

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

Corresponding Author: Dr Lucy Van Kleunen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This manuscript proposes a novel method to predict unobserved interactions in food webs. As such, it participates to the recently growing body of literature on this question (for food webs and for other types of networks). Consequently, one may perceive this effort as incremental. However, this contribution stands out on several aspects.

Firstly, while the choice of model stacking is by itself not a breakthrough (one may indeed expect that the combination of the results of different existing models will be superior to the results of any individual model), it is based here on a large number of simple predictors that were, to my knowledge and for most of them, not used before for this purpose. Also, the information used is not only based on the network itself (structure-only models), but also integrates external information on the nodes, with a sensible distinction between assortative and disassortative descriptors.

Secondly, the work is really extensive, by the use of a large number of predictors, by the careful test of the approach with synthetic data, and by the application of the method to a large dataset of 290 published food webs.

Thirdly, the analyses of the results are comprehensive: which predictor and which type of information perform best; how these results differ between the different types of networks (grouped according to the habitat type); how model performance correlates with food web properties.

Finally, the discussion of the results with regard to the existing literature is interesting and exhaustive, which is very useful for future work on this topic. There is however one topic that should be treated in this section. You mention that food webs sometimes come with additional complexity, among them edge weights (Li, 876). By giving the same importance to all edges in qualitative networks, a full description of a food web will obscure rather than clarify its structure (see e.g. Tylianakis et al. 2007, Nature 445, 202–205, for a practical example). Consequently, edges of food webs described extensively should ideally be weighted (with relevant metrics). In this respect, it is useful to indicate if, and how, your method could predict edge weights (a priori, it should be straightforward to adapt the predictors to quantitative information).

I have a single technical reservation on the work. I do not understand the choice of a linear regression coefficient for the analysis of the relationships between food web properties and performance (ROC-AUC and PR-AUC). From Fig. S4 and S5, it is clearly inappropriate since values larger than 1 can be predicted. Also, the distribution of the points may not be linear (e.g. in Fig. S4d). I recommend replacing the slopes with non-parametric correlation coefficients (Kendall or Spearman), which are perfectly adequate here. This requires rerunning a lot of analyses (Figs. 7 and S6 to S12), but this should be straightforward to adapt the code for this purpose. Also, I would replace the regression lines in Fig. S4 and S5 by lines obtained e.g., by moving average or generalized additive models.

I have some small comments:

Li. 16: ... to represent the complexity of large ecological systems. ("large" rather than "complex")

Li. 139: I would just make clear here that stacking was not applied to the existing ecological models just mentioned before (this possibility is aptly discussed on lines 707-721).

Li. 184-188: This description is quite vague here. I would just mention: "(see details section IV)".

Figure 3a: What are the blue and green triangles?

Legend of Fig.3.: I would indicate the difference between both network representations (upper and middle rows). Also, I would indicate that the lower row represents the adjacency matrices (if I am correct, with black indicating no interaction?). Also, correct "examaple" to "example".

Li. 307-311: This explanation is elusive to me.
Li. 380: "Fig. 4A,B" not "4A,C".
Li. 387: "Fig. 4C,D" not "4B,D".
Li. 395: It is not clear to me how this "relative importance" can be estimated for a particular food web.
Li. 458: Why do you have to "train" a univariate regression (but see my comment above).
Li. 715: I guess reference 88 should be removed from the list since it is mentioned on line 716.
Li. 875 and 877-878: what is meant by "multiple interactions"? This is not clear to me.
Li. 1110: Remove "neighbors" after the parenthesis.
Li. 1658: The name of the journal is missing for reference 106.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In this manuscript, the authors present a stacked model approach to predict missing links in food webs. The novelty lies in the use of stacked generalization, which appears to improve predictive performance by optimally combining information from both structural models (which predict links based solely on network topology) and trait-based models (which use species-level information). To adapt this methodology to ecological networks, the authors thoughtfully modify established network metrics, taking into account the specific characteristics of food webs such as directionality and trophic hierarchy. They convincingly argue that this combined approach outperforms previous methods in missing link prediction.

I find this article interesting and timely, as improving the resolution of ecological networks is becoming increasingly important for advancing both theoretical and empirical studies in community ecology. While I am not a specialist in machine learning, I found the methodology to be clearly described, well executed, and the results effectively communicated. In particular, I appreciate the authors' effort to explain how standard predictors were adapted to the ecological context; I thank them for the clarity of writing throughout the manuscript.

Although I cannot fully assess the methodological novelty from a machine learning perspective, I believe the manuscript will be of broad interest to researchers working with ecological networks.

I, however had difficulty locating the data/code accessibility statement, so I recommend to include it more clearly in the manuscript. Also, consider improving the readme file to more clearly explain the workflow, as it is very difficult to follow it how it is included now.

That said, I only have some minor comments:

Reference 106 and Reference 3: Missing journal information—please complete.

Line 105: There is an unmatched quotation mark (").

Line 359: I recommend briefly explaining what ROC-AUC represents the first time it appears.

Line 465: Indicate more precisely where in the Supplementary Information the empirical distributions can be found (e.g., "Fig. S17–S20").

Fig. S17: The sentence "(b) Counts of node movement type across the 5562 unique nodes, with multiple movement types either within or between webs for 83 nodes" is unclear. It is not clear what "nodes" refers to in the second instance—do the authors mean "webs" or something else? Please clarify.

Lines 576 & 601: Consider repeating the figure number where the referenced result is shown to help guide the reader.

Figure 'Top features by importance': It is unclear whether some predictors appear exclusively in the combined model or also in the individual ones. For instance, does the "log mass ratio" feature appear only in the trait-based model (blue) or also in the combined model (orange)? Please clarify this in the figure legend or main text.

Line 644: The phrase "may be in alignment with prior work" is vague. Either confirm the alignment (e.g., "These results are consistent with...") or rephrase cautiously (e.g., "These results appear to follow similar patterns to those reported in...").

Line 896: The sentence with "respectively" is incomplete. Please revise it to clarify what is being compared.

(Remarks on code availability)

The readme file from the code is not sufficiently well written, which makes it not accessible to the reader/user. I was not able to use the code.

Reviewer #3

(Remarks to the Author)

Please see attached review.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Dear Authors.

I read in detail the corrected manuscript and was happy to see that my comments were adequately addressed. I have a single editorial remark: on lines 137-139, I would change "work on missing link prediction in food webs has found that there is no universally best predictor for missing links in food webs" to "work on missing link prediction in food webs has found that there is no universally best predictor for such missing links"

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have satisfactorily addressed most of the issues I raised in my initial review, and I appreciate their thoughtful revisions to the manuscript.

However, I still have concerns regarding data and code accessibility.

First, I would like to thank the authors for including an example notebook to illustrate the procedure in more detail. This is a valuable addition and greatly helps in understanding the workflow. That said, I was initially unable to run the code, as some of the required Python libraries (e.g. imblearn) seemed not to be available in Anaconda (or I was not able to find them).

I strongly recommend that the authors include a complete list of required packages, ideally along with a conda environment file (in .yaml format) or a spec file. These would allow users to recreate the computational environment reliably and locally (See <https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html>)

After installing it via pip, I still was not able to run the Python notebook as the files

'./Processed_Data_Disaggregated_Lifestage/fw_attr_types.pickle' and

'./Processed_Data_Disaggregated_Lifestage/fw_ids.pickle'

where missing (or I was not able to find them).

As a consequence, I was not able to test the notebook.

Additionally, I encountered a technical issue when attempting to open the notebook in VS Code: an error was thrown regarding the initialization of the webview. Although the notebook opens without issues in JupyterLab, it may be helpful to check compatibility.

While I appreciate the authors' effort in expanding the README and including the notebook, I still find the overall workflow unclear, particularly with regard to the data preprocessing steps. Specifically:

What is the structure of the input data used for the link prediction step? ¿What is the difference between the aggregated and disaggregated lifestage networks? (I guess the number of nodes? Is this lifestage structure relevant? Which ones are used in the text? I have not seen any mention of "disaggregated" networks in the main text, for example. Why are they included here?

How is the pre-processing carried out, and what is its purpose?

In what order are the scripts or notebooks meant to be executed?

Clarifying these elements would greatly improve the reproducibility of the work.

(Remarks on code availability)

I have not been able to review the code as I was not able to run it.

Reviewer #3

(Remarks to the Author)

The authors have made significant improvements to address the reviewers' comments. Specifically, my concerns and comments have been well-addressed and the authors have added more discussions to the manuscript to improve the clarity. I'm satisfied with current version.

(Remarks on code availability)

Though I didn't try to reproduce the results, the code looks pretty self-contained with a detailed README file and all data used to generate the results.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Reviewers,

Thank you for your detailed consideration of our manuscript and the opportunity to address your comments with a revision. We appreciate that reviewers highlighted the novelty, timeliness, and scope of this work as well as identified areas in which we could improve the analysis and its presentation. We have carefully considered and addressed each comment, with our responses laid out below in blue. We also enclose a document indicating what changes we have made to the original manuscript.

Supplementary figure numbers in this document refer to the updated numbers after re-ordering supplementary figures in the revision and are labelled as S{X}. Line numbers refer to line numbers in the revised manuscript without track changes.

We have also made some minor changes to address the *Nature Communications* formatting instructions and editorial policy checklist:

- Corresponding authors identified with an asterisk and email addresses provided for corresponding authors
- Abstract reduced to 150 words or fewer
- Final sentence of the Abstract revised per the formatting instructions to summarize the major results and conclusions of the work in present tense
- Introduction revised per the formatting instructions so that the final paragraph is the only one that discusses the results and conclusions of the current study in the form of a summary
- Final subheading of the Results reduced to be less than 60 characters
- "Methods & Material" changed to "Methods"
- Added a Data Availability section after the Methods
- Added a Code Availability section after the Methods
- References revised to match the required format
- Reference list reduced from 117 to 70 to comply with the limit. This involved moving some minor methodological details to the supplement and some minor reduction of tangential discussion of related work in the main text.
- The reference that was repeated in figure captions was removed so that references are listed in the order they are found in the non-caption text.
- Added definitions of ROC-AUC and PR-AUC abbreviations in the figure legend captions
- Added definitions of RGG and SBM abbreviations in Fig. S22.
- Symbol descriptions added to the captions in Figs. 3, 7, S12-S20, S22
- Replaced all feature importance figures (Figs. 6, S9, S10) which were previously bar graphs with box and whisker plots (with box plot descriptions in the caption)
- Added box plot descriptions to captions of Figs. 4, 5, S8, S11, and S2
- Added whisker descriptions to captions of Figs. S14-S19

- Supplementary Figures, Notes, and Tables labelled as "Supplementary Fig./Table/Note X" in the supplementary information and referenced as such (rather than as Fig. S1, etc.) in the main text, with lowercase a, b, c to refer to subplots.
- Food web database description in the Supplementary Information labelled as Supplementary Note 2 (previous Supplementary Note 2 moved to Note 3).
- Fig. 3 plot labels changed to lowercase a, b
- Supplementary References labelled a Supplementary References and formatted according to guidelines
- Minor Fig. 2 revision for clarity that the links are directed
- Supplementary Data file legends moved from the Supplementary Information to the Cover Letter and labelled as "Supplementary Data X" there and when referenced.
- Noted use of betareg package for statistical analysis (Line 1221)
- Acknowledgements, Author Contributions, and Competing Interest Statement have been moved after the references

As well as with the reporting summary:

- We explicitly list the sample size as $N=X$ for all figures with the derivations, which also makes clear whether values are aggregated over distinct or separate samples. This led to small updates to the legends for Figs. 3-7, S2, S8-S20, S22.
- We add a supplementary table providing the details of the test statistic, confidence intervals, effect sizes, degrees of freedom, and p value for the pairwise two-sided t-tests shown in Fig. 4 and Fig. 5 (Supplementary Tables 4,5).
- We add Supplementary Data 1, which provides comprehensive details of the Beta regression results shown in Fig. 7 (mean coefficient estimate, standard error, 95% confidence interval, z value, and adjusted p value).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript proposes a novel method to predict unobserved interactions in food webs. As such, it participates to the recently growing body of literature on this question (for food webs and for other types of networks). Consequently, one may perceive this effort as incremental. However, this contribution stands out on several aspects.

Firstly, while the choice of model stacking is by itself not a breakthrough (one may indeed expect that the combination of the results of different existing models will be superior to the results of any individual model), it is based here on a large number of simple predictors that were, to my knowledge and for most of them, not used before for this purpose. Also, the information used is not only based on the network itself (structure-only models), but also integrates external information on the nodes, with a sensible distinction between assortative and disassortative descriptors.

Secondly, the work is really extensive, by the use of a large number of predictors, by the careful test of the approach with synthetic data, and by the application of the method to a large dataset of 290 published food webs.

Thirdly, the analyses of the results are comprehensive: which predictor and which type of information perform best; how these results differ between the different types of networks (grouped according to the habitat type); how model performance correlates with food web properties.

Thank you for your review of our work, and for pointing out its novelty in terms of the set of network and attribute-based predictors that we include in the model stacking procedure as well as the scope of our evaluation on synthetic and empirical food webs.

We have revised the Introduction to better describe how our study expands the methodological literature in ecology (second to last paragraph of the Introduction, Line 162) based on these comments.

Finally, the discussion of the results with regard to the existing literature is interesting and exhaustive, which is very useful for future work on this topic. There is however one topic that should be treated in this section. You mention that food webs sometimes come with additional complexity, among them edge weights (Li. 876). By giving the same importance to all edges in qualitative networks, a full description of a food web will obscure rather than clarify its structure (see e.g. Tylianakis et al. 2007, Nature 445, 202–205, for a practical example). Consequently, edges of food webs described extensively should ideally be weighted (with relevant metrics). In this respect, it is useful to indicate

if, and how, your method could predict edge weights (a priori, it should be straightforward to adapt the predictors to quantitative information).

The reviewer makes a good point. We agree that it is important to consider the weighted structure of food webs. In future work, the approach we suggest could be adapted to (a) consider edge weights in the calculation of predictors included in the ensemble by adapting network and attribute-based predictors to take these edge weights into account and (b) predict edge weights rather than (or in addition to) binary interactions. We have revised the Discussion to add a paragraph about this topic as suggested (Line 807).

Future work in this area would need to develop new edge-predictor features beyond those considered in our current work and address data-availability challenges. From a methodological perspective, adapting the model stacking procedure and set of predictors that we include in the models to take edge weights into account in a way that makes sense considering ecological theory should be possible, but would require some care and theoretical effort. For example, this would include adapting the custom ecological predictors based on common neighbors to also consider edge weight. Additionally, further work in this direction is limited by the lack of publicly available food web datasets with detailed edge weight information (Pringle & Hutchinson 2020, <https://doi.org/10.1146/annurev-ecolsys-110218-024908>).

I have a single technical reservation on the work. I do not understand the choice of a linear regression coefficient for the analysis of the relationships between food web properties and performance (ROC-AUC and PR-AUC). From Fig. S4 and S5, it is clearly inappropriate since values larger than 1 can be predicted. Also, the distribution of the points may not be linear (e.g. in Fig. S4d). I recommend replacing the slopes with non-parametric correlation coefficients (Kendall or Spearman), which are perfectly adequate here. This requires rerunning a lot of analyses (Figs. 7 and S6 to S12), but this should be straightforward to adapt the code for this purpose. Also, I would replace the regression lines in Fig. S4 and S5 by lines obtained e.g., by moving average or generalized additive models.

Thank you for pointing this out. We have replaced the univariable linear regression analyses comparing food web properties and ROC/PR-AUC results with univariable Beta regression analyses using a logit link, as these are more appropriate for the ROC/PR-AUC outcomes, which are bounded in the range [0,1]. These changes are now reflected in Fig. 7 and in Figs. S12-S20. Beta regression mean coefficients can be similarly interpreted as the linear regression coefficients in our previous analysis in terms of their relative direction and size of effect. The effects are now shown on a log-odds scale due to the logit link function used. We have also replaced the linear regression lines in Figs. S12, S13 (previously S4, S5) with Beta regression lines.

"Univariate" was also changed to "univariable" to clarify that each of the regressions of food web properties to outcomes were run separately.

We have reviewed the text and the high-level results that we focus on in the paper hold up with the updated analysis. In the Results section, we have adjusted some of the text to reflect small changes that we observed in the beta regression results. We saw some changes in the plots split up by ecosystem type – for example, in the previous linear regression results we saw a stronger negative trend for terrestrial belowground food webs for the attribute-only model between higher taxonomic level proportion and ROC-AUC performance. We adjusted the paragraph describing how taxonomic resolution results are ecosystem type dependent (Line 528). Additionally, we now see that for the average degree results, there was a significant correlation after controlling for the number of nodes for all three models and both metrics. We have updated this finding in the text (Line 551). Additionally, the trend for modularity with ROC-AUC is now only significant for all three models after controlling for the number of nodes. We have updated this finding in the text as well (Line 562).

We have also made a small revision to the text in this section of the Results to improve the clarity of the presentation, not due to changes from the beta regression analyses - in the sentence "For both ROC-AUC and PR-AUC,..." changing "that all three model performed better" to "better performance," as it is not the case that we see better performance in all three models in Figs. S12A, S13A.

I have some small comments:

Li. 16: ... to represent the complexity of large ecological systems. ("large" rather than "complex")

We have accepted this suggestion (Line 27).

Li. 139: I would just make clear here that stacking was not applied to the existing ecological models just mentioned before (this possibility is aptly discussed on lines 707-721).

We have adjusted this paragraph to make it clear that in this case we ensemble over a set of simple predictors many of which have not been used previously for ecological link prediction (Line 150).

Li. 184-188: This description is quite vague here. I would just mention: "(see details section IV)".

We have removed the details of which predictors were the most important from the Introduction, as we agree that we don't have enough space there to provide necessary context for the reader.

Figure 3a: What are the blue and green triangles?

We thank the reviewer for asking about this. The blue and green triangles were generated for Fig. S22 to show the curves of the theoretical ROC-AUC maximum performance for missing link prediction in RGG and SBM networks generated separately under varying ρ (in contrast to the theoretical maximum curve for combined networks shown with black triangles), but they were included in the main text Fig 3 by mistake and therefore not defined in the caption. In the revised version, we have removed them.

Legend of Fig.3.: I would indicate the difference between both network representations (upper and middle rows). Also, I would indicate that the lower row represents the adjacency matrices (if I am correct, with black indicating no interaction?). Also, correct "examaple" to "example".

We have added text to the Fig. 3 caption to describe these three visual representations of the networks are, including, yes, the adjacency matrices in the third row. And we have fixed that typo.

Li. 307-311: This explanation is elusive to me.

In this sentence, we were discussing why the performance of the structure-only model (Fig. 3, green) is high both at $\rho=0$ (SBM network) and at $\rho=1$ (pure RGG model), forming a U shape. This behavior is in contrast to what we observed for the attribute-based model (Fig. 3 blue), where the performance decreased as moving from $\rho=1$ to $\rho=0$. This behavior can be explained by the fact that the pure RGG networks have structural regularities (e.g., clustering of (dis)similar nodes) which can be picked up on by the structure-based model to predict missing links (using common neighbor predictors, etc.). We thank the reviewer for asking about this explanation, as we agree that we did not explain this adequately in the original text. In the revised version, we have revised this paragraph to make this point more clear (Line 330).

Li. 380: "Fig. 4A,B" not "4A,C".

We have fixed this (Line 405).

Li. 387: "Fig. 4C,D" not "4B,D".

We have fixed this (Line 413).

Li. 395: It is not clear to me how this "relative importance" can be estimated for a particular food web.

Here we were trying to make the point that the full model learns which predictors to focus on for a specific food web, with no *a priori* knowledge. However, reflecting on the reviewer's comment, we think that this point could be better covered in the Discussion, so we have moved it there in the revised version (Line 708).

Li.458: Why do you have to "train" a univariate regression (but see my comment above).

Per the earlier point, we have revised this analysis to instead perform beta regressions, and we have adjusted this text to replace "train" with "perform" to emphasize that this is not a predictive analysis.

Li. 715: I guess reference 88 should be removed from the list since it is mentioned on line 716.

We have adjusted this.

Li.875 and 877-878: what is meant by "multiple interactions"? This is not clear to me.

In this sentence we are referring to cases in which more than one trophic interaction is listed between the same two species in the same food web dataset. We observed this to occur occasionally in the empirical food webs (307 interactions across 16 food webs) that we used in this study, and we only included a single interaction between the two species in these cases before further analysis. It is possible that these are data errors or these could be ecologically meaningful, for example due to multiple interactions recorded by research teams between the same two species in different seasons, different tidal zones, etc. To make this idea more clear in this sentence, in the revised version, we have added detail: " between the same two nodes (e.g., across multiple seasons or tidal zones)".

Li. 1110: Remove "neighbors" after the parenthesis.

We have fixed this.

Li. 1658: The name of the journal is missing for reference 106.

We have fixed this.

Reviewer #2 (Remarks to the Author):

In this manuscript, the authors present a stacked model approach to predict missing links in food webs. The novelty lies in the use of stacked generalization, which appears to improve predictive performance by optimally combining information from both structural models (which predict links based solely on network topology) and trait-based models (which use species-level information). To adapt this methodology to ecological networks, the authors thoughtfully modify established network metrics, taking into account the specific characteristics of food webs such as directionality and trophic hierarchy. They convincingly argue that this combined approach outperforms previous methods in missing link prediction.

I find this article interesting and timely, as improving the resolution of ecological networks is becoming increasingly important for advancing both theoretical and empirical studies in community ecology. While I am not a specialist in machine learning, I found the methodology to be clearly described, well executed, and the results effectively communicated. In particular, I appreciate the authors' effort to explain how standard predictors were adapted to the ecological context; I thank them for the clarity of writing throughout the manuscript.

Although I cannot fully assess the methodological novelty from a machine learning perspective, I believe the manuscript will be of broad interest to researchers working with ecological networks.

Thank you for the positive assessment of our work, and for noting the timeliness of this work for advancing research within community ecology. We revised a line in the introduction to explicitly mention contributing to empirical and theoretical studies in community ecology to better explain our contribution: "More accurate methods for predicting missing links in food webs would benefit both empirical and theoretical studies in community ecology by increasing the efficiency of collecting species interaction data in the field and providing more reliable insights into questions about ecosystem stability, conservation efforts, and tests of ecological theories." (Line 98)

I, however had difficulty locating the data/code accessibility statement, so I recommend to include it more clearly in the manuscript. Also, consider improving the readme file to more clearly explain the workflow, as it is very difficult to follow it how it is included now.

We apologize for the difficulty. We have added Data and Code accessibility statements after the Methods and before the References.

And, we appreciate the suggestion to improve the usability of the GitHub codebase – we have completely revised the README.md file to explain how to run the code, as one of our goals is for others in the community to be able to use the methods we develop

and evaluate in this manuscript for their own projects. The new readme file can be found at: https://github.com/lucyvkk/food_web_link_prediction. Minor changes were also made to the comments within the code in the repository to improve usability, in particular in the file `Real_Networks/link_prediction_food_webs.py`.

That said, I only have some minor comments:

Reference 106 and Reference 3: Missing journal information—please complete.

We have fixed reference 106 (now Ref 11). Reference 3 was removed in revisions.

Line 105: There is an unmatched quotation mark (").

We have fixed this (Line 112).

Line 359: I recommend briefly explaining what ROC-AUC represents the first time it appears.

We have moved the sentence explaining what the ROC-AUC measures from the section on performance on empirical food webs to the section on the performance on synthetically generated networks so that it shows up when the ROC-AUC first appears in the text (Line 299).

Line 465: Indicate more precisely where in the Supplementary Information the empirical distributions can be found (e.g., "Fig. S17–S20").

We have added this and slightly adjusted the count of features for the metadata and data processing features to mention only the 11 numeric features here for which regression analyses were performed (excluding ecosystem type, which is listed in Table S6, but was used to split up the data into categories rather than as a regression covariate).

Fig. S17: The sentence "(b) Counts of node movement type across the 5562 unique nodes, with multiple movement types either within or between webs for 83 nodes" is unclear. It is not clear what "nodes" refers to in the second instance—do the authors mean "webs" or something else? Please clarify.

In Fig. S3b (previously S17b), the goal is to show the distribution of movement types across the 5562 unique nodes (determined by unique node names) in the food web database in order to provide information on the scope of species traits considered in our analysis.

As described in the supplemental section on the food web data processing, we treated movement type as a categorical node attribute that was transformed into binary attributes for each possible movement type value in a given web (a representation that is sometimes called a "one-hot encoding"). Species could have multiple movement

types within the same food web if they were listed with different movement types in different interactions in the food web. Unique nodes (same species name) also sometimes had different movement types in different food webs. Overall, this allowance for multiple movement types per unique node increased the count by 83 over the number of unique nodes (5645, as listed previously in the plot title vs. 5562, the number of unique nodes).

We agree that this was not clear in the original version of the plot, so we have revised the Fig. S3b title to list "out of 5562" rather than "out of 5645" to make it clear this is over the same set of unique nodes as Fig. S3a. We have also simplified the caption to make this point more clear: "(b) Counts of node movement type across the 5562 unique nodes, taking into account that unique nodes can have multiple movement types."

Lines 576 & 601: Consider repeating the figure number where the referenced result is shown to help guide the reader.

We have added figure references at the end of the sentences starting with "We also find that..." (Fig. 5, we also adjusted this slightly to also call out terrestrial aboveground as the lower performing ecosystem type with marine food webs, line 628), "However, we had the fewest..." (Fig. S1a), line 634, and "Our feature importance results also support this conclusion... (Figs. 6, S10), line 657" - in this case we have highlighted the permutation importance rather than Gini importance supplemental results as it is here that we see log mass showing up as one of the top 10 features across multiple ecosystem types and the relative importance of log mass isn't evident in Fig. S9 apart from in marine food webs.

Figure 'Top features by importance': It is unclear whether some predictors appear exclusively in the combined model or also in the individual ones. For instance, does the "log mass ratio" feature appear only in the trait-based model (blue) or also in the combined model (orange)? Please clarify this in the figure legend or main text.

Thank you for asking about this. To clarify, Fig. 6 ('Top features by importance') shows the top features based on feature importance scores from only the combined / full model, with results averaged across folds and food webs, and then combined across ecosystem types as described (taking the union of the top 10 predictors from each ecosystem type). While we mention this already in the caption "Across 106 predictors in the full model," we realize that this could be made more clear. We have added additional detail to the Fig. 6 caption to explain this. The predictors included in each model are explained in the section on the synthetic network tests, and we have revised this to list, as in the Fig. 6 caption, the number of predictors included in each model, so that this connection to Fig. 6 is clearer (Line 311).

Line 644: The phrase “may be in alignment with prior work” is vague. Either confirm the alignment (e.g., “These results are consistent with...”) or rephrase cautiously (e.g., “These results appear to follow similar patterns to those reported in...”).

We have rephrased this cautiously as suggested (Line 699).

Line 896: The sentence with “respectively” is incomplete. Please revise it to clarify what is being compared.

We have revised this sentence as follows (Line 965):

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

Reviewer #2 (Remarks on code availability):

The readme file from the code is not sufficiently well written, which makes it not accessible to the reader/user.

I was not able to use the code.

We agree that the usability of the GitHub codebase could be improved and we have thus revised the README file to make it more clear how to run the code.

Reviewer #3 (Remarks to the Author):

Review Report for “Predicting missing links in food webs using stacked models and species traits”

The authors developed a network-based method to predict missing links in food webs. The method, which is an extension of the stacking model from Ghasemian et. al. (2020), optimally combines both structural and trait-based predictions with ecological grounding. It is shown that stacking predictors from both structure-level and the attribute-level outperforms the models that consider only one of those predictors. The authors demonstrated a comparison of the stacking models with a simulated dataset and further applied the methods to real food webs.

Overall, this is an applied paper that extends an existing state-of-the-art model in missing-edge prediction to an important application in food webs discovering in the ecological context. Although the manuscript is well-written and easy to follow, I feel the contribution was not highlighted and some technical detail needs further clarification. My detailed comments are below:

Thank you for taking the time to review our manuscript and providing suggestions as to how to improve the presentation of our work. We have revised the paper to better describe our technical contributions and clarify the points that you have highlighted as needing further attention, as described below.

1. It is not clear what is the novelty of this manuscript. The author applied an existing model with some adaption to the ecological-domain knowledges. The contribution appears marginal to me from the perspective of statistical modeling.

Thank you for pointing out that the technical novelty of the adaptation of the published stacking model to the ecological context could be better highlighted and emphasized in the manuscript.

The technical contributions towards network science method development that we made were as follows:

- Ghasemian et al. did not consider node attributes as predictors, which is a significant limitation for ecology. In ecological networks, node attributes (species traits) are a critical source of information about interactions, and often when collecting a new ecological network, some species trait information is known even when little is known about species interactions. Hence, extending the Ghasemian et al. work in this direction is essential for developing better ecological networks for both empirical and theoretical work in ecology.
- We introduced ~29 new predictors to the stacking model for each network as part of adapting the stacking model to consider node attributes, as described in the Materials & Methods sections on "Attribute-based predictors."

- We adapted the training and testing steps of model stacking algorithm to work with directed networks, which required several technical innovations. As described in the Methods & Materials section "Training and testing in model stacking for missing link prediction," these innovations required altering the training and test datasets to consider directed rather than undirected edges in the machine learning task for the meta-level model.
- We adapted the set of predictors themselves to include directed variants of many of the undirected predictors used in Ghasemian et al. 2020, as described in the Materials & Methods section on "Structural predictors," e.g., defining version for each of the in- and out- edge directions.
- We grounded the set of link predictors more firmly in ecological theory by defining 6 new topological predictors that are based on theoretical assumptions or predictions about the expected global hierarchy of food webs. The details of this contribution are discussed in the Materials & Methods as well as in the beginning of the Results section in Figs. 1, 2.
- We adapted K-nearest-neighbor predictors to the stacking model context and added 12 such predictors to the ensemble, as described in the Material & Methods section on "Nearest-neighbor predictors."

Taken together, these innovations constitute a significant advance in the state of the art for the specific scientific setting of food webs, one that is better grounded in ecological theory, as highlighted by Reviewer 1 and by Reviewer 2, compared to the approach proposed in Ghasemian et al. 2020.

We thank the reviewer for pushing on these points, and we have revised the manuscript to better articulate these methodological contributions in the following ways:

- We have made these methodological contributions clearer in the paragraph of the Introduction starting with "Here, we comprehensively adapt..." (line 162)
- Materials & Methods - We have adjusted some language to make it more clear what we are proposing in this manuscript that is a change from the model used in Ghasemian et al. 2020. (Line 957, "in our adaptation", Line 1008, changes in the paragraph starting with "The structure-based models")
- Discussion – While not our focus in this work, we have added a short paragraph noting these methodological contributions and how they are potentially applicable to missing link prediction in other fields for directed, attributed, and/or hierarchical networks, which could be explored in future work. (Line 851)

2. The authors are setting up synthetic random networks in part II.A to replicate the real food webs as discussed in part IV.B by having similar network statistics. However, the performance on the synthetic networks does not match the ones of the real food webs (part II.B). (ROC-AUC of 0.7 ~ 0.9 in the synthetic data v.s. >

0.9 in the real data). The difference here suggests that the real food webs are not in the random network family discussed in part II.A, which makes the simulation in II.A less relevant to the real food webs.

We thank the reviewer for asking about this point, and we acknowledge that the original manuscript did not explain the value of the synthetic network tests sufficiently clearly. In the revised manuscript, we now elaborate on the specific value of those tests in the synthetic network Results section (line 242, “These tests allowed us to evaluate the stacking model performance on a difficult task for the model, and to see whether the model could learn to predict missing links using network structure, node attributes, or a combination.”).

Specifically, the purpose of these tests was to quantify the degree to which the full stacking model could detect whether to predict missing links based on network structure or independent node attribute patterns (or some combination) in challenging food web cases. That is, the synthetic networks were not meant to necessarily include realistic food webs as special cases, but rather they provide a simple way to tune the degree to which an edge exists “because” of trait vs. structure information. In that way, by varying the relative amount of the two models (SBM and RGG), we can make the synthetic network more or less challenging for a structure-only or trait-only model, and evaluate the degree to which the “full” model can adapt to whatever patterns are available in the observed network. That is, the purpose of the synthetic tests is to build confidence that the method is both capable of and accurate for predicting missing links in networks where edges exist for some variable (and unknown) amount of trait vs. structural reasons.

We agree that it is interesting that we found better performance when the model was applied to empirical networks than in the synthetic tests cases. To highlight this point for the reader, we have added a sentence to the Discussion to note this as an area for future exploration (line 620). Our own hypothesis for the origin of this difference is that the empirical networks likely showed more correlation between structure and traits than those networks in the most difficult mid- ρ range of our mixed network tests, and thus the full stacking model performed on average better on the empirical networks.

3. The edges(links) in the food webs is supposed to be sparse. Therefore the classification problem of identifying missing links has way much smaller proportions of positives (missing links). For such a case, ROC-AUC could be misleading and PR-AUC is more appropriate. The authors also discussed this and considered both metrics in Fig. 4. But then in Fig. 5, only ROC-AUC is discussed. I think the authors should tie to PR-AUC for the sparse food webs.

We appreciate the reviewer’s perspective. However, we also note that ROC-AUC is a standard measure of accuracy in the link prediction literature and hence it remains an important measure to use in order for studies to be comparable to each other. In

addition, ROC-AUC is mathematically equivalent to a Mann-Whitney test for class distinguishability, and hence as a measure for comparing algorithms, it is quite useful. We say this even as we agree that a high ROC-AUC may not correlate with performance under other measures, such as the classic “predictions on a budget” measure of top-k or even PR-AUC. This is not a failure of ROC-AUC, but rather highlights that different accuracy statistics measure different things, each useful in its own way.

That all said, we agree that it can be useful to consider PR-AUC along with ROC-AUC. We thus considered both metrics throughout the empirical food web results. We already provide the same information as Fig. 5 (link prediction performance by food web ecosystem type) in Fig. S8 for PR-AUC. Based on the size of the figures and not wanting to overly clutter the main text, we chose to show only one of these as Fig. 5 there. And, we chose the ROC-AUC for the main text in this case because, as stated above, it is a standard classification metric in the literature for comparing methods and their ability to tell true positives from true negatives.

In our revised manuscript, we have specifically mentioned the class imbalance consideration with sparse networks when the PR-AUC is first defined (line 385), as well as elevated the Fig. S8 PR-AUC results into the Results text when we discuss results across ecosystem types, as suggested (line 432).

4. Fig. 6 reveals that all of the several leading predictors is the KNN predictors. And the KNN predictors make a lot of ecological sense to me (e.g. identifying similar species and categories). I would wonder how a simpler model built only from those KNN predictor would perform.

This is an interesting idea, thank you for the suggestion. To find out, we have run a supplemental analysis to evaluate the overall performance across the food web database if the ensemble for the full and structure-only models is restricted to only KNN predictors (no KNN predictors are included in the attribute-only model). We found that predictive performance slightly decreased compared to the models with all predictors but remained high, indicating that these predictors alone could encode much of the information necessary for predicting missing links (new supplemental figure S11, mentioned in the main text results). We agree that KNN predictors make ecological sense and can be related to ecological theory. To illustrate these ideas and the supplemental results for the reader, we have added the supplemental result to the manuscript (line 478) and a sentence on these results to the Discussion (line 756).

5. In reporting the AUC numbers, more digits are needed to see how the difference compared to the standard deviations. Also, is “stddev” the sample standard deviation or the standard error for the sample mean?

In the "Performance on empirical food webs" section we have reported all standard deviations in the text now with 2 significant figures. `stddev` is the population standard deviation, calculated with `numpy.std()` over the 290 average performance results across the food web database (for each food web, the average performance metric is obtained by averaging over 5 folds x 5 iterations per food web). (See code here, line 266 - https://github.com/lucyvk/food_web_link_prediction/blob/main/Real_Networks/summarize_results_food_webs.py)

Dear Reviewers,

Thank you for the review of our revised submission. All reviewers indicated that we resolved all or the majority of their initial comments. We have provided responses to outstanding comments (largely from Reviewer #2 related to code availability) below. All updates are to our public GitHub repository: https://github.com/lucyvk/food_web_link_prediction. Figure and table numbers refer to the updated numbers in the final version. A track changes version of the main manuscript file has been included.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Dear Authors.

I read in detail the corrected manuscript and was happy to see that my comments were adequately addressed. I have a single editorial remark: on lines 137-139, I would change "work on missing link prediction in food webs has found that there is no universally best predictor for missing links in food webs" to "work on missing link prediction in food webs has found that there is no universally best predictor for such missing links"

Thank you for noting that we addressed all of your comments. We have made the suggested change.

Reviewer #2 (Remarks to the Author):

The authors have satisfactorily addressed most of the issues I raised in my initial review, and I appreciate their thoughtful revisions to the manuscript.

Thank you for pointing out that we addressed most of your initial comments.

However, I still have concerns regarding data and code accessibility.

First, I would like to thank the authors for including an example notebook to illustrate the procedure in more detail. This is a valuable addition and greatly helps in understanding the

workflow. That said, I was initially unable to run the code, as some of the required Python libraries (e.g. imblearn) seemed not to be available in Anaconda (or I was not able to find them).

I strongly recommend that the authors include a complete list of required packages, ideally along with a conda environment file (in .yaml format) or a spec file. These would allow users to recreate the computational environment reliably and locally (See <https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html>)

Thank you for these comments. We would like to ensure that the repository is easily usable. We have added an environment.yml file as suggested which can be used to configure a conda environment with the required python packages.

After installing it via pip, I still was not able to run the Python notebook as the files `'./Processed_Data_Disaggregated_Lifestage/fw_attr_types.pickle'` and `'./Processed_Data_Disaggregated_Lifestage/fw_ids.pickle'` where missing (or I was not able to find them).

As a consequence, I was not able to test the notebook.

We have provided the two missing .pickle files in the repository. These files are produced from the data pre-processing notebooks, but this was not made clear previously in the documentation and we agree that it would be better if the link prediction demonstration notebook could be run immediately after downloading the repository rather than only after running the data pre-processing notebooks.

Additionally, I encountered a technical issue when attempting to open the notebook in VS Code: an error was thrown regarding the initialization of the webview. Although the notebook opens without issues in JupyterLab, it may be helpful to check compatibility.

We have tested with a fresh clone of the repository and confirmed that the notebooks can be opened and run in VS Code (version 1.106.2).

While I appreciate the authors' effort in expanding the README and including the notebook, I still find the overall workflow unclear, particularly with regard to the data preprocessing steps. Specifically:

Thank you for reviewing the usability of the README file, we have added additional information per your questions below:

What is the structure of the input data used for the link prediction step?

We have added a new "Quick start" section to the README file which provides a description of the expected format for the input food web data. We have changed the names of the folders in the repository to "Input"_Data_(Dis)Aggregated_Lifestage (formerly "Processed") to make it clear that we have provided the data directly in the format used in the link prediction analyses in the repository.

¿What is the difference between the aggregated and disaggregated lifestage networks? (I guess the number of nodes?

In the aggregated lifestage networks the nodes in the food webs are considered to be at a species level, with nodes aggregated across multiple lifestages of the same species if they exist within the same food web dataset, whereas for the disaggregated lifestage networks a node is considered at species-lifestage level, so nodes are split up across multiple lifestages present in the same food web dataset. This is explained in supplementary Note 2 "Food web data pre-processing" (second paragraph) and we have added additional detail to explain this in the repository README as well (under the heading "Empirical food web pre-processing"). Supplementary note 2 was formally called "Food web data processing," but we changed the title to match the repository.

Is this lifestage structure relevant?

As noted in the supplementary Note 2 and now in the README, lifestage disaggregation changed the set of nodes for 25 out of the 290 food webs (491 nodes added in total with disaggregation). Thus, this only changes the set of nodes for ~9% of the food webs. We found nearly identical missing link prediction performance with lifestage aggregation, which is shown in Supplementary Figure 3 (details in Supplementary Table 7).

Which ones are used in the text?

In the main text we show results for the disaggregated version. Traits might differ between lifestages for the same species, so this was a more relevant approach given our goal of using traits and network structure for missing link prediction. However, we noted that food web datasets are sometimes pre-processed differently, so we performed a sensitivity test to see if the results looked different specifically for that set of 25 food webs that showed a change in their set of nodes based on this decision.

I have not seen any mention of "disaggregated" networks in the main text, for example. Why are they included here?

Thank you for pointing this out. We realized that this decision and the relevant sensitivity test results shown in Supplementary figure 3 were not discussed in the main text, so we have added this: "We considered nodes at the species-lifestage level given traits might

differ across lifestages for the same species, though we found that aggregating traits across all lifestages of the same species in a food web dataset did not change results (Supplementary Fig. 3, Supplementary Table 7)"

How is the pre-processing carried out, and what is its purpose?

The pre-processing is carried out by running the Jupyter notebooks found in the Real_Networks/Data_PreProcessing_Code_(Dis)aggregated_Lifestage folders in the repository (we have added the "Pre" to "Processing" for consistency). These notebooks transform the data from the form in which they were originally downloaded from <https://idata.idiv.de/ddm/Data/ShowData/283?version=3> (CSV file listing all interactions) into individual edge and node lists per food web, with associated traits per node also listed in the node list. The end-to-end data pre-processing is described in detail in supplementary Note 2. We have also added an additional description in the README "Empirical food web pre-processing" section. We have changed the titles to say "pre"-processing to make it clear that this is a step for formatting and cleaning the data that happens before the core analysis.

In what order are the scripts or notebooks meant to be executed?

We agree that this was not formerly clear in the README. We have added a Quick start guide in the README as well as revised the descriptions throughout to make it clear the purpose of each script and in what order we ran files. Some users might be interested in partially or fully reproducing our work, while others might be interested in simply applying the link prediction method to their own datasets, so the updates should make it clear how to accomplish the desired goal.

Clarifying these elements would greatly improve the reproducibility of the work.

Reviewer #2 (Remarks on code availability):

I have not been able to review the code as I was not able to run it.

We have addressed the issues noted above, which we believe has improved the usability of the code repository.

Reviewer #3 (Remarks to the Author):

The authors have made significant improvements to address the reviewers' comments. Specifically, my concerns and comments have been well-addressed and the authors have added more discussions to the manuscript to improve the clarity. I'm satisfied with current version.

Thank you for reviewing the manuscript again and noting that we addressed your comments.

Reviewer #3 (Remarks on code availability):

Though I didn't try to reproduce the results, the code looks pretty self-contained with a detailed README file and all data used to generate the results.

Thank you for highlighting that we have provided all code and data necessary to reproduce our results. We hope that we have improved the clarity of our documentation even further per the updates that we have noted.

Review Report for

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

The authors developed a network-based method to predict missing links in food webs. The method, which is an extension of the stacking model from Ghasemian et. al. (2020), optimally combines both structural and trait-based predictions with ecological grounding. It is shown that stacking predictors from both structure-level and the attribute-level outperforms the models that consider only one of these predictors. The authors demonstrated a comparison of the stacking models with a simulated dataset and further applied the methods to real food webs.

Overall, this is an applied paper that extends an existing state-of-the-art model in missing-edge prediction to an important application in food webs discovering in the ecological context. Although the manuscript is well-written and easy to follow, I feel the contribution was not highlighted and some technical detail needs further clarification. My detailed comments are below:

1. It is not clear what is the novelty of this manuscript. The author applied an existing model with some adaption to the ecological-domain knowledges. The contribution appears marginal to me from the perspective of statistical modeling.
2. The authors are setting up synthetic random networks in part II.A to replicate the real food webs as discussed in part IV.B by having similar network statistics. However, the performance on the synthetic networks does not match the ones of the real food webs (part II.B). (ROC-AUC of $0.7 \sim 0.9$ in the synthetic data v.s. > 0.9 in the real data). The difference here suggests that the real food webs are not in the random network family discussed in part II.A, which makes the simulation in II.A less relevant to the real food webs.
3. The edges(links) in the food webs is supposed to be sparse. Therefore the classification problem of identifying missing links has way much smaller proportions of positives (missing links). For such a case, ROC-AUC could be misleading and PR-AUC is more appropriate. The authors also discussed this and considered both metrics in Fig. 4. But then in Fig. 5, only ROC-AUC is discussed. I think the authors should tie to PR-AUC for the sparse food webs.
4. Fig. 6 reveals that all of the several leading predictors is the KNN predictors. And the KNN predictors make a lot of ecological sense to me (e.g. identifying similar species and categories). I would wonder how a simpler model built only from those KNN predictor would perform.
5. In reporting the AUC numbers, more digits are needed to see how the difference compared to the standard deviations. Also, is “stddev” the sample standard deviation or the standard error for the sample mean?