

Guidelines for the supervised learning of species interactions

Timothée Poisot^{1,2}

¹ Université de Montréal ² Québec Centre for Biodiversity Sciences

Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

This work is released by its authors under a CC-BY 4.0 license



Last revision: *December 8, 2021*

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 assessment 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.

- 1 example on diagnostic test: rare events are hard to detect even with really good models
- 2 summary of model challenges for networks - Strydom et al. (2021) importance of drawing on traits +
- 3 validation is challenging - Whalen et al. (2021) machine learning from genomics
- 4 introduction to the confusion matrix
- 5 list of problems to solve - baseline values and response to bias - effect of training set bias on performance -
- 6 which models need the least amount of interactions to work
- 7 summary of the results

8 **Baseline values**

9 **Confusion matrix with skill and bias**

10 In this section, we will assume a network of connectance ρ , *i.e.* having ρS^2 interactions (where S is the
 11 species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the
 12 network is a column vector $\mathbf{o}^T = [\rho(1 - \rho)]$ (we can safely drop the S^2 terms, as we will work on the
 13 confusion matrix, which ends up expressing *relative* values).

14 In order to write the values of the confusion matrix for a hypothetical classifier, we need to define two
 15 characteristics: its skill, and its bias. Skill, here, refers to the propensity of the classifier to get the correct
 16 answer (*i.e.* to assign interactions where they are, and to not assign them where they are not). A no-skill
 17 classifier guesses at random, *i.e.* it will guess interactions with a probability ρ . The predictions of a no-skill
 18 classifier can be expressed as a row vector $\mathbf{p} = [\rho(1 - \rho)]$. The confusion matrix \mathbf{M} for a no-skill classifier
 19 is given by the element-wise product of these vectors $\mathbf{o} \odot \mathbf{p}$, *i.e.*

$$\mathbf{M} = \begin{pmatrix} \rho^2 & \rho(1 - \rho) \\ (1 - \rho)\rho & (1 - \rho)^2 \end{pmatrix}.$$

20 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 21 to s , and off-diagonal elements equal to $(1 - s)$, and re-express the skill-adjusted confusion matrix as
 22 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

$$\begin{pmatrix} \rho^2 & \rho(1-\rho) \\ (1-\rho)\rho & (1-\rho)^2 \end{pmatrix} \odot \begin{pmatrix} s & (1-s) \\ (1-s) & s \end{pmatrix}.$$

23 Note that when $s = 0$, $\text{Tr}(\mathbf{M}) = 0$ (the classifier is *always* wrong), when $s = 0.5$, the classifier is no-skill
 24 and guesses at random, and when $s = 1$, the classifier is perfect.

25 The second element we can adjust in this hypothetical classifier is its bias, specifically its tendency to
 26 over-predict interactions. Like above, we can do so by defining a bias matrix \mathbf{B} , where interactions are
 27 over-predicted with probability b , and express the final classifier confusion matrix as $\mathbf{M} \odot \mathbf{S} \odot \mathbf{B}$, *i.e.*

$$\begin{pmatrix} \rho^2 & \rho(1-\rho) \\ (1-\rho)\rho & (1-\rho)^2 \end{pmatrix} \odot \begin{pmatrix} s & (1-s) \\ (1-s) & s \end{pmatrix} \odot \begin{pmatrix} b & b \\ (1-b) & (1-b) \end{pmatrix}.$$

28 The final expression for the confusion matrix in which we can regulate the skill and the bias is

$$\mathbf{C} = \begin{pmatrix} s \times b \times \rho^2 & (1-s) \times b \times \rho(1-\rho) \\ (1-s) \times (1-b) \times (1-\rho)\rho & s \times (1-b) \times (1-\rho)^2 \end{pmatrix}.$$

29 **What are the baseline values of performance measures?**

30 In this section, we will change the values of b and s for a given value of ρ , and see how the values of
 31 common performance measures for binary classification are affected.

32 **Numerical experiments**

33 In the following section, we will generate random networks, and train four binary classifiers (as well as an
 34 ensemble model using the sum of the outputs) on 30% of the interaction data. Networks are generated by
 35 picking random generality g and vulnerability v traits for $S = 200$ species uniformly on the unit interval,
 36 and assigning an interaction from species i to species j if $0.2g_i - \xi \leq v_j \leq 0.2g_i + \xi$, where ξ is a constant
 37 regulating the connectance of the networks, and varies uniformly in $[5 \times 10^{-3}, 10^{-1}]$. This model gives
 38 fully interval networks that are close analogues to the niche model (Williams & Martinez, 2000), but has

39 the benefit of only relying on two features (g_i, v_j) , and having the exact same rule for all interactions. It is,
40 therefore, a simple case which most classifiers should be able to learn.

41 **Effect of training set on performance**

42 **Required amount of positives to get the best performance**

43 **Guidelines for prediction**

44 **References**

- 45 Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz,
46 N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap
47 towards predicting species interaction networks (across space and time). *Philosophical Transactions of*
48 *the Royal Society B: Biological Sciences*, 376(1837), 20210063.
49 <https://doi.org/10.1098/rstb.2021.0063>
- 50 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine
51 learning in genomics. *Nature Reviews Genetics*, 1–13.
52 <https://doi.org/10.1038/s41576-021-00434-9>
- 53 Williams, R., & Martinez, N. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.
54 <http://userwww.sfsu.edu/>