

# Guidelines for the supervised learning of species interactions

Timothée Poisot<sup>1,2</sup>

<sup>1</sup> Université de Montréal; <sup>2</sup> Québec Centre for Biodiversity Sciences

## Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

1. The prediction of species interaction networks is gaining momentum as a way to circumvent limitations in data volume. Yet, ecological networks are challenging to predict because they are typically small and sparse. Dealing with extreme class imbalance is a challenge for most binary classifiers, and there are currently no guidelines as to how predictive models can be trained.
2. Using simple mathematical arguments and numerical experiments in which a variety of classifiers (for supervised learning) are trained on simulated networks, we develop a series of guidelines related to the choice of measures to use for model selection, and the degree of unbiasing to apply to the training dataset.
3. Classifier accuracy and the ROC-AUC are not informative measures for the performance of interaction prediction. PR-AUC is a fairer assesment of performance. In some cases, even standard measures can lead to selecting a more biased classifier because the effect of connectance is strong. The amount of correction to apply to the training dataset depends as a function of the classifier and the network connectance.
4. These results reveal that training machines to predict networks is a challenging task, and that in virtually all cases, the composition of the training set needs to be experimented on before performing the actual training. We discuss these consequences in the context of the low volume of data.

## Keywords:

species interaction networks  
binary classifiers  
machine learning  
regression  
supervised learning

example on diagnostic test: rare events are hard to detect even with really good models

summary of model challenges for networks

list of problems to solve - baseline values and response to bias - effect of training set bias on performance  
- which models need the least amount of interactions to work

summary of the results

1

## Baseline values

In this section, we will assume a network of connectance  $\rho$ , *i.e.* having  $\rho S^2$  interactions (where  $S$  is the species richness), and  $(1 - \rho)S^2$  non-interactions. Therefore, the vector describing the *true* state of the network is  $\vec{o} = [\rho, (1 - \rho)]$ .

2 \_\_\_\_\_

## **Numerical experiments**

2.1. Effect of training set on performance

2.2. Required amount of positives to get the best performance

3 \_\_\_\_\_

## **Guidelines for prediction**

4 \_\_\_\_\_

## **References**