in determining which edges exist in food webs as predators tend to be larger than prey and body size correlates with many species properties including locomotion, mortality, and abundance[13,15,18,39], and hence is likely to be broadly important in link prediction, regardless of ecosystem type[16,40]. Our feature importance results also support this conclusion (Fig. 6, Supplementary Fig. 11). At the same time, body mass is not a universal determinant of species interactions, a fact exemplified by parasitic and terrestrial herbivorous interactions, which tend to invert the usual direction of body mass's influence, among other caveats[37,41,42]. Feeding interaction prediction is typically improved by including other trait information along with body size[9,13,26,27]. Hence, augmenting species interactions by collecting maximally detailed species trait information is likely to improve the prediction of missing links for both trait-based models[15,20,43] and the joint trait- and structure-based models we study here.

Previous work has also used information on species phylogenetic relationships as a proxy for missing traits[14–17,20,28,44], or inferred links based on species spatial co-occurrence[39,45] (but see Refs. 46,47 for discussions of methodological limitations). In particular, we do not consider any traits related to phenology or habitat overlap between species, which we would expect to constrain interactions. In adapting the model stacking approach to consider node attributes, we build upon particularly successful previous approaches that have used flexible machine learning methods to infer trait-matching rules[15,16,20,43]. Machine learning methods will likely benefit from such additional trait information to learn useful trait-matching rules directly from data, shedding new light on the underlying ecology that structures species interactions.

Our results also reinforce and refine the relationship between species traits and network structure, showing variation across ecosystem types in the relative performance of the attribute-based model, as well as which attribute-based predictors are the most important. For example, our results show a smaller gap between the average performance of the attribute-based model and the structure-based model in

marine food webs vs. terrestrial food webs. Collecting additional species traits beyond the three included in our modeling thus would be most important for missing link prediction in terrestrial food webs. We also see high feature importance for $\log_{10}$ mass ratio in marine food webs. These results appear to follow similar patterns to those reported in prior work showing marine food webs are more highly structured by body size than terrestrial food webs, a result with implications for theoretical models of food webs and stability analyses[48], and they highlight how link prediction with model stacking can be used to understand mechanistic differences in food-web link formation between ecosystem types.

Stacking models like those considered here provide a further advantage by learning for a specific food web how to combine trait information with structural patterns to predict missing links with no a priori knowledge. We find that this combined learning approach produces superior performance compared to using structure alone or traits alone. In this way, we synthesize prior work on trait-matching in ecology with prior work on network structure[2,11]. If network structural predictors improve missing link prediction beyond that possible using node attributes, this means that there are other latent rules driving link formation that are not captured by available node attribute data. Our results on both synthetically generated networks and empirical food webs show that a model stacking approach can be used to learn how to combine these predictors for a given network without knowing whether its link formation is driven by node attributes, by other latent rules, or by some combination of the two. Beyond improvements in prediction performance, model stacking can also facilitate analyzing the relative importance of different individual trait- or structure-based predictors[2,32], in a way that is analogous to examining coefficients in a regression model. Insights about the features that drive better prediction performance can then be used to test specific ecological hypotheses or develop new ecological theories. For instance, a strong effect of structural predictors or phylogeny-based predictors, if included, relative to trait-based predictors, might suggest the existence of predator and prey adaptation not fully captured by the other species traits, such as exoskeleton hardness in coleopterans[9] or web-building vs. non web-building spider species[39].

Our feature importance results recapitulate ecological theory that the relative masses of two species and interactions between generalist consumers and generalist resources are important predictors of feeding links across all ecosystem types (Fig. 6). For instance, these patterns are key parts of the niche model for food-web structure, in which species are ranked by a niche parameter, which is often associated with body size, and feed on species with lower values in this niche hierarchy within a range determined by a fundamental generality parameter[18,21,23]. Our results also indicate the broad utility of KNN or nearest-neighbor predictors in food webs[15,20], which assume that closely related species will share traits and interaction partners, which is another phenomenon incorporated into models of food-web structure to represent phylogenetic constraints on interactions[22,49,50]. A model restricted to only these KNN predictors still performed quite well for missing link prediction. The existence of compartments of highly interacting nodes, for example due to habitat or seasonality constraints[22], could also improve the performance of KNN predictors. Our results also show variation in the most important predictors in each ecosystem type, which supports investigating different generative models to best approximate food webs from different ecosystem types[48].

Our approach also has limitations. While we included many topological predictors in our stacked models, some of which have been used in prior work in ecological networks[15,24,25], there are other structure-based link prediction methods that we did not include as predictors within the ensemble. These include the probabilistic niche model[27], the allometric diet breadth model[51], linear filtering[29], large scale models of grouping or hierarchical structure[1–3,22], and matching-centrality latent trait models[25]. Future work could evaluate whether adding these as predictors within a stacking framework further improves predictions, or not, and whether using a different meta-learning model, such as a neural network approach[45] might improve performance.

While predictors based on low-rank approximations were among the most important across food webs, we did not consider other link predictors based on learning a node embedding in a latent space—a technique that underlies many deep learning techniques[52,53]—although recent work suggests that model stacking is often superior to these techniques for missing link prediction[2]. Instead, we focused on an ensemble of easy-to-compute topological and trait-based predictors that encode empirically-grounded aspects of food-web structure, including hierarchical, clustering, and latent trait structure. Predictors like these have the advantage of not requiring complex or computationally expensive model fitting or node embedding procedures, and can be more easily interpreted in light of ecological mechanisms.

In the 290 food webs we analyzed here, we removed species interactions that were parasitic, cannibalistic, and repeated, following standard practice in the food webs literature. However, these interactions are increasingly believed to represent important information about ecosystems[54], and a useful direction of future work may be to incorporate, and predict, the type of species interaction[55], perhaps using predictors for multi-level food webs (see ref. 56). For instance, some past work suggests that incorporating parasitic links can improve the prediction of missing non-parasitic interactions[57].

We also considered all interactions to be binary, while ideally food webs would be studied as weighted networks. Lack of publicly available food-web datasets with detailed interaction strength information limits current studies[58]. However, future work could build on our approach by developing a model stacking approach that predicts edge weights between species and by taking edge weights into account in the calculation of topological predictors.

In addition, future work could explore more ecologically realistic, non-uniform patterns of link missingness, perhaps by modeling the characteristics that lead some links to be easier or harder to observe in the field, e.g., due to taxonomic or geographic biases of sampling[30], rather than simulating missing links by removing them uniformly at random[59]. Recent work suggests that non-uniform missingness functions can lead to substantially different results compared to the uniform assumption[59], even as model stacking tends to perform well even when links are missing in non-uniform ways[60].

In our predictive setting, we assumed that all nodes have the same set of attributes or species traits, which is not always the case, even for standard traits like body size. For example, in pre-processing the food-web database, we had to make an assumption about how to assign placeholder body mass values for nodes like basal plants that lacked an observed value[20]. If more detailed trait data is collected, mismatches between sets of traits on some species vs. others will become more important to resolve. Finally, when evaluating missing link predictive performance we standardized on removing 20% of links. However, real-world ecological networks could be undersampled to far greater degrees, with the most extreme case being to construct a network from list of species and their traits alone.

Two other use cases for link prediction algorithms represent promising future directions: (i) to guide the collection of data in the field, e.g., under an active learning framework[61], in which the model iteratively selects informative pairs of species to be observed by researchers, or (ii) to leverage more common data from some ecosystems to make predictions about interactions in another ecosystem, as under a transfer learning framework[44,53,62] for which additional network-level features could be included[63].

Additionally, while we did not evaluate performance of our approach outside of the food-web context, our methodological adaptations to model stacking for the food-web context could have

broader applicability for predicting missing links in networks from other domains with node attributes, directed links, or global hierarchical structure, and these further investigations are an interesting direction for future work.

Our results demonstrate the scientific utility of model stacking for predicting missing links in ecological networks, and their ability to learn how to combine both structural and trait-based information to improve predictions, regardless of whether traits exhibit an assortative or disassortative pattern with links. For future work in this area, an advantage of model stacking will be its easy extensibility: as new, theoretically grounded predictors are developed or as new data on traits or interactions becomes available, a stacked model can easily incorporate these new predictors in combination with existing techniques to produce more accurate predictions. Similarly, these techniques can adapt to more realistic models of missingness[60], can be used to predict future interactions in dynamic network settings[64], and may potentially be useful in guiding the collection of food-web data in the field. We look forward to these many potential applications, and the benefits to ecological science they will bring.

Methods

Training and testing in model stacking for missing link prediction

In the missing link prediction problem, we assume that there is a true network $G = (V, E)$ with a set of nodes V and a set of edges E; however, we only have access to an incomplete or observed network $G' = (V, E')$, in which $E' \subset E$ is the incompletely observed set of edges. Let $X = V \times V - E'$, the set of unconnected pairs of nodes in G', denote the set of possible missing links, and define $Y = E - E'$ as the set of missing links (the positive class). We note that $Y \subset X$, and $X - Y$ is the set of pairs of nodes that are unconnected in G, i.e., the true non-links (the negative class).

At a high level, the stacking model from Ref. 2 uses information from G' to learn how to identify the pairs $i, j \in Y$, i.e., unconnected pairs in G', that are in fact missing links $i, j \in X$ relative to G. In order to train the stacking model, example missing links are produced from G' via uniform removal to produce a training network G''. Following Ref. 2, $1 - \alpha$ is the probability that a link is removed to create this training dataset; thus, we remove $(1 - \alpha)E'$ links. In our experiments, we set $1 - \alpha = 0.2$. For model training, the training dataset is composed of all unconnected pairs in G''; removed links are the positive class and all non-links in G' are taken as the negative class. Note that this negative class for training is noisy because it contains the true missing links from G (i.e., the set Y), but this effect on the model training is expected to be negligible for sparse networks[2]. In sparse networks like food webs, these classes are imbalanced with the negative class being much larger than the positive class. To better balance the classes before model training, we randomly up-sample with replacement the positive class, and for large networks we reduced the number of negative examples to 10,000 random examples of non-links to speed up model training[2].

A stacking model learns how to best combine the outputs of individual missing link prediction methods into an aggregate prediction for a given node pair. Features for the training data are generated for each candidate missing link by applying missing link prediction methods (detailed below). Each such method takes G'' as input and produces a score or probability for each unconnected node pair i, j. In our stacking models, a supervised machine learning algorithm is used as the meta-learner and is trained on this dataset to differentiate between non-links (the negative class) and missing links (the positive class). This trained model is then used to make predictions for the test dataset (all pairs in X), with corresponding features generated from G'. The performance of the stacking model is then evaluated by comparing the ranking of pairs $i, j \in X$ and whether they are missing links $i, j \in Y$. Here, we use a random forest[65] as the meta-learning model, with hyper-parameters chosen via 5-fold cross validation and optimized by selecting the best PR-AUC performance on average on the held-out fold, following advice in Ref. 31. In our initial experiments, this choice slightly improved the downstream performance on food webs compared to using the F1 statistic, as in Ref. 2. We note that there is some flexibility in the missing link prediction task regarding how one designs the validation, training, and test link sets, as well as in the choice of meta-level model and metric used for hyper-parameter optimization on the training set.

Importantly, food webs are typically represented as directed networks, with links pointing from resources (prey or primary producers) to consumers (predators, herbivores)[23], and sometimes have additional complexities, such as multiple interactions between the same two nodes (e.g., across multiple seasons or tidal zones), self-loops, and edge weights. We consider food webs as simple directed networks, and remove all multiple interactions between species and all self-loops, which represent cannibalistic interactions. We leave a consideration of these complexities for future work. We note that for simple directed networks, the size of the set of potential missing links for a set of nodes V doubles from $|V|(|V| - 1)/2$ as in Ref. 2 for undirected networks to $|V|(|V| - 1)$ in our adaption, because i, j and j, i are independently potentially missing links. Hence, both directions are independently scored by the link prediction algorithm in our setting. This increases the size of the training set in G'' from $|V|(|V|-1)/2-|E''|$ to $|V|(|V|-1)-|E''|$. Similarly, the size of the testing set in G' increases from $|V|(|V| - 1)/2 - |E'|$ to $|V|(|V| - 1) - |E'|$.

Synthetic network model parameters

We chose interaction probabilities for the SBM anchor networks and threshold values for the RGG anchor networks such that the median directed connectance value over 1000 generated networks matched that of the large database of food webs we explored in our empirical results (0.125). The resulting median directed connectance values were 0.126 for the SBM networks, 0.127 for the assortative RGG networks, and 0.127 for the disassortative RGG networks.

We generated SBM networks with 45 nodes and 3 equally-sized groups of 15 nodes with a high probability ($p = 0.544$) of interaction between nodes in groups 1 and 2, and between nodes in groups 2 and 3, and no probability of interaction ($p = 0$) among nodes between or within groups otherwise. This represents an extreme case of a food web with no omnivory, and is similar to model rectangular food webs with no omnivory used for modeling in prior work, e.g., as in ref. 66. The direction of the generated edges were chosen to replicate expected hierarchical structure in food webs by always pointing links from a lower group number to a higher group number; however, in a uniformly random 2% of cases (chosen to be similar to a large food-web database, with a median value of 1.3%), edges were reciprocated (pointed in both directions).

To parameterize the RGG model, a random attribute vector was generated for each node consisting of four attributes: two numeric traits in the range [0,1] and two binary traits on {0, 1}. These traits were chosen to align with typical node attribute data for empirical food webs[33,37] while also testing the ability to simultaneously consider both scalar and categorical traits in missing link prediction. The probability that two nodes are connected was then given by a simple step function, in which a pair of nodes is connected if the Euclidean distance between their attribute vectors was under a threshold for assortative networks [$d(i, j) < 1.00375$] or over a threshold for disassortative networks [$d(i, j) > 1.425$], and otherwise were not connected. Undirected edges were then converted to directed edges by randomly choosing an edge direction for each edge with equal probability, and again reciprocating edges in a uniformly random 2% of cases.

Structural predictors

The structure-based models include 47 topological predictors for missing link prediction in food webs (see additional details in

Supplementary Table 1). 34 node and node-pair level topological predictors for undirected networks originally proposed in Ref. 2 were adapted for use in food webs. Of these, 18 were updated for directed networks by computing the same predictor but with a directed network rather than an undirected network as input and 10 were duplicated in the in- and out- directions:

- Six predictors based on singular value decomposition, a strategy which has been shown to be good at predicting links in ecological networks[24,29,53], using a directed version of the adjacency matrix: low-rank approximation, low rank dot product, low rank mean, low-rank approximation (approx), low rank dot product (approx), low rank mean (approx).
- Five predictors based on shortest directed paths: shortest path, load centrality node *i*, load centrality node *j*, betweenness centrality node *i*, betweenness centrality node *j*.
- Four predictors based on the number of local directed triangles: local clustering coefficient node *i*, local clustering coefficient node *j*, local triangles node *i*, local triangles node *j*.
- Three page rank predictors, with a directed network as input: personalized page rank, page rank node *i*, page rank node *j*.
- Ten predictors were duplicated to calculate separate scores for both in- and out- directions:

– average neighbor in degree node *i*, average neighbor in degree node *j*, average neighbor out degree node *i*, average neighbor out degree node *j*;
– closeness centrality (in) node *i*, closeness centrality (in) node *j*, closeness centrality (out) node *i*, closeness centrality (out) node *j*;
– in-degree centrality node *i*, in-degree centrality node *j*, out-degree centrality node *i*, out-degree centrality node *j*;
– eigenvector centrality (in) node *i*, eigenvector centrality (in) node *j*, eigenvector centrality (out) node *i*, eigenvector centrality (out) node *j* and
– Katz centrality (in) node *i*, Katz centrality (in) node *j*, Katz centrality (out) node *i*, Katz centrality (out) node *j*.

Additionally, six of these topological predictors were adapted to our setting based on known topological properties of food webs. Food webs have directed links pointing from resource to consumer species and globally are hierarchically structured with links generally pointing in the direction of the flow of energy from lower to higher trophic levels. Food webs typically display many links between trophic levels and relatively fewer links within and across trophic levels, although links across a single trophic level can happen with omnivory. To account for this pattern of ecological organization, we adapted the preferential attachment (PA) link prediction method, which represents the intuition that two nodes with many links are more likely to share missing links, to an ecological preferential attachment score (EPA) between nodes *i* and *j* by considering the product between the out-degree of species *i* and the in-degree of species *j*, thus predicting a higher likelihood of missing links between generalist resources and generalist consumers (see Eq. (1) and (2), where deg(*i*) represents the degree of *i*, $\deg^-(i)$ represents the in-degree of *i* and $\deg^+(i)$ represents the out-degree of *i*).

$$\mathrm{PA(i,j)} = \deg(\mathrm{i}) \times \deg(\mathrm{j}) \tag{1}$$

$$\mathrm{EPA}(i,j) = \deg^+(i) \times \deg^-(j) \tag{2}$$

We also define the concept of a set of ecological common neighbors (ECNS, Eq. (4)) between a pair of species *i* and *j*. In undirected networks, the set of common neighbors (CNS, Eq. (3)) denotes the nodes that are connected to both *i* and *j*, and $\Gamma(i)$ gives the neighbor set of node *i*. Common neighbor count (CN, Eq. (5)) predictors encode that we predict missing links by closing triangles of interactions[67]; however, this assumption is not appropriate for food webs because of their directed and hierarchical nature, and the rarity of loops[40]. Instead, the ecological common neighbor count (ECN, Eq. (6)) predictor encodes that we expect to close omnivory motifs[68] (Fig. 1), where $\Gamma^+(i)$ represents the out-neighbor set of species *i* and $\Gamma^-(i)$ represents the in-neighbor set of species *i*.

We replace CNS with ECNS in 5 topological predictor calculations. For example, the Jaccard coefficient predictor (JC, Eq. (9)) which represents the number of common neighbors between two nodes (CN) divided by their number of total neighbors (TN, Eq. (7)) becomes the ecological Jaccard coefficient predictor (EJC, Eq. (10)), in which we consider ecologically relevant common (ECN) and total (ETN) neighbors. Three other predictors are similarly updated and we add an additional predictor (ecological common neighbor scores) based on this concept (see Supplementary Note 3). Finally, given the importance of trophic level for determining species interaction patterns and thus contextualizing other predictors, we added two indicators of approximate trophic level of a species calculated from the network structure (trophic level node *i* and trophic level node *j*).

$$\mathrm{CNS}(i,j) = \Gamma(i) \cap \Gamma(j) \tag{3}$$

$$\mathrm{ECNS}(i,j) = \Gamma^+(i) \cap \Gamma^-(j) \tag{4}$$

$$\mathrm{CN}(i,j) = |\mathrm{CNS}(i,j)| \tag{5}$$

$$\mathrm{ECN}(i,j) = |\mathrm{ECNS}(i,j)| \tag{6}$$

$$\mathrm{TN}(i,j) = |\Gamma(i) \cup \Gamma(j)| \tag{7}$$

$$\mathrm{ETN}(i,j) = |\Gamma^+(i) \cup \Gamma^-(j)| \tag{8}$$

$$\mathrm{JC}(i,j) = \mathrm{CN}(i,j) \,/\, \mathrm{TN}(i,j) \tag{9}$$

$$\mathrm{EJC}(i,j) = \mathrm{ECN}(i,j) \,/\, \mathrm{ETN}(i,j) \tag{10}$$

Some of these predictors, such as the Jaccard index, common neighbors, and degree product, have been previously explored for missing link prediction for food webs[25] and some have been noted as ecologically interpretable, for example the degree product can be interpreted as relating to the generality of the two species[28].

Attribute-based predictors

We additionally move beyond the set of predictors proposed in Ref. 2 by including predictors based on node attributes (species traits). We assume that all nodes have the same set of attributes *A*, which includes a subset of numeric attributes *N* and binary attributes *B*, $N \cap B = \varnothing$, $A = N \cup B$. The synthetic and empirical food webs we consider vary in their attribute sets per node, and the number of predictors added for a given food web is a function of the size of these sets.

We add $2|A|$ attribute features to each potential link simply by including attribute values for each of the nodes *i*, *j* in a pair, ordered based on the direction of the link between the two nodes (e.g., log(-mass) of node *i*, log(mass) of node *j*), with categorical features first transformed into binary features via one-hot encoding.

Additionally, we add derived features based on the insight that many networks have assortative or disassortative structure based on node attributes (Fig. 2). In networks with assortative structure, interactions occur between nodes with similar attributes and in

disassortative networks interactions occur between nodes with dissimilar attributes[69]. These relationships can be incorporated into model stacking by adding distance metrics between the vectors of node attribute values. If $|N| > 0$, we include the Euclidean distance, Manhattan distance, cosine distance, and dot product between the numeric parts of the attribute vectors of node pairs (numeric attributes are first min-max normalized to the range [0, 1]). If $|B| > 0$, we also include the Hamming distance and Jaccard distance between the binary parts of the attribute vectors of node pairs. If $|A| \neq |N|$ and $|A| \neq |B|$ (i.e., the nodes have a mix of numeric and binary attributes), then we also include the Euclidean distance, Manhattan distance, cosine distance, and dot product between the full attribute vectors.

Additionally, there may be some ratio between two numeric attributes that relates to the probability of a link. For example, trophic interactions have been shown to have typical log body mass ratios[37]. We thus add the ratio between each of the numeric attributes of the two nodes, adding $|N|$ additional predictors, taking into account the direction of a link (e.g., log(mass) ratio = $\log(\text{mass})_i / \log(\text{mass})_j$).

In total, we add $2|A|$ attributes, up to 10 predictors measuring assortative or disassortative structure, and $|N|$ attribute ratio predictors (Supplementary Table 2). For example, with $|A| = 9$, $|N| = 1$ and $|B| = 8$, representing a typical case for the food webs we considered, we added 29 attribute-based predictors. As explained in Supplementary Note 2, we also added an additional binary attribute predictor for each node indicating whether the mass value was originally missing.

Nearest-neighbor predictors

Finally, we adapt the stacking model to include 12 K-Nearest Neighbor (KNN) predictors (Supplementary Table 3). KNN predictors are based on the assumption that nodes that are similar have similar sets of interaction partners. KNN has performed well for link prediction in food webs in previous work[15,20]. KNN predictors are used to recommend new interaction partners to a node. For example, in a food web with $K = 2$, a KNN predictor for species i would find the two most similar species to species i in the food web and then recommend resource species for species i from the resource sets of these two similar species with those resources found in both resource sets recommended first. The same approach can also be taken to recommend consumers to resources.

We adapt KNN predictors for the stacking context by calculating for a node pair i, j the fraction of node i's K-nearest neighbors in the food web that are in-neighbors (resources) of node j, as well as the fraction of node j's K-nearest neighbors in the food web that are out-neighbors (consumers) of node i. These predictors assume that if none of the nodes most similar to node i interact with node j and none of the nodes most similar to node j interact with node i, it is unlikely node i and node j interact whereas if all or most of the nodes similar to node i interact with node j, and the inverse, it is more likely they interact.

The nearness of neighbors can be determined by applying distance metrics to the attribute vectors or neighbor sets of two nodes. We used six distance metrics to produce 12 KNN predictors total based on both in- and out- directions (Supplementary Fig. 22):

- D1: the Euclidean distance between the full normalized attribute vectors,
- D2: the Manhattan distance between the full normalized attribute vectors,
- D3: the Manhattan distance between the binary part of the attribute vectors,
- D4: the Jaccard distance between the binary part of the attribute vectors, following ref. 15
- D5: the Jaccard distance between in-neighbor sets (i.e., resource sets, also following ref. 15)
- D6: the Jaccard distance between out-neighbor sets (i.e., consumer sets).

D5 and D6 are calculated by subtracting the Jaccard similarity from 1 (Eqs. (11) and (12)). In cases of comparing two empty sets, for which this value is undefined, we set the Jaccard distance to 0 as we would expect two nodes that both don't have resources (e.g., basal species) or two nodes that both don't have consumers (e.g., top predator species) to be similar. The predictors based on D5 and D6 only consider network structure, while the predictors based on D1, D2, D3, and D4 consider network structure and node attributes in combination. We implemented all KNN predictors with $K = 3$.

$$\mathrm{D5} = 1 - |\Gamma^-(i) \cap \Gamma^-(j)| \,/\, |\Gamma^-(i) \cup \Gamma^-(j)|| \tag{11}$$

$$\mathrm{D6} = 1 - |\Gamma^+(i) \cap \Gamma^+(j)| \,/\, |\Gamma^+(i) \cup \Gamma^+(j)| \tag{12}$$

Statistical analysis

Beta regressions were performed with R version 4.4.3 using the betareg package version 3.2-3 with a logit link function[70].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The empirical food-web datasets used in this study are available in the GATEWAy database (GlobAl daTabasE of traits and food Web Architecture[33,37], version 3, accessible at https://doi.org/10.25829/idiv.283-3-756). The data was pre-processed as described in Supplementary Note 2. The pre-processed data are available at https://doi.org/10.5281/zenodo.18026669. Supplementary Data 1 includes food web features across the datasets and summary statistics for food web features split by ecosystem type. Supplementary Data 2 includes details of Beta regression results.

Code availability

The code for this paper is available at https://doi.org/10.5281/zenodo.18026669 ref. 71. Data pre-processing and missing link prediction scripts were run with Python 3.6.3 on CentOS Linux 7 using High Performance Computing resources supported by BioFrontiers IT. Results visualization scripts were run on Windows 11 with Python 3.12.4. A configuration file is provided in the code repository for reproducing a Python environment with the necessary packages for running the scripts.

References

1. Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).
2. Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M. & Clauset, A. Stacking models for nearly optimal link prediction in complex networks. *Proc. Natl. Acad. Sci. USA* **117**, 23393–23400 (2020).
3. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* **106**, 22073–22078 (2009).
4. Zhou, T. Progresses and challenges in link prediction. *iScience* **24**, 103217 (2021).
5. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
6. Dunne, J. A., Brose, U., Williams, R. J. & Martinez, N. D. Modeling food-web dynamics: complexity-stability implications. In *Aquatic food webs: an ecosystem approach*, 117–129 (Oxford, 2005).
7. McDonald-Madden, E. et al. Using food-web theory to conserve ecosystems. *Nat. Commun.* **7**, 10245 (2016).

8. Kamenova, S. et al. Invasions Toolkit. In *Advances in Ecological Research*, vol. 56, 85–182 (Elsevier, 2017). https://linkinghub.elsevier.com/retrieve/pii/S0065250416300587.
9. Van De Walle, R. et al. Arthropod food webs predicted from body size ratios are improved by incorporating prey defensive properties. *J. Anim. Ecol.* **92**, 913–924 (2023).
10. Jordano, P. Sampling networks of ecological interactions. *Funct. Ecol.* **30**, 1883–1893 (2016).
11. Terry, J. C. D. & Lewis, O. T. Finding missing links in interaction networks. *Ecology* **101**, e03047 (2020).
12. Novak, M. et al. Predicting community responses to perturbations in the face of imperfect knowledge and network complexity. *Ecology* **92**, 836–846 (2011).
13. Eklöf, A. et al. The dimensionality of ecological networks. *Ecol. Lett.* **16**, 577–583 (2013).
14. Morales-Castilla, I., Matias, M. G., Gravel, D. & Araújo, M. B. Inferring biotic interactions from proxies. *Trends Ecol. Evol.* **30**, 347–356 (2015).
15. Desjardins-Proulx, P., Laigle, I., Poisot, T. & Gravel, D. Ecological interactions and the Netflix problem. *PeerJ* **5**, e3644 (2017).
16. Laigle, I. et al. Species traits as drivers of food web structure. *Oikos* **127**, 316–326 (2018).
17. Pomeranz, J. P. F., Thompson, R. M., Poisot, T. & Harding, J. S. Inferring predator-prey interactions in food webs. *Methods Ecol. Evol.* **10**, 356–367 (2019).
18. Gravel, D., Poisot, T., Albouy, C., Velez, L. & Mouillot, D. Inferring food web structure from predator-prey body size relationships. *Methods Ecol. Evol.* **4**, 1083–1090 (2013).
19. Nagelkerke, L. A. J. & Rossberg, A. G. Trophic niche-space imaging, using resource and consumer traits. *Theor. Ecol.* **7**, 423–434 (2014).
20. Wootton, K. L. et al. Layer-specific imprints of traits within a plant-herbivore-predator network - complementary insights from complementary methods. *Ecography* **2024**, e07028 (2024).
21. Digel, C., Riede, J. O. & Brose, U. Body sizes, cumulative and allometric degree distributions across natural food webs. *Oikos* **120**, 503–509 (2011).
22. Allesina, S. & Pascual, M. Food web models: a plea for groups. *Ecol. Lett.* **12**, 652–662 (2009).
23. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).
24. Dalla Riva, G. V. & Stouffer, D. B. Exploring the evolutionary signature of food webs' backbones using functional traits. *Oikos* **125**, 446–456 (2016).
25. Rohr, R. P., Naisbit, R. E., Mazza, C. & Bersier, L.-F. *Matching-centrality* decomposition and the forecasting of new links in networks. *Proc. R. Soc. B Biol. Sci.* **283**, 20152702 (2016).
26. Rohr, R., Scherer, H., Kehrli, P., Mazza, C. & Bersier, L. Modeling food webs: exploring unexplained structure using latent traits. *Am. Naturalist* **176**, 170–177 (2010).
27. Williams, R. J., Anandanadesan, A. & Purves, D. The probabilistic niche model reveals the niche structure and role of body size in a complex food web. *PLoS ONE* **5**, e12092 (2010).
28. Pearse, I. S. & Altermatt, F. Predicting novel trophic interactions in a non-native world. *Ecol. Lett.* **16**, 1088–1094 (2013).
29. Stock, M., Poisot, T., Waegeman, W. & De Baets, B. Linear filtering reveals false negatives in species interaction data. *Sci. Rep.* **7**, 45908 (2017).
30. Papadogeorgou, G., Bello, C., Ovaskainen, O. & Dunson, D. B. Covariate-informed latent interaction models: addressing geographic & taxonomic bias in predicting bird-plant interactions. *J. Am. Stat. Assoc.* **118**, 2250–2261 (2023).
31. Poisot, T. Guidelines for the prediction of species interactions through binary classification. *Methods Ecol. Evol.* **14**, 1333–1345 (2023).
32. Guimerà, R. One model to rule them all in network science? *Proc. Natl. Acad. Sci. USA* **117**, 25195–25197 (2020).
33. Brose, U. GlobAL daTabasE of traits and food Web Architecture (GATEWAy) version 1.0 (2018). https://idata.idiv.de/ddm/Data/ShowData/283?version=3.
34. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: first steps. *Soc. Netw.* **5**, 109–137 (1983).
35. Penrose, M. *Random geometric graphs* (Oxford University Press, 2003). OCLC: 271204794.
36. Ling, C. X., Huang, J. & Zhang, H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In Xiang, Y. & Chaib-draa, B. (eds.) *Advances in Artificial Intelligence*, vol. 2671, 329–341 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003). http://link.springer.com/10.1007/3-540-44886-1_25. Series Title: Lecture Notes in Computer Science.
37. Brose, U. et al. Predator traits determine food-web architecture across ecosystems. *Nat. Ecol. Evol.* **3**, 919–927 (2019).
38. Brimacombe, C., Bodner, K., Michalska-Smith, M., Poisot, T. & Fortin, M.-J. Shortcomings of reusing species interaction networks created by different sets of researchers. *PLoS Biol.* **21**, e3002068 (2023).
39. Bohan, D. A., Caron-Lormier, G., Muggleton, S., Raybould, A. & Tamaddoni-Nezhad, A. Automated discovery of food webs from ecological data using logic-based machine learning. *PLoS ONE* **6**, e29028 (2011).
40. Stouffer, D. B., Rezende, E. L. & Amaral, L. A. N. The role of body mass in diet contiguity and food-web structure: body mass and food-web structure. *J. Anim. Ecol.* **80**, 632–639 (2011).
41. Huxham, M., Beaney, S. & Raffaelli, D. Do parasites reduce the chances of triangulation in a real food web? *Oikos* **76**, 284 (1996).
42. Brose, U. et al. Consumer-resource body size relationships in natural food webs. *Ecology* **87**, 2411–2417 (2006).
43. Pichler, M., Boreux, V., Klein, A., Schleuning, M. & Hartig, F. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods Ecol. Evol.* **11**, 281–293 (2020).
44. Caron, D. et al. Trait-matching models predict pairwise interactions across regions, not food web properties. *Glob. Ecol. Biogeogr.* **33**, e13807 (2024).
45. Strydom, T. et al. A roadmap towards predicting species interaction networks (across space and time). *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20210063 (2021).
46. Barner, A. K., Coblentz, K. E., Hacker, S. D. & Menge, B. A. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology* **99**, 557–566 (2018).
47. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* **23**, 1050–1063 (2020).
48. Li, J. et al. A size-constrained feeding-niche model distinguishes predation patterns between aquatic and terrestrial food webs. *Ecol. Lett.* **26**, 76–86 (2023).
49. Cattin, M.-F., Bersier, L.-F., Banašek-Richter, C., Baltensperger, R. & Gabriel, J.-P. Phylogenetic constraints and adaptation explain food-web structure. *Nature* **427**, 835–839 (2004).
50. Ives, A. & Godfray, H. Phylogenetic analysis of trophic associations. *Am. Naturalist* **168**, E1–E14 (2006).
51. Petchey, O. L., Beckerman, A. P., Riede, J. O. & Warren, P. H. Size, foraging, and food web structure. *Proc. Natl. Acad. Sci. USA* **105**, 4191–4196 (2008).
52. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864 (ACM, 2016). https://doi.org/10.1145/2939672.2939754.
53. Strydom, T. et al. Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations. *Methods Ecol. Evol.* **14**, 2917–2930 (2023).

54. Lafferty, K. D. et al. Parasites in food webs: the ultimate missing links. *Ecol. Lett.* **11**, 533–546 (2008).
55. Michalska-Smith, M. J. & Allesina, S. Telling ecological networks apart by their structure: a computational challenge. *PLOS Comput. Biol.* **15**, e1007076 (2019).
56. Bar-Hen, A., Barbillon, P. & Donnet, S. Block models for generalized multipartite networks: applications in ecology and ethnobiology. *Stat. Model.* **22**, 273–296 (2022).
57. Jacobs, A. Z., Dunne, J. A., Moore, C. & Clauset, A. Untangling the roles of parasites in food webs with generative network models (2015). Preprint, arxiv:1505.04741.
58. Pringle, R. M. & Hutchinson, M. C. Resolving food-web structure. *Annu. Rev. Ecol. Evol. Syst.* **51**, 55–80 (2020).
59. Tabouy, T., Barbillon, P. & Chiquet, J. Variational inference for stochastic block models from sampled data. *J. Am. Stat. Assoc.* **115**, 455–466 (2020).
60. He, X. et al. Link prediction accuracy on real-world networks under non-uniform missing-edge patterns. *PLoS ONE* **19**, e0306883 (2024).
61. Settles, B. From theories to queries: Active learning in practice. In Guyon, I., Cawley, G., Dror, G., Lemaire, V. & Statnikov, A. (eds.) *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, vol. 16 of *Proceedings of Machine Learning Research*, 1–18 (PMLR, 2011). https://proceedings.mlr.press/v16/settles11a.html.
62. Biton, B., Puzis, R. & Pilosof, S. Inductive link prediction facilitates the discovery of missing links and enables cross-community inference in ecological networks. *Nature Ecology & Evolution* 1–10 (Springer, 2025).
63. Botella, C., Dray, S., Matias, C., Miele, V. & Thuiller, W. An appraisal of graph embeddings for comparing trophic network architectures. *Methods Ecol. Evol.* **13**, 203–216 (2022).
64. He, X., Ghasemian, A., Lee, E., Clauset, A. & Mucha, P. J. Sequential stacking link prediction algorithms for temporal networks. *Nat. Commun.* **15**, 1364 (2024).
65. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
66. Eklöf, A. & Ebenman, B. Species loss and secondary extinctions in simple and complex model communities. *J. Anim. Ecol.* **75**, 239–246 (2006).
67. Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).
68. Cirtwill, A. R. & Wootton, K. L. Stable motifs delay species loss in simulated food webs. *Oikos***2022** (2022). https://onlinelibrary.wiley.com/doi/10.1111/oik.09436.
69. Newman, M. E. J. & Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **7**, 11863 (2016).
70. Cribari-Neto, F. & Zeileis, A. Beta regression in r. *J. Stat. Softw.* **34**, 1–24 (2010).
71. Van Kleunen, L., Dee, L. E., Wootton, K. L., Massol, F. & Clauset, A. Predicting missing links in food webs using stacked models and species traits (2025). https://doi.org/10.5281/zenodo.18026669.

Acknowledgements

The authors thank the Brose lab for help with the empirical data, A. Ghasemian for helpful conversations about stacking models for link prediction, and B. Singh, D.B. Larremore and E. Bradley for helpful discussions and feedback. This work was supported in part by the National Science Foundation Division of Ocean Sciences (NSF OCE 2049360), the Eppley Foundation for Research, and the Chateaubriand Fellowship of the Office for Science & Technology of the Embassy of France in the United States. The stacking model code used here is adapted for use in directed, attributed, hierarchical networks from Ghasemian et al. (2020). The authors acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources supported by BioFrontiers IT.

Author contributions

L.V., L.D., and A.C. conceptualized the study. L.D. and A.C. supervised the study. L.V. and A.C. designed the methodology and performed the mathematical analysis. L.V., L.D., and K.W. planned and validated the data pre-processing steps. L.V. wrote the code to pre-process the data, performed the computational analysis, and visualized results. L.D., K.W., F.M., and A.C. provided feedback to validate and improve the computational analysis and visualizations. L.V. and A.C. wrote an initial paper draft and all authors contributed to final paper writing, review, and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-68769-7.

Correspondence and requests for materials should be addressed to Lucy B. Van Kleunen or Aaron Clauset.

Peer review information *Communications Materials* thanks Chencheng Cai, Virginia Dominguez-Garcia, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.