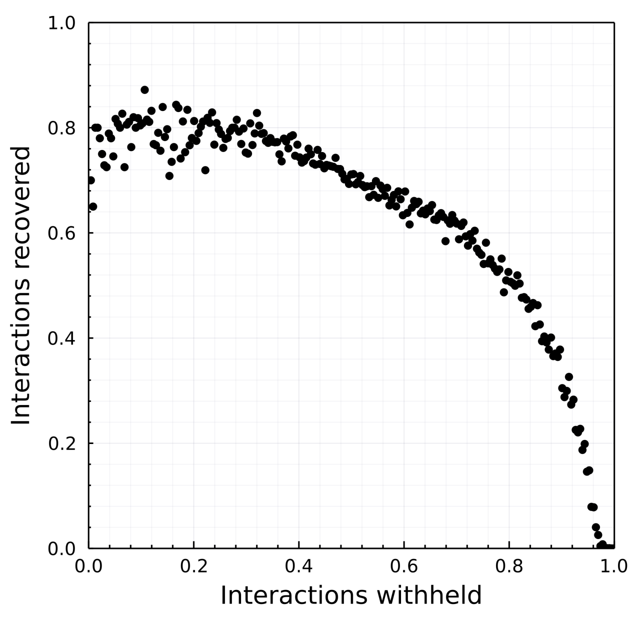
S.M.1 - SVD does not overfit on the European network

In order to ensure that the creation of the RDPG on the European network does not lead to overfitting, we performed the following numerical experiment. First, we extracted $\mathcal{L}$ and $\mathcal{R}$, the left and right subspace of the entire network, at rank 12. Then, for every number n of interactions between 10 and $\text{links}(M)-1$ (where M is the European metaweb), we define m as a network in which n interactions have been randomly removed. We then define $\updownarrow$ and ∇ as the left and right subspaces coming from the rank-12 RDPG applied to this network, and compare the original network M to the one that was reconstructed after thresholding $\updownarrow\nabla$ by picking the cutoff that maximizes Youden's J measure.

This experiment allows masuring the response of various measures of fit of the binary classifier to incomplete sampling. We are specifically interested in (i) the ability of RDPG to identify removed interactions, (ii) the ability of RDPG to function as a performant classifier in the presence of uncertainty in the original data, and (iii) the ability of RDPG to reconstruct biologically realistic data when interactions are withheld.

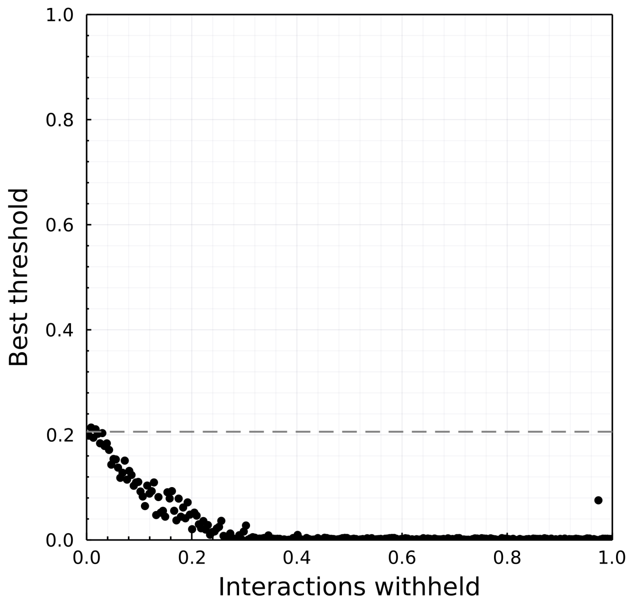
RDPG recovers withheld interactions

When removing up to 40% of interactions in the European metaweb, the RDPG was able to correctly recover 80% of these interactions:



The stochasticity in the proportion of recovered interactions is larger when a small number of interactions are withheld, which makes sense as the *number* of interactions is far smaller (compared to the overall network size).

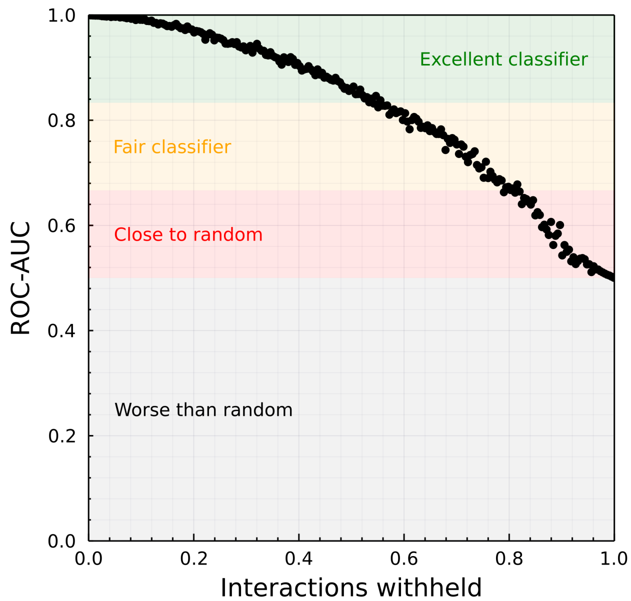
It is interesting to note that the threshold “adapts” to the amount of missing information - the dashed line corresponds to the threshold we used in the manuscript.



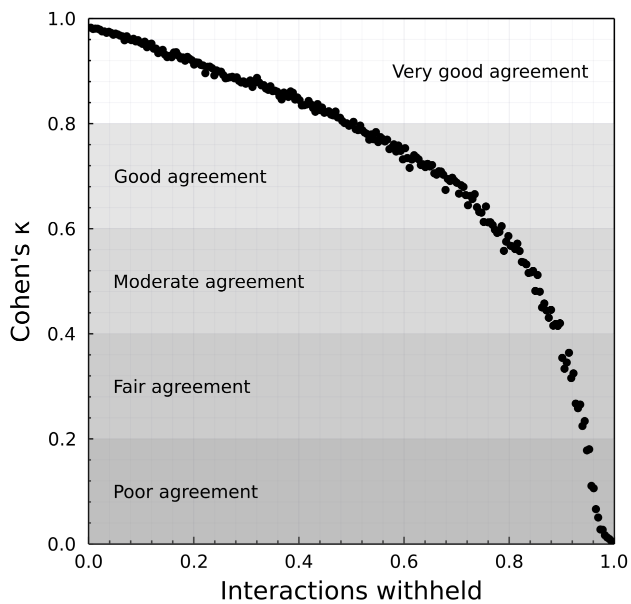
The important consequence of this result is that training the RDPG on a subsample of the network would result in a lower threshold, thereby creating more false positives when applied to the new data.

RDPG yields an accurate classifier

More important than the recovery of removed interaction is the fact that the classifier should not create too much false positives. One measure to assess this is the area under the receiving operator characteristic curve, or ROC-AUC. By this measure, the RDPG remains an excellent classifier even if 50% of interactions are withheld.



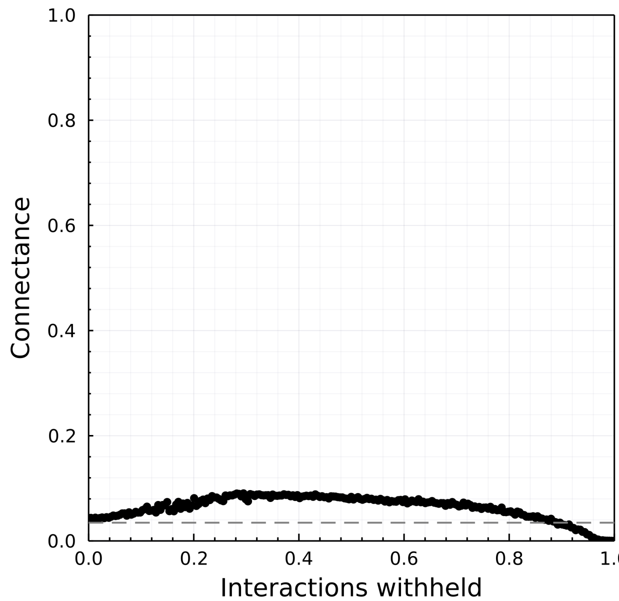
The overall agreement between a classifier and the actual data can be measured by Cohen's κ, which gives a similar result.



These two diagnostic figures reveal that, although we used a probably exhaustive list of interactions to do the initial RDPG, there are chances that the approach would work on less complete datasets.

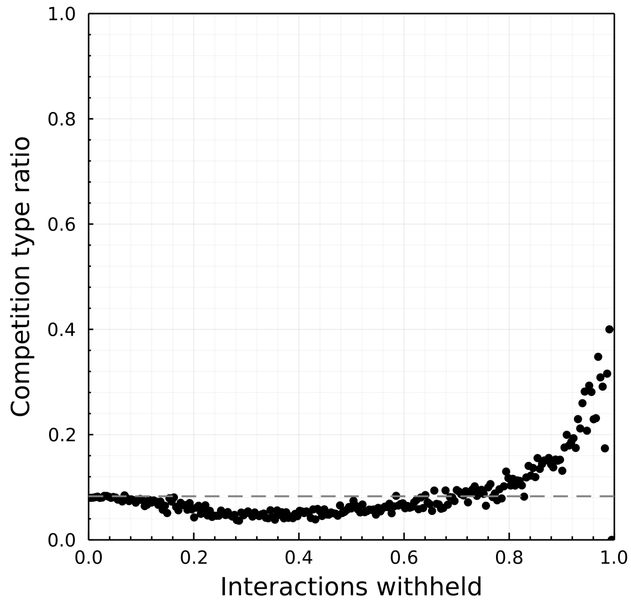
RDPG creates ecologically realistic networks

In this section, we present the relationship between the empirical measure of the network structure (dashed line) and the reconstructed estimate based on RDPG after the optimal threshold has been applied. We focus on connectance (for its broad relevant to food web structure) first:



Connectance increases slightly when initial information is incomplete, but saturates at a value of around 0.12 – this is still within the bounds of connectances expected for food webs.

Next, we look at the ratio between direct competition ($a \rightarrow (b, c)$) and apparent competition ($(a, b) \rightarrow c$) motifs, as motifs are known to be conserved blocks in food webs:



This ratio remains close to the real one up until 75% of initial interactions are lost.

Consequences

Based on these results, applying RDPG on the entire European network is reasonable, especially since the threshold is adapting to the amount of withheld interactions. Interestingly, the RDPG remains an excellent binary classifier even in the face of strong data deficiencies, which suggests that our framework can be used even in the absence of a complete metaweb.