

# 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](mailto:timothee.poisot@umontreal.ca)

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



Last revision: *December 11, 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.

- example on diagnostic test: rare events are hard to detect even with really good models
- summary of model challenges for networks
- Strydom et al. (2021) importance of drawing on traits + validation is challenging + comparing across space

Binary classifiers are usually assessed by measuring properties of their confusion matrix, *i.e.* the contingency table reporting true/false positive/negative hits. A confusion matrix is laid out as

$$\begin{pmatrix} \text{tp} & \text{fp} \\ \text{fn} & \text{tn} \end{pmatrix},$$

wherein tp is the number of interactions predicted as positive, tn is the number of non-interactions predicted as negative, fp is the number of non-interactions predicted as positive, and fn is the number of interactions predicted as negative. Almost all measures based on the confusion matrix express rates of error or success as proportions, and therefore the values of these components matter in a *relative* way. At a coarse scale, a classifier is *accurate* when the trace of the matrix divided by the sum of the matrix is close to 1, with other measures focusing on different ways in which the classifier is wrong.

The same approach is used to evaluate *e.g.* species distribution models (SDMs). Indeed, the training and evaluation of SDMs as binary classifiers suffers from the same issue of low prevalence. In a previous work, Allouche et al. (2006) suggested that  $\kappa$  was a better test of model performance than the True Skill Statistic (TSS), which we will refer to as Youden's informedness (or  $J$ ); these conclusions were later criticized by Somodi et al. (2017), who emphasized that informedness' relationship to prevalence depends on assumptions about bias in the model, and therefore recommend the use of  $\kappa$  as a validation of classification performance. Although this work offers recommendations about the comparison of models, it doesn't establish baselines or good practices for training on imbalanced ecological data. Within the context of networks, there are three specific issues that need to be addressed. First, what values of performance measures are we expecting for a classifier that has poor performance? This is particularly important as it can evaluate whether low prevalence can lull us into a false sense of predictive accuracy. Second, independently of the question of model evaluation, is low prevalence an issue for *training*, and can we remedy it? Finally, because the low amount of data on interaction makes a lot of imbalance correction methods (see *e.g.* Branco et al., 2015) hard to apply, which indicators can be optimized with the

27 least amount of positive interaction data?

28 In addition to the literature on SDMs, most of the research on machine learning application to life  
29 sciences is focused on genomics (which has very specific challenges, see a recent discussion by Whalen et  
30 al., 2021); this sub-field has generated largely different recommendations. Chicco & Jurman (2020)  
31 suggest using Matthews correlation coefficient (MCC) over  $F_1$ , as a protection against over-inflation of  
32 predicted results; Delgado & Tibau (2019) advocate against the use of Cohen's  $\kappa$ , again in favor of MCC, as  
33 the relative nature of  $\kappa$  means that a worse classifier can be picked over a better one; similarly, Boughorbel  
34 et al. (2017) recommend MCC over other measures of performance for imbalanced data, as it has more  
35 desirable statistical properties. More recently, Chicco et al. (2021) temper the apparent supremacy of the  
36 MCC, by suggesting it should be replaced by Youden's informedness (also known as  $J$ , bookmaker's  
37 accuracy, and the True-Skill Statistic) when the imbalance in the dataset may not be representative  
38 (Jordano, 2016a, which is the case as networks are under-sampled; 2016b), when classifiers need to be  
39 compared across different datasets (for example when predicting a system in space, where undersampling  
40 varies locally; McLeod et al., 2021), and when comparing the results to a no-skill (baseline) classifier is  
41 important. As these conditions are likely to be met with network data, there is a need to evaluate which  
42 measures of classification accuracy respond in a desirable way.

43 A lot of binary classifiers are built by using a regressor (whose task is to guess the value of the interaction,  
44 and can therefore return somethings considered to be a pseudo-probability); in this case, the optimal value  
45 below which predictions are assumed to be negative (*i.e.* the interaction does not exist) can be determined  
46 by picking a threshold maximizing some value on the ROC curve or the PR curve. The area under these  
47 curves (ROC-AUC and PR-AUC henceforth) give ideas on the overall goodness of the classifier. Saito &  
48 Rehmsmeier (2015) established that the ROC-AUC is biased towards over-estimating performance for  
49 imbalanced data; on the contrary, the PR-AUC is able to identify classifiers that are less able to detect  
50 positive interactions correctly, with the additional advantage of having a baseline value equal to  
51 prevalence. Therefore, it is important to assess whether these two measures return different results when  
52 applied to ecological network prediction.

53 We establish that due to the low prevalence of interactions, even poor classifiers applied to food web data  
54 will reach a high accuracy; this is because the measure is dominated by the accidental correct predictions  
55 of negatives. The  $F_1$  score and positive predictive values are less sensitive to bias, but **TODO**

## 56 Baseline values

### 57 Definition of the performance measures

58  $\kappa$

59  $F_\beta$

60 informedness

61 MCC

### 62 Confusion matrix with skill and bias

63 In this section, we will assume a network of connectance  $\rho$ , *i.e.* having  $\rho S^2$  interactions (where  $S$  is the  
64 species richness), and  $(1 - \rho)S^2$  non-interactions. Therefore, the vector describing the *true* state of the  
65 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  
66 confusion matrix, which ends up expressing *relative* values).

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

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

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

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

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

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

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

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

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

82 In all further simulations, the confusion matrix  $\mathbf{C}$  is transformed so that it sums to 1.

### 83 **What are the baseline values of performance measures?**

84 In this section, we will change the values of  $b$ ,  $s$ , and  $\rho$ , and report how the main measures discussed in  
 85 the introduction (MCC,  $F_1$ ,  $\kappa$ , and informedness) are responding to issues with the classifier. Before we do  
 86 so, it is important to explain why we will not focus on accuracy too much. Accuracy is the number of  
 87 correct predictions ( $\text{Tr}(\mathbf{C})$ ) divided by the sum of the confusion matrix. For a no-skill, no-bias classifier,  
 88 accuracy is equal to  $\rho^2 + (1-\rho)^2$ ; for  $\rho = 0.05$ , this is  $\approx 0.90$ , and for  $\rho = 0.01$ , this is equal to  $\approx 0.98$ . In  
 89 other words, the values of accuracy are expected to be so high that they are not really informative (this is  
 90 simply explained by the fact that for  $\rho$  small,  $\rho^2 \ll (1-\rho)^2$ ). More concerning is the fact that introducing  
 91 bias changes the response of accuracy in unexpected ways. Assuming a no-skill classifier, the numerator  
 92 of accuracy becomes  $b\rho^2 + (1-b)(1-\rho)^2$ , which increases when  $b$  is low, which specifically means that at  
 93 equal skill, a classifier that under-predicts interactions will have higher accuracy than an un-biased  
 94 classifier. These issues are absent from balanced accuracy, but should nevertheless lead us to not report  
 95 accuracy as the primary measure of network prediction success; moving forward, we will focus on other  
 96 measures.

97 In order to examine how MCC,  $F_1$ ,  $\kappa$ , and informedness change w.r.t. the imbalance, skill, and bias, we  
98 performed a grid exploration of the values of  $\text{logit}(s)$  and  $\text{logit}(b)$  linearly from  $-10$  to  $10$ , of  $\rho$  linearly in  
99  $[0, 0.5]$ , which is within the range of usually observed connectance values for empirical food webs. Note  
100 that at this point, there is no food web model to speak of; rather, the confusion matrix we discuss can be  
101 obtained for any classification task. Based on the previous discussion, the desirable properties for a  
102 measure of classifier success should be: an increase with classifier skill, especially at low bias; a  
103 hump-shaped response to bias, especially at high skill, and ideally center around  $\text{logit}(b) = 0$ ; an increase  
104 with prevalence up until equiprevalence is reached.

105 [Figure 1 about here.]

106 In fig. 1, we show that none of the four measures satisfy all the considerations at once:  $F_1$  increases with  
107 skill, and increases monotonously with bias; this is because  $F_1$  does not account for true negatives, and the  
108 increase in positive detection masks the over-prediction of interactions. Informedness varies with skill,  
109 reaching 0 for a no-skill classifier, but is entirely unsensitive to bias. Both MCC and  $\kappa$  have the same  
110 behavior, whereby they increase with skill.  $\kappa$  peaks at increasing values of bias for increasing skill, *i.e.* is  
111 likely to lead to the selection of a classifier that over-predicts interactions. By contract, MCC peaks at the  
112 same value, regardless of skill, but this value is not  $\text{logit}(b) = 0$ : unless at very high classifier skill, MCC  
113 risks leading to a model that over-predicts interactions. In fig. 2, we show that all measures except  $F_1$  give  
114 a value of 0 for a no-skill classifier, and are forced towards their correct maximal value when skill changes  
115 (*i.e.* a more connected networks will have higher values for a skilled classifier, and lower values for a  
116 classifier making mostly mistakes).

117 [Figure 2 about here.]

118 These two analyses point to the following recommendations: MCC is indeed more appropriate than  $\kappa$ , as  
119 although sensitive to bias, it is sensitive in a consistent way. Informedness is appropriate at discriminating  
120 between different skills, but confounded by bias.  $F_1$  is increasing with bias, and should not be prioritized  
121 to evaluate the performance of the model. The discussion of sensitivity to bias should come with a  
122 domain-specific caveat: although it is likely that interactions documented in ecological networks are  
123 correct, a lot of non-interactions are simply unobserved; as predictive models are used for data-inflation

(i.e. the prediction of new interactions), it is not necessarily a bad thing to select models that predict more interactions than the original dataset, because the original dataset misses some interactions.

## Numerical experiments

In the following section, we will generate random networks, and train four binary classifiers (as well as an ensemble model using the sum of the outputs) on 30% of the interaction data. Networks are generated by picking random generality  $g$  and vulnerability  $v$  traits for  $S = 200$  species uniformly on the unit interval, and assigning an interaction from species  $i$  to species  $j$  if  $0.2g_i - \xi \leq v_j \leq 0.2g_i + \xi$ , where  $\xi$  is a constant regulating the connectance of the networks, and varies uniformly in  $[5 \times 10^{-3}, 10^{-1}]$ . This model gives fully interval networks that are close analogues to the niche model (Williams & Martinez, 2000), but has the benefit of only relying on two features  $(g_i, v_j)$ , and having the exact same rule for all interactions. It is, therefore, a simple case which most classifiers should be able to learn.

The training sample is composed of 30% of the  $4 \times 10^4$  possible entries in the network, i.e.  $n = 12000$ . Out of these interactions, we pick a proportion  $\nu$  (the training set bias) to be positive, so that the training set has  $\nu n$  interactions, and  $(1 - \nu)n$  non-interactions. We vary  $\nu$  uniformly in  $]0, 1[$ . This allows to evaluate how the measures of binary classification performance respond to artificially rebalanced dataset for a given network connectance. Note that both  $\xi$  and  $\nu$  are sampled from a distribution rather than being picked on a grid; this is because there is no direct relationship between the value of  $\xi$  and the connectance of the simulated network, and therefore the precise value of  $\xi$  is not relevant for the analysis of the results.

The dataset used for numerical experiments is composed of 20000 such  $(\xi, \nu)$  pairs, on which four learners are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random forest regressor. All models were taken from the `MLJ.jl` package (Blaom et al., 2020; Blaom & Vollmer, 2020) in Julia 1.7 (Bezanson et al., 2017). In order to pick the best adjacency matrix for a given learner, we performed a thresholding approach using 500 steps on predictions from the testing set, and picking the threshold that maximized Youden's informedness, which is usually the optimized target for imbalanced classification. During the thresholding step, we measured the area under the receiving-operator characteristic (ROC-AUC) and precision-recall (PR-AUC) curves, as measures of overall performance over the range of returned values. We report the ROC-AUC and PR-AUC, as well as a suite of other measures as introduced in the next section, for the best threshold. The ensemble model was generated by summing the



152 predictions of all component models on the testing set (ranged in  $[0, 1]$ ), then put through the same  
153 thresholding process. The complete code to run the simulations is given as an appendix.

154 After the simulations were completed, we removed all runs (*i.e.* pairs of  $\xi$  and  $\nu$ ) for which at least one of  
155 the following conditions was met: the accuracy was 0, the true positive or true negative rates were 0, the  
156 connectance was larger than 0.2. This removes both the obviously failed model runs, and the networks  
157 that are more densely connected compared to the connectance of empirical food webs (and are therefore  
158 less difficult to predict, being less imbalanced).

### 159 **Effect of training set bias on performance**

160 In fig. 3, we present the response of MCC and informedness to (i) four levels of network connectance and  
161 (ii) a gradient of training set bias, for the four component models as well as the ensemble. All models  
162 reached a higher performance on more connected networks, and using more biased training sets (with the  
163 exception of ridge regression, whose informedness decreased in performance with training set bias). In all  
164 cases, informedness was extremely high, which is an expected consequence of the fact that this is the  
165 value we optimized to determine the cutoff. MCC increased with training set bias, although this increase  
166 became less steep with increasing connectance. Interestingly, the ensemble almost always outclassed its  
167 component models.

168 [Figure 3 about here.]

169 In fig. 4, we present the same information as fig. 3, this time using ROC-AUC and PR-AUC. ROC-AUC is  
170 always high, and does not vary with training set bias. On the other hand, PR-AUC shows very strong  
171 responses, increasing with training set bias. It is notable here that two classifiers that seemed to be  
172 performing well (Decision Tree and Random Forest) based on their MCC are not able to reach a high  
173 PR-AUC even at higher connectances. As in fig. 3, the ensemble outperforms its component models.

174 [Figure 4 about here.]

175 Based on the results presented in fig. 3 and fig. 4, it seems that informedness and ROC-AUC are not  
176 necessarily able to discriminate between good and bad classifiers (although this result may be an artifact  
177 for informedness, as it has been optimized when thresholding). On the other hand, MCC and PR-AUC  
178 show a strong response to training set bias, and may therefore be more useful at model comparison.

## Required amount of positives to get the best performance

The previous results revealed that the measure of classification performance responds both to the bias in the training set *and* to the connectance of the network; from a practical point of view, assembling a training set requires to withhold positive information, which in ecological networks are very scarce (and typically more valuable than negatives, on which there is a doubt). For this reason, across all values of connectance, we measured the training set bias that maximized a series of performance measures. When this value is high, the training set needs to skew positive in order to get a good model; when this value is about 0.5, the training set needs to be artificially balanced to optimize the model performance. These results are presented in fig. 5.

[Figure 5 about here.]

Interestingly, as long as the connectance of the network was above  $\approx 0.1$ , the optimal prevalence in the training set is 0.5, *i.e.* as many positives as negatives. Low connectance is usually achieved for very large networks, due to the scaling relationship between richness and links (MacDonald et al., 2020). Therefore, larger networks may require *more* biasing of the training set in order to be optimally predicted, whereas smaller, more connected networks may not. It is worth noting that the optimal bias for the training set stabilizes at 0.5 regardless of connectance *and* model *and* measure of model evaluation.

[Figure 6 about here.]

## Guidelines for prediction

The results presented here highlight an interesting paradox: larger networks more bias, smaller networks less bias but fewer int to begin with: it's difficult either way

INF because we trust positives more than negative, but check with MCC

PR-AUC

usually default to training set bias of 0.5

## References

- Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>
- Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, 5(55), 2704. <https://doi.org/10.21105/joss.02704>
- Blaom, A. D., & Vollmer, S. J. (2020, December 31). *Flexible model composition in machine learning and its implementation in MLJ*. <http://arxiv.org/abs/2012.15505>
- Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS One*, 12(6), e0177678. <https://doi.org/10.1371/journal.pone.0177678>
- Branco, P., Torgo, L., & Ribeiro, R. (2015, May 13). *A Survey of Predictive Modelling under Imbalanced Distributions*. <http://arxiv.org/abs/1505.01658>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14, 13. <https://doi.org/10.1186/s13040-021-00244-z>
- Delgado, R., & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PloS One*, 14(9), e0222916. <https://doi.org/10.1371/journal.pone.0222916>
- Jordano, P. (2016a). Chasing Ecological Interactions. *PLOS Biol*, 14(9), e1002559. <https://doi.org/10.1371/journal.pbio.1002559>
- Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*.

229 <https://doi.org/10.1111/1365-2435.12763>

230 MacDonald, A. A. M., Banville, F., & Poisot, T. (2020). Revisiting the Links-Species Scaling Relationship in  
 231 Food Webs. *Patterns*, 1(0). <https://doi.org/10.1016/j.patter.2020.100079>

232 McLeod, A., Leroux, S. J., Gravel, D., Chu, C., Cirtwill, A. R., Fortin, M.-J., Galiana, N., Poisot, T., & Wood,  
 233 S. A. (2021). Sampling and asymptotic network properties of spatial multi-trophic networks. *Oikos*,  
 234 n/a(n/a). <https://doi.org/10.1111/oik.08650>

235 Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot  
 236 When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432.  
 237 <https://doi.org/10.1371/journal.pone.0118432>

238 Somodi, I., Lepesi, N., & Botta-Dukát, Z. (2017). Prevalence dependence in model goodness measures with  
 239 special emphasis on true skill statistics. *Ecology and Evolution*, 7(3), 863–872.  
 240 <https://doi.org/10.1002/ece3.2654>

241 Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz,  
 242 N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap  
 243 towards predicting species interaction networks (across space and time). *Philosophical Transactions of*  
 244 *the Royal Society B: Biological Sciences*, 376(1837), 20210063.  
 245 <https://doi.org/10.1098/rstb.2021.0063>

246 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine  
 247 learning in genomics. *Nature Reviews Genetics*, 1–13.  
 248 <https://doi.org/10.1038/s41576-021-00434-9>

249 Williams, R., & Martinez, N. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.  
 250 <http://userwww.sfsu.edu/>

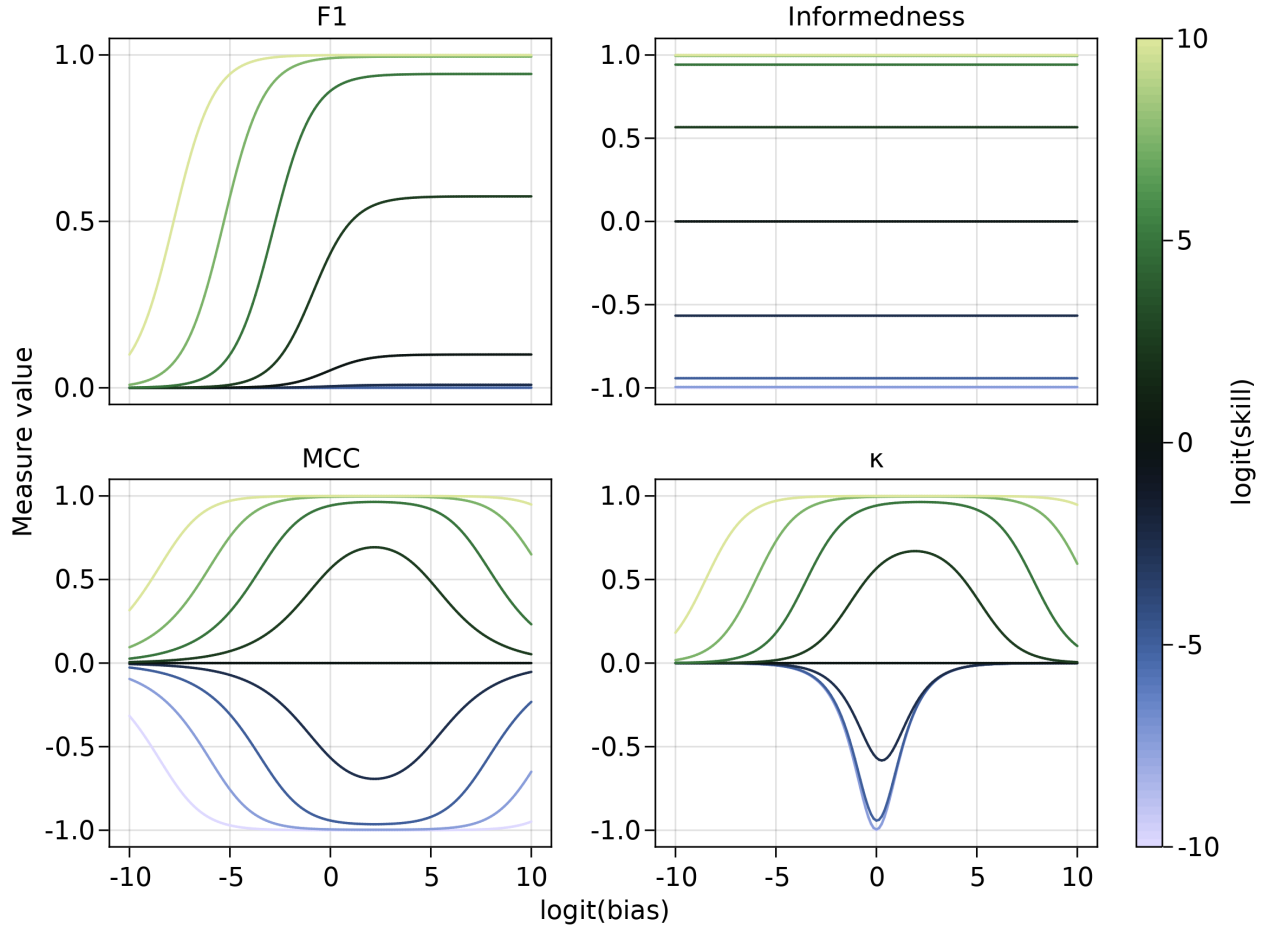


Figure 1: Consequences of changing the classifier skills ( $s$ ) and bias ( $b$ ) for a connectance  $\rho = 0.15$ , on accuracy,  $F_1$ , positive predictive value, and  $\kappa$ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions. The  $F_1$  score increases with skill but also increases when the bias tends towards estimating *more* interactions; PPV behaves in the same way. Interestingly,  $\kappa$  responds as expected to skill (being negative whenever  $s < 0.5$ ), and peaks for values of  $b \approx 0.5$ ; nevertheless, the value of bias for which  $\kappa$  is maximized is *not*  $b = 0.5$ , but instead increases with classifier skill. In other words, at equal skill, maximizing  $\kappa$  would lead to select a *more* biased classifier.

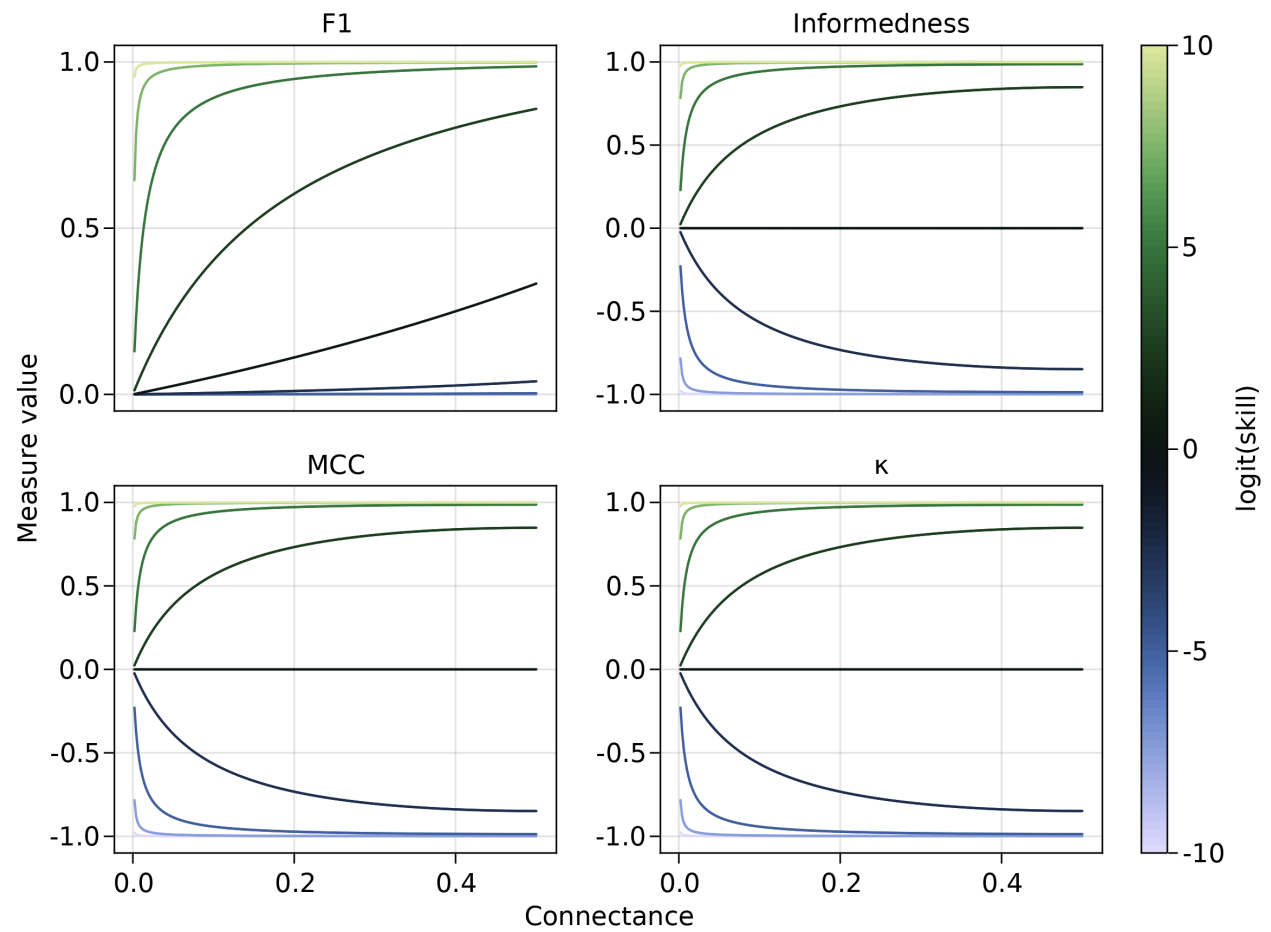


Figure 2: TODO

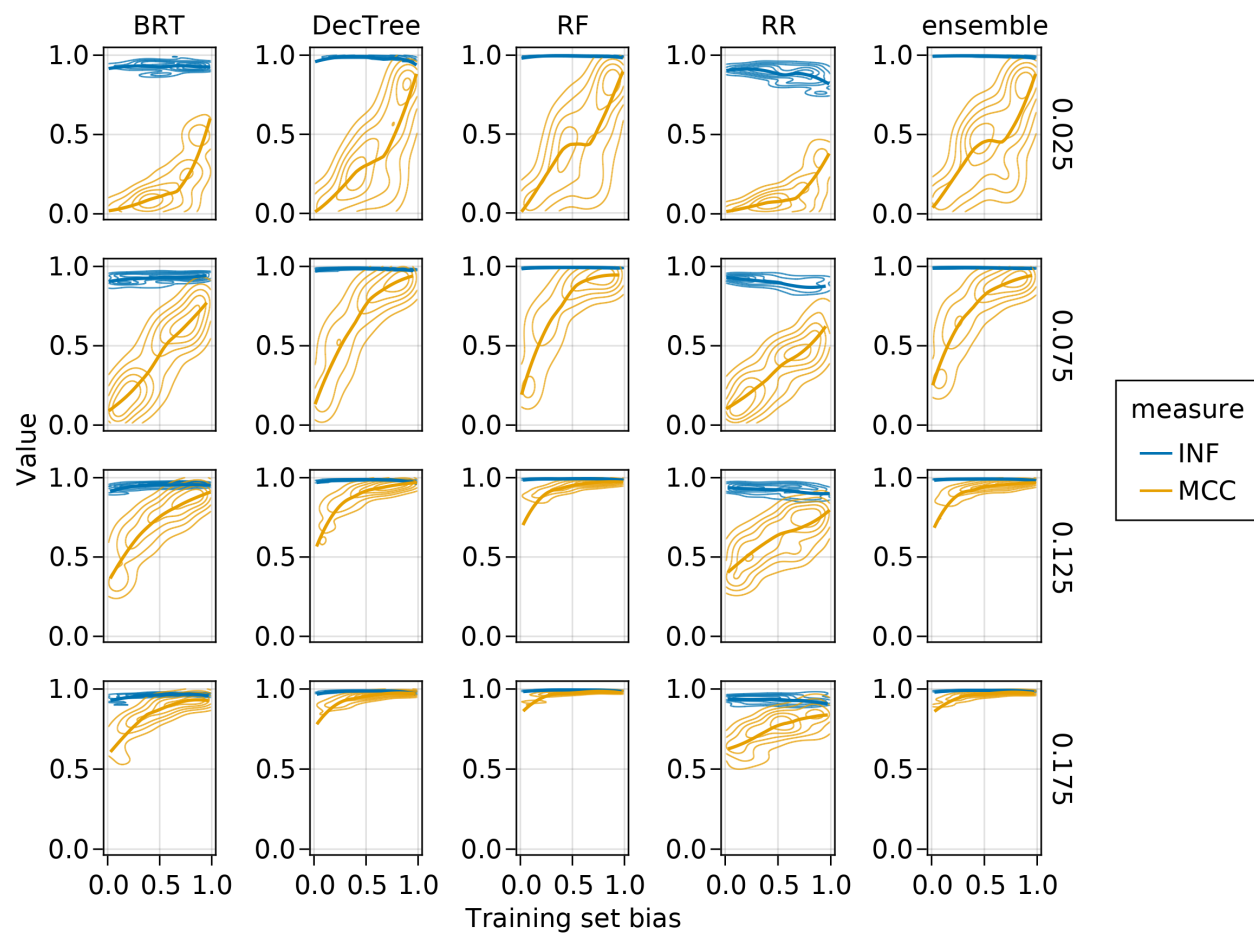


Figure 3: TODO

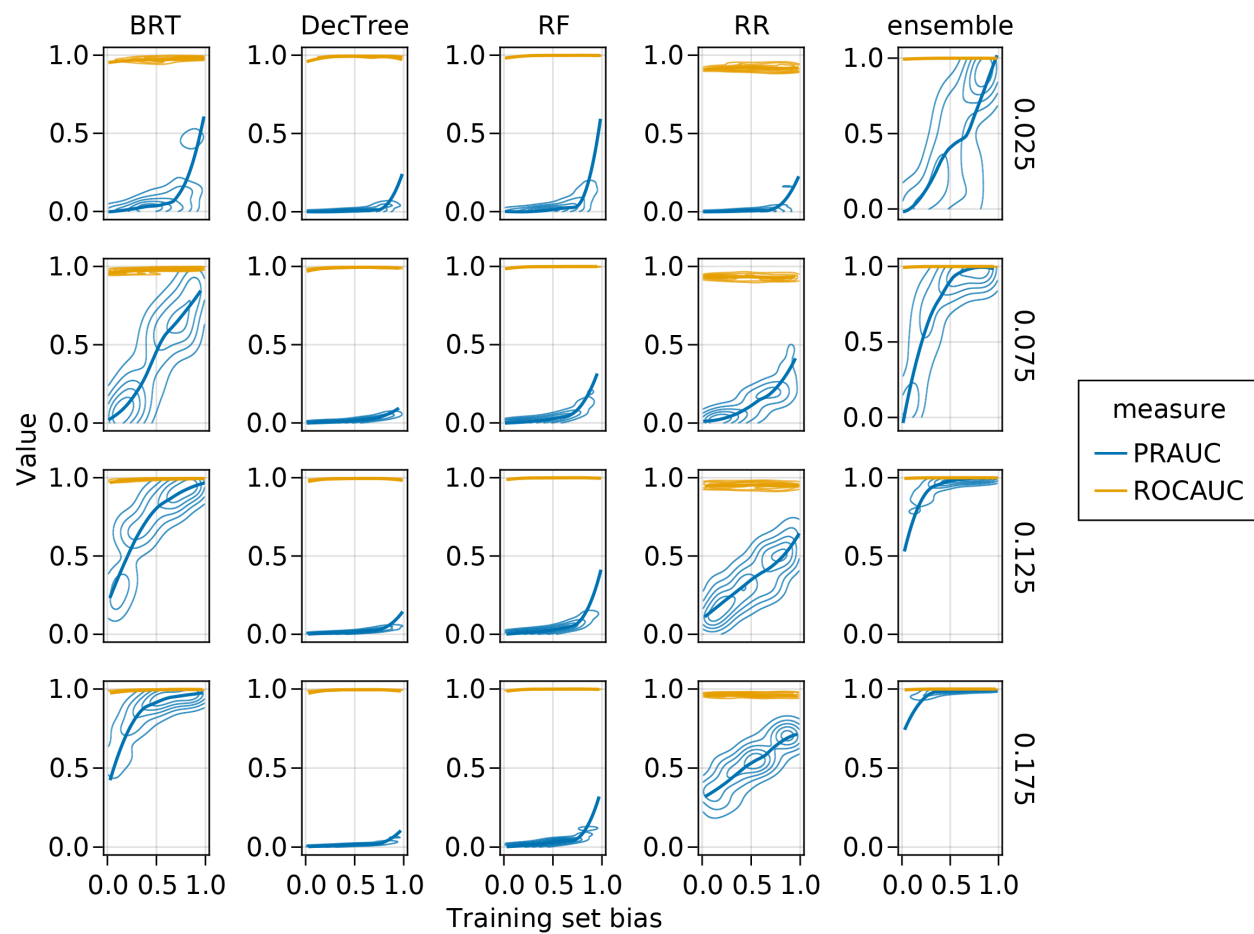


Figure 4: TODO



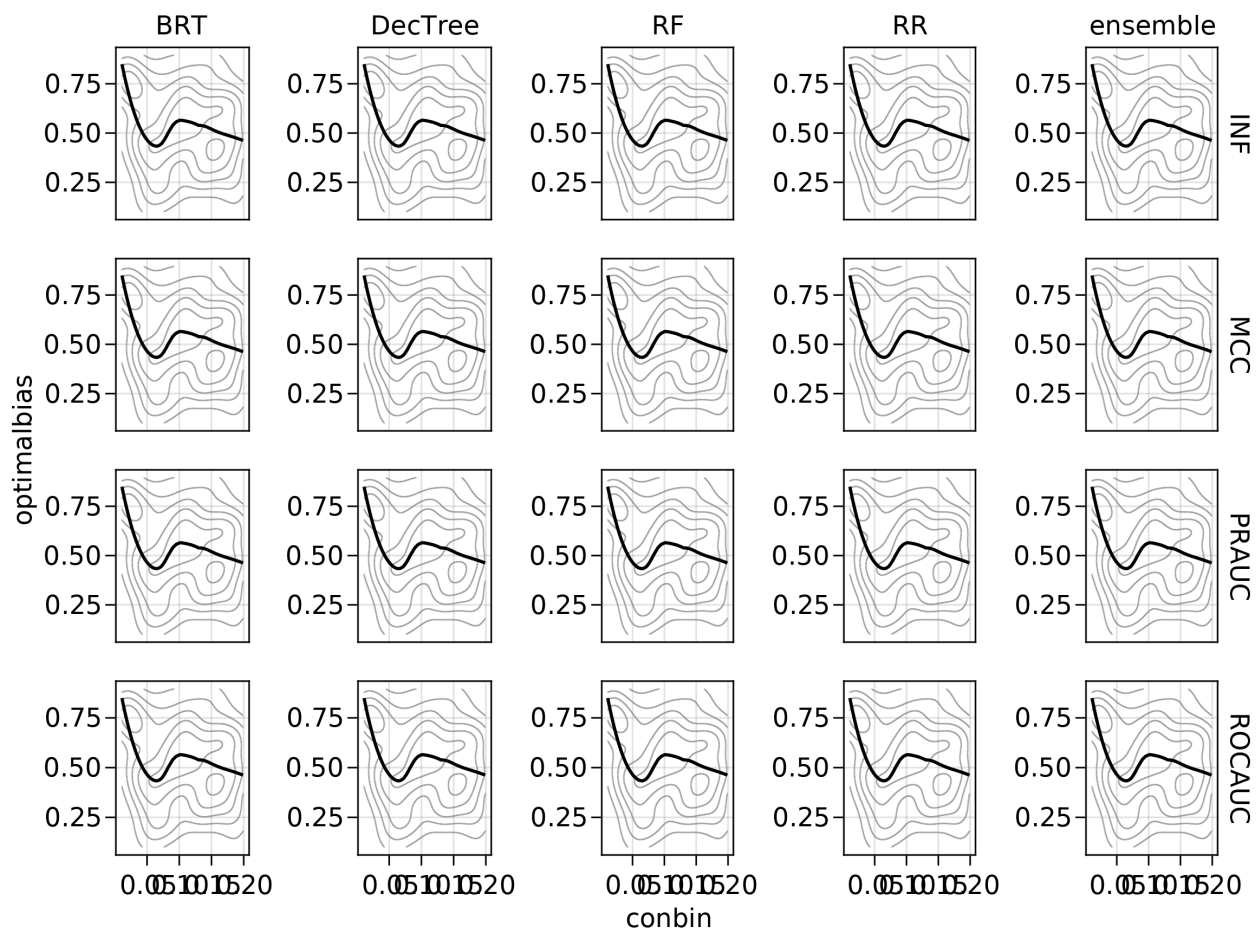


Figure 5: TODO

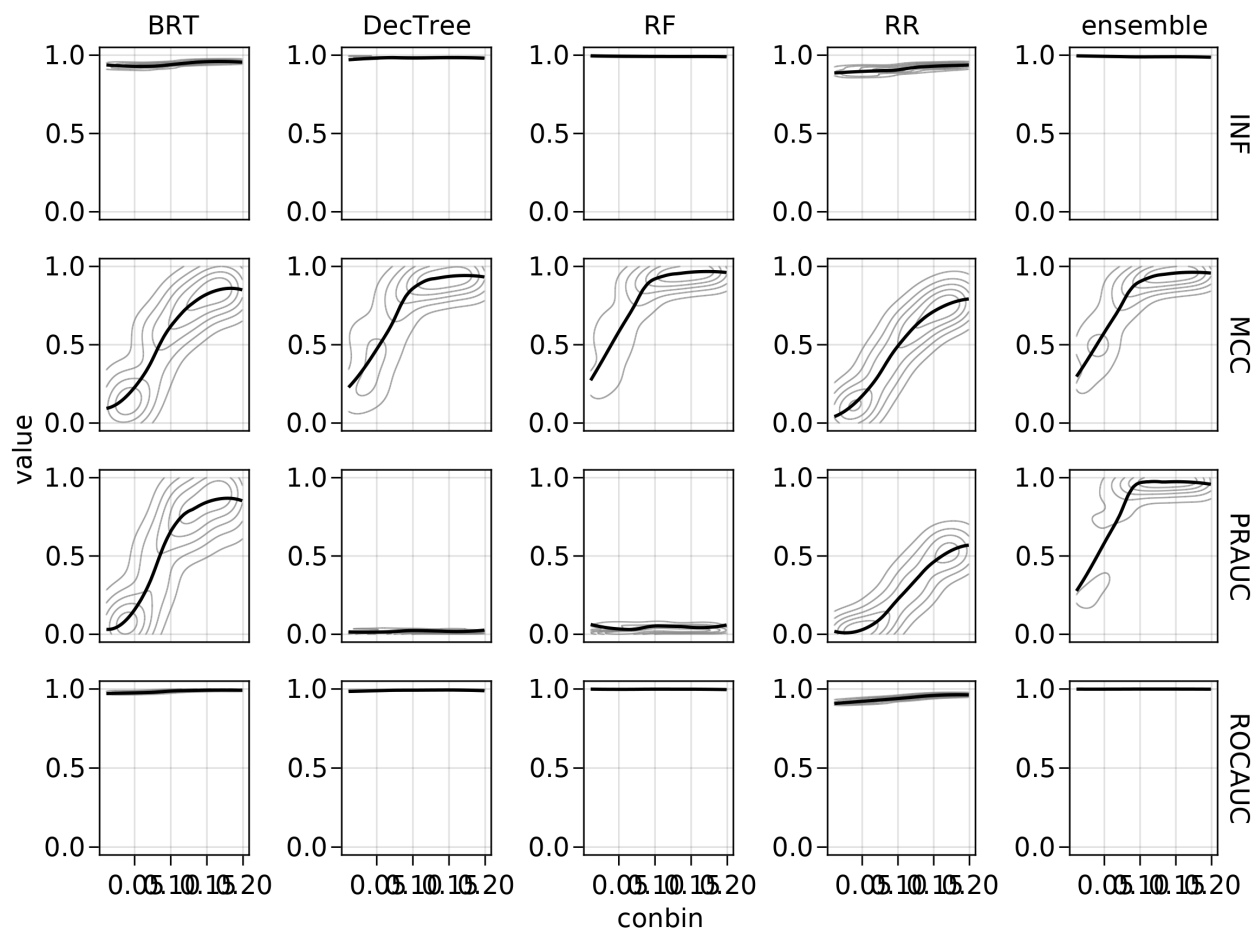


Figure 6: TODO