

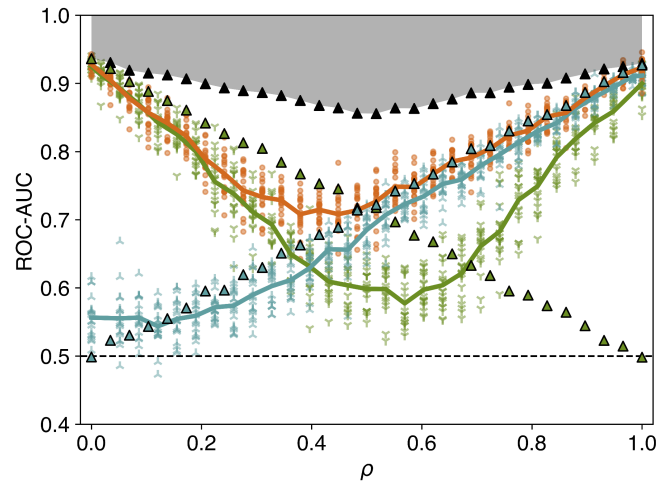
# Supplementary Information:

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

### SUPPLEMENTARY NOTE 1. CALCULATING THEORETICAL ROC-AUC MAXIMUM PERFORMANCE

We contextualized our results on synthetic networks by calculating the theoretical maximum prediction performance in terms of the commonly used ROC-AUC statistic (Area Under the Receiver Operating Characteristic Curve) [1] for each value of  $\rho$  using Monte Carlo sampling as in Ref. [2]. Mathematically, the ROC-AUC statistic is the probability that a higher score is assigned to a true positive (in this case, a missing link) than to a true negative (in this case, a non-link). At each value of  $\rho$ , we randomly sampled 10,000 pairs of true positives (edges) and true negatives (non-edges). For each pair, we then calculated the score as their probability of being connected from the known generative model (SBM, RGG, or a mixed SBM and RGG, all with an adjustment for the choice of edge direction). Our theoretical maximum ROC-AUC estimate was then the proportion of times out of these 10,000 samples that the true positive score was higher than the true negative score.

Further, we could examine partial theoretical maximum curves based on the SBM and RGG models individually by calculating the SBM and RGG probabilities of connecting edges separately. These partial curves are visualized for the case of an assortative RGG anchor network in Supplementary Fig. 1.



**Supplementary Fig. 1.** Performance of missing link prediction on synthetic networks mixing SBM (stochastic block model) and assortative RGG (random geometric graph model) networks with varying the mixing parameter  $\rho$  as shown in main text Fig. 3A, with partial theoretical ROC-AUC (Area Under the Receiver Operating Characteristic Curve) maximum curves shown for the RGG model (blue triangles) and the SBM model (green triangles). Performance is shown for the structure-only model (green Ys) the attribute-only model (blue inverted Ys), and the full model (orange circles) as a function of the mixing parameter  $\rho$  for 20 iid synthetic networks evaluated at each of 30  $\rho$  values, in each case averaged across 5-fold cross validation ( $N=3000$  results for each of the 3 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  $\rho$  value. The theoretical maximum ROC-AUC curve for the mixed networks is shown with black triangles, and the baseline ROC-AUC performance is shown at 0.5 with a dashed line.

### SUPPLEMENTARY NOTE 2. FOOD WEB DATA PRE-PROCESSING

The GlobAl daTabasE of traits and food Web Architecture (GATEWAY; [3, 4]; version 3, accessible at [doi.org/10.25829/idi.v.283-3-756](https://doi.org/10.25829/idi.v.283-3-756)) includes 222,151 consumer-resource interactions across 290 food web datasets spanning terrestrial above and below-ground, lake, marine, and stream ecosystem types (Supplementary Fig. 2A, Supplementary Table 4). The consumer and resource nodes involved in each interaction are identified by the `con.taxonomy` and `res.taxonomy` columns respectively, and each interaction is associated with information for both the consumer and

resource species on their respective lifestage, taxonomic aggregation level, movement type, metabolic type, and mass. Our data pre-processing code to format the GATEWAY data for missing link prediction are available in the public code repository associated with this manuscript.

In constructing food webs from this list of interactions, we chose to additionally disaggregate food web nodes by lifestage to ensure the accuracy of trait values assigned to each node, e.g., we treated the adult and larval stages of the same species as separate nodes. Thus, unique nodes were identified by a combination of the `con.taxonomy` or `res.taxonomy` columns and the respective lifestage column. Lifestage disaggregation changed the set of nodes for 25 out of the 290 food webs (491 nodes added in total). We found nearly identical missing link prediction performance with or without lifestage disaggregation for these food webs (Supplementary Fig. 3).

In food web construction, we included all interactions labeled as predacious, herbivorous, fungivorous, bacterivorous and detritivorous (Supplementary Fig. 2B). We excluded links labeled as parasitic or parasitoid, which applied to 2853 interactions in 20 food webs. In this database, the food webs containing parasitic interactions were mostly terrestrial aboveground food webs (Supplementary Fig. 2C). We made this choice because parasites have been documented to change food web structure [5–9], and our ecologically-oriented topological predictors were not designed with parasitic interaction patterns in mind. Parasitic interactions are also not often systematically included in food webs, and if they are, their interactions are often inferred rather than directly observed [10]. We additionally excluded 1920 interactions with NA interaction type in the Carpinteria food web [11] as, upon inspection, these interactions involved parasites, though we included the interactions labeled as NA interaction type in the food webs from Refs. [12] and [13], as these links appeared to be non-parasitic trophic interactions. We included all interaction classifications: individual-based intact and attack (`ibi`), interactions with groups such as filter feeding or grazing (`nibi`), and NA. We combined the possible values for node taxonomic level into 4 categories: species (species, subspecies, variety, form), genus (genus), family+ (family, order, class, phylum, kingdom), and unclassified (NA, unranked). There was only one case of a taxonomic level conflict between repeated records for the same unique node (node name *Gomphonema*, lifestage NA, taxonomic level records as species and genus), which we manually set to genus (Supplementary Fig. 2D).

We additionally excluded cannibalistic interactions (self-loops), which applied to 3275 interactions (1 of which was also parasitic) across 271 food webs, because we set up our missing link prediction approach for networks without self-loops, although we note this as an area for future work. We also trimmed any remaining multi-edges, or interactions recorded twice for the same two nodes in a food web, which applied to 307 interactions across 16 food webs. After trimming interactions, the dataset included 213,797 interactions among 19077 nodes (5562 unique node names).

For each node, we included three species traits as node attributes: metabolic type, movement type, and log mass. We filled in values for 12 nodes with metabolic type NA as listed in Supplementary Table 5, leaving 9 possible values for metabolic type across the database (Supplementary Fig. 4A). We manually replaced movement type values for non-living nodes (node names: `benthic detritus_NA`, `carrion-detritus_NA`, `CPOM_NA`, `Detritus_adults`, `Detritus_NA`, `FPOM_NA`, `Marine detritus_NA`, `POM_NA`, `POM (detritus)_NA`, `terrestrial detritus_NA`, `Sediment_NA`, `Unidentified detritus_NA`, `Phytodetritus_NA`, `dead organic matter_NA`, `detritus_NA`, `leaf litter_NA`) with a new `other_nonliving` category, and otherwise did not investigate between-web differences in movement type values as we were more concerned with in-web consistency due to our approach of predicting missing links in individual food webs. We filled in values for 43 living nodes with movement type NA as listed in Supplementary Table 6, leaving 7 possible values for movement type in the database (Supplementary Fig. 4B). Most food webs had 2-5 distinct metabolic types present (Supplementary Fig. 4C) and 3 or 4 distinct movement types present (Supplementary Fig. 4D).

We included all interaction records for a given food web, including those that were trimmed, when collecting trait values for a node in that food web. Metabolic type and movement type traits were categorical and thus turned into binary attributes for each category, with only those types present in a given food web included as potential values for a node. If movement type records differed for a species within a food web we listed a 1 in all relevant columns for that node after removing NA values because there might be biological reasons for which a species differed in movement type between interactions (56 instances across 28 food webs).

If mass records for a node differed within a food web (18 nodes across 14 food webs) we saved as the mass value the mean across all mass records for that node in the food web, thus representing an expected value for that node in that food web. Mass values were also completely missing within a food web for 1371 nodes (279 unique names) across 143 food webs. For 579 of these nodes with missing mass and an unclassified taxonomic level, we replaced their mass with the minimum mass observed in their food web. These nodes were generally exceptional in that they tended to represent groups of species where mass might more appropriately be treated as a NA value (for example: roots, leaf litter, or algae). However, using our link prediction approach, missing numeric attributes need to be imputed for all nodes before they are used as features in prediction and we did not want to lose topological information about feeding interactions in the food web by excluding these nodes. Thus, we chose to replace these with the minimum mass because we observed that these were generally primary producers or detritus nodes that we expected to be at the bottom of the food web, and thus we assigned them a low mass for predicting missing links following body mass ratio expectations for consumer-resource links [4]. We assumed trait-matching rules learned by the meta-level model

would likely learn to point links from nodes of lower body mass to nodes of higher body mass. We excluded nodes with metabolic type invertebrate and taxonomic level unclassified from this replacement, instead using the approach below.

For the remaining 792 nodes missing mass, we filled in their mass values using guesses informed by prior knowledge, by averaging the mass values of similar nodes found in the database. This strategy was employed to represent how missing trait values might be realistically assigned to species in a food web dataset (e.g., as in Ref. [14]). Node mass values were replaced (i) first with the mean mass value of nodes with the same name (typically, those of the same species) seen in other food webs (110 replacements), then (ii) with the mean mass value of species with the same first part of the node name (splitting node names by spaces, thus typically these were nodes of the same genus seen in other food webs, 317 replacements), then (iii) with the mean value of nodes in the same food web with the same movement and metabolic type (75 replacements), then (iv) with the mean value of nodes in the same food web with the same metabolic type (188 replacements), then (v) with the mean value of nodes in the same food web with the same movement type (10 replacements), then (vi) with the mean value of nodes from other webs with the same metabolic type (89 replacements), and then finally (vii) with the mean value of nodes from other webs with the same movement type (3 replacements).

Because we used assumptions to fill in missing mass values, we further provided information to the predictive model by providing a binary indicator for each node indicating whether its mass was originally missing.

The distribution of final mass values across all nodes in all the food webs was extremely skewed with many low mass nodes—the mean value in the dataset across nodes was 7954.16g but the median was 0.0005g. The majority of species mass values were under 1g (17,127 nodes out of 19,077), with the smallest mass value in the dataset for bacteria ( $6.8 \times 10^{-21}$ g) in the Lough Hyne food web and the largest value (52,395,350g) for the *Balaenoptera acutorostrata* (common minke whale) node in the Kongsfjorden food web. Because of this characteristic skewed nature of body mass distributions driven by evolutionary processes [15], we thus included log (base 10) mass as a node attribute rather than raw mass values (Supplementary Fig. 5) following prior work [4].

To contextualize predictive performance results across the database we saved features related to (i) metadata and data processing (12 features, Supplementary Table 10), (ii) global network topology (5 features, Supplementary Table 11), and (iii) network assortativity (7 features, Supplementary Table 12) for each food web. The values for these 24 features across the 290 food webs, along with the reciprocated edge fraction, is provided in Supplementary Data 1 and the distributions across the database are visualized in Supplementary Fig. 6, Supplementary Fig. 7, and Supplementary Fig. 8.

The median percent of nodes originally missing mass in a food web was 0%, but this value was up to 64.8% (in the Sutton Stream food web) (Supplementary Fig. 6A). Most food webs had many nodes that were resolved at higher than the species level — the median percent of nodes with a higher taxonomic level was 25%, though this value was down to 0% (Lake Malawi food web, Caribbean Reef food web) and up to 63.5% (Dutch Microfauna food web PlotB) (Supplementary Fig. 6B). It was also common for food webs to have some nodes with an unclassified taxonomic level, with the median percent of nodes with an unclassified taxonomic level at 12.2% (ranging from 0% in the Tuesday Lake 1984, Tuesday Lake 1986, and Scottish Lake food webs to an outlier of 100% in the Lake Malawi food web) (Supplementary Fig. 6C). Most food webs did not have any parasitic interactions removed (median 0%), with the outlier of the Carpinteria food web at 69.4% (Supplementary Fig. 6D). Most food webs had a handful of cannibalistic interactions removed (median 2.4%), with the largest value (26.9%) for the PC2P5 food web (Supplementary Fig. 6E). The final number of attributes for nodes in a food web varied from 5 to 15, with most food webs having around 8 attributes per node (Supplementary Fig. 6F) — this was equal to the number of unique movement types in the food web (binary indicators) plus the number of unique metabolic types in the food web (binary indicators), an additional binary indicator in the case of mass being originally missing for any nodes in the food web, and one numeric attribute for log mass. The distributions for minimum, maximum, mean, and median log mass values across the food webs reflected the observation that most node masses were very small (Supplementary Fig. 6G-J). Universally positive mass skewness values (median 0.6722) also indicated that all food webs had right skewed mass distributions, with the least skewed distribution (0.1884) for the Lough Hyne food web and the most skewed distribution (1.6973) for the Bylot food web (Supplementary Fig. 6K).

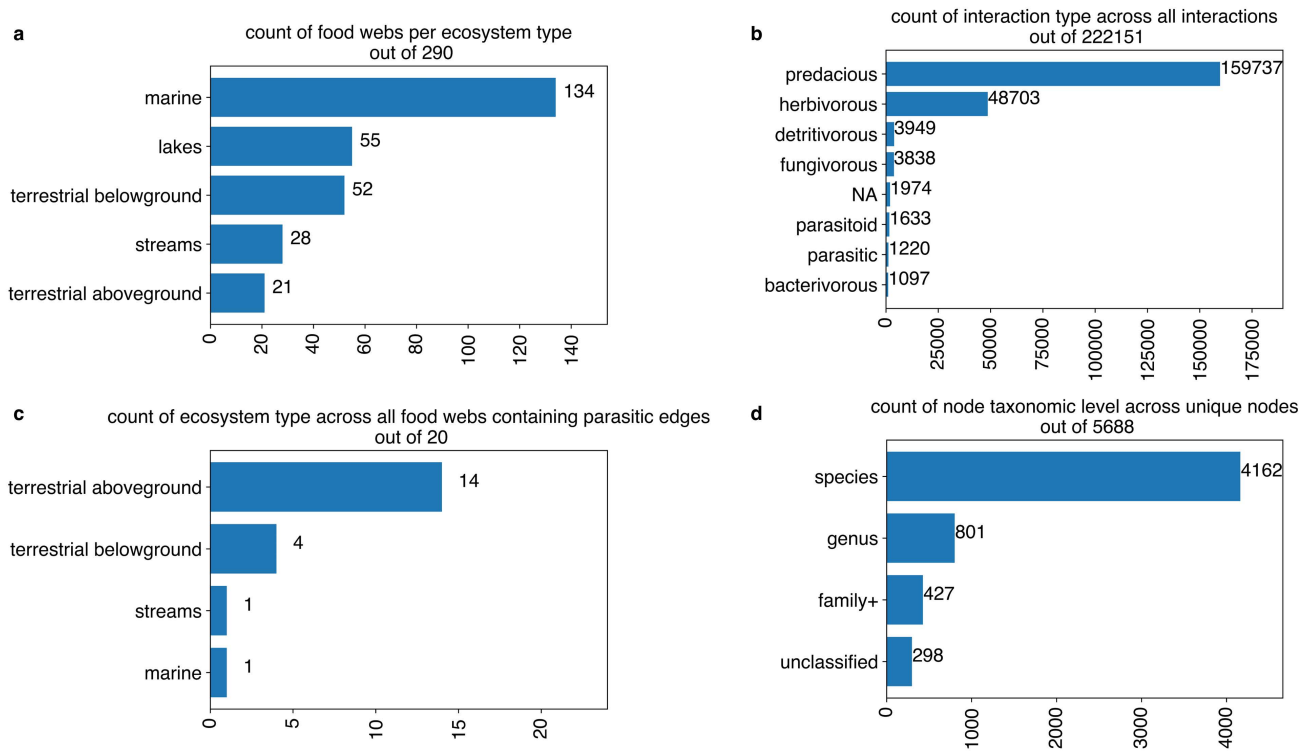
Global network topology features varied across the database. The median number of nodes was 43.5 (median log nodes 1.64), with the lowest number of nodes (10) in the SF1M2 food web and the highest number (606) in the Chesapeake Bay food web (Supplementary Fig. 7A). The median average degree was 5.28, with the lowest average degree (1.15) in the AP1 food web and the highest average degree (32.63) in the Weddell Sea food web (Supplementary Fig. 7B). The distribution of directed connectance values appeared approximately normal with a median connectance value of 0.1253, the minimum connectance (0.0231) in the Kongsfjorden food web and the maximum connectance (0.2975) in the Skipwith Pond food web (Supplementary Fig. 7C). The clustering coefficients of undirected versions of the food webs varied dramatically, with a median clustering coefficient value of 0.2245, a minimum (0), indicating no undirected triangles in 12 food webs (Brook trout lake, Deep lake, Gull lake north, Indian lake, Owl Lake, Rock

Lake, South Lake, Twin Lake East, Twin Lake West, Wolf Lake, Svalbard, and PP112), and a maximum (0.674) in the SEW18 food web (Supplementary Fig. 7D). The distribution of modularity values appeared approximately normal with a median value of 0.217, a minimum value (0.015) in the Iceland stream IS7 August 2008 food web, and a maximum value (0.457) in the Kongsfjorden food web (Supplementary Fig. 7E). As 0.3 is considered a modularity value above which there is significant community structure [16], this indicates that most food webs had a low to moderate level of community structure, with variation across the database.

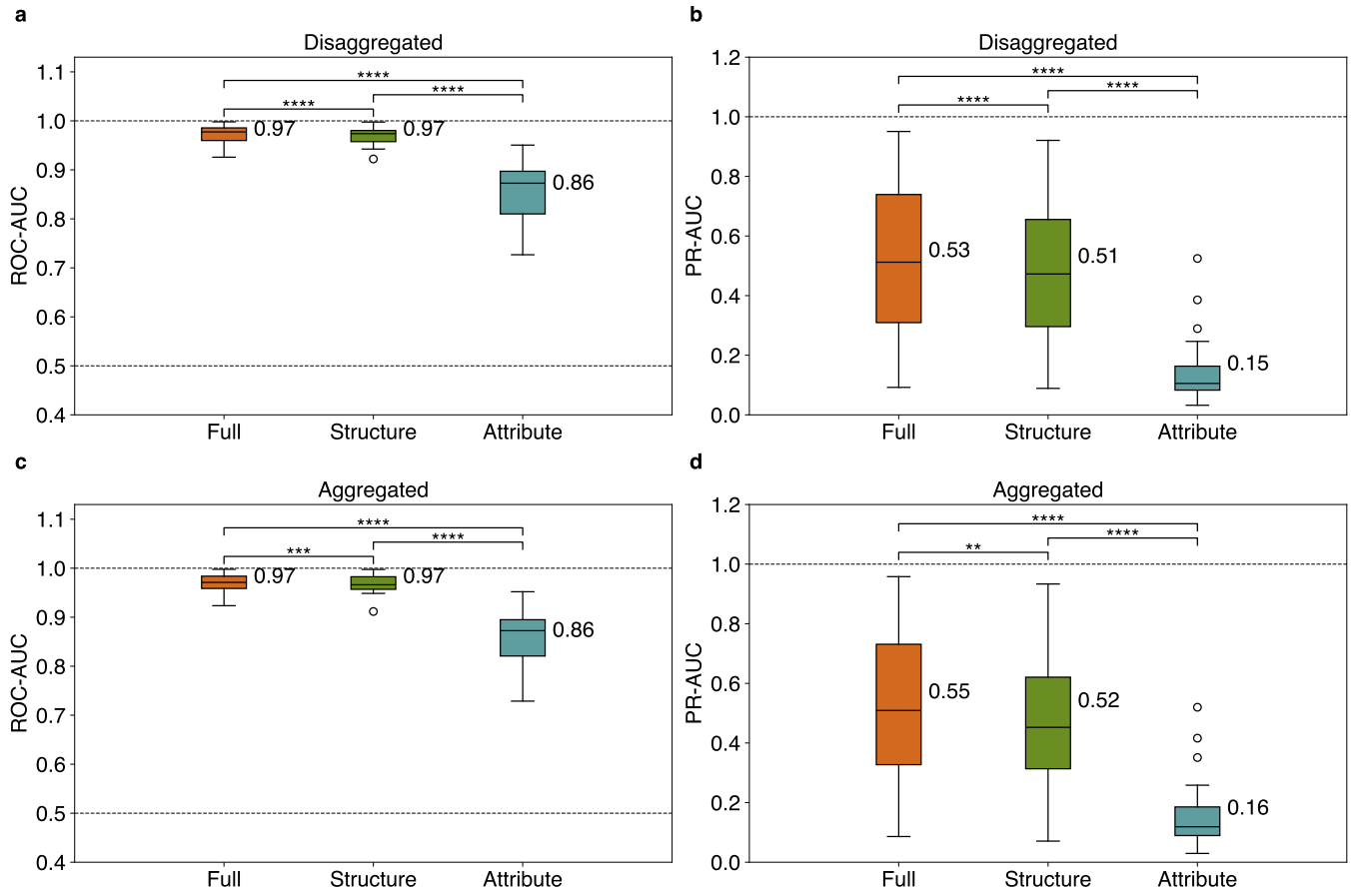
Assortativity coefficients measure to what extent nodes with similar attributes or degree tend to interact more than those with differing attributes and are 1 for perfect assortative mixing (nodes only connect to those with the same trait), 0 for no assortative mixing, and -1 for perfect disassortative mixing (nodes only connect to those with different traits) [17]. There was variety across the dataset in terms of whether food webs were assortative or disassortative by node attribute and degree.

Assortativity values based on bin logged mass showed a bimodal distribution, with most food webs showing no to slightly positive assortativity based on mass and another group of food webs showing somewhat larger assortativity values, with the maximum value (0.223) for the HEW22 food web. Some food webs were also disassortative by mass, with the lowest value (-0.172) for the CGP1 food web (Supplementary Fig. 8A). Food webs generally displayed no to weakly negative assortativity based on metabolic type (median -0.0001) and movement type (median -0.003) (Supplementary Fig. 8B,C). There was a tail of food webs for movement type assortativity showing higher assortativity values, with the largest value (0.404) for the Dutch Microfauna food web Plot B. There were wide ranges of assortativity values for assortativities based on node degree. The median degree assortativity in to out was negative (-0.1612) as was the median degree assortativity out to in (-0.2706) (Supplementary Fig. 8D,E). This indicates that generalist consumers tended to link to specialist resources and generalist resources tended to link to specialist consumers, however this differed across the dataset, for example the maximum assortativity value out to in was quite large (0.9289 in the Iceland stream IS8 August 2008 food web). Conversely, assortativity values were generally slightly positive for degree assortativity out to out (median 0.094) and in to in (median 0.1859) (Supplementary Fig. 8F,G), indicating that generalist resources slightly tended to link with generalist resources and generalist consumers slightly tended to link to generalist consumers, although there were many disassortative examples as well.

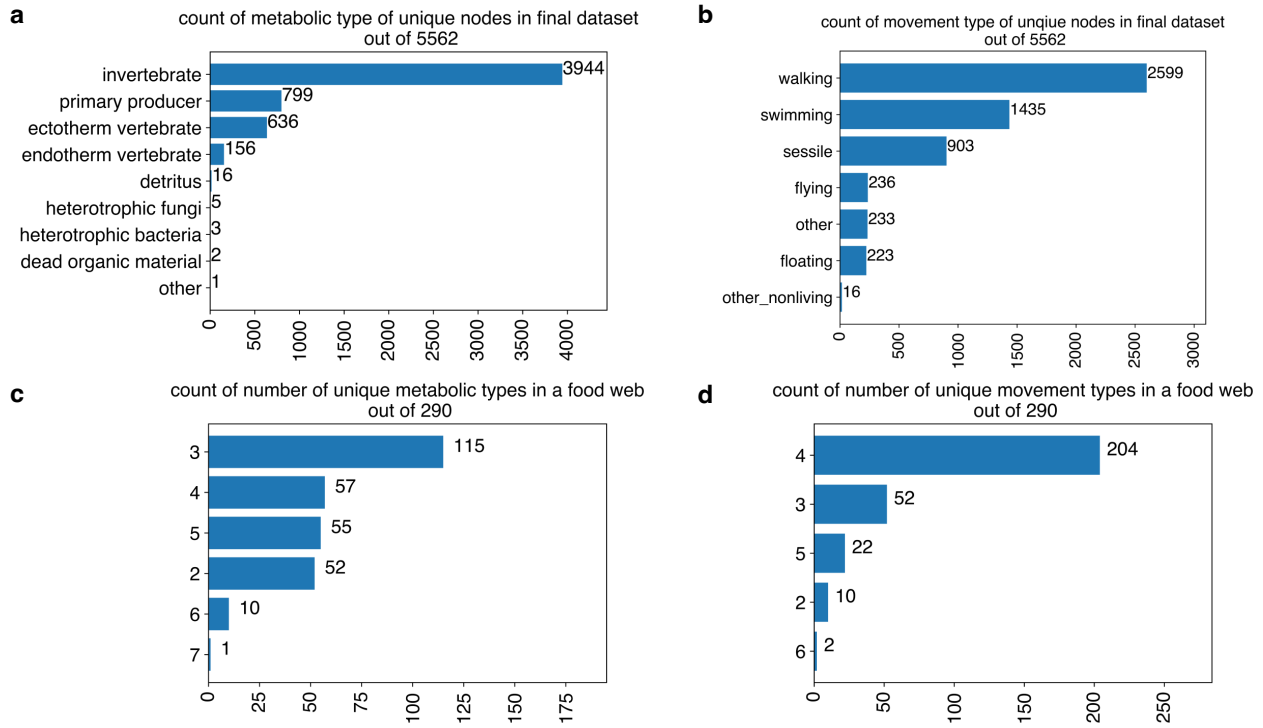
Finally, to contextualize our results, we further investigated the distributions of each of these 25 features across the database partitioned into 5 sets of food webs based on ecosystem type. Notable patterns we observed included that aquatic food webs in the dataset were generally smaller (streams: median 1.93 log nodes; lakes: median 1.57 log nodes; marine: median 1.45 log nodes) than the terrestrial food webs in the dataset (terrestrial aboveground: median 2.09 log nodes; terrestrial belowground: median 2.09 log nodes). Terrestrial belowground food webs were more clustered (median clustering coefficient 0.582) than food webs of the other types, and they were also more assortative by mass (median mass assortativity of 0.1735) than food webs of other types. Marine food webs had more dramatically right-skewed mass distributions (median skewness 0.8763) than food webs of other ecosystem types. Further details about how these distributions differed across ecosystem types are provided in Supplementary Data 1.



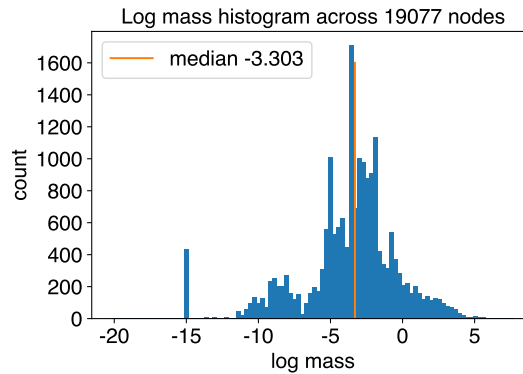
**Supplementary Fig. 2.** Food web database details. (a) Ecosystem type counts across the 290 food webs. (b) Interaction type across the original 222,151 interactions. (c) Ecosystem type counts across the 20 food webs containing parasitic interactions. (d) Node taxonomic level counts across the 5688 unique node names in the original database.



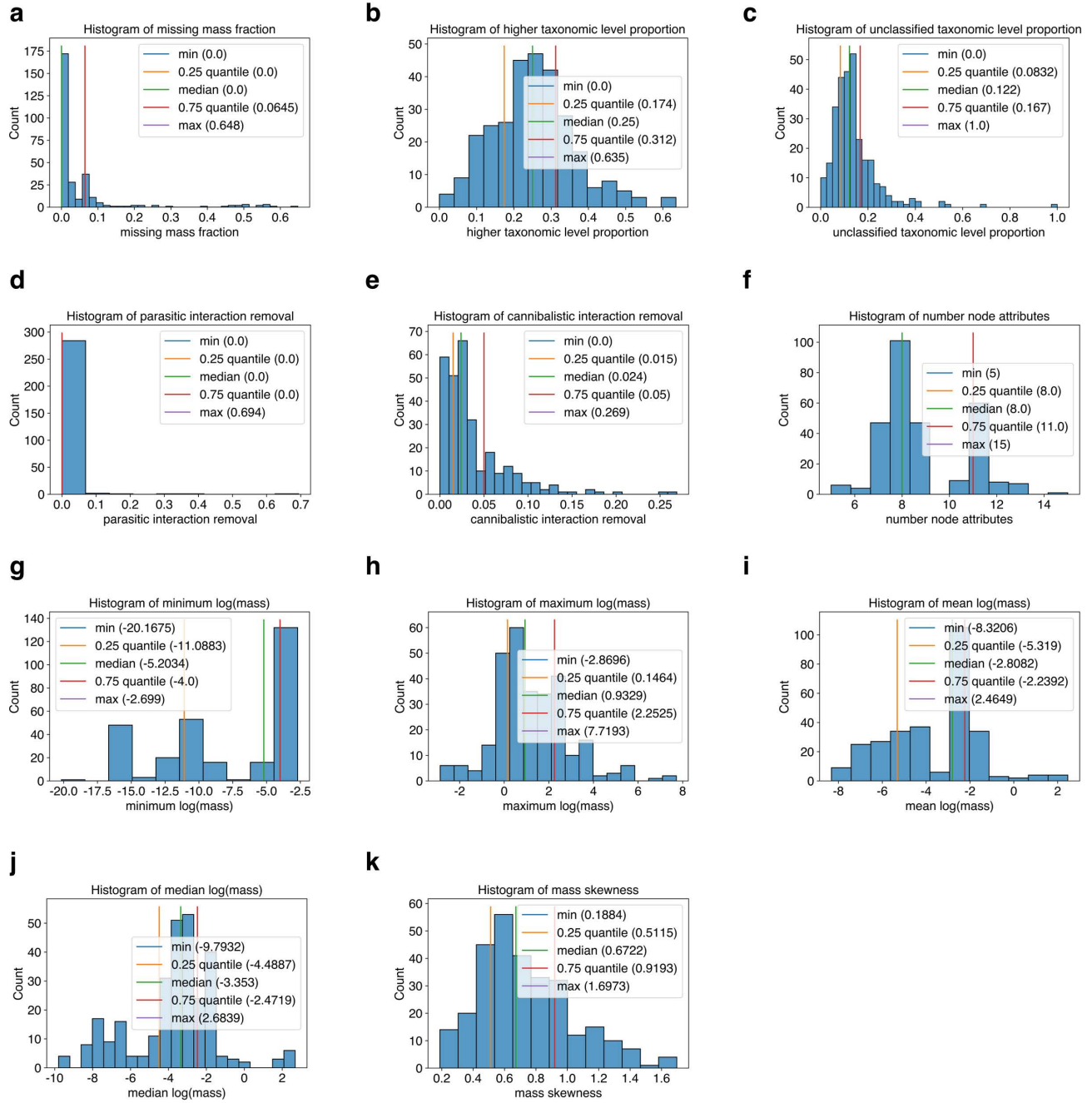
**Supplementary Fig. 3.** Aggregated results for the 25 food webs whose node sets change with lifestage aggregation for stacked models using structure-only predictors (Structure), attribute-only predictors (Attribute), and both (Full), (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) results with nodes disaggregated by lifestage as in the main text results, (c) ROC-AUC and (d) PR-AUC results if nodes for the same species are aggregated across lifestages. Disaggregated results are averaged, as in the main text, over 5 iterations of evaluation across 5 unique folds (25 results per food web per model, N=1875 results across 5 iterations, 5 folds, 25 food webs, and 3 models). Aggregated results are averaged across 5 unique folds (5 results per food web, N=375 results across 1 iteration, 5 folds, 25 food webs, and 3 models). Significant differences in mean model performance based on false discovery rate (FDR) adjusted (Benjamini-Hochberg method [18]) within-subjects pair-wise two-sided t-tests are shown where \*\*\*\* indicates a p value<0.0001, \*\*\* p value<0.001, and \*\* 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 7.



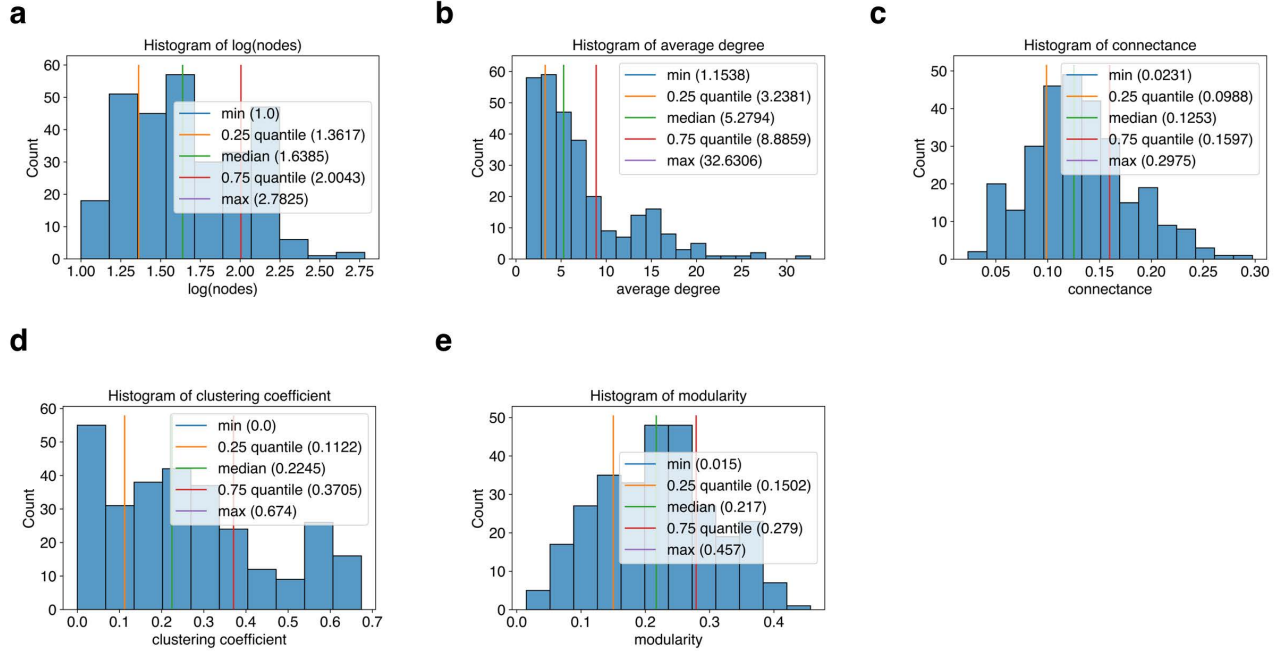
**Supplementary Fig. 4.** Metabolic type and movement type details across the food web database. (a) Counts of node metabolic type across the 5562 unique nodes. (b) Counts of node movement type across the 5562 unique nodes, taking into account that unique nodes can have multiple movement types. (c) Counts of the number of final unique metabolic types in a given food web, across the 290 food webs. (d) Counts of the number of final unique movement types in a given food web, across the 290 food webs.



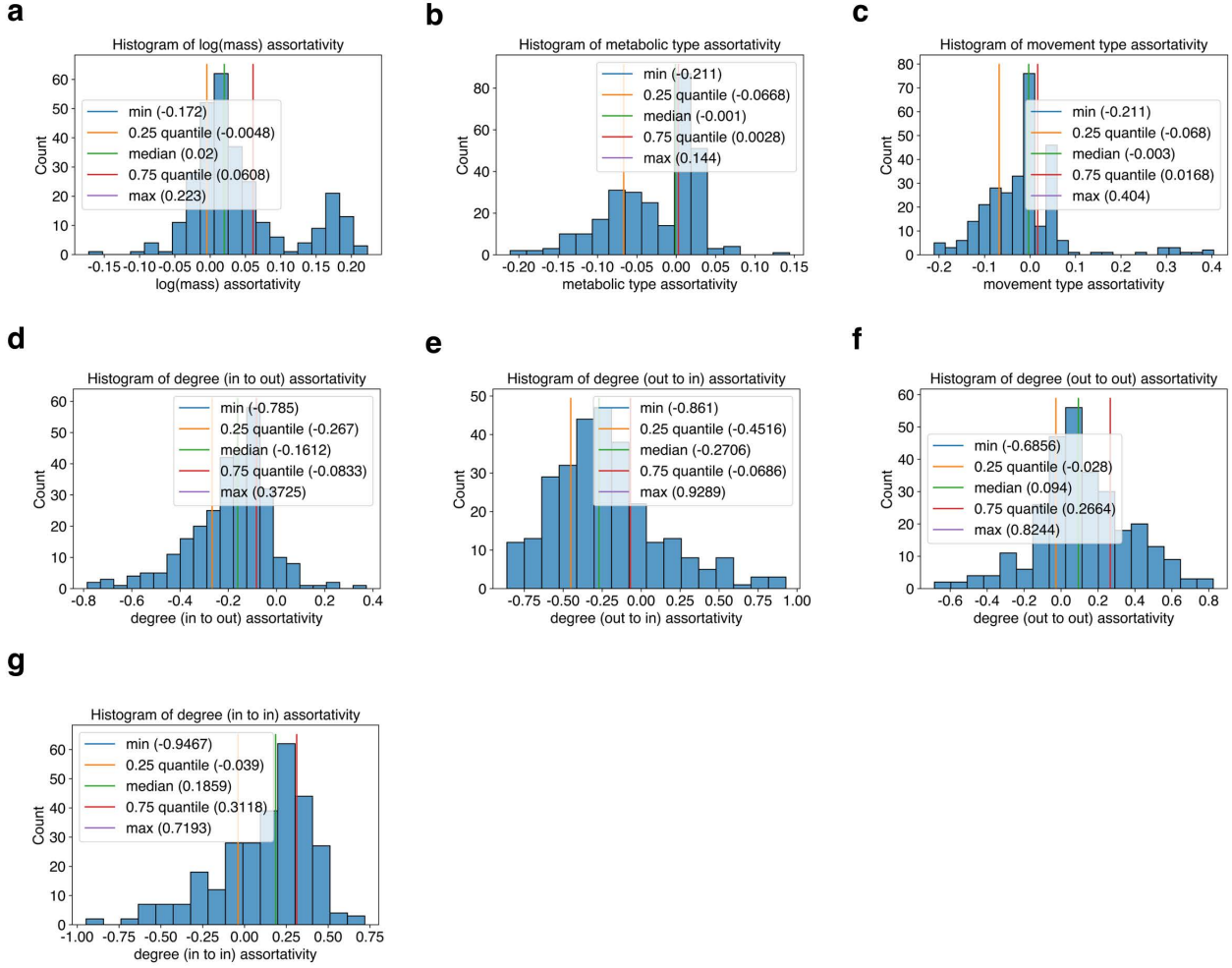
**Supplementary Fig. 5.** Log mass histogram. Distribution of log (base 10) mass across the 19,077 nodes in the pre-processed food web database.



**Supplementary Fig. 6.** Food web metadata and data processing feature distributions across the database. Histograms of (a) the missing mass fraction, (b) the higher taxonomic level proportion, (c) the unclassified taxonomic level proportion, (d) parasitic interaction removal, (e) cannibalistic interaction removal, (f) number node attributes, (g) minimum log mass, (h) maximum log mass, (i) mean log mass, (j) median log mass, and (k) mass skewness across the 290 food webs in the pre-processed database. Minimum, 0.25 quantile, median, 0.75 quantile, and maximum values are indicated, with vertical lines drawn at the 0.25 quantile, median, and 0.75 quantile.



**Supplementary Fig. 7.** Global network topology feature distributions across the database. Histograms of (a) log number of nodes, (b) average degree, (c) directed connectance, (d) clustering coefficient, and (e) modularity across the 290 food webs in the pre-processed database. Minimum, 0.25 quantile, median, 0.75 quantile, and maximum values are indicated, with vertical lines drawn at the 0.25 quantile, median, and 0.75 quantile.



**Supplementary Fig. 8.** Assortativity coefficient feature distributions across the database. Histograms of (a) log body mass assortativity, (b) metabolic type assortativity, (c) movement type assortativity, (d) degree (in to out) assortativity, (e) degree (out to in) assortativity, (f) degree (out to out) assortativity, and (g) degree (in to in) assortativity across the 290 food webs in the pre-processed database. Minimum, 0.25 quantile, median, 0.75 quantile, and maximum values are indicated, with vertical lines drawn at the 0.25 quantile, median, and 0.75 quantile.

### SUPPLEMENTARY NOTE 3. CUSTOMIZING ECOLOGICAL TOPOLOGICAL PREDICTORS

We adapt the classic preferential attachment (PA, Eq. (1)) and common neighbor (CN, Eq. (5)) predictors to an ecological setting, to define the EPA (Eq. (2)), and ECN (Eq. (6)) predictors, where  $\deg(i)$  represents the degree of node  $i$ ,  $\deg^-(i)$  represents the in-degree of node  $i$ ,  $\deg^+(i)$  represents the out-degree of node  $i$ ,  $I^-(i)$  represents the neighbor set of node  $i$ ,  $I^+(i)$  represents the out-neighbor set of node  $i$ , and  $I^-(i)$  represents the in-neighbor set of node  $i$ . The EPA predictor assumes that interactions are more likely between generalist consumers and generalist resources. The ECN predictor is updated by substituting the common neighbor set (CNS) between nodes  $i$  and  $j$

(Eq. (3)) with the ecological common neighbor set (ECNS, Eq. (4), see Fig. 1).

$$\text{PA}(i, j) = \deg(i) \times \deg(j) \quad (1)$$

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

$$\text{CNS}(i, j) = \Gamma(i) \cap \Gamma(j) \quad (3)$$

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

$$\text{CN}(i, j) = |\text{CNS}(i, j)| \quad (5)$$

$$\text{ECN}(i, j) = |\text{ECNS}(i, j)| \quad (6)$$

We update the Leight-Holme-Newman index (LHN, Eq. (7)) [19] to the ecological Leight-Holme-Newman index (ELHN, Eq. (8)), by replacing its classic quantities with ecologically-oriented versions of those same quantities.

$$\text{LHN}(i, j) = \text{CN}(i, j) / \text{PA}(i, j) \quad (7)$$

$$\text{ELHN}(i, j) = \text{ECN}(i, j) / \text{EPA}(i, j) \quad (8)$$

We update the Adamic Adar index (AA, Eq. (9)) [20] to the ecological Adamic Adar index (EAA, Eq. (10)) in the same way, replacing its classic quantity CNS with the ecologically-oriented version ECNS. This predictor indicates that common neighbors of lower degree are potentially more impactful for predicting missing links.

$$\text{AA}(i, j) = \sum_{w \in \text{CNS}} \frac{1}{\ln(\deg(w))} \quad (9)$$

$$\text{EAA}(i, j) = \sum_{w \in \text{ECNS}} \frac{1}{\ln(\deg(w))} \quad (10)$$

We update the Resource Allocation index (RA, Eq. (11)) [21] to the ecological Resource Allocation index (ERA, Eq. (12)) in the same way.

$$\text{RA}(i, j) = \sum_{w \in \text{CNS}} \frac{1}{\deg(w)} \quad (11)$$

$$\text{ERA}(i, j) = \sum_{w \in \text{ECNS}} \frac{1}{\deg(w)} \quad (12)$$

We also add an additional score derived from all potential sets of directed common neighbors between two nodes ( $\text{ECN}_{\text{score}}$ , Eq. (16)).

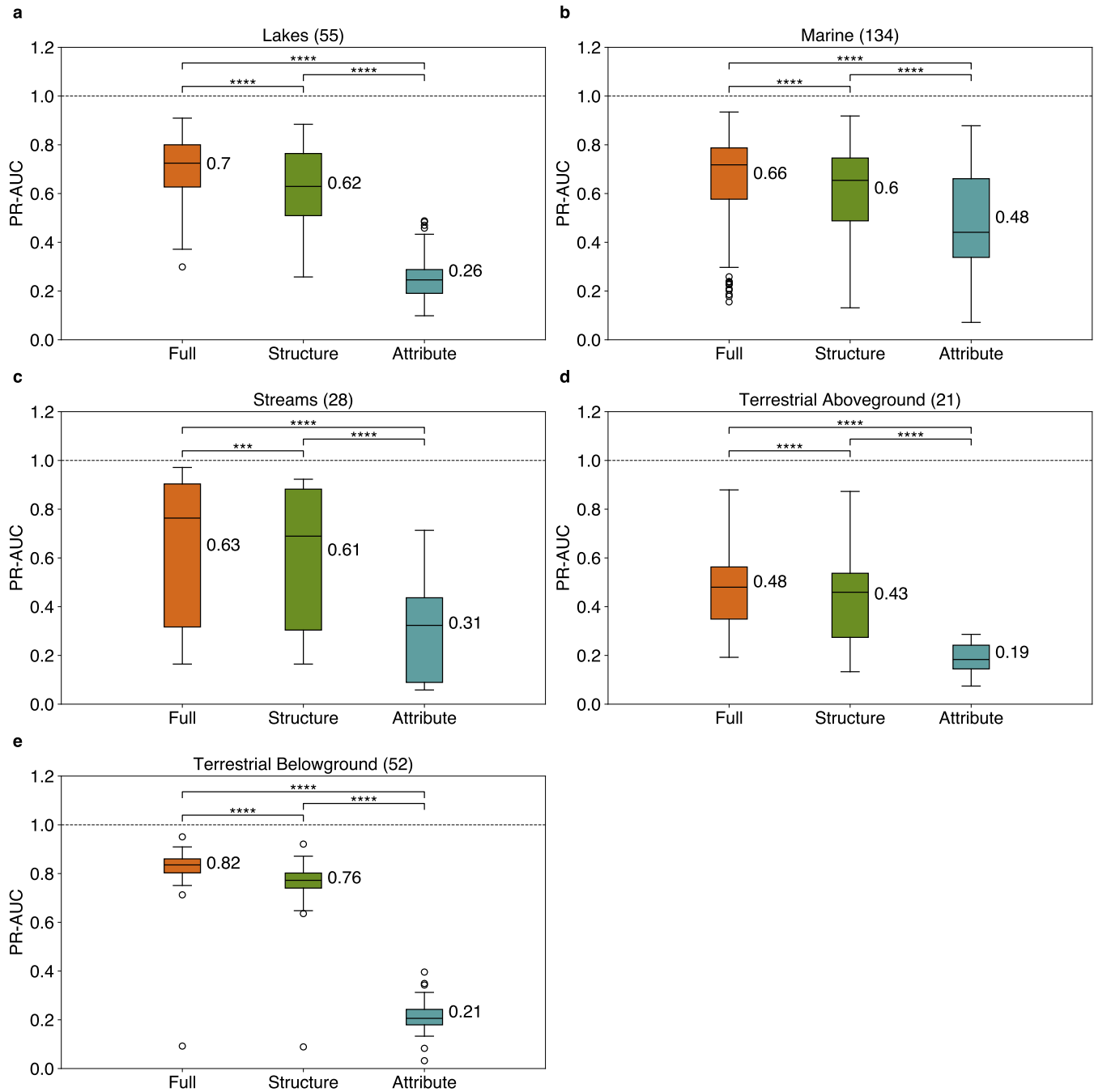
$$\text{C2}(i, j) = |\Gamma^-(i) \cap \Gamma^+(j)| \quad (13)$$

$$\text{C3}(i, j) = |\Gamma^+(i) \cap \Gamma^+(j)| \quad (14)$$

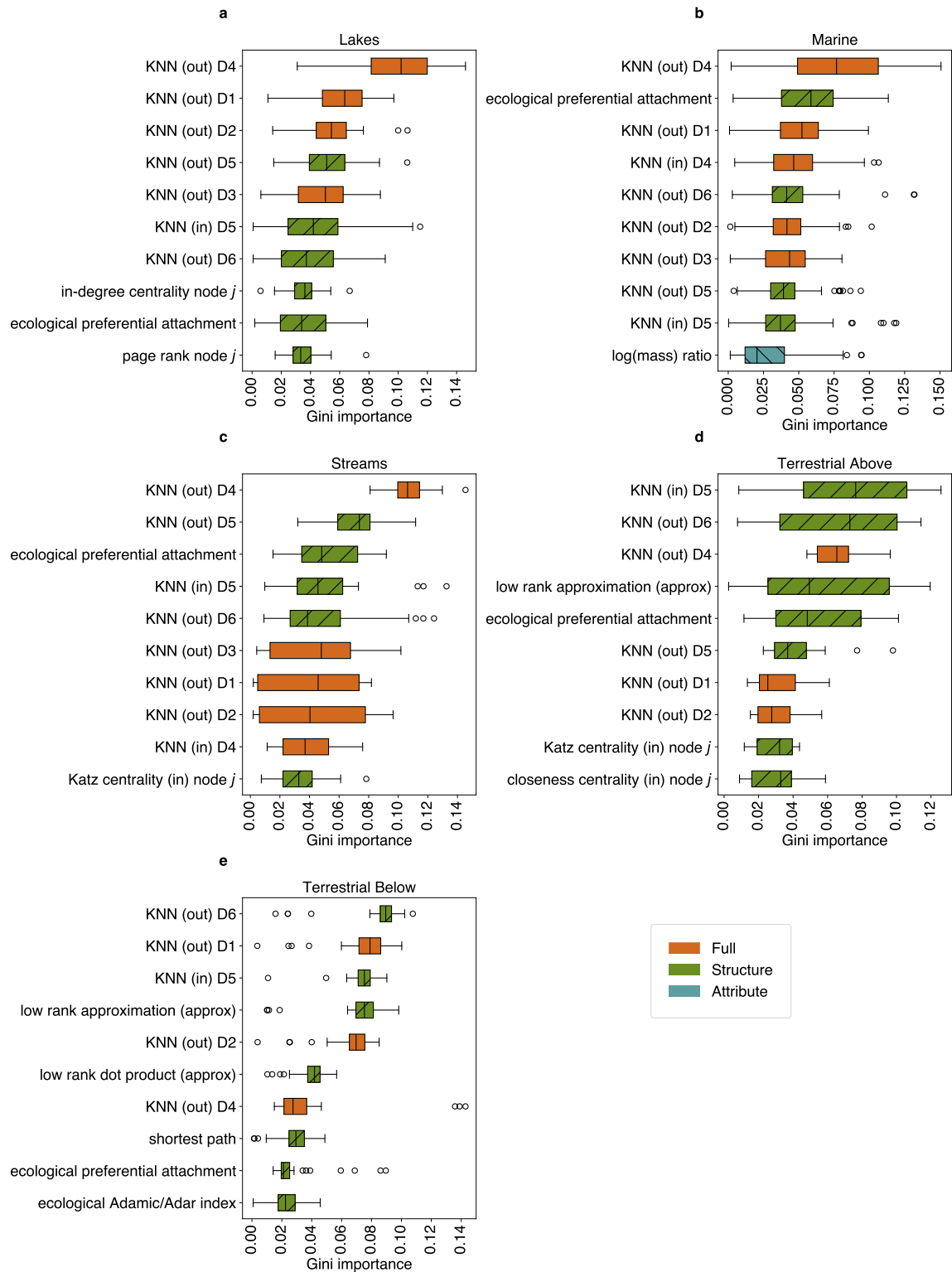
$$\text{C4}(i, j) = |\Gamma^-(i) \cap \Gamma^-(j)| \quad (15)$$

$$\text{ECN}_{\text{score}}(i, j) = \text{ECN}(i, j) - [\text{C2}(i, j) + \text{C3}(i, j) + \text{C4}(i, j)] \quad (16)$$

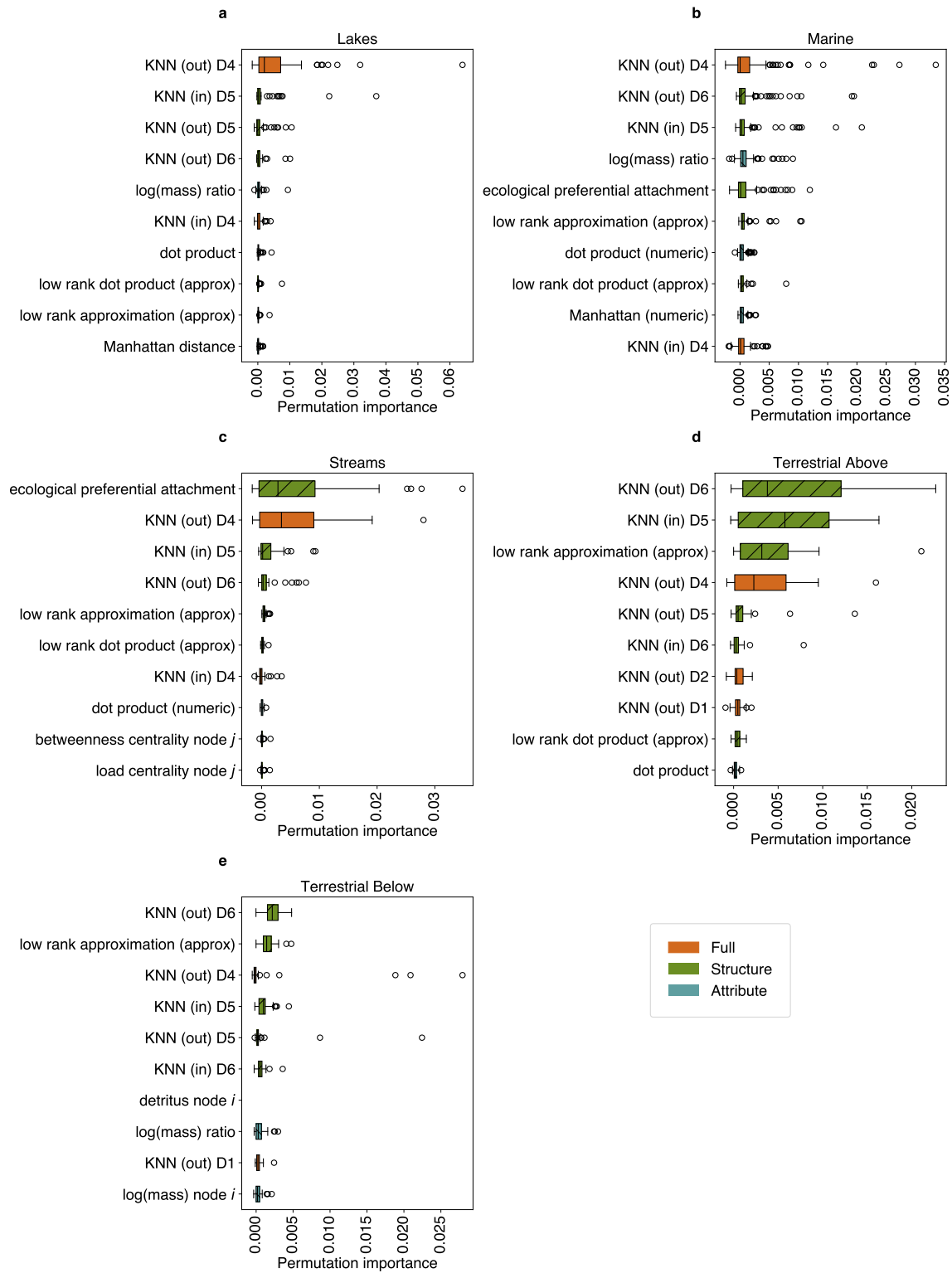
## SUPPLEMENTARY FIGURES



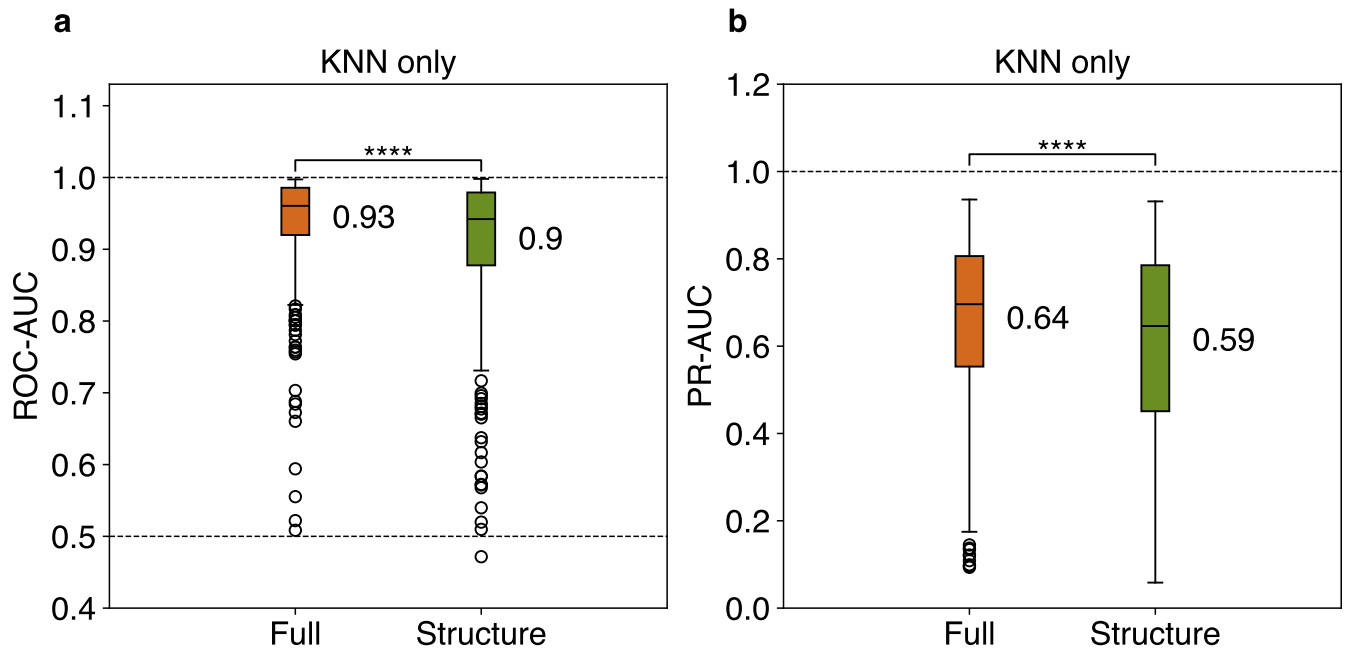
**Supplementary Fig. 9.** Average link prediction performance for stacked models using structure-only predictors (Structure), attribute-only predictors (Attribute), and both (Full), measured by Area Under the Precision-Recall Curve (PR-AUC), separated by ecosystem type: (a) Lakes, (b) Marine, (c) Streams, (d) Terrestrial Aboveground, and (e) Terrestrial Belowground. Performance for each of the 3 models is calculated for each food web by averaging across 5 independent iterations of evaluating across 5 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, and significant differences in mean model performance based on false discovery rate (FDR) adjusted (Benjamini-Hochberg method [18]) within-subjects pair-wise two-sided t-tests are shown where \*\*\*\* indicates a p value<0.0001 and \*\*\* indicates a p value<0.001. 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.



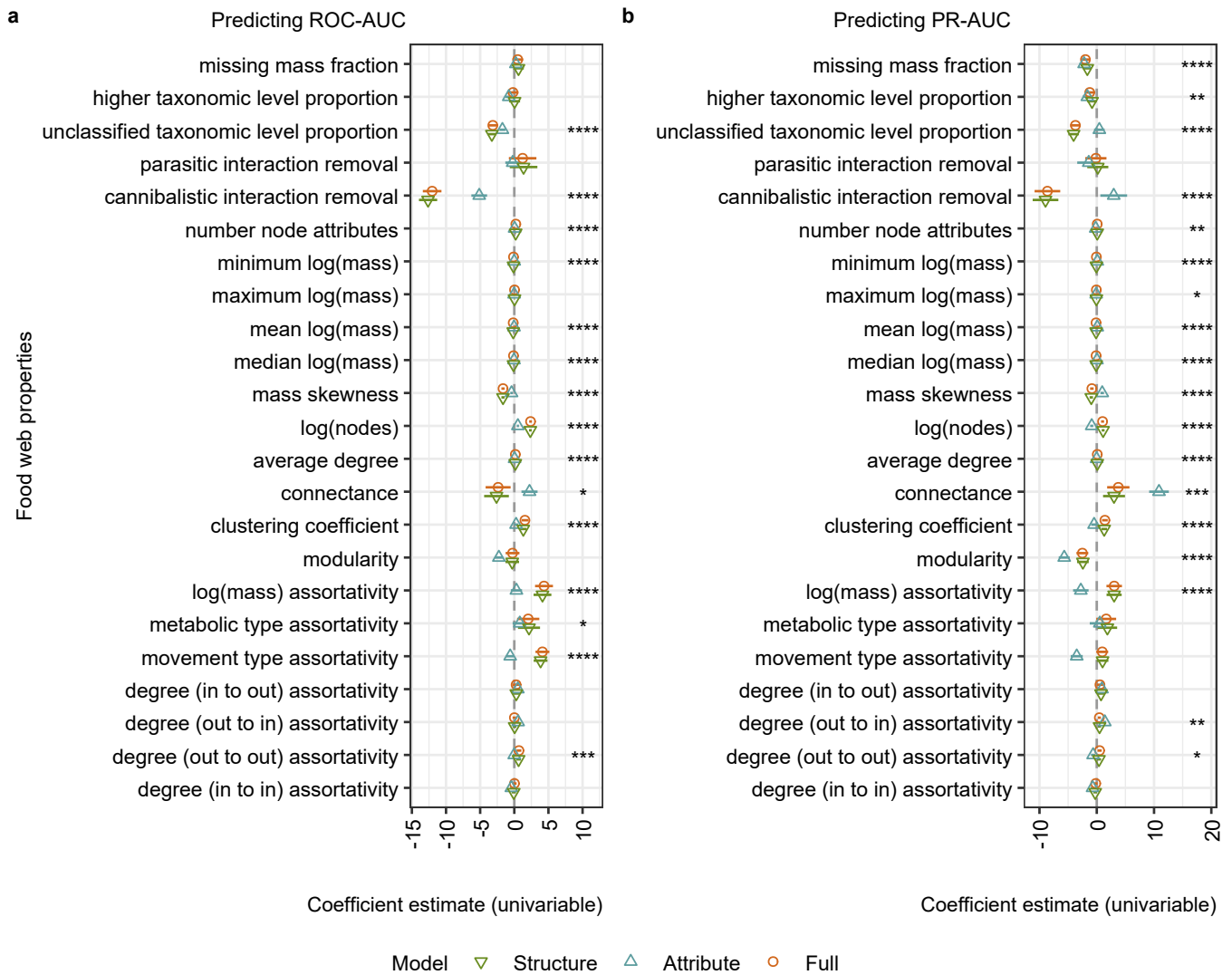
**Supplementary Fig. 10.** Top 10 features ranked by Gini importance averaged across iterations and folds for food webs belonging to each ecosystem type: (a) Lakes, (b) Marine, (c) Streams, (d) Terrestrial Aboveground, and (e) Terrestrial Belowground. Structure predictors (out of 51) are based only on 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. Node  $j$  refers to the consumer species. KNN predictors are based on K-Nearest Neighbors. 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. Results are taken from the full model containing both structure and attribute predictors (106 predictors in total,  $N=7250$  results per feature across 5 folds, 5 iterations, and 290 food webs). 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).



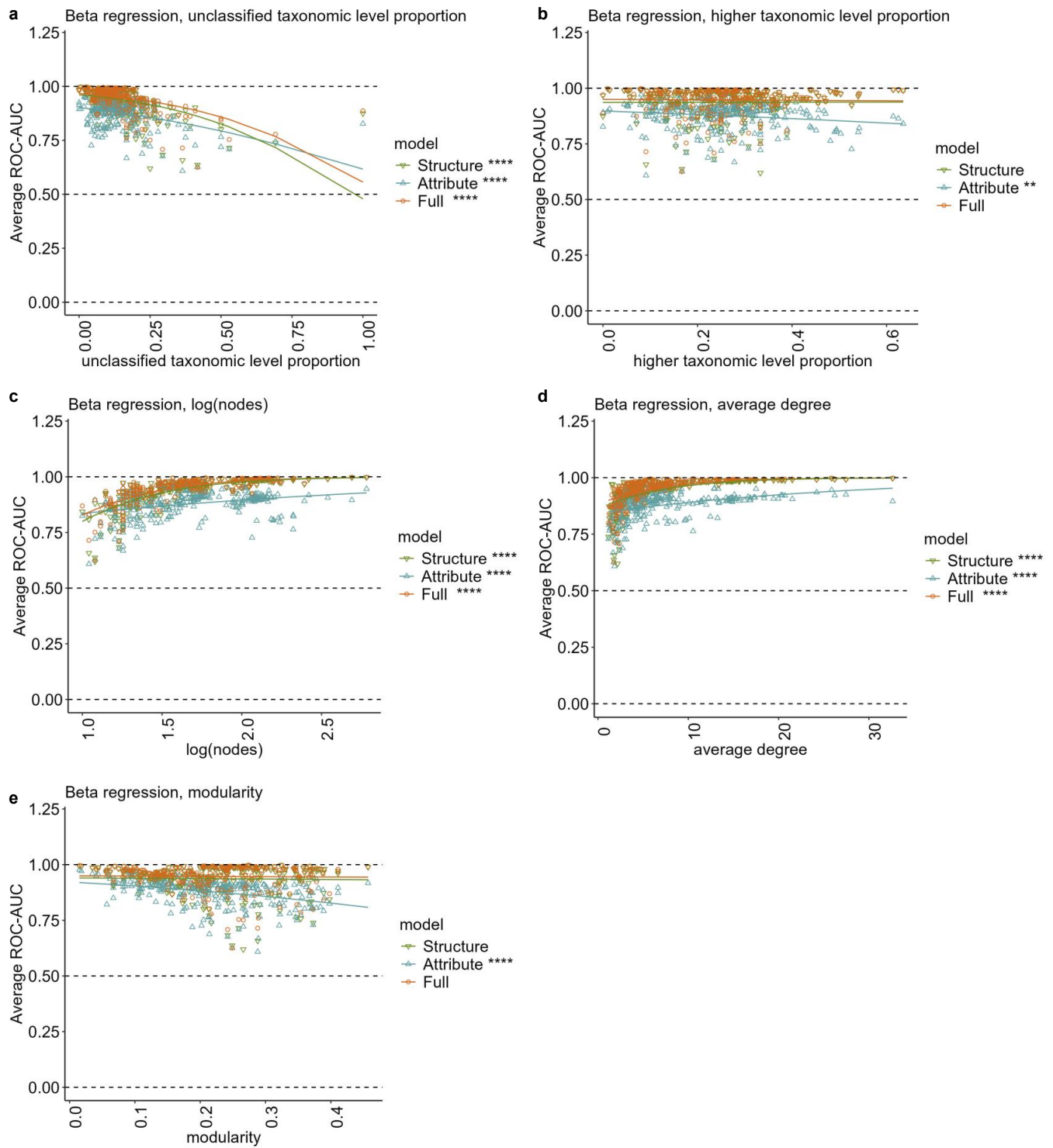
**Supplementary Fig. 11.** Top 10 features ranked by permutation importance averaged across iterations and folds for food webs belonging to each ecosystem type: (a) Lakes, (b) Marine, (c) Streams, (d) Terrestrial Aboveground, and (e) Terrestrial Belowground. Structure predictors (out of 51) are based 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. Node  $j$  refers to the consumer species and node  $i$  refers to the resource species. KNN predictors are based on K-Nearest Neighbors. 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. Results are taken from the full model containing both structure and attribute predictors (106 predictors in total,  $N=7250$  results per feature across 5 folds, 5 iterations, and 290 food webs). 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).



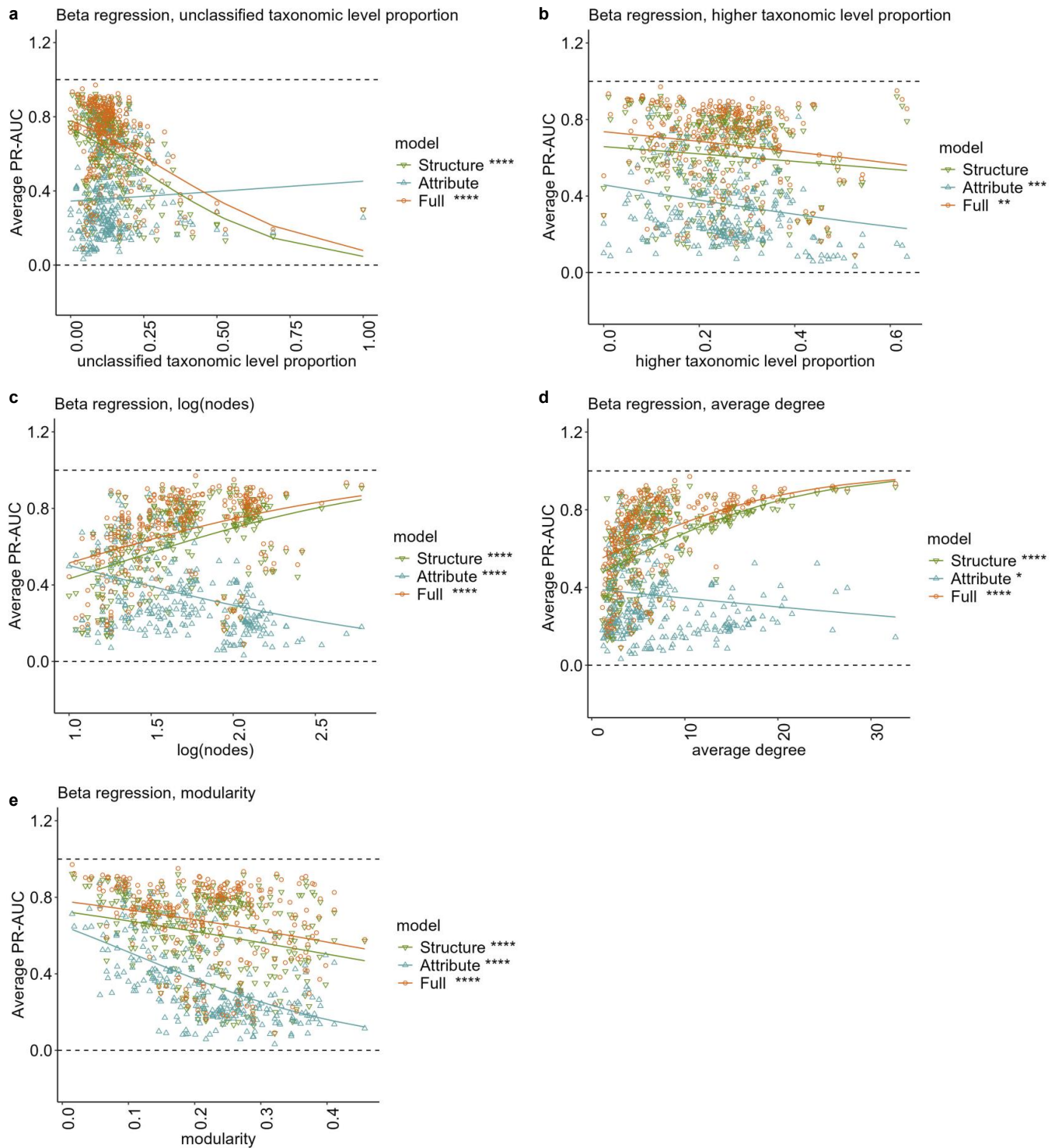
**Supplementary Fig. 12.** Performance of the structure-only and full models restricted to only KNN (K-Nearest Neighbors) predictors. Average (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) results. Results are averaged across 5 unique folds (5 results per food web, N=2900 results across 1 iteration, 5 folds, 290 food webs, and 2 models). Significant difference in mean model performance based on within-subjects pair-wise two-sided t-test where \*\*\*\* indicates a p value < 0.0001. For ROC-AUC, the full vs. structure results are: t-statistic: 11.252, confidence interval: (0.024, 0.035), cohen's d: 0.306, degrees of freedom: 289, adjusted p value: 0.0. For PR-AUC the full vs. structure results are: t-statistic: 11.658, confidence interval: (0.047, 0.066), cohen's d: 0.26, degrees of freedom: 289, adjusted p value: 0.0. Horizontal dashed lines indicate the threshold for better than random predictive performance (ROC-AUC=0.5) and the maximum possible performance (ROC-AUC, PR-AUC=1). 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).



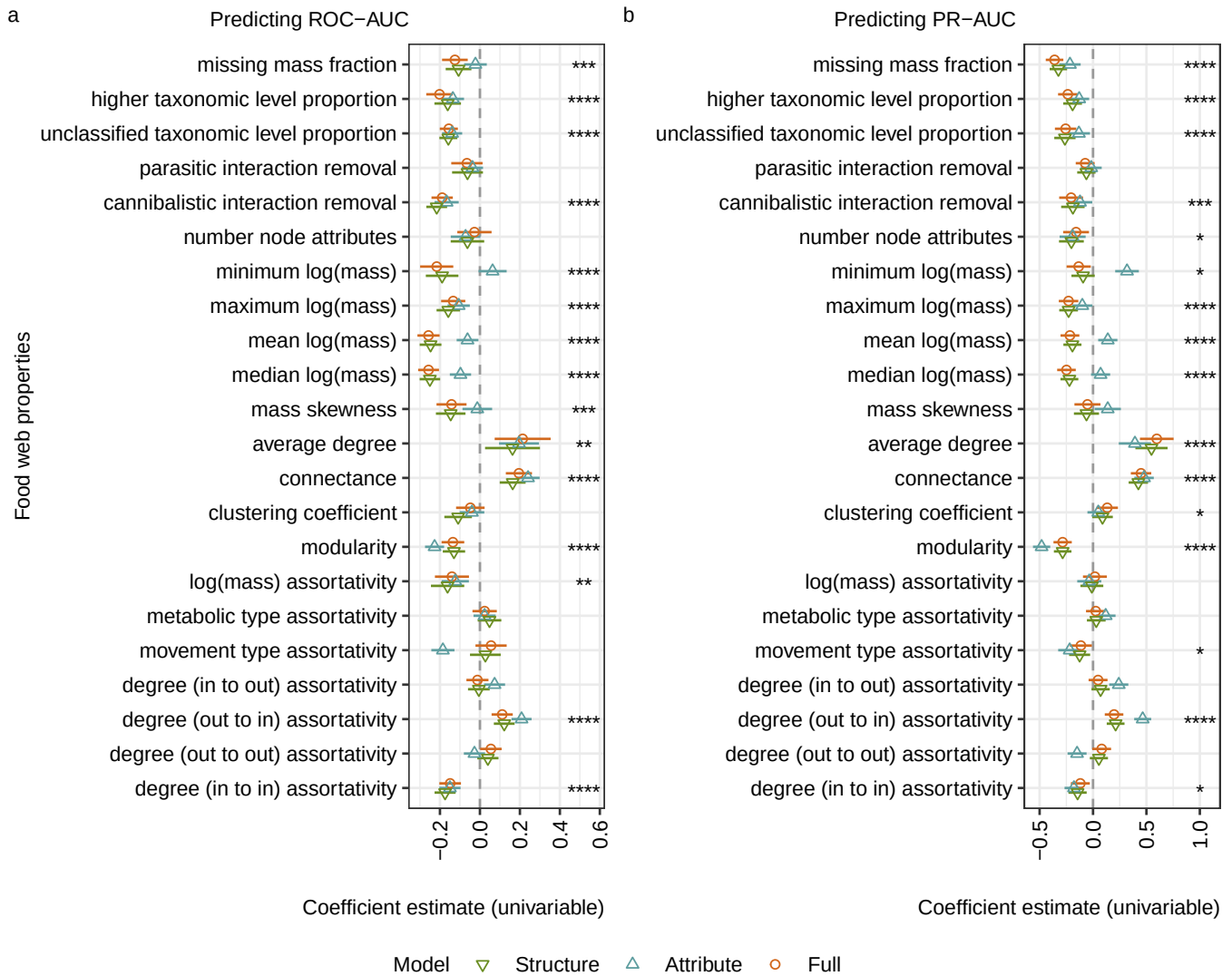
**Supplementary Fig. 13.** Mean coefficient estimates in univariable beta regressions between 23 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, with covariates un-scaled ( $N=290$  samples for each regression,  $N=138$  regressions across 23 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). 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. For the Full model,  $p \text{ value} \leq 0.0001$ : \*\*\*\*,  $p \text{ value} \leq 0.001$ : \*\*\*,  $p \text{ value} \leq 0.01$ : \*\*,  $p \text{ value} \leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Details are in Supplementary Data 2.



**Supplementary Fig. 14.** Performance of the structure-only (green inverted triangles), attribute-only (blue triangles), and full (orange circles) models contextualized by select dataset characteristics (N=290 results per model). Lines are shown for the beta regression fit for each model. Average ROC-AUC (Area Under the Receiver Operating Characteristic Curve) results by (a) unclassified taxonomic level proportion, (b) higher taxonomic level proportion, (c) log(nodes), (d) average degree, and (e) modularity. Significance of coefficients in univariable beta regressions fit to these relationships are indicated by level (p value $\leq$ 0.0001: \*\*\*\*, p value $\leq$ 0.001: \*\*\*, p value $\leq$ 0.01: \*\*, p value $\leq$ 0.05: \*, FDR adjusted, Benjamini-Hochberg method [18]). Horizontal dashed lines indicate the threshold for better than random predictive performance (ROC-AUC=0.5) and the maximum possible performance (ROC-AUC=1). Details are in Supplementary Data 2.

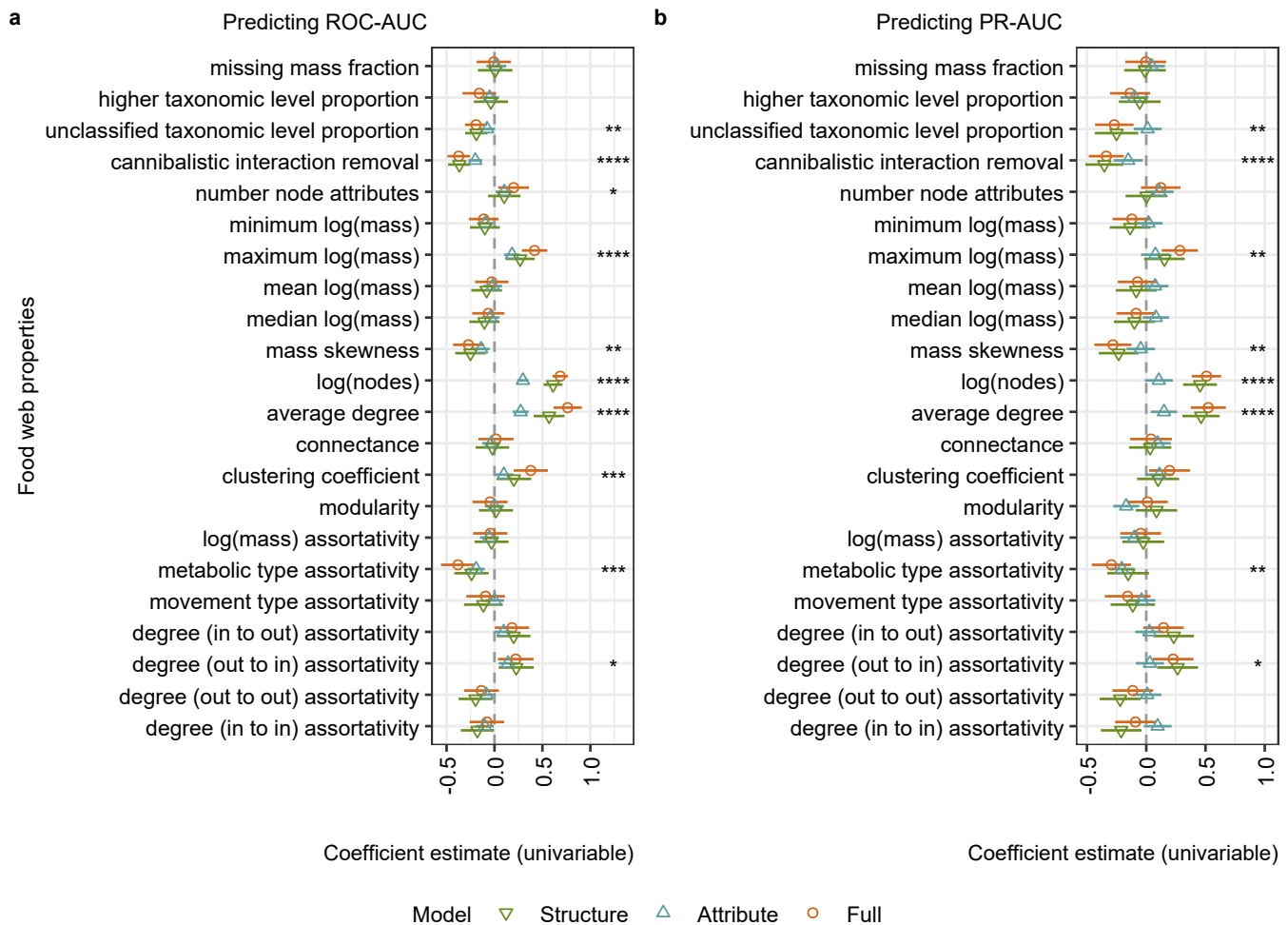


**Supplementary Fig. 15.** Performance of the structure-only (green inverted triangles), attribute-only (blue triangles), and full models (orange circles) contextualized by select dataset characteristics (N=290 results per model). Lines are shown for the beta regression fit for each model. Average PR-AUC (Area Under the Precision-Recall Curve) results by (a) unclassified taxonomic level proportion, (b) higher taxonomic level proportion, (c) log(nodes), (d) average degree, and (e) modularity. Significance of coefficients in univariable beta regressions fit to these relationships are indicated by level (p value  $\leq 0.0001$ : \*\*\*\*, p value  $\leq 0.001$ : \*\*\*, p value  $\leq 0.01$ : \*\*, p value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]). Horizontal dashed lines indicate the minimum possible performance (PR-AUC=0) and the maximum possible performance (PR-AUC=1). Details are in Supplementary Data 2.



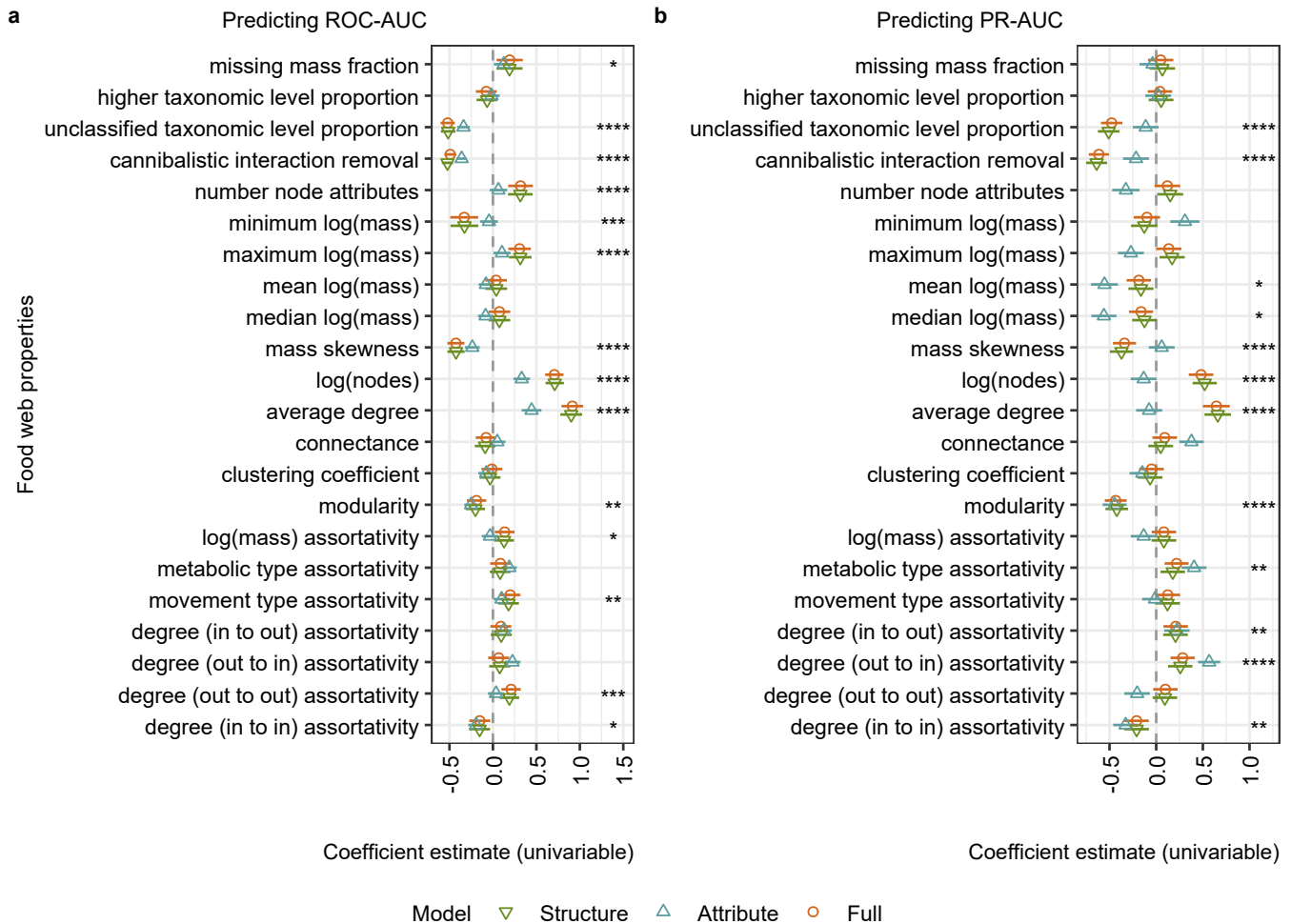
**Supplementary Fig. 16.** Mean coefficient estimates in beta regressions between 22 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, when the log of the number of nodes in the food web is also included as a covariate ( $N=290$  samples for each regression,  $N=132$  regressions across 22 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model,  $p$  value  $\leq 0.0001$ : \*\*\*\*,  $p$  value  $\leq 0.001$ : \*\*\*,  $p$  value  $\leq 0.01$ : \*\*,  $p$  value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Details are in Supplementary Data 2.

## Lakes



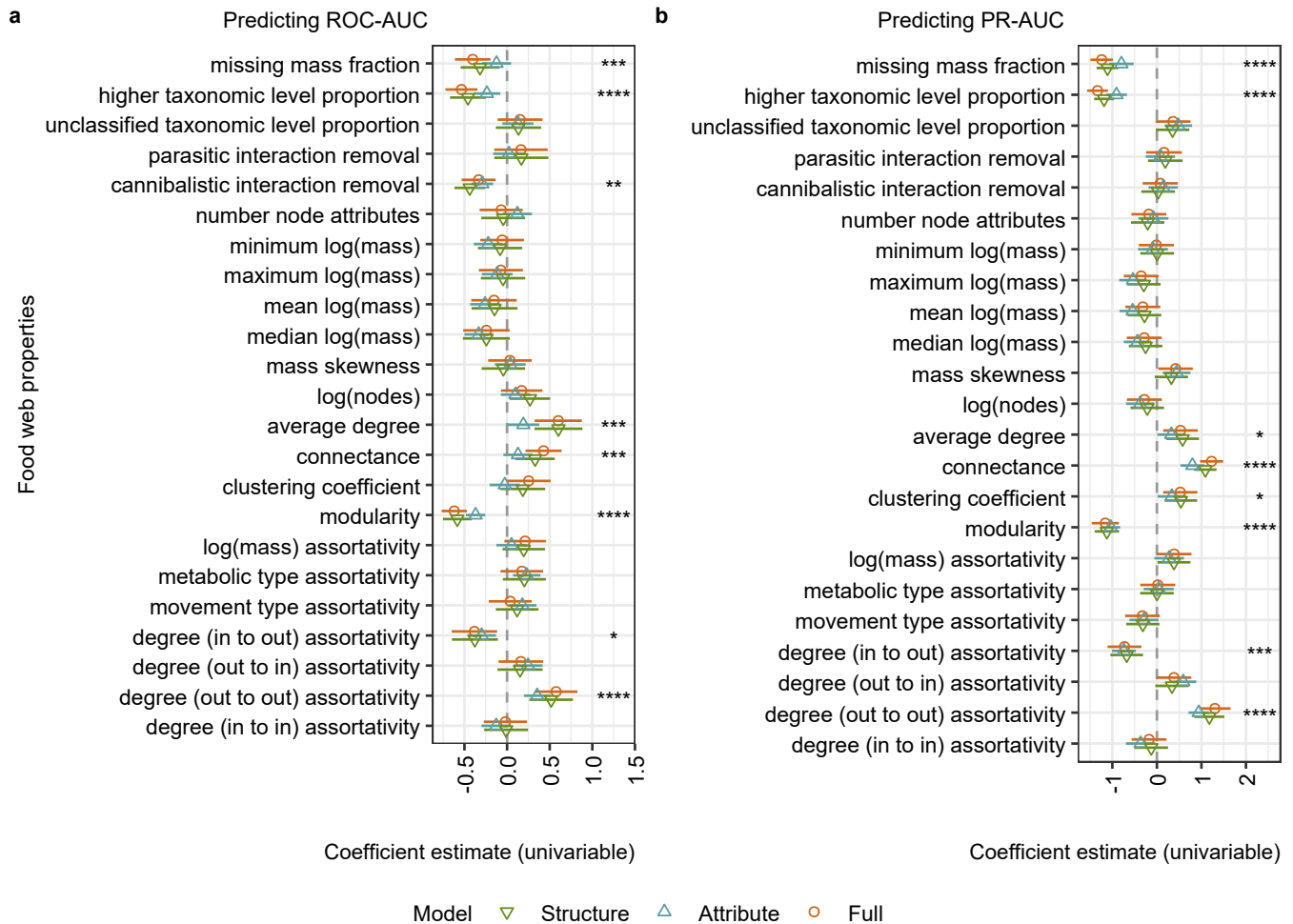
**Supplementary Fig. 17.** Mean coefficient estimates in univariable beta regressions between 22 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, for only lake food webs (N=55 samples for each regression, N=132 regressions across 22 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model, p value  $\leq 0.0001$ : \*\*\*\*, p value  $\leq 0.001$ : \*\*\*, p value  $\leq 0.01$ : \*\*, p value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Parasitic interaction removal is not included as a property as no lake food webs had parasitic interactions removed. Details are in Supplementary Data 2.

Marine



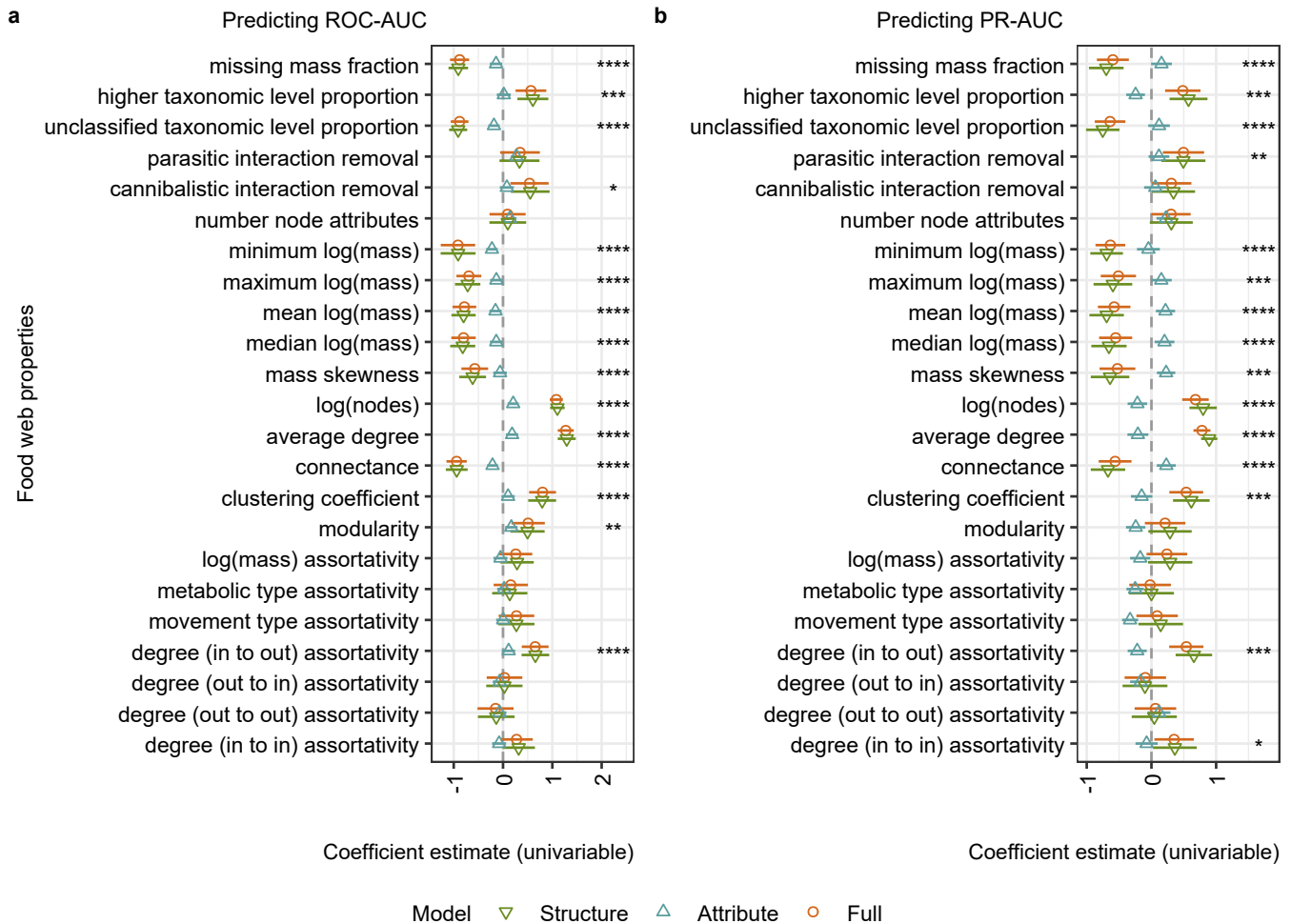
**Supplementary Fig. 18.** Mean coefficient estimates in univariable beta regressions between 22 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, for only marine food webs (N=134 samples for each regression, N=132 regressions across 22 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model, p value  $\leq 0.0001$ : \*\*\*\*, p value  $\leq 0.001$ : \*\*\*, p value  $\leq 0.01$ : \*\*, p value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Parasitic interaction removal is not included as property as no marine food webs had parasitic interactions removed. Details are in Supplementary Data 2.

## Streams



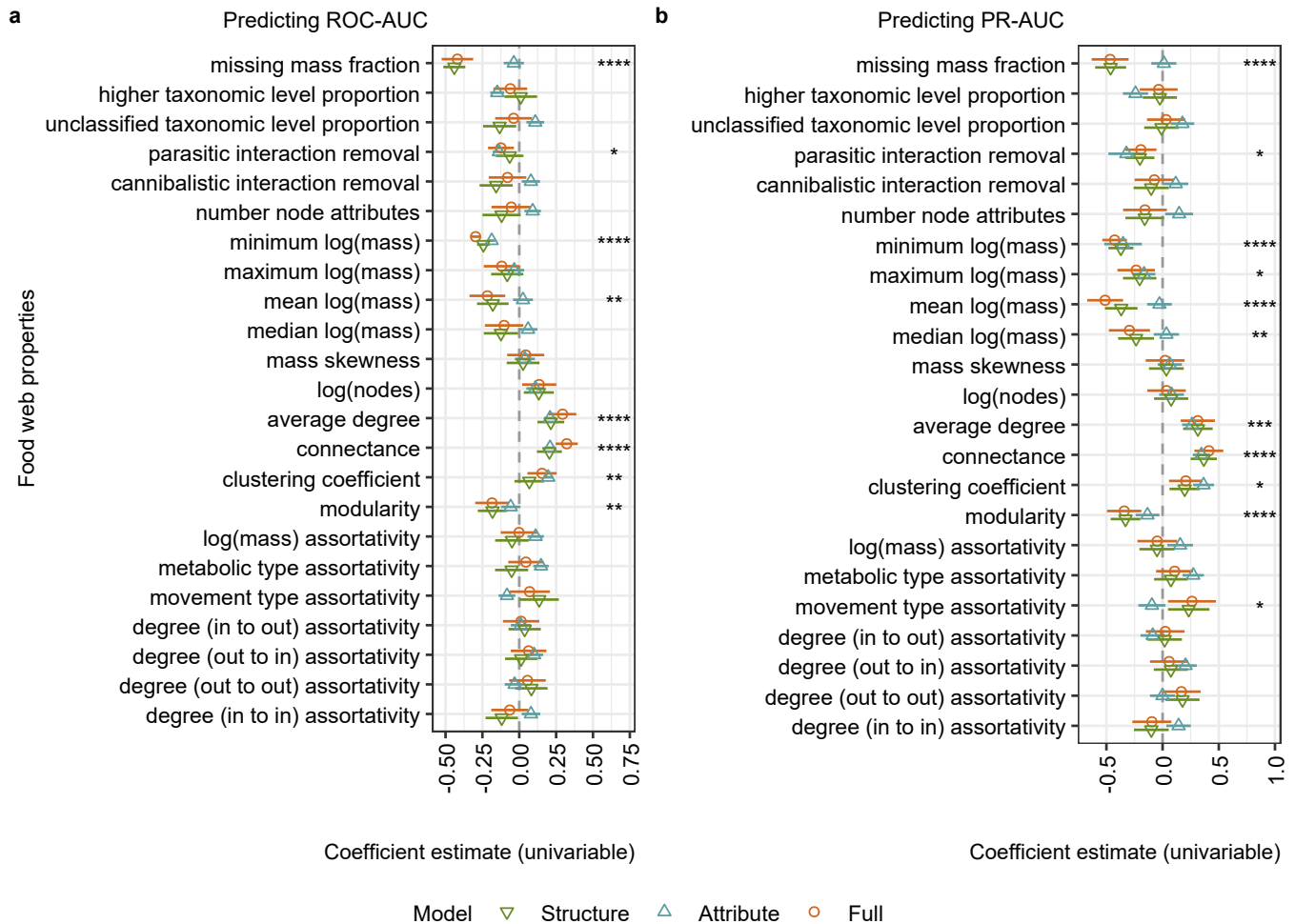
**Supplementary Fig. 19.** Mean coefficient estimates in univariable beta regressions between 23 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, for only stream food webs (N=28 samples for each regression, N=138 regressions across 23 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model, p value  $\leq 0.0001$ : \*\*\*\*, p value  $\leq 0.001$ : \*\*\*, p value  $\leq 0.01$ : \*\*, p value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Details are in Supplementary Data 2.

## Terrestrial Above

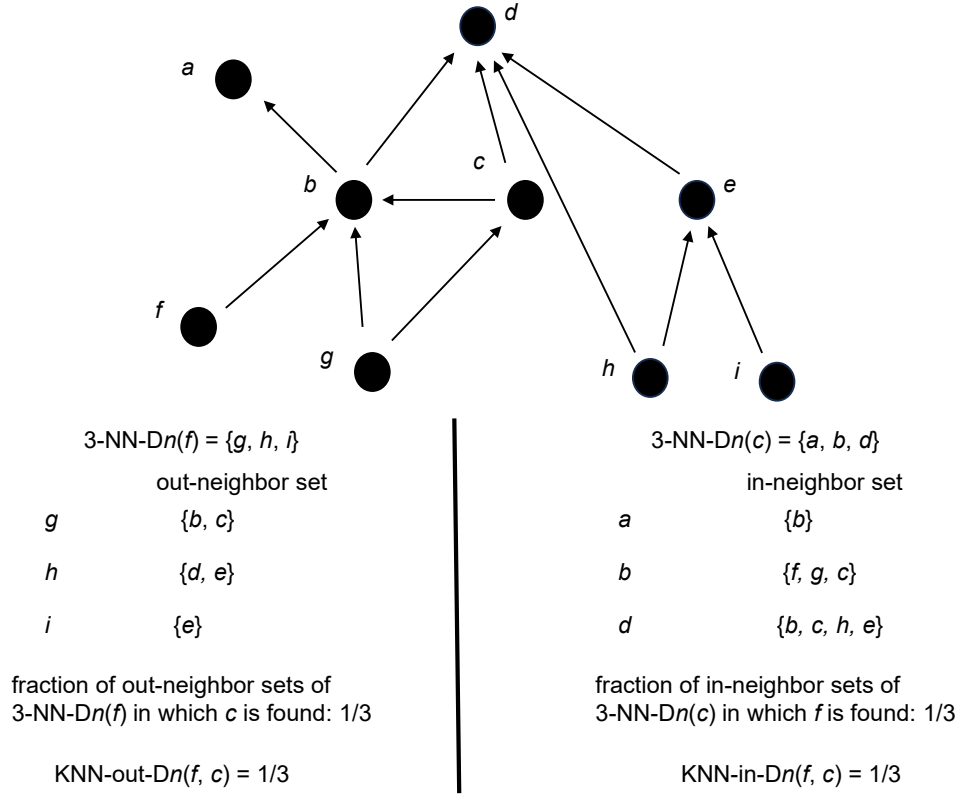


**Supplementary Fig. 20.** Mean coefficient estimates in univariable beta regressions between 23 food web properties and (a) ROC-AUC (Area Under the Receiver Operating Characteristic Curve) and (b) PR-AUC (Area Under the Precision-Recall Curve) performance for missing link prediction, for only terrestrial aboveground food webs (N=21 samples for each regression, N=138 regressions across 23 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model, p value  $\leq 0.0001$ : \*\*\*\*, p value  $\leq 0.001$ : \*\*\*, p value  $\leq 0.01$ : \*\*, p value  $\leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Details are in Supplementary Data 2.

## Terrestrial Below



**Supplementary Fig. 21.** Mean coefficient estimates in univariable beta regressions between 23 food web properties and (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, for only terrestrial belowground food webs (N=52 samples for each regression, N=138 regressions across 23 properties, 2 performance metrics, and 3 models). Estimates are shown based on the performance of the structure-only model (green inverted triangles), the attribute-only model (blue triangles), and the full model (orange circles). All food web properties were first z score normalized. 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. For the Full model,  $p \text{ value} \leq 0.0001$ : \*\*\*\*,  $p \text{ value} \leq 0.001$ : \*\*\*,  $p \text{ value} \leq 0.01$ : \*\*,  $p \text{ value} \leq 0.05$ : \*, FDR adjusted, Benjamini-Hochberg method [18]. Details are in Supplementary Data 2.



**Supplementary Fig. 22.** KNN (K-Nearest Neighbors) predictors. With  $K = 3$ , two KNN predictors are evaluated for in and out directions for each distance metric  $D_n$ . Here they are evaluated for the directed node pair  $(f, c)$  in an example food web with a hypothetical distance metric  $D_n$ . After the 3 nearest neighbors are identified for each of  $f$  and  $c$  (3-NN- $D_n(f)$ , 3-NN- $D_n(c)$ ), the out and in-neighbor sets are evaluated respectively to calculate a missing link prediction score based on how often  $c$  or  $f$  interact with the 3 nearest neighbors.

## SUPPLEMENTARY TABLES

**Supplementary Table 1:** List of structure predictors. All of these predictors are included in the structure-only model and the full model. Node  $i$  refers to the resource species and node  $j$  refers to the consumer species.

| Predictor Category     | Name                                    | Details                                                                                                               |
|------------------------|-----------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| Structure (pairwise)   | Ecological common neighbor score        | See Supplemental Equation (16), $ECN_{score}(i, j)$ .                                                                 |
| Structure (pairwise)   | Ecological preferential attachment      | See Main Text Equation (2), $EPA(i, j)$ .                                                                             |
| Structure (pairwise)   | Personalized page rank                  | $j$ th entry of the personalized page rank of node $i$ , with input network treated as directed. Code: NetworkX [22]. |
| Structure (node-level) | Page rank node $i$ , Page rank node $j$ | Page rank values for nodes $i$ and $j$ , with a directed network as input. Code: NetworkX [22].                       |
| Structure (pairwise)   | Ecological Jaccard coefficient          | See Main Text Equation (10), $EJC(i, j)$ .                                                                            |
| Structure (pairwise)   | Ecological Adamic/Adar index            | See Supplemental Equation (10), $EAA(i, j)$ .                                                                         |
| Structure (pairwise)   | Ecological resource allocation index    | See Supplemental Equation (12), $ERA(i, j)$ .                                                                         |

|                        |                                                                                  |                                                                                                               |
|------------------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Structure (pairwise)   | Shortest path                                                                    | Shortest directed path between nodes $i$ and $j$ . Ref: [23], Code: NetworkX [22].                            |
| Structure (node-level) | Average neighbor in degree node $i$ ,<br>Average neighbor in degree node $j$     | Code: NetworkX [22].                                                                                          |
| Structure (node-level) | Average neighbor out degree node $i$ ,<br>Average neighbor out degree node $j$   | Code: NetworkX [22].                                                                                          |
| Structure (node-level) | In-degree centrality node $i$ , In-degree centrality node $j$                    | Code: NetworkX [22].                                                                                          |
| Structure (node-level) | Out-degree centrality node $i$ ,<br>Out-degree centrality node $j$               | Code: NetworkX [22].                                                                                          |
| Structure (node-level) | Eigenvector centrality (in) node $i$ ,<br>Eigenvector centrality (in) node $j$   | Eigenvector centralities in the in-direction for nodes $i$ and $j$ . Code: NetworkX [22].                     |
| Structure (node-level) | Eigenvector centrality (out) node $i$ ,<br>Eigenvector centrality (out) node $j$ | Eigenvector centralities in the out-direction for nodes $i$ and $j$ . Code: NetworkX [22].                    |
| Structure (node-level) | Katz centrality (in) node $i$ , Katz centrality (in) node $j$                    | Katz centralities in the in-direction for nodes $i$ and $j$ . Code: NetworkX [22].                            |
| Structure (node-level) | Katz centrality (out) node $i$ , Katz centrality (out) node $j$                  | Katz centralities in the out-direction for nodes $i$ and $j$ . Code: NetworkX [22].                           |
| Structure (node-level) | Betweenness centrality node $i$ ,<br>Betweenness centrality node $j$             | Shortest-path (directed) betweenness centralities for nodes $i$ and $j$ . Code: NetworkX [22].                |
| Structure (node-level) | Closeness centrality (in) node $i$ ,<br>Closeness centrality (in) node $j$       | Closeness centralities based on in-paths for nodes $i$ and $j$ . Code: NetworkX [22].                         |
| Structure (node-level) | Closeness centrality (out) node $i$ ,<br>Closeness centrality (out) node $j$     | Closeness centralities based on out-paths for nodes $i$ and $j$ . Code: NetworkX [22].                        |
| Structure (node-level) | Load centrality node $i$ , Load centrality node $j$                              | Load centralities for nodes $i$ and $j$ based on shortest directed paths. Code: NetworkX [22].                |
| Structure (pairwise)   | Ecological Leicht-Holme-Newman index                                             | See Supplemental Equation (8), $ELHN(i, j)$ .                                                                 |
| Structure (pairwise)   | Ecological common neighbor count                                                 | See Supplemental Equation (6), $ECN(i, j)$ .                                                                  |
| Structure (node-level) | Local clustering coefficient node $i$ ,<br>Local clustering coefficient node $j$ | Local clustering coefficients based on directed triangles for nodes $i$ and $j$ . Code: NetworkX [22].        |
| Structure (node-level) | Local triangles node $i$ , Local triangles node $j$                              | Local number of directed triangles for nodes $i$ and $j$ . Code: NetworkX [22].                               |
| Structure (pairwise)   | Low rank approximation                                                           | Entry $i, j$ in low rank approximation (LRA) via singular value decomposition (SVD). Ref: [24], Code: [2].    |
| Structure (pairwise)   | Low rank dot product                                                             | Dot product of columns $i$ and $j$ in LRA via SVD for each pair of nodes $i, j$ (dLRA). Ref: [24], Code: [2]. |
| Structure (pairwise)   | Low rank mean                                                                    | Average of entries $i$ and $j$ 's neighbors in LRA (mLRA). Ref: [24], Code: [2].                              |
| Structure (pairwise)   | Low rank approximation (approx)                                                  | An approximation of LRA. Ref: [24], Code: [2].                                                                |

|                        |                                                 |                                                 |
|------------------------|-------------------------------------------------|-------------------------------------------------|
| Structure (pairwise)   | Low rank dot product (approx)                   | An approximation of dLRA. Ref: [24], Code: [2]. |
| Structure (pairwise)   | Low rank mean (approx)                          | An approximation of mLRA. Ref: [24], Code: [2]. |
| Structure (node-level) | Trophic level node $i$ , Trophic level node $j$ | Ref: [25], Code: NetworkX [22].                 |

**Supplementary Table 2:** List of attribute predictors. All of these predictors are included in the attribute-only model and the full model. Node  $i$  refers to the resource species and node  $j$  refers to the consumer species.

| Predictor Category     | Name                                | Details                                                                                         |
|------------------------|-------------------------------------|-------------------------------------------------------------------------------------------------|
| Attribute (node-level) | $\{a\}$ node $i$ , $\{a\}$ node $j$ | Raw attribute values (repeated for all $ A $ attributes $a$ ).                                  |
| Attribute (pairwise)   | Euclidean distance                  | Euclidean distance between attribute vectors.                                                   |
| Attribute (pairwise)   | Manhattan distance                  | Manhattan distance between attribute vectors.                                                   |
| Attribute (pairwise)   | Cosine distance                     | Cosine distance between attribute vectors.                                                      |
| Attribute (pairwise)   | Dot product                         | Dot product between attribute vectors.                                                          |
| Attribute (pairwise)   | Euclidean (numeric)                 | Euclidean distance between numeric part of attribute vectors.                                   |
| Attribute (pairwise)   | Manhattan (numeric)                 | Manhattan distance between numeric part of attribute vectors.                                   |
| Attribute (pairwise)   | Cosine (numeric)                    | Cosine distance between numeric part of attribute vectors.                                      |
| Attribute (pairwise)   | Dot product (numeric)               | Dot product between numeric part of attribute vectors.                                          |
| Attribute (pairwise)   | Hamming (binary)                    | Hamming distance (equivalent to Manhattan distance) between binary part of attribute vectors.   |
| Attribute (pairwise)   | Jaccard (binary)                    | Jaccard distance (equivalent to Tanimoto coefficient) between binary part of attribute vectors. |
| Attributes (pairwise)  | $\{a\}$ ratio                       | $\frac{a_i}{a_j}$ repeated for all $ N $ numeric attributes $a$ .                               |

**Supplementary Table 3:** List of nearest neighbor predictors. All of these predictors are the in full model and a subset based on  $D5$  and  $D6$  are in the structure-only model. Node  $i$  refers to the resource species and node  $j$  refers to the consumer species.

| Predictor Category             | Abbreviation | Definition                                                                                                                                                            |
|--------------------------------|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Structure+Attribute (pairwise) | KNN (out) D1 | The fraction of out- neighbor sets of KNN_D1( $i$ ) in which $j$ is found, where D1 is the Euclidean distance between full min-max normalized node attribute vectors. |
| Structure+Attribute (pairwise) | KNN (in) D1  | The fraction of in- neighbor sets of KNN_D1( $j$ ) in which $i$ is found, where D1 is the Euclidean distance between full min-max normalized node attribute vectors.  |

|                                |              |                                                                                                                                                                      |
|--------------------------------|--------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Structure+Attribute (pairwise) | KNN (out) D2 | The fraction of out- neighbor sets of $KNN\_D2(i)$ in which $j$ is found, where D2 is the Manhattan distance between full min-max normalized node attribute vectors. |
| Structure+Attribute (pairwise) | KNN (in) D2  | The fraction of in- neighbor sets of $KNN\_D2(j)$ in which $i$ is found, where D2 is the Manhattan distance between full min-max normalized node attribute vectors.  |
| Structure+Attribute (pairwise) | KNN (out) D3 | The fraction of out- neighbor sets of $KNN\_D3(i)$ in which $j$ is found, where D3 is the Manhattan distance between the binary part of node attribute vectors.      |
| Structure+Attribute (pairwise) | KNN (in) D3  | The fraction of in- neighbor sets of $KNN\_D3(j)$ in which $i$ is found, where D3 is the Manhattan distance between the binary part of node attribute vectors.       |
| Structure+Attribute (pairwise) | KNN (out) D4 | The fraction of out- neighbor sets of $KNN\_D4(i)$ in which $j$ is found, where D4 is the Jaccard distance between the binary part of node attribute vectors.        |
| Structure+Attribute (pairwise) | KNN (in) D4  | The fraction of in- neighbor sets of $KNN\_D4(j)$ in which $i$ is found, where D4 is the Jaccard distance between the binary part of node attribute vectors.         |
| Structure (pairwise)           | KNN (out) D5 | The fraction of out- neighbor sets of $KNN\_D5(i)$ in which $j$ is found, where D5 is the Jaccard similarity based on in-neighbor sets.                              |
| Structure (pairwise)           | KNN (in) D5  | The fraction of in- neighbor sets of $KNN\_D5(j)$ in which $i$ is found, where D5 is the Jaccard similarity based on in-neighbor sets.                               |
| Structure (pairwise)           | KNN (out) D6 | The fraction of out- neighbor sets of $KNN\_D6(i)$ in which $j$ is found, where D6 is the Jaccard similarity based on out-neighbor sets.                             |
| Structure (pairwise)           | KNN (in) D6  | The fraction of in- neighbor sets of $KNN\_D6(j)$ in which $i$ is found, where D6 is the Jaccard similarity based on out-neighbor sets.                              |

**Supplementary Table 4:** Food web details. Names, ecosystem types, and references for the 290 food webs in the GATEWAY database [3, 4]; (version 3, accessible at [doi.org/10.25829/individ.283-3-756](https://doi.org/10.25829/individ.283-3-756)).

| Food web name                   | Ecosystem type          | Ref  |
|---------------------------------|-------------------------|------|
| Grand Caricaie marsh Clmown1    | terrestrial aboveground | [26] |
| Grand Caricaie marsh Clmown2    | terrestrial aboveground | [26] |
| Grand Caricaie marsh ClControl1 | terrestrial aboveground | [26] |
| Grand Caricaie marsh ClControl2 | terrestrial aboveground | [26] |
| Grand Caricaie marsh Scmown1    | terrestrial aboveground | [26] |
| Grand Caricaie marsh Scmown2    | terrestrial aboveground | [26] |

|                                 |                         |          |
|---------------------------------|-------------------------|----------|
| Grand Caricaie marsh ScControl1 | terrestrial aboveground | [26]     |
| Grand Caricaie marsh ScControl2 | terrestrial aboveground | [26]     |
| Ythan Estuary                   | marine                  | [27]     |
| AEW01                           | terrestrial belowground | [28]     |
| AEW02                           | terrestrial belowground | [28]     |
| AEW03                           | terrestrial belowground | [28]     |
| AEW04                           | terrestrial belowground | [28]     |
| AEW05                           | terrestrial belowground | [28]     |
| AEW06                           | terrestrial belowground | [28]     |
| AEW07                           | terrestrial belowground | [28]     |
| AEW08                           | terrestrial belowground | [28]     |
| AEW09                           | terrestrial belowground | [28]     |
| AEW11                           | terrestrial belowground | [28]     |
| AEW17                           | terrestrial belowground | [28]     |
| AEW18                           | terrestrial belowground | [28]     |
| AEW25                           | terrestrial belowground | [28]     |
| AEW27                           | terrestrial belowground | [28]     |
| AEW30                           | terrestrial belowground | [28]     |
| AEW49                           | terrestrial belowground | [28]     |
| HEW01                           | terrestrial belowground | [28]     |
| HEW02                           | terrestrial belowground | [28]     |
| HEW03                           | terrestrial belowground | [28]     |
| HEW04                           | terrestrial belowground | [28]     |
| HEW05                           | terrestrial belowground | [28]     |
| HEW06                           | terrestrial belowground | [28]     |
| HEW10                           | terrestrial belowground | [28]     |
| HEW11                           | terrestrial belowground | [28]     |
| HEW12                           | terrestrial belowground | [28]     |
| HEW13                           | terrestrial belowground | [28]     |
| HEW16                           | terrestrial belowground | [28]     |
| HEW17                           | terrestrial belowground | [28]     |
| HEW21                           | terrestrial belowground | [28]     |
| HEW22                           | terrestrial belowground | [28]     |
| HEW36                           | terrestrial belowground | [28]     |
| HEW47                           | terrestrial belowground | [28]     |
| SEW01                           | terrestrial belowground | [28]     |
| SEW02                           | terrestrial belowground | [28]     |
| SEW03                           | terrestrial belowground | [28]     |
| SEW04                           | terrestrial belowground | [28]     |
| SEW05                           | terrestrial belowground | [28]     |
| SEW06                           | terrestrial belowground | [28]     |
| SEW07                           | terrestrial belowground | [28]     |
| SEW08                           | terrestrial belowground | [28]     |
| SEW09                           | terrestrial belowground | [28]     |
| SEW18                           | terrestrial belowground | [28]     |
| SEW35                           | terrestrial belowground | [28]     |
| SEW36                           | terrestrial belowground | [28]     |
| SEW37                           | terrestrial belowground | [28]     |
| SEW41                           | terrestrial belowground | [28]     |
| SEW43                           | terrestrial belowground | [28]     |
| SEW48                           | terrestrial belowground | [28]     |
| Kongsfjorden                    | marine                  | [29]     |
| Alford lake                     | lakes                   | [30, 31] |
| Balsam lake                     | lakes                   | [30, 31] |
| Beaver lake                     | lakes                   | [30, 31] |

|                               |        |                      |
|-------------------------------|--------|----------------------|
| Big hope lake                 | lakes  | [30, 31]             |
| Brandy lake                   | lakes  | [30, 31]             |
| Bridge brook lake             | lakes  | [30, 31]             |
| Brook trout lake              | lakes  | [30, 31]             |
| Buck pond                     | lakes  | [30, 31]             |
| Burntbridge lake              | lakes  | [30, 31]             |
| Cascade lake                  | lakes  | [30, 31]             |
| Chub lake                     | lakes  | [30, 31]             |
| Chub pond                     | lakes  | [30, 31]             |
| Connera lake                  | lakes  | [30, 31]             |
| Constable lake                | lakes  | [30, 31]             |
| Deep lake                     | lakes  | [30, 31]             |
| Emerald lake                  | lakes  | [30, 31]             |
| Falls lake                    | lakes  | [30, 31]             |
| Fawn lake                     | lakes  | [30, 31]             |
| Federation lake               | lakes  | [30, 31]             |
| Goose lake                    | lakes  | [30, 31]             |
| Grass lake                    | lakes  | [30, 31]             |
| Gull lake                     | lakes  | [30, 31]             |
| Gull lake north               | lakes  | [30, 31]             |
| Helldiver pond                | lakes  | [30, 31]             |
| High pond                     | lakes  | [30, 31]             |
| Hoel lake                     | lakes  | [30, 31]             |
| Horseshoe lake                | lakes  | [30, 31]             |
| Indian lake                   | lakes  | [30, 31]             |
| Little rainbow                | lakes  | [30, 31]             |
| Long lake                     | lakes  | [30, 31]             |
| Loon lake                     | lakes  | [30, 31]             |
| Lost lake                     | lakes  | [30, 31]             |
| Lost lake east                | lakes  | [30, 31]             |
| Lower sister lake             | lakes  | [30, 31]             |
| Oswego lake                   | lakes  | [30, 31]             |
| Owl lake                      | lakes  | [30, 31]             |
| Rat lake                      | lakes  | [30, 31]             |
| Razorback lake                | lakes  | [30, 31]             |
| Rock lake                     | lakes  | [30, 31]             |
| Russian lake                  | lakes  | [30, 31]             |
| Safford lake                  | lakes  | [30, 31]             |
| Sand lake                     | lakes  | [30, 31]             |
| South lake                    | lakes  | [30, 31]             |
| Squaw lake                    | lakes  | [30, 31]             |
| Stink lake                    | lakes  | [30, 31]             |
| Twelfth tee lake              | lakes  | [30, 31]             |
| Twin lake east                | lakes  | [30, 31]             |
| Twin lake west                | lakes  | [30, 31]             |
| Whipple lake                  | lakes  | [30, 31]             |
| Wolf lake                     | lakes  | [30, 31]             |
| Weddell Sea                   | marine | [32]                 |
| Tuesday Lake 1984             | lakes  | [33]                 |
| Tuesday Lake 1986             | lakes  | [33]                 |
| Chilean Intertidal Curaumilla | marine | [12]                 |
| Chilean Intertidal El Quisco  | marine | [12]                 |
| Chilean Intertidal Las Cruces | marine | [12]                 |
| Chilean Intertidal Los Molles | marine | [12]                 |
| Chesapeake Bay                | marine | (Kroll, unpublished) |

|                                |                         |          |
|--------------------------------|-------------------------|----------|
| Carpinteria                    | terrestrial aboveground | [11]     |
| Scottish lake                  | lakes                   | [34]     |
| Afon Hafren 2005               | streams                 | [34]     |
| Allt a Mharcaidh               | streams                 | [34]     |
| Broadstone                     | streams                 | [34]     |
| Dargall Lane                   | streams                 | [34]     |
| Duddon Pike Beck               | streams                 | [34]     |
| Hardknott Gill                 | streams                 | [34]     |
| Mill Stream                    | streams                 | [34]     |
| Mosendale Beck                 | streams                 | [34]     |
| Old Lodge                      | streams                 | [34]     |
| Alert                          | terrestrial aboveground | [35]     |
| Bylot                          | terrestrial aboveground | [35]     |
| Herschel                       | terrestrial aboveground | [35]     |
| Nenetsky                       | terrestrial aboveground | [35]     |
| Svalbard                       | terrestrial aboveground | [35]     |
| Yamal                          | terrestrial aboveground | [35]     |
| Zackenberg                     | terrestrial aboveground | [35]     |
| Gearagh                        | terrestrial aboveground | [36]     |
| Dutch Detrital food web PlotA  | terrestrial aboveground | [37, 38] |
| Dutch Detrital food web PlotB  | terrestrial aboveground | [37, 38] |
| Dutch Detrital food web PlotC  | terrestrial aboveground | [37, 38] |
| Bure Stream                    | streams                 | [39, 40] |
| Loddon Stream                  | streams                 | [39, 40] |
| Lyde Stream                    | streams                 | [39, 40] |
| Test Stream                    | streams                 | [39, 40] |
| Wensum Stream                  | streams                 | [39, 40] |
| Lake Malawi                    | lakes                   | [41]     |
| Iceland stream IS7 April 2009  | streams                 | [42]     |
| Iceland stream IS7 August 2008 | streams                 | [42]     |
| Iceland stream IS8 April 2008  | streams                 | [42]     |
| Iceland stream IS8 August 2008 | streams                 | [42]     |
| Caribbean Reef                 | marine                  | [13]     |
| FloridaIslandE1                | terrestrial aboveground | [43, 44] |
| FloridaIslandE2                | terrestrial aboveground | [43, 44] |
| FloridaIslandE3                | terrestrial aboveground | [43, 44] |
| FloridaIslandE7                | terrestrial aboveground | [43, 44] |
| FloridaIslandE9                | terrestrial aboveground | [43, 44] |
| Lough Hyne                     | marine                  | [45]     |
| Blackrock Stream               | streams                 | [46]     |
| Broad Stream                   | streams                 | [46]     |
| Canton Creek                   | streams                 | [46]     |
| Dempsters Stream               | streams                 | [46]     |
| German Creek                   | streams                 | [46]     |
| Healy Creek                    | streams                 | [46]     |
| Kye Burn                       | streams                 | [46]     |
| Little Kye Burn                | streams                 | [46]     |
| Stony Stream                   | streams                 | [46]     |
| Sutton Stream                  | streams                 | [46]     |
| AP1                            | marine                  | [47]     |
| AP2                            | marine                  | [47]     |
| AP3                            | marine                  | [47]     |
| AP4                            | marine                  | [47]     |
| BP1                            | marine                  | [47]     |
| BP2                            | marine                  | [47]     |

|       |        |    |
|-------|--------|----|
| BP3   | marine | 47 |
| CGP1  | marine | 47 |
| CGP2  | marine | 47 |
| CGP3  | marine | 47 |
| CR1P1 | marine | 47 |
| CR1P2 | marine | 47 |
| CR1P3 | marine | 47 |
| CR1P4 | marine | 47 |
| CR2P1 | marine | 47 |
| CR2P2 | marine | 47 |
| CR2P3 | marine | 47 |
| CR2P4 | marine | 47 |
| F1P1  | marine | 47 |
| F1P2  | marine | 47 |
| F1P3  | marine | 47 |
| F1P4  | marine | 47 |
| F2P1  | marine | 47 |
| F2P2  | marine | 47 |
| F2P3  | marine | 47 |
| F2P4  | marine | 47 |
| FP1   | marine | 47 |
| FXAP1 | marine | 47 |
| FXAP2 | marine | 47 |
| FXAP3 | marine | 47 |
| FXAP4 | marine | 47 |
| FXBP1 | marine | 47 |
| FXBP2 | marine | 47 |
| FXBP3 | marine | 47 |
| FXBP4 | marine | 47 |
| GJAP1 | marine | 47 |
| GJAP2 | marine | 47 |
| GJAP3 | marine | 47 |
| GJAP4 | marine | 47 |
| GJBP1 | marine | 47 |
| GJBP2 | marine | 47 |
| GJBP3 | marine | 47 |
| GJBP4 | marine | 47 |
| L1P1  | marine | 47 |
| L1P2  | marine | 47 |
| L1P3  | marine | 47 |
| L1P4  | marine | 47 |
| L2P1  | marine | 47 |
| L2P2  | marine | 47 |
| L2P3  | marine | 47 |
| L2P4  | marine | 47 |
| L3P1  | marine | 47 |
| L3P2  | marine | 47 |
| L3P3  | marine | 47 |
| L3P4  | marine | 47 |
| L4P1  | marine | 47 |
| L4P2  | marine | 47 |
| L4P3  | marine | 47 |
| L4P4  | marine | 47 |
| MBP1  | marine | 47 |
| MBP2  | marine | 47 |

|        |        |    |
|--------|--------|----|
| MBP3   | marine | 47 |
| MBP4   | marine | 47 |
| PC1P1  | marine | 47 |
| PC1P2  | marine | 47 |
| PC1P3  | marine | 47 |
| PC1P4  | marine | 47 |
| PC2P1  | marine | 47 |
| PC2P2  | marine | 47 |
| PC2P3  | marine | 47 |
| PC2P4  | marine | 47 |
| PC2P5  | marine | 47 |
| PGSBP1 | marine | 47 |
| PGSBP2 | marine | 47 |
| PGUBP1 | marine | 47 |
| PGUBP2 | marine | 47 |
| PGUBP3 | marine | 47 |
| PGUBP4 | marine | 47 |
| PP1I1  | marine | 47 |
| PP1I2  | marine | 47 |
| PP1I3  | marine | 47 |
| PP1I4  | marine | 47 |
| PP2I1  | marine | 47 |
| PP2I2  | marine | 47 |
| PP2I3  | marine | 47 |
| PP2I4  | marine | 47 |
| PP2M1  | marine | 47 |
| PP2M2  | marine | 47 |
| PP2M3  | marine | 47 |
| PP2M4  | marine | 47 |
| RMP1   | marine | 47 |
| RMP2   | marine | 47 |
| RMP3   | marine | 47 |
| RMP4   | marine | 47 |
| RMP5   | marine | 47 |
| RV1P1  | marine | 47 |
| RV1P2  | marine | 47 |
| RV1P3  | marine | 47 |
| RV1P4  | marine | 47 |
| RV2P1  | marine | 47 |
| RV2P2  | marine | 47 |
| RV2P3  | marine | 47 |
| RV2P4  | marine | 47 |
| SF1I1  | marine | 47 |
| SF1I2  | marine | 47 |
| SF1I3  | marine | 47 |
| SF1I4  | marine | 47 |
| SF1M1  | marine | 47 |
| SF1M2  | marine | 47 |
| SF1M3  | marine | 47 |
| SF1M4  | marine | 47 |
| SF2I1  | marine | 47 |
| SF2I2  | marine | 47 |
| SF2I3  | marine | 47 |
| SF2I4  | marine | 47 |
| SF2M1  | marine | 47 |

|               |        |      |
|---------------|--------|------|
| SF2M2         | marine | [47] |
| SF2M3         | marine | [47] |
| SF2M4         | marine | [47] |
| SP1           | marine | [47] |
| WP1           | marine | [47] |
| WP2           | marine | [47] |
| WP3           | marine | [47] |
| WP4           | marine | [47] |
| Skipwith Pond | lakes  | [48] |

**Supplementary Table 5.** Filling in NA metabolic type values in the food web database. Node names are written [taxonomy]\_[lifestage]

| Food web                                                          | Node name                      | Filled metabolic type value |
|-------------------------------------------------------------------|--------------------------------|-----------------------------|
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown1 | Trichoptera_larvae             | invertebrate                |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown2 | Trichoptera_larvae             | invertebrate                |
| Chilean Intertidal Curaumilla                                     | plankton_NA                    | primary producer            |
| Chilean Intertidal El Quisco                                      | plankton_NA                    | primary producer            |
| Chilean Intertidal Las Cruces                                     | plankton_NA                    | primary producer            |
| Chilean Intertidal Los Molles                                     | plankton_NA                    | primary producer            |
| Chesapeake Bay                                                    | Carcinus sp._adults            | invertebrate                |
| Chesapeake Bay                                                    | Portunus spinimanus_adults     | invertebrate                |
| Chesapeake Bay                                                    | Leiostomus xanthurus_juveniles | ectotherm vertebrate        |
| Chesapeake Bay                                                    | Plecoptera_larvae              | invertebrate                |
| Chesapeake Bay                                                    | Trichoptera_larvae             | invertebrate                |
| Skipwith Pond                                                     | detritus_NA                    | detritus                    |

**Supplementary Table 6:** Filling in NA movement type values in the food web database. Node names are written [taxonomy]\_[lifestage].

| Food web                                                                 | Node name                | Filled movement type value |
|--------------------------------------------------------------------------|--------------------------|----------------------------|
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown1        | Trichoptera_larvae       | other                      |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown1        | Limoniidae_larvae        | other                      |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown1        | Pipinculus sp_NA         | flying                     |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown2        | Trichoptera_larvae       | other                      |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown2        | Limoniidae_larvae        | other                      |
| Grand Caricaie marsh dominated by Cladietum marisci, mown Clmown2        | Pipinculus sp_NA         | flying                     |
| Grand Caricaie marsh dominated by Cladietum marisci, not mown ClControl1 | Phlaeothripidae phyto_NA | flying                     |
| Grand Caricaie marsh dominated by Cladietum marisci, not mown ClControl1 | Phlaeothripidae pred_NA  | flying                     |
| Grand Caricaie marsh dominated by Cladietum marisci, not mown ClControl2 | Phlaeothripidae_NA       | flying                     |
| Grand Caricaie marsh dominated by Schoenus nigricans, mown Scmown1       | Insecta_larvae           | other                      |

|                                                                           |                                |          |
|---------------------------------------------------------------------------|--------------------------------|----------|
| Grand Caricaie marsh dominated by Schoenus nigricans, mown Scmown1        | Limoniidae_larvae              | other    |
| Grand Caricaie marsh dominated by Schoenus nigricans, mown Scmown2        | Limoniidae_larvae              | other    |
| Grand Caricaie marsh dominated by Schoenus nigricans, mown Scmown2        | Pipinculus sp_NA               | flying   |
| Grand Caricaie marsh dominated by Schoenus nigricans, not mown ScControl1 | Limoniidae_larvae              | other    |
| Grand Caricaie marsh dominated by Schoenus nigricans, not mown ScControl2 | Limoniidae_larvae              | other    |
| Chesapeake Bay                                                            | Carcinus sp._adults            | walking  |
| Chesapeake Bay                                                            | Portunus spinimanus_adults     | walking  |
| Chesapeake Bay                                                            | Leiostomus xanthurus_juveniles | swimming |
| Chesapeake Bay                                                            | Plecoptera_larvae              | other    |
| Chesapeake Bay                                                            | Trichoptera_larvae             | other    |
| Carpinteria                                                               | Epipellic flora_adults         | sessile  |
| Carpinteria                                                               | Monanthocloe littoralis_adults | sessile  |
| Alert                                                                     | arthropods_NA                  | other    |
| Bylot                                                                     | arthropods_NA                  | other    |
| Herschel                                                                  | arthropods_NA                  | other    |
| Nenetsky                                                                  | arthropods_NA                  | other    |
| Svalbard                                                                  | arthropods_NA                  | other    |
| Yamal                                                                     | arthropods_NA                  | other    |
| Zackenbergl                                                               | arthropods_NA                  | other    |
| Gearagh                                                                   | Bacteria_NA                    | other    |
| Iceland stream IS7 April 2009                                             | Bacteria_NA                    | other    |
| Iceland stream IS7 August 2008                                            | Bacteria_NA                    | other    |
| Iceland stream IS8 April 2009                                             | Bacteria_NA                    | other    |
| Iceland stream IS8 August 2008                                            | Bacteria_NA                    | other    |
| FloridaIslandE1                                                           | Lasioseius sp._NA              | walking  |
| FloridaIslandE1                                                           | Poecilopa rhizophorae_NA       | walking  |
| FloridaIslandE1                                                           | Lapisma sp._NA                 | walking  |
| FloridaIslandE1                                                           | Orgygia detrita_NA             | flying   |
| FloridaIslandE3                                                           | Lasioseius sp._NA              | walking  |
| FloridaIslandE3                                                           | Poecilopa rhizophorae_NA       | walking  |
| FloridaIslandE3                                                           | Lapisma sp._NA                 | walking  |
| FloridaIslandE3                                                           | Orgygia detrita_NA             | flying   |
| Kye Burn                                                                  | Meiofauna_NA                   | floating |

**Supplementary Table 7.** Details of pairwise two-sided t-tests comparing performance between different versions of the model on the food webs whose node sets change with lifestage aggregation. P values are false discovery rate adjusted (Benjamini-Hochberg method).

| Comparison              | Lifestage Aggregation | Performance Metric | T-statistic | Confidence Interval | Cohen d avg | Degrees of freedom | p value |
|-------------------------|-----------------------|--------------------|-------------|---------------------|-------------|--------------------|---------|
| Structure vs. Attribute | Aggregated            | ROC-AUC            | 11.467      | (0.089, 0.128)      | 2.769       | 24                 | 0.0     |
| Full vs. Attribute      | Aggregated            | ROC-AUC            | 11.263      | (0.091, 0.131)      | 2.862       | 24                 | 0.0     |
| Full vs. Structure      | Aggregated            | ROC-AUC            | 4.07        | (0.001, 0.004)      | 0.159       | 24                 | 0.0     |
| Structure vs. Attribute | Aggregated            | PR-AUC             | 8.896       | (0.271, 0.435)      | 1.82        | 24                 | 0.0     |
| Full vs. Attribute      | Aggregated            | PR-AUC             | 8.809       | (0.292, 0.471)      | 1.85        | 24                 | 0.0     |
| Full vs. Structure      | Aggregated            | PR-AUC             | 3.675       | (0.013, 0.046)      | 0.113       | 24                 | 0.001   |
| Structure vs. Attribute | Disaggregated         | ROC-AUC            | 9.979       | (0.09, 0.137)       | 2.656       | 24                 | 0.0     |
| Full vs. Attribute      | Disaggregated         | ROC-AUC            | 10.138      | (0.092, 0.14)       | 2.713       | 24                 | 0.0     |
| Full vs. Structure      | Disaggregated         | ROC-AUC            | 5.346       | (0.001, 0.003)      | 0.131       | 24                 | 0.0     |
| Structure vs. Attribute | Disaggregated         | PR-AUC             | 9.073       | (0.281, 0.446)      | 1.888       | 24                 | 0.0     |
| Full vs. Attribute      | Disaggregated         | PR-AUC             | 9.095       | (0.298, 0.472)      | 1.914       | 24                 | 0.0     |
| Full vs. Structure      | Disaggregated         | PR-AUC             | 4.695       | (0.012, 0.031)      | 0.083       | 24                 | 0.0     |

**Supplementary Table 8.** Details of pairwise two-sided t-tests comparing performance between different versions of the model on the entire food web database. P values are false discovery rate adjusted (Benjamini-Hochberg method).

| Comparison              | Performance Metric | T-statistic | Confidence Interval | Cohen d avg | Degrees of freedom | p value |
|-------------------------|--------------------|-------------|---------------------|-------------|--------------------|---------|
| Structure vs. Attribute | ROC-AUC            | 23.658      | (0.055, 0.065)      | 0.965       | 289                | 0.0     |
| Full vs. Attribute      | ROC-AUC            | 30.477      | (0.066, 0.075)      | 1.215       | 289                | 0.0     |
| Full vs. Structure      | ROC-AUC            | 11.584      | (0.009, 0.012)      | 0.172       | 289                | 0.0     |
| Structure vs. Attribute | PR-AUC             | 19.537      | (0.24, 0.293)       | 1.302       | 289                | 0.0     |
| Full vs. Attribute      | PR-AUC             | 25.209      | (0.303, 0.355)      | 1.624       | 289                | 0.0     |
| Full vs. Structure      | PR-AUC             | 22.194      | (0.057, 0.068)      | 0.308       | 289                | 0.0     |

**Supplementary Table 9:** Details of pairwise two-sided t-tests comparing performance between different versions of the model split by ecosystem type. P values are false discovery rate adjusted (Benjamini-Hochberg method).

| Comparison              | Performance Metric | Ecosystem Type | T-statistic | Confidence Interval | Cohen d avg | Degrees of freedom | p value |
|-------------------------|--------------------|----------------|-------------|---------------------|-------------|--------------------|---------|
| Structure vs. Attribute | ROC-AUC            | Lakes          | 21.251      | (0.073, 0.088)      | 1.848       | 54                 | 0.0     |
| Full vs. Attribute      | ROC-AUC            | Lakes          | 24.662      | (0.083, 0.098)      | 2.234       | 54                 | 0.0     |
| Full vs. Structure      | ROC-AUC            | Lakes          | 4.187       | (0.005, 0.015)      | 0.26        | 54                 | 0.0     |

|                         |         |                         |        |                |       |     |       |
|-------------------------|---------|-------------------------|--------|----------------|-------|-----|-------|
| Structure vs. Attribute | PR-AUC  | Lakes                   | 16.218 | (0.32, 0.41)   | 2.746 | 54  | 0.0   |
| Full vs. Attribute      | PR-AUC  | Lakes                   | 23.26  | (0.409, 0.486) | 3.686 | 54  | 0.0   |
| Full vs. Structure      | PR-AUC  | Lakes                   | 7.893  | (0.062, 0.104) | 0.532 | 54  | 0.0   |
| Structure vs. Attribute | ROC-AUC | Marine                  | 12.405 | (0.026, 0.036) | 0.44  | 133 | 0.0   |
| Full vs. Attribute      | ROC-AUC | Marine                  | 20.859 | (0.04, 0.049)  | 0.682 | 133 | 0.0   |
| Full vs. Structure      | ROC-AUC | Marine                  | 9.624  | (0.011, 0.017) | 0.205 | 133 | 0.0   |
| Structure vs. Attribute | PR-AUC  | Marine                  | 7.256  | (0.085, 0.148) | 0.57  | 133 | 0.0   |
| Full vs. Attribute      | PR-AUC  | Marine                  | 12.632 | (0.154, 0.211) | 0.91  | 133 | 0.0   |
| Full vs. Structure      | PR-AUC  | Marine                  | 19.658 | (0.059, 0.072) | 0.338 | 133 | 0.0   |
| Structure vs. Attribute | ROC-AUC | Streams                 | 15.512 | (0.053, 0.07)  | 1.765 | 27  | 0.0   |
| Full vs. Attribute      | ROC-AUC | Streams                 | 13.33  | (0.056, 0.077) | 1.999 | 27  | 0.0   |
| Full vs. Structure      | ROC-AUC | Streams                 | 2.847  | (0.001, 0.009) | 0.21  | 27  | 0.008 |
| Structure vs. Attribute | PR-AUC  | Streams                 | 12.133 | (0.243, 0.342) | 1.161 | 27  | 0.0   |
| Full vs. Attribute      | PR-AUC  | Streams                 | 12.261 | (0.265, 0.372) | 1.229 | 27  | 0.0   |
| Full vs. Structure      | PR-AUC  | Streams                 | 4.271  | (0.014, 0.039) | 0.091 | 27  | 0.0   |
| Structure vs. Attribute | ROC-AUC | Terrestrial Aboveground | 8.653  | (0.095, 0.156) | 1.78  | 20  | 0.0   |
| Full vs. Attribute      | ROC-AUC | Terrestrial Aboveground | 11.879 | (0.113, 0.162) | 2.216 | 20  | 0.0   |
| Full vs. Structure      | ROC-AUC | Terrestrial Aboveground | 3.195  | (0.004, 0.02)  | 0.149 | 20  | 0.005 |
| Structure vs. Attribute | PR-AUC  | Terrestrial Aboveground | 4.907  | (0.142, 0.353) | 1.648 | 20  | 0.0   |
| Full vs. Attribute      | PR-AUC  | Terrestrial Aboveground | 6.345  | (0.195, 0.386) | 2.107 | 20  | 0.0   |
| Full vs. Structure      | PR-AUC  | Terrestrial Aboveground | 7.116  | (0.03, 0.056)  | 0.222 | 20  | 0.0   |
| Structure vs. Attribute | ROC-AUC | Terrestrial Belowground | 30.093 | (0.082, 0.094) | 4.162 | 51  | 0.0   |
| Full vs. Attribute      | ROC-AUC | Terrestrial Belowground | 32.035 | (0.087, 0.098) | 4.364 | 51  | 0.0   |
| Full vs. Structure      | ROC-AUC | Terrestrial Belowground | 14.94  | (0.003, 0.004) | 0.419 | 51  | 0.0   |
| Structure vs. Attribute | PR-AUC  | Terrestrial Belowground | 41.972 | (0.517, 0.569) | 6.089 | 51  | 0.0   |
| Full vs. Attribute      | PR-AUC  | Terrestrial Belowground | 45.395 | (0.577, 0.63)  | 6.608 | 51  | 0.0   |
| Full vs. Structure      | PR-AUC  | Terrestrial Belowground | 19.258 | (0.054, 0.066) | 0.548 | 51  | 0.0   |

**Supplementary Table 10.** Food web metadata and data processing metrics used to understand predictive performance across the food web database.

| Property                                                           | Type                  | Definition                                                                                                   |
|--------------------------------------------------------------------|-----------------------|--------------------------------------------------------------------------------------------------------------|
| Ecosystem type                                                     | Categorical           | Ecosystem type in the database.                                                                              |
| Missing mass fraction                                              | Numeric (range [0-1]) | Fraction of nodes in the network originally missing mass trait data.                                         |
| Higher taxonomic level proportion                                  | Numeric (range [0-1]) | Fraction of nodes in the network at a taxonomic level higher than species (genus or family+).                |
| Unclassified taxonomic level proportion                            | Numeric (range [0-1]) | Fraction of nodes in the network with unclassified taxonomic level (originally NA or unranked).              |
| Parasitic interaction removal (labeled as parasitic or parasitoid) | Numeric (range [0-1]) | The number of parasitic interactions removed divided by the final number of interactions in the network.     |
| Cannibalistic interaction removal                                  | Numeric (range [0-1]) | The number of cannibalistic interactions removed divided by the final number of interactions in the network. |
| Number node attributes                                             | Numeric               | The number of node attributes for this food web.                                                             |
| Minimum log mass                                                   | Numeric               | The minimum log mass value in the food web.                                                                  |
| Maximum log mass                                                   | Numeric               | The maximum log mass value in the food web.                                                                  |
| Mean log mas                                                       | Numeric               | The mean log mass value across nodes in the food web.                                                        |
| Median log mass                                                    | Numeric               | The median log mass value across nodes in the food web.                                                      |
| Mass skewness                                                      | Numeric               | Pearsons 2nd coefficient of skewness of the raw mass values.                                                 |

**Supplementary Table 11.** Global network topology metrics used to understand predictive performance across the food web database. Topological properties were calculated using functions from the NetworkX library version 2.5.1 [22].

| Property               | Definition                                                                                                                                                                                                                                                                      |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Log(nodes)             | Where $n$ is the number of nodes, $\log_{10}(n)$ .                                                                                                                                                                                                                              |
| Average degree         | Where $m$ is the number of directed edges, $m/n$ .                                                                                                                                                                                                                              |
| Connectance            | Number of directed edges divided by number of possible simple directed edges. Note that the denominator is adjusted such that cannibalistic interactions are not allowed.<br>$m/(n * (n - 1))$                                                                                  |
| Clustering coefficient | The fraction of possible triangles present in the network. This calculation is performed on an undirected version of the network, thus ignoring directed structure of triangles.                                                                                                |
| Modularity             | With communities found via greedy modularity maximization, this metric measures to what extent these communities have more connections within the same community as opposed to between communities. This calculation is performed on an undirected version of the network. [16] |

**Supplementary Table 12.** Assortativity coefficients used to understand predictive performance across the food web database. Calculated using the NetworkX library version 2.5.1 [22], trait assortativity metrics are based on [17] and in- and out- degree assortativity coefficients are based on [49]. All assortativity coefficients take the directed structure of the network into account. In the rare cases of multiple movement types for a node in a network, one was randomly chosen.

| Property                          | Definition                                                                                                                  |
|-----------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| Log body mass assortativity       | Assortativity coefficient based on the log mass attribute split into 4 bins of equal size.                                  |
| Metabolic type assortativity      | Assortativity coefficient based on metabolic type categorical attribute.                                                    |
| Movement type assortativity       | Assortativity coefficient based on movement type categorical attribute.                                                     |
| Degree (in to out) assortativity  | Tendency of nodes with high in-degree (generalist consumers) to link to nodes with high out-degree (generalist resources).  |
| Degree (out to in) assortativity  | Tendency of nodes with high out-degree (generalist resources) to link to nodes with high in-degree (generalist consumers).  |
| Degree (out to out) assortativity | Tendency of nodes with high out-degree (generalist resources) to link to nodes with high out-degree (generalist resources). |
| Degree (in to in) assortativity   | Tendency of nodes with high in-degree (generalist consumers) to link to nodes with high in-degree (generalist consumers).   |

## SUPPLEMENTARY REFERENCES

- 
- [1] 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). URL [http://link.springer.com/10.1007/3-540-44886-1\\_25](http://link.springer.com/10.1007/3-540-44886-1_25). Series Title: Lecture Notes in Computer Science.
  - [2] Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airolidi, E. M. & Clauset, A. Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences* **117**, 23393–23400 (2020). URL <https://pnas.org/doi/full/10.1073/pnas.1914950117>.
  - [3] Brose, U. GlobAL daTabasE of traits and food Web Architecture (GATEWAY) version 1.0 (2018). URL <https://idata.idiv.de/ddm/Data/ShowData/283?version=3>.
  - [4] Brose, U. *et al.* Predator traits determine food-web architecture across ecosystems. *Nature Ecology & Evolution* **3**, 919–927 (2019). URL <http://www.nature.com/articles/s41559-019-0899-x>.
  - [5] Cirtwill, A. R. & Stouffer, D. B. Concomitant predation on parasites is highly variable but constrains the ways in which parasites contribute to food web structure. *Journal of Animal Ecology* **84**, 734–744 (2015). URL <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2656.12323>.
  - [6] Dunne, J. A. *et al.* Parasites Affect Food Web Structure Primarily through Increased Diversity and Complexity. *PLoS Biology* **11**, e1001579 (2013). URL <https://dx.plos.org/10.1371/journal.pbio.1001579>.
  - [7] Lafferty, K. D. & Kuris, A. M. Parasites reduce food web robustness because they are sensitive to secondary extinction as illustrated by an invasive estuarine snail. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 1659–1663 (2009). URL <https://royalsocietypublishing.org/doi/10.1098/rstb.2008.0220>.
  - [8] 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.
  - [9] Michalska-Smith, M. J., Sander, E. L., Pascual, M. & Allesina, S. Understanding the role of parasites in food webs using the group model. *Journal of Animal Ecology* **87**, 790–800 (2018). URL <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2656.12782>.
  - [10] Morton, D. N. *et al.* A food web including parasites for kelp forests of the Santa Barbara Channel, California. *Scientific Data* **8**, 99 (2021). URL <https://www.nature.com/articles/s41597-021-00880-4>.
  - [11] Lafferty, K. D., Dobson, A. P. & Kuris, A. M. Parasites dominate food web links. *Proceedings of the National Academy of Sciences* **103**, 11211–11216 (2006). URL <https://pnas.org/doi/full/10.1073/pnas.0604755103>.
  - [12] Kéfi, S. *et al.* Network structure beyond food webs: mapping non-trophic and trophic interactions on Chilean rocky shores. *Ecology* **96**, 291–303 (2015). URL <https://esajournals.onlinelibrary.wiley.com/doi/10.1890/13-1424.1>.
  - [13] Opitz, S. Trophic interactions in caribbean coral reefs. Tech. Rep. 43, ICLARM (1996).
  - [14] 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). URL <https://nsojournals.onlinelibrary.wiley.com/doi/10.1111/ecog.07028>.
  - [15] Clauset, A. & Erwin, D. H. The Evolution and Distribution of Species Body Size. *Science* **321**, 399–401 (2008). URL <https://www.science.org/doi/10.1126/science.1157534>.

- [16] Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Physical Review E* **70**, 066111 (2004). URL <http://arxiv.org/abs/cond-mat/0408187>. ArXiv:cond-mat/0408187.
- [17] Newman, M. E. J. Mixing patterns in networks. *Physical Review E* **67**, 026126 (2003). URL <http://arxiv.org/abs/cond-mat/0209450>. ArXiv:cond-mat/0209450.
- [18] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
- [19] Leicht, E. A., Holme, P. & Newman, M. E. J. Vertex similarity in networks. *Physical Review E* **73**, 026120 (2006). URL <https://link.aps.org/doi/10.1103/PhysRevE.73.026120>.
- [20] Adamic, L. A. & Adar, E. Friends and neighbors on the Web. *Social Networks* **25**, 211–230 (2003). URL <https://linkinghub.elsevier.com/retrieve/pii/S0378873303000091>.
- [21] Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *The European Physical Journal B* **71**, 623–630 (2009). URL <http://link.springer.com/10.1140/epjb/e2009-00335-8>.
- [22] Hagberg, A., Swart, P. & Chult, D. Exploring network structure, dynamics, and function using NetworkX. Tech. Rep. No. LA-UR-08-05495; LA-UR-08-5495., Los Alamos National Lab (LANL), Los Alamos, NM (United States) (2008).
- [23] Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **58**, 1019–1031 (2007). URL <https://onlinelibrary.wiley.com/doi/10.1002/asi.20591>.
- [24] Cukierski, W., Hamner, B. & Yang, B. Graph-based features for supervised link prediction. In *The 2011 International Joint Conference on Neural Networks*, 1237–1244 (IEEE, San Jose, CA, USA, 2011). URL <http://ieeexplore.ieee.org/document/6033365/>.
- [25] Levine, S. Several measures of trophic structure applicable to complex food webs. *Journal of Theoretical Biology* **83**, 195–207 (1980). URL <https://linkinghub.elsevier.com/retrieve/pii/002251938090288X>.
- [26] Cattin Blandenier, M.-F. *Food web ecology: models and application to conservation*. PhD Thesis, Verlag nicht ermittelbar (2004).
- [27] Cohen, J. E., Schittler, D. N., Raffaelli, D. G. & Reuman, D. C. Food webs are more than the sum of their tritrophic parts. *Proceedings of the National Academy of Sciences* **106**, 22335–22340 (2009). URL <https://pnas.org/doi/full/10.1073/pnas.0910582106>.
- [28] Digel, C., Curtsdotter, A., Riede, J., Klarner, B. & Brose, U. Unravelling the complex structure of forest soil food webs: higher omnivory and more trophic levels. *Oikos* **123**, 1157–1172 (2014). URL <https://onlinelibrary.wiley.com/doi/10.1111/oik.00865>.
- [29] Eklöf, A. *et al.* The dimensionality of ecological networks. *Ecology Letters* **16**, 577–583 (2013). URL <https://onlinelibrary.wiley.com/doi/10.1111/ele.12081>.
- [30] Sutherland, J. Adirondack Biota Project. Tech. Rep. (1989).
- [31] Havens, K. Scale and Structure in Natural Food Webs. *Science* **257**, 1107–1109 (1992). URL <https://www.science.org/doi/10.1126/science.257.5073.1107>.
- [32] Jacob, U. *et al.* The Role of Body Size in Complex Food Webs. In *Advances in Ecological Research*, vol. 45, 181–223 (Elsevier, 2011). URL <https://linkinghub.elsevier.com/retrieve/pii/B9780123864758000058>.
- [33] Jonsson, T., Cohen, J. E. & Carpenter, S. R. Food webs, body size, and species abundance in ecological community description. *Advances in ecological research* **36**, 1–84 (2005).
- [34] Layer, K., Hildrew, A., Monteith, D. & Woodward, G. Long-term variation in the littoral food web of an acidified mountain lake. *Global Change Biology* **16**, 3133–3143 (2010). URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2010.02195.x>.
- [35] Legagneux, P. *et al.* Arctic ecosystem structure and functioning shaped by climate and herbivore body size. *Nature Climate Change* **4**, 379–383 (2014). URL <https://www.nature.com/articles/nclimate2168>.
- [36] McLaughlin, B., Jonsson, T. & Emmerson, M. C. Temporal Variability in Predator–Prey Relationships of a Forest Floor Food Web. In *Advances in Ecological Research*, vol. 42, 171–264 (Elsevier, 2010). URL <https://linkinghub.elsevier.com/retrieve/pii/B9780123813633000046>.
- [37] Mulder, C. & Elser, J. J. Soil acidity, ecological stoichiometry and allometric scaling in grassland food webs. *Global Change Biology* **15**, 2730–2738 (2009). URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2009.01899.x>.
- [38] Sechi, V., Brussaard, L., De Goede, R. G. M., Rutgers, M. & Mulder, C. Choice of Resolution by Functional Trait or Taxonomy Affects Allometric Scaling in Soil Food Webs. *The American Naturalist* **185**, 142–149 (2015). URL <https://www.journals.uchicago.edu/doi/10.1086/678962>.
- [39] Gray, C. *et al.* Joining the dots: An automated method for constructing food webs from compendia of published interactions. *Food Webs* **5**, 11–20 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S2352249615300045>.
- [40] Thompson, M. S. A. *et al.* Large woody debris “rewilding” rapidly restores biodiversity in riverine food webs. *Journal of Applied Ecology* **55**, 895–904 (2018). URL <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13013>.
- [41] Nsiku, E. *Changes in the fisheries of Lake Malawi, 1976-1996: Ecosystem-based analysis*. PhD Thesis, University of British Columbia (1999).
- [42] O’Gorman, E. J. *et al.* Impacts of Warming on the Structure and Functioning of Aquatic Communities. In *Advances in Ecological Research*, vol. 47, 81–176 (Elsevier, 2012). URL <https://linkinghub.elsevier.com/retrieve/pii/B9780123983152000028>.
- [43] Simberloff, D. S. & Wilson, E. O. Experimental Zoogeography of Islands: The Colonization of Empty Islands. *Ecology* **50**, 278–296 (1969). URL <https://esajournals.onlinelibrary.wiley.com/doi/10.2307/1934856>.

- [44] Piechnik, D., Lawler, S. P. & Martinez, N. D. Food-web assembly during a classic biogeographic study species trophic breadth corresponds to colonization order. *Oikos* **117**, 665–674 (2008).
- [45] Jacob, U. *et al.* Valuing Biodiversity and Ecosystem Services in a Complex Marine Ecosystem. In *Aquatic Functional Biodiversity*, 189–207 (Elsevier, 2015). URL <https://linkinghub.elsevier.com/retrieve/pii/B9780124170155000086>.
- [46] Townsend *et al.* Disturbance, resource supply, and food-web architecture in streams. *Ecology Letters* **1**, 200–209 (1998). URL <https://onlinelibrary.wiley.com/doi/10.1046/j.1461-0248.1998.00039.x>.
- [47] Mendonça, V. *et al.* What’s in a tide pool? Just as much food web network complexity as in large open ecosystems. *PLOS ONE* **13**, e0200066 (2018). URL <https://dx.plos.org/10.1371/journal.pone.0200066>.
- [48] Warren, P. H. Spatial and Temporal Variation in the Structure of a Freshwater Food Web. *Oikos* **55**, 299 (1989). URL <https://www.jstor.org/stable/3565588?origin=crossref>.
- [49] Foster, J. G., Foster, D. V., Grassberger, P. & Paczuski, M. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences* **107**, 10815–10820 (2010). URL <https://pnas.org/doi/full/10.1073/pnas.0912671107>.