Reviewer 1

Major points

Validation (246 – 280): This section could benefit from comparison of main performance metrics – as used in ML and subfields, as they are often missing from ecological papers. For instance ROC-AUC, PR-ROC-AUC, F-score, K-statistics, Cohen's kappa to mention a few. A way to go about it would be discussing differences between measures calculated in the sensitivity-specificity space (e.g., ACC, AUC, TSS), and those calculated in the precision-recall space (e.g., PR-AUC, F-score, etc). The former cannot capture the intricacies of imbalanced class/sparse network problem. You can have a very high TSS and very low F-Score, for instance. The choice for which metric to tune a predictive model for/select between different models/algorithms depends on the application of the prediction exercise.

> We thank the reviewer for bringing up this important point. We have expanded the section of validation, to not only include more validation measures, but to provide guidance on how these measures should be interpreted in the context of predicting binary species interactions. In the context of accommodating the two reviewers concerns, we have extensively reworked the proof of concept, and how it fits in the broader manuscript narrative, and so the name of figures have changed.

These discussions need to be added to this section to make it correct (see below), and provide an explanation of these metrics and their utilisation that I think would be invaluable to the community at large.

> We have now included a table in the validation section showcasing the various metrics, what is considered the 'best' outcome and a brief description of the metric. We have also expanded the discussion around accuracy, sensitivity, and specificity. This should specifically help in addressing the second point raised below. We are convinced that this revised section is providing actionable guidance for ecologists to employ validation measures going forward.

Considering the above, validation of proof of concept/and Figure 4 must be updated to reflect the different values/natures of these metrics (e.g., in addition to TSS, a metric from precision-recall space must be shown). The discussion around validation of proof of concept needs to be expanded with: 1) the additional metrics as per above;

> The additional measures have been added in a table, and the precision-recall curve is now presented alongside the ROC curve. Both show extremely good performance given the relative paucity of data in the starting dataset.

2) backing up of why a TSS around 0.5/50% is actually any good; yes, it

> should undoubtedly be used instead of accuracy, but what does a roughly around random TSS means?
>
> We emphasize the use of Youden's J, and adopt a more holistic view of the validation by looking at a broader range of measures. That being said, we want to correct a misconception that the reviewer introduces: a TSS of 0.5 is not "roughly around random". TSS takes values in [-1, 1], and a random classifier would have a TSS close to 0. The usual guidelines around interpretation suggest that a TSS above 0.5 indicate "moderate to strong" agreement. Regardless, we substituted TSS by Cohen's κ, and obtain a value of about 0.6, similarly indicating an acceptable performance of the model.

The authors make very strong claims with regards to their proof of concept which need to be backed up. For instance, when measured, proof of concept might yield good performance on the PR-metrics, which could strengthen the argument for it.

> We thank the reviewer for requesting the inclusion of these metrics, as they do strengthen the claims we make based on the proof of concept. So as to be clear, our only claim with regards to the proof of concept is that it is *possible* to apply AI/ML to sparse interaction data – the revised section contextualizes how this *can* be done a little more, but this is not intended as a normative example, and we have edited the text in a few places to make this clearer.

Uncertainties and variations in models' outcomes must be discussed more prominently. This is becoming increasingly important, particularly with DL/ML algorithms being prone to underspecification. For instance, we might train proof of concept 1000 times, changes in random seeds/sampling/training data might lead to equally-performing trained models (including on test sets) but which might yield wildly/somewhat different predictions. Providing means to quantifying this uncertainty is crucial for any predictive (including on networks) models and it must be discussed. In addition, mentions of different subsampling techniques and their role/effect on models and their predictions needs to be mentioned: for instance what might happen if we were to over-sample/under-sample at random, or within the restrictions of a given network/interaction type/biases in the underlying data. How can subsampling/bootstrapping be used effectively? And how can you quantify (to an extent) the uncertainties in underlying data/network structure?

> We somewhat agree with the reviewer on this point, but have decided not to make extensive modifications to the text. There are a few reasons for this. First, the reviewer is essentially asking for a methodological study of deep learning best practices for ecological networks predictions – this does sound like a very exciting and important paper, but not one we intend to write within the context of this piece. Second, all of the approaches identified by the reviewer make the

> implicit assumption that the sample size is large; this assumption is entirely unreasonable for ecological networks, and whatever estimate of uncertainty we would obtain would be pseudo-replication. Our code now specifies a seed so that the results are reproducible; we also use a mini-batching strategy that increases the effective sample size for training. All that being said, this case study, and indeed most ecological network inference studies, will be working very close to the edge of "not enough data", and re-using the data for bootstrapping would look technically correct, but would be in direct contradiction with ecological good practices of avoiding pseudo-replication.

Lines 87-8: in addition, a claim is made that a slight inflation of positive interactions would overcome existing biases (in relation to imbalanced nature of the network). This needs to be described in more details, and qualified with evidence (e.g., sampling metrics), also a discussion must be made in relation to various subsampling/instance-synthesis techniques, and their effect on the predictions (as per above). Furthermore, clarification needs to be made if performance metrics were derived from the raw or the slightly inflated data.

> We now refer to other work that outline the importance of balancing the dataset in terms of postive/negative interactions: creating balanced datasets allowed to reach high predictive value in a species interaction models, even when the initial amount of data was much lower (in terms of connectance, at least). We further add a reference to an article showing that setting threshold for balance within the training batches is a well established practice.

Becker et al 2020: I am aware of this study and its limitations. Here, it is presented in the same way peer-reviewed studies are (see below), without discussing those limitations, particularly:

> We want to make a point very clear: we discuss Becker et al. not "in the same way peer-reviewed studies are", but in the same way studies we have critically read, assessed, and looked at were. The reviewer may have issues with the Becker et al. article, and a different assessment of this publication that we do, but our manuscript is not the place to raise these concerns – there is, indeed, nothing we can do to remedy them. Nevertheless, as the editors will see, we have responded to the comments on this study that we can address.

1) Study design – the overall performance of the final ensemble is significantly worse than some of components.

> We agree with the reviewer, and we have highlighted this fact as a more general recommendation of being careful about the actual performance of the ensemble. We disagree with the reviewer that this is a "study design" issue – this is a result, that we can highlight and take as a call to pay more attention to the validity of ensembles for specific problems. It is, indeed, likely that many studies will

> suffer from the issue of having ensembles that underperform their best component models.

2) Lack of any quantification of uncertainties (e.g., as per discussion above).

> We have no control on this part of the Becker et al. preprint (though there is a discussion of the disagreement between models in Becker et al., which also relates to the point the reviewer made about the ensemble); we have identified no changes to make.

Overall, the authors cite only two pre-prints, to the exclusion of other pre-prints that could be useful in this discussion. Therefore, the authors must either: 1) justify why only those two preprints were used, 2) expand selection to include additional pre-prints; or 3) remove these citations. Furthermore, there are few cases where the aforementioned work has been cited to the exclusion of other peer-reviewed work (e.g., in discussion of node-embedding: line 365).

> In the absence of specific suggestions by the reviewer, and seeing that the point of this manuscript is not to be an exhaustive review, we have made no changes to the manuscript in response to this specific comment. We also want to emphasize that some manuscripts we cited as preprints have since been published.

How do we predict how species that we have never observed together will interact (355-374): there are other ways to incorporate network-structure into models to predict interactions within a given network, such as calculating network-based features, they need to be mentioned here for completeness.

> We agree with the reviewer – in fact, there was in the original submission a section of "What network properties should we use to inform our predictions of interactions?". Although not a review of other models (a lot of which are cited elsewhere in the text, some of which we added during the revisions, and the ones we missed we would happily cite would the reviewer unambiguously identify them), this section is a discussion of using network properties to penalize other statistical models. We specifically discuss the potential and limitations of using network properties for prediction. In the former section we argue that the network-based feature most likely to provide useful predictions is connectance, as most network properties highly covary with it. We have added a mention to this section on using other network properties (modularity and embedding) for prediction, in order to better guide the readers, and we thank the reviewer for pointing out the lack of clarity.

Minor points:

> We have introduced the concept of different interaction types earlier in the manuscript (ca. line 30) as well as expanding on how network type (uni-, bi- k,) plays a role in the modelling process (ca. line 480)

Lines 318-340: There are few other models that connect networks/their structures to predict interactions within the networks/subset of nodes in networks. They need to be cited here.

Previously covered as one of the major points.

Lines 375-397: this section could benefit from discussion of various types of flow in networks, for instance there are few examples in the literature (needs to be cited), were authors looked at the concept of flow in subtypes of ecological networks, and its meaning. Particularly for unipartite ecological networks – flow can be misleading in some scenarios, and very powerful in others.

> We have now included references to studies that have addressed network flows and network flow prediction as part of this section. As the reviewer has not explicited which specific work they were thinking about, we cannot guarantee they will be satisfied by the revision, but we are convinced that the editors will recognize this as a good faith attempt.

In addition to interaction strength, weighted networks need to be mentioned somewhere, even if only for completeness.

> Weighted networks are the focus of the section on interaction strength, and we have clarified the text at several places to make this clear.

Reviewer 2

> As per our discussion with the editorial team, the review from reviewer 2 has not been communicated to graduate students authors of this manuscript, as it contained several personal attacks, and a concerning number of comments that felt far outside the bounds of professional, or indeed acceptable, behavior. We want to unambiguously point out that, when confronted with a review as aggressive as review 2 was, it is very unclear which of the points are made in good faith or out of spite (this is true in this specific instance as authors, but we also hold this to be true in our collective experiences as reviewers and editors). Therefore, we have veered on the size of caution, and discarded most of the points, including all of the minor ones, which were written in the most acerbic way. The main comments in this section are summaries of the second review made by the last author, and all replies and in-text modifications have been made by the last author. We want to clarify that we have read the review several times over, and made changes to the manuscript when appropriate including based on the minor comments.

Comment 1: the aim of the manuscript is unclear

> We are sure that the editors will appreciate that this comment

might stem from the reviewer apparent anger at the manuscript. We would like to point out that the manuscript opens on a very strong rationale for a holistic view of predicting interactions: we will not be able to document all of them, co-occurrence issues are necessarily limiting what data we can gather, and yet we know that knowledge of interactions is key to understanding ecosystem properties – therefore, we need some sort of research agenda for prediction. These points are made, almost verbatim, in the first two paragraphs of the manuscript. Should these be insufficient, we welcome editorial advice about how to best emphasize the aims of the manuscript.

Comment 2: the manuscript lacks substance and needs more references

As we have clarified in the responses to reviewer 1, and made more explicit in the text, this manuscript is not a review; it is a conceptual roadmap that is meant to illustrate our opinion (informed by our active involvement in this area of research), and therefore does not require the same level of exhaustiveness. Reviewer 2 identified articles that we could have cited, but that did not made it into the final version. They have all been covered, some in depth, in recent reviews on ecological networks, some of which were written by co-authors of this manuscript. We do want to point out that the original submission had about 100 references, and the revised version has more; claims that the manuscript is insufficiently referenced are unsupported.

Comment 3: the manuscript focuses both on understanding and prediction

This is correct, and widely regarded as a best practice in predictive ecology. We have added a sentence at the end of the introduction to make explicit the fact that we think of prediction as being fundamentally a component of understanding, and vice versa. We do sincerely hope that readers will get the message that prediction and understanding advance at the same pace.

Comment 4: the flow of the manuscript is nonsensical

We are attributing this comment to an unfair reading of the text. The structure of the manuscript is guided by the conceptual figure, and the structure of the sections are outlined within each section. Based on the detailed comments by reviewer 1, we have identified areas where the flow may be improved, and made the relevant edits.

Comment 5: the proof of concept does not bring new information

Reviewer 1 commented on the proof of concept, and we have expanded this section. Clearly reviewer 2 had a lot of knowledge on this topic and may have gained less from reading the case study, but we do not think that this is necessarily going to be true of the field of network ecology as a whole. For this reason, we have extended the case study and merged it some more into the main text, and notably developed a table to guide the interpretation of the various validation measures.