

Validating machine learning predictions 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 21, 2021*

1. The prediction of species interactions 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 for this specific problem.
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 on network connectance, on the measure to be optimized, and only weakly on the classifier.
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 The accuracy paradox is the basis of a number of problems in statistical education, and lies in the fact that,
2 when the desired class is rare, a model that gets less and less performant will become more and more
3 accurate and useful, simply by (i) underpredicting true positive cases and (ii) over-predicting false
4 negatives. In other words, accuracy, defined as the proportion of predictions that are correct, is often
5 useless as a measure of how predictive a model is. This is particularly true in ecological networks; the
6 desired class (presence of an interaction between two species) is the one we care most about, and by far
7 the least common. Herein lies the core challenge of predicting species interactions: the extreme
8 imbalance between classes makes the training of predictive models difficult. In a recent contribution,
9 Strydom et al. (2021) highlight that predictive models of interactions can likely be improved by adding
10 information (in the form of, *e.g.* traits), but that we do not have robust guidelines as to how the predictive
11 ability of these models should be evaluated, nor about how the models should be trained. Here, by relying
12 on simple derivations and a series of simulations, we formulate a number of such guidelines, specifically
13 for the case of binary classifiers derived from thresholded values.

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

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

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

22 There is an immense amount of measures to evaluate the performance of classification tasks (Ferri et al.,
23 2009). Here we will focus on five of them with high relevance for imbalanced learning (He & Ma, 2013);
24 three threshold metrics (κ , informedness, and the Matthews Correlation Coefficient), and two ranking
25 metrics (the ROC-AUC and PR-AUC). The κ measure of agreement (Landis & Koch, 1977) establishes the
26 extent to which two observers (here the data and the prediction) agree, and is measured as

$$2 \frac{tp \times tn - fn \times fp}{(tp + fp) \times (fp + tn) + (tn + fp) \times (tn + fn)}.$$

27 Informedness (Youden, 1950) (also known as bookmaker informedness or the True Skill Statistic) is
 28 $TPR + TNR - 1$, where $TPR = tp/(tp + fn)$ and $TNR = tn/(tn + fp)$; informedness can be used to find
 29 the optimal cutpoint in thresholding analyses (Schisterman et al., 2005). The MCC is defined as

$$\frac{tp \times tn - fn \times fp}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}}.$$

30 Finally, F_1 is the harmonic mean of precision (the chance that a positive even was correctly classified) and
 31 sensitivity (the ability to correctly classify positive events), and is defined as

$$2 \frac{tp}{2 \times tp + fp + fn}.$$

32 A lot of binary classifiers are built by using a regressor (whose task is to guess the value of the interaction,
 33 and can therefore return somethins considered to be a pseudo-probability); in this case, the optimal value
 34 below which predictions are assumed to be negative (*i.e.* the interaction does not exist) can be determined
 35 by picking a threshold maximizing some value on the ROC curve or the PR curve. The area under these
 36 curves (ROC-AUC and PR-AUC henceforth) give ideas on the overall goodness of the classifier. Saito &
 37 Rehmsmeier (2015) established that the ROC-AUC is biased towards over-estimating performance for
 38 imbalanced data; on the contrary, the PR-AUC is able to identify classifiers that are less able to detect
 39 positive interactions correctly, with the additional advantage of having a baseline value equal to
 40 prevalence. Therefore, it is important to assess whether these two measures return different results when
 41 applied to ecological network prediction. The ROC curve is defined by the false positive rate on the x axis,
 42 and the true positive rate on the y axis, and the PR curve is defined by the true positive rate on the x axis,
 43 and the positive predictive value on the y axis. By comparison with the previous paragraph, it is obvious
 44 that F_1 has ties to the PR curve (being close to the expected PR-AUC), and that informedness has ties to
 45 the ROC curve (whereby the threshold maximizing informedness is also the point of maximal inflection
 46 on the ROC curve). One important difference between ROC and PR is that the later does not prominently
 47 account for the size of the true negative compartments: in short, it is more sensitive to the correct positive
 48 predictions. In a context of strong imbalance, PR-AUC is therefore a more stringent test of model

49 performance.

50 The same approach is used to evaluate *e.g.* species distribution models (SDMs). Indeed, the training and
51 evaluation of SDMs as binary classifiers suffers from the same issue of low prevalence. In previous work,
52 Allouche et al. (2006) suggested that κ was a better test of model performance than the True Skill Statistic
53 (TSS; which we refer to as Youden's informedness); these conclusions were later criticized by Somodi et al.
54 (2017), who emphasized that informedness' relationship to prevalence depends on assumptions about bias
55 in the model, and therefore recommend the use of κ as a validation of classification performance.
56 Although this work offers recommendations about the comparison of models, it doesn't establishes
57 baselines or good practices for training on imbalanced ecological data. Within the context of networks,
58 there are three specific issues that need to be addressed. First, what values of performance measures are we
59 expecting for a classifier that has poor performance? This is particularly important as it can evaluate
60 whether low prevalence can lull us into a false sense of predictive accuracy. Second, independently of the
61 question of model evaluation, is low prevalence an issue for *training*, and can we remedy it? Finally,
62 because the low amount of data on interaction makes a lot of imbalance correction methods (see *e.g.*
63 Branco et al., 2015) hard to apply, which indicators can be optimized with the least amount of positive
64 interaction data?

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

79 measures of classification accuracy respond in a desirable way.

80 We establish that due to the low prevalence of interactions, even poor classifiers applied to food web data
81 will reach a high accuracy; this is because the measure is dominated by the accidental correct predictions
82 of negatives. On simulated confusion matrices with ranges of imbalance that are credible for ecological
83 networks, MCC had the most desirable behavior, and informedness is a linear measure of classifier skill.
84 By performing simulations with four models and an ensemble, we show that informedness and ROC-AUC
85 are consistently high on network data, and that MCC and PR-AUC are more accurate measures of the
86 effective performance of the classifier. Finally, by measuring the structure of predicted networks, we
87 highlight an interesting paradox: the models with the best performance measures are not the models with
88 the closest reconstructed network structure. We discuss these results in the context of establishing
89 guidelines for the prediction of ecological interactions.

90 **Baseline values**

91 In this section, we will assume a network of connectance ρ , *i.e.* having ρS^2 interactions (where S is the
92 species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the
93 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
94 confusion matrix, which ends up expressing *relative* values). We will apply skill and bias to this matrix,
95 and measure how a selection of performance metrics respond to changes in these values, in order to assess
96 their suitability for model evaluation.

97 **Confusion matrix with skill and bias**

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

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

104 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 105 to s , and off-diagonal elements equal to $(1-s)$, and re-express the skill-adjusted confusion matrix as
 106 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

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

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

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

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

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

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

114 What are the baseline values of performance measures?

115 In this section, we will change the values of b , s , and ρ , and report how the main measures discussed in
 116 the introduction (MCC, F_1 , κ , and informedness) are responding to issues with the classifier. Before we do
 117 so, it is important to explain why we will not focus on accuracy too much. Accuracy is the number of

correct predictions ($\text{Tr}(\mathbf{C})$) divided by the sum of the confusion matrix. For a no-skill, no-bias classifier, accuracy is equal to $\rho^2 + (1 - \rho)^2$; for $\rho = 0.05$, this is ≈ 0.90 , and for $\rho = 0.01$, this is equal to ≈ 0.98 . In other words, the values of accuracy are expected to be so high that they are not really informative (this is simply explained by the fact that for ρ small, $\rho^2 \ll (1 - \rho)^2$). More concerning is the fact that introducing bias changes the response of accuracy in unexpected ways. Assuming a no-skill classifier, the numerator of accuracy becomes $b\rho^2 + (1 - b)(1 - \rho)^2$, which increases when b is low, which specifically means that at equal skill, a classifier that under-predicts interactions will have higher accuracy than an un-biased classifier. These issues are absent from balanced accuracy, but should nevertheless lead us to not report accuracy as the primary measure of network prediction success; moving forward, we will focus on other measures.

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

[Figure 1 about here.]

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

147 classifier making mostly mistakes).

148 [Figure 2 about here.]

149 These two analyses point to the following recommendations: MCC is indeed more appropriate than κ , as
150 although sensitive to bias, it is sensitive in a consistent way. Informedness is appropriate at discriminating
151 between different skills, but confounded by bias. F_1 is increasing with bias, and should not be prioritized
152 to evaluate the performance of the model. The discussion of sensitivity to bias should come with a
153 domain-specific caveat: although it is likely that interactions documented in ecological networks are
154 correct, a lot of non-interactions are simply unobserved; as predictive models are used for data-inflation
155 (*i.e.* the prediction of new interactions), it is not necessarily a bad thing in practice to select models that
156 predict more interactions than the original dataset, because the original dataset misses some interactions.
157 Furthermore, the weight of positive interactions could be adjusted if some information about the extent of
158 undersampling exists (*e.g.* Branco et al., 2015).

159 **Numerical experiments on training strategy**

160 In the following section, we will generate random bipartite networks (this works without loss of generality
161 on unipartite networks), and train four binary classifiers (as well as an ensemble model using the sum of
162 ranged outputs from the component models) on 30% of the interaction data. Networks are generated by
163 picking a random infectiousness trait v_i for 100 species (from a $B(6, 8)$ distribution), and a resistance trait
164 h_j for 100 species (from a $B(2, 8)$ distribution). There is an interaction between i and j when
165 $v_i - \xi/2 \leq h_j \leq v_i + \xi/2$, where ξ is a constant regulating the connectance of the network (there is an
166 almost 1:1 relationship between ξ and connectance), and varies uniformly in $[0.05, 0.35]$. This model gives
167 fully interval networks that are close analogues to the bacteria–phage model of Weitz et al. (2005), with
168 both a modular structure and a non-uniform degree distribution. This model is easy to learn: when
169 trained with features $[v_i, h_j, \text{abs}(v_i, h_j)]^T$ to predict the interactions between i and j , all four models
170 presented below were able to reach almost perfect predictions all the time (data not presented here) – this
171 is in part because the rule is fixed for all interactions. In order to make the problem more difficult to solve,
172 we use $[v_i, h_j]$ as a feature vector, and therefore the models will have to uncover that the rule for
173 interaction is $\text{abs}(v_i, h_j) \leq \xi$.

174 The training sample is composed of 30% of the 10^4 possible entries in the network, *i.e.* $n = 3000$. Out of
175 these interactions, we pick a proportion ν (the training set bias) to be positive, so that the training set has
176 νn interactions, and $(1 - \nu)n$ non-interactions. We vary ν uniformly in $]0, 1[$. This allows to evaluate how
177 the measures of binary classification performance respond to artificially rebalanced dataset for a given
178 network connectance. The rest of the dataset ($n = 7000$ pairs of species) is used as a testing set, on which
179 all further measures are calculated. Note that although the training set is balanced, the testing set is not,
180 and retains (part of) the imbalance of the original data.

181 The dataset used for numerical experiments is composed of 64000 such (ξ, ν) pairs, on which four
182 machines are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random
183 forest regressor. All models were taken from the MLJ.jl package (Blaom et al., 2020; Blaom & Vollmer,
184 2020) in Julia 1.7 (Bezanson et al., 2017). All machines use the default parameterization; this is an obvious
185 deviation from best practices, as the hyperparameters of any machine require training before its
186 application on a real dataset. As we use 64000 such datasets, this would require 256000 unique instances
187 of tweaking the hyperparameters, which is not realistic. Therefore, we assume that the default
188 parameterizations are comparable across networks. All machines return a quantitative prediction, usually
189 (but not necessarily) in $[0, 1]$, which is proportional (but not necessarily linearly) to the probability of an
190 interaction between i and j .

191 In order to pick the best adjacency matrix for a given trained machine, we performed a thresholding
192 approach using 500 steps on predictions from the testing set, and picking the threshold that maximized
193 Youden's informedness, which is usually the optimized target for imbalanced classification. During the
194 thresholding step, we measured the area under the receiving-operator characteristic (ROC-AUC) and
195 precision-recall (PR-AUC) curves, as measures of overall performance over the range of returned values.
196 We report the ROC-AUC and PR-AUC, as well as a suite of other measures as introduced in the next
197 section, for the best threshold. The ensemble model was generated by summing the predictions of all
198 component models on the testing set (ranged in $[0, 1]$), then put through the same thresholding process.
199 The complete code to run the simulations is given as an appendix; running the final simulation required
200 4.8 core days (approx. 117 hours).

201 After the simulations were completed, we removed all runs (*i.e.* pairs of ξ and ν) for which at least one of
202 the following conditions was met: the accuracy was 0, the true positive or true negative rates were 0, the
203 connectance was larger than 0.25. This removes both the obviously failed model runs, and the networks

that are more densely connected compared to the connectance of empirical food webs (and are therefore less difficult to predict, being less imbalanced; preliminary analyses of data with a connectance larger than 3 revealed that all machines reached consistently high performance).

Effect of training set bias on performance

In fig. 3, we present the response of MCC and informedness to (i) five levels of network connectance and (ii) a gradient of training set bias, for the four component models as well as the ensemble. All models reached a higher performance on more connected networks, and using more biased training sets (with the exception of ridge regression, whose informedness decreased in performance with training set bias). In all cases, informedness was extremely high, which is an expected consequence of the fact that this is the value we optimized to determine the cutoff. MCC increased with training set bias, although this increase became less steep with increasing connectance. Interestingly, the ensemble almost always outclassed its component models. In a few cases, both MCC and informedness started decreasing when the training set bias got too close to one, which suggests that it is possible to over-correct the imbalance.

[Figure 3 about here.]

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

[Figure 4 about here.]

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

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

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

237 [Figure 5 about here.]

238 The more “optimistic” measures (ROC-AUC and informedness) required a biasing of the dataset from
239 about 0.4 to 0.75 to be maximized, with the amount of bias required decreasing only slightly with the
240 connectance of the original network. MCC and PR-AUC required values of training set bias from 0.75 to
241 almost 1 to be optimized, which is in line with the results of the previous section, *i.e.* they are more
242 stringent tests of model performance.

243 [Figure 6 about here.]

244 When trained at their optimal training set bias, performance still had a significant impact on the
245 performance of some machines fig. 6. Notably, Decision Tree, Random Forest, and Ridge Regression had
246 low values of PR-AUC. In all cases, the Boosted Regression Tree was reaching very good predictions
247 (especially for connectances larger than 0.1), and the ensemble was almost always scoring perfectly. This
248 suggests that all the models are biased in different ways, and that the averaging in the ensemble is able to
249 correct these biases. We do not expect this last result to have any generality, and provide a discussion of a
250 recent exemple in which the ensemble was performing worse than its components models.

251 **Do better classification accuracy result in more realistic networks?**

252 In this last section, we generate a network using the same model as before, with $S_1, S_2 = 50, 80$ species, a
253 connectance of ≈ 0.16 ($\xi = 0.19$), and a training set bias of 0.7. The prediction made on the complete

dataset is presented in fig. 7. Visualizing the results this way highlights the importance of exploratory data analysis: whereas all models return a network with interactions laying mostly on the diagonal (as expected), the Ridge Regression is quite obviously biased. Despite this, we can see that the ensemble is close to the initial dataset.

[Figure 7 about here.]

The trained models were then thresholded (again by optimising informedness), and their predictions transformed back into networks for analysis; specifically, we measured the connectance, nestedness (REF), and modularity (REF). This process was repeated 250 times, and the results are presented in tbl. 1. The random forest model is an interesting instance here: it produces the network that looks the most like the original dataset, despite having a very low PR-AUC, suggesting it hits high recall at the cost of low precision. Although the ensemble was about to reach a very high PR-AUC (and a very high ROC-AUC), this did not necessarily translate into more accurate reconstructions of the structure of the network. This result bears elaborating. Measures of model performance capture how much of the interactions and non-interactions are correctly identified. As long as these predictions are not perfect, some interactions will be predicted at the “wrong” position in the network; these measures cannot describe the structural effect of these mistakes. On the other hand, measures of network structure can have the same value with interactions that fall at drastically different positions; this is in part because a lot of these measures covary with connectance, and in part because as long as these values are not 0 or their respective maximum, there is a large number of network configurations that can have the same value. That ROC-AUC is consistently larger than PR-AUC may be a case of this measure masking models that are not, individually, strong predictors (Jeni et al., 2013).

Table 1: Values of four performance metrics, and three network structure metrics, for 250 independent predictions similar to the ones presented in fig. 7. The values in **bold** indicate the best value for each column (including ties). Because the values have been rounded, values of 1.0 for the ROC-AUC column indicate an average ≥ 0.99 .

Model	MCC	Inf.	ROC-AUC	PR-AUC	Conn.	η	Q
Decision tree	0.85	0.92	0.97	0.12	0.21	0.76	0.31
BRT	0.90	0.90	0.98	0.86	0.23	0.82	0.27
Random Forest	0.90	0.96	1.00	0.27	0.20	0.72	0.32

	Model	MCC	Inf.	ROC-AUC	PR-AUC	Conn.	η	Q
	Ridge Regression	0.80	0.91	0.95	0.58	0.24	1.0	0.18
	Ensemble	0.88	0.94	1.00	0.96	0.20	0.75	0.31
	Data					0.18	0.66	0.34

275 Guidelines for the assesment of network predictive models

276 The results presented here highlight an interesting paradox: although the Random Forest was ultimately
277 able to get a correct estimate of network structure tbl. 1, it ultimately remains a poor classifier, as
278 evidenced by its low PR-AUC. This suggests that the goal of predicting *interactions* and predicting
279 *networks* may not be solvable in the same way – of course a perfect classifier of interactions would make a
280 perfect network prediction; but even the best scoring predictor of interactions (the ensemble model) had
281 not necessarilly the best prediction of network structure. The tasks of predicting networks structure and of
282 predicting interactions within networks are essentially two different ones. For some applications (*e.g.*
283 comparison of network structure across gradients), one may care more about a robust estimate of the
284 structure, at the cost at putting some interactions at the wrong place. For other applications (*e.g.*
285 identifying pairs of interacting species), one may conversely care more about getting as many pairs right,
286 even though the mistakes accumulate in the form of a slightly worse estimate of network structure. How
287 these two approaches can be reconciled is undoubtedly a task for further research. Despite this apprent
288 tension at the heart of the predictive exercise, we can use the results presented here to suggest a number of
289 guidelines.

290 First, because we should have more trust in reported interactions than in reported absences of interactions,
291 we can draw on previous literature to recommend informedness as a measure to decide on a threshold
292 (Chicco et al., 2021); this being said, because informedness is insensitive to bias, the model performance is
293 better evaluated through the use of MCC fig. 3. Because F_1 is monotonously sensitive to classifier bias
294 fig. 1 and network connectance fig. 2, MCC should be prefered as a measure of model evaluation.

295 Second, because the PR-AUC responds more to network connectance fig. 6 and training set imbalance
296 fig. 4, it should be used as a measure of model performance over the ROC-AUC. This is not to say that
297 ROC-AUC should be discarded (in fact, a low ROC-AUC is a sign of an issue with the model), but that its

298 interpretation should be guided by the PR-AUC value. Specifically, a high ROC-AUC is not informative, as
299 it can be associated to a low PR-AUC (see *e.g.* Random Forest in tbl. 1) This again echoes
300 recommendations from other fields (Jeni et al., 2013; Saito & Rehmsmeier, 2015).

301 Thirdly, regardless of network connectance, maximizing informedness required a training set bias of about
302 0.5, and maximizing the MCC required a training set bias of 0.7 and more. This has an important
303 consequence in ecological networks, for which the pool of positive cases (interactions) to draw from is
304 typically small: the most parsimonious measure (*i.e.* the one requiring to discard the least amount of
305 information to train the model) will give the best validation potential, and is probably the informedness
306 (maximizing informedness is the generally accepted default for imbalanced classification; Schisterman et
307 al., 2005).

308 Finally, it is noteworthy that the ensemble model was systematically better than the component models;
309 even when the models were individually far from perfect, the ensemble was able to leverage the different
310 biases expressed by the models to make an overall more accurate prediction. We do not expect that
311 ensembles will *always* be better than single models. In a recent multi-model comparison, Becker et al.
312 (2021) found that the ensemble was *not* the best model. There is no general conclusion to draw from this
313 besides reinforcing the need to be pragmatic about which models should be included in the ensemble, or
314 whether to use an ensemble at all. In a sense, the surprising performance of the ensemble model should
315 form the basis of the last recommendation: optimal training set bias and its interaction with connectance
316 and binary classifier is, in a sense, an hyperparameter that should be assessed. The distribution of results
317 in fig. 5 and fig. 6 show that there are variations around the trend; furthermore, networks with different
318 structures than the one we simulated here may respond in different ways.

319 **Acknowledgements:** We acknowledge that this study was conducted on land within the traditional
320 unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and
321 Omàmiwininiwak nations. This research was enabled in part by support provided by Calcul Québec
322 (www.calculquebec.ca) and Compute Canada (www.computeCanada.ca) through the Narval general
323 purpose cluster. TP is supported by a NSERC Discovery Grant and Discovery Acceleration Supplement,
324 and by a grant from the Institut de Valorisation des Données (IVADO).

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>
- Becker, D., Albery, G. F., Sjödin, A. R., Poisot, T., Bergner, L., Dallas, T., Eskew, E. A., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., Stock, M., Teeling, E. C., & Carlson, C. J. (2021). Optimizing predictive models to prioritize viral discovery in zoonotic reservoirs. *bioRxiv*, 2020.05.22.111344. <https://doi.org/10.1101/2020.05.22.111344>
- 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>

353 Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance
 354 measures for classification. *Pattern Recognition Letters*, 30(1), 27–38.
 355 <https://doi.org/10.1016/j.patrec.2008.08.010>

356 He, H., & Ma, Y. (Eds.). (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications* (1st
 357 edition). Wiley-IEEE Press.

358 Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing Imbalanced Data—Recommendations for the Use
 359 of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent*
 360 *Interaction*, 245–251. <https://doi.org/10.1109/ACII.2013.47>

361 Jordano, P. (2016a). Chasing Ecological Interactions. *PLOS Biol*, 14(9), e1002559.
 362 <https://doi.org/10.1371/journal.pbio.1002559>

363 Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*.
 364 <https://doi.org/10.1111/1365-2435.12763>

365 Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data.
 366 *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>

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

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

373 Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal Cut-point and Its Corresponding
 374 Youden Index to Discriminate Individuals Using Pooled Blood Samples. *Epidemiology*, 16(1), 73–81.
 375 <https://doi.org/10.1097/01.ede.0000147512.81966.ba>

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

379 Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz,
 380 N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap

381 towards predicting species interaction networks (across space and time). *Philosophical Transactions of*
382 *the Royal Society B: Biological Sciences*, 376(1837), 20210063.
383 <https://doi.org/10.1098/rstb.2021.0063>

384 Weitz, J. S., Hartman, H., & Levin, S. A. (2005). Coevolutionary arms races between bacteria and
385 bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27),
386 9535–9540. <https://doi.org/10.1073/pnas.0504062102>

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

390 Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
391 [https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3)

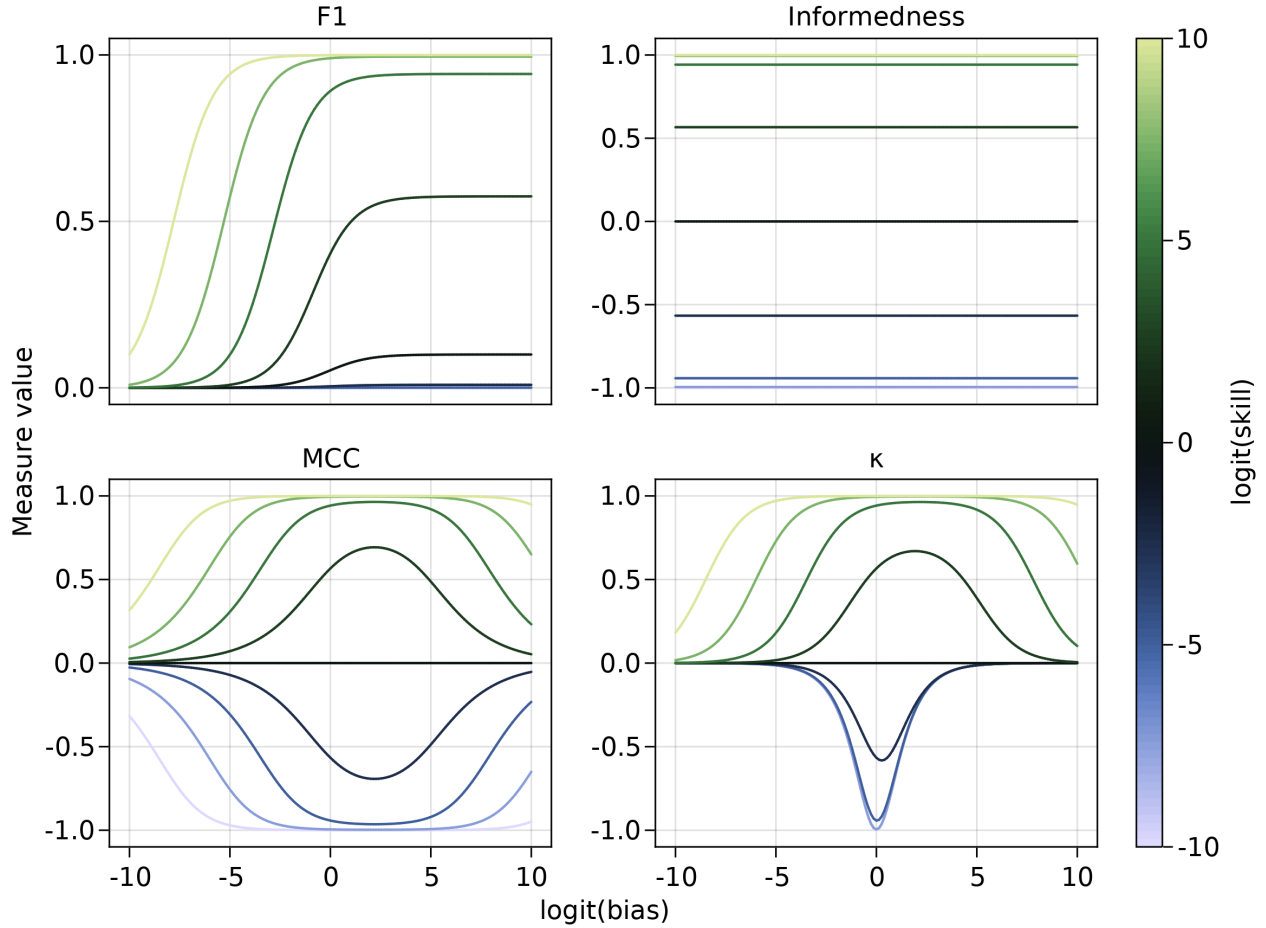


Figure 1: Consequences of changing the classifier skills (s) and bias (b) for a connectance $\rho = 0.15$, on accuracy, F_1 , positive predictive value, and κ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions. The F_1 score increases with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

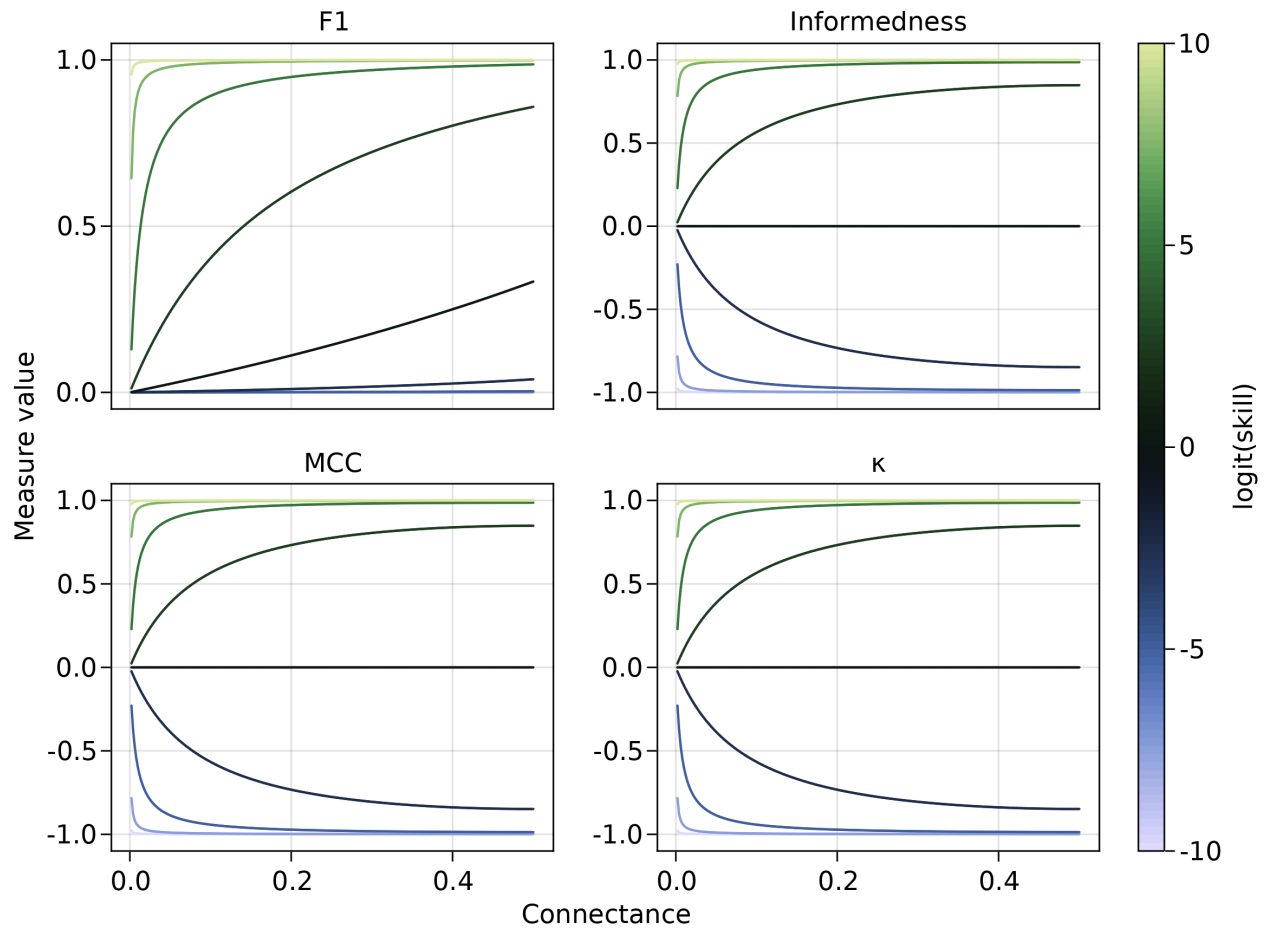


Figure 2: As in fig. 1, consequences of changing connectance for different levels of classifier skill, assuming no classifier bias. Informedness, κ , and MCC do increase with connectance, but only when the classifier is not no-skill; by way of contrast, a more connected network will give a higher F_1 value even with a no-skill classifier.

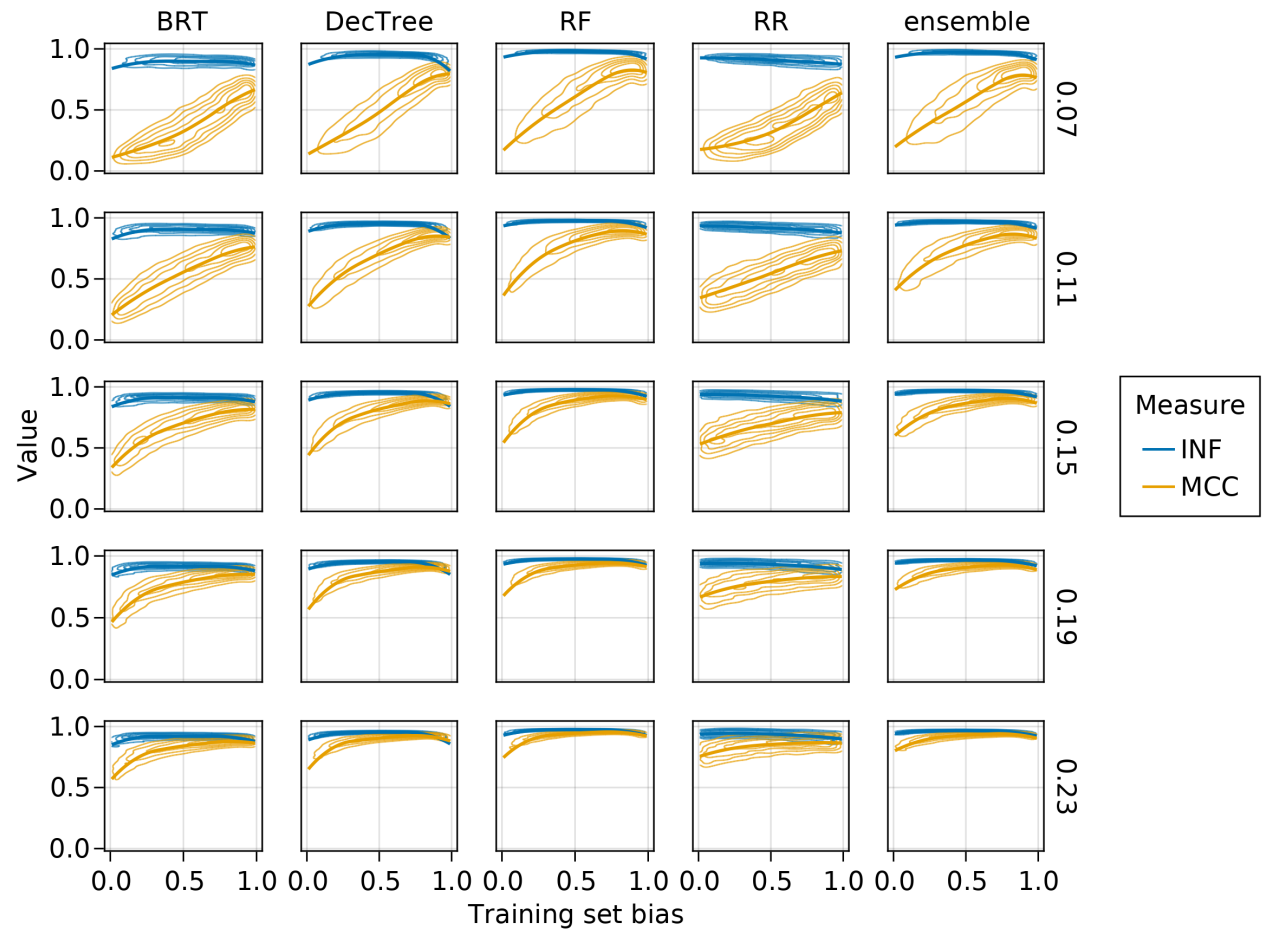


Figure 3: TODO

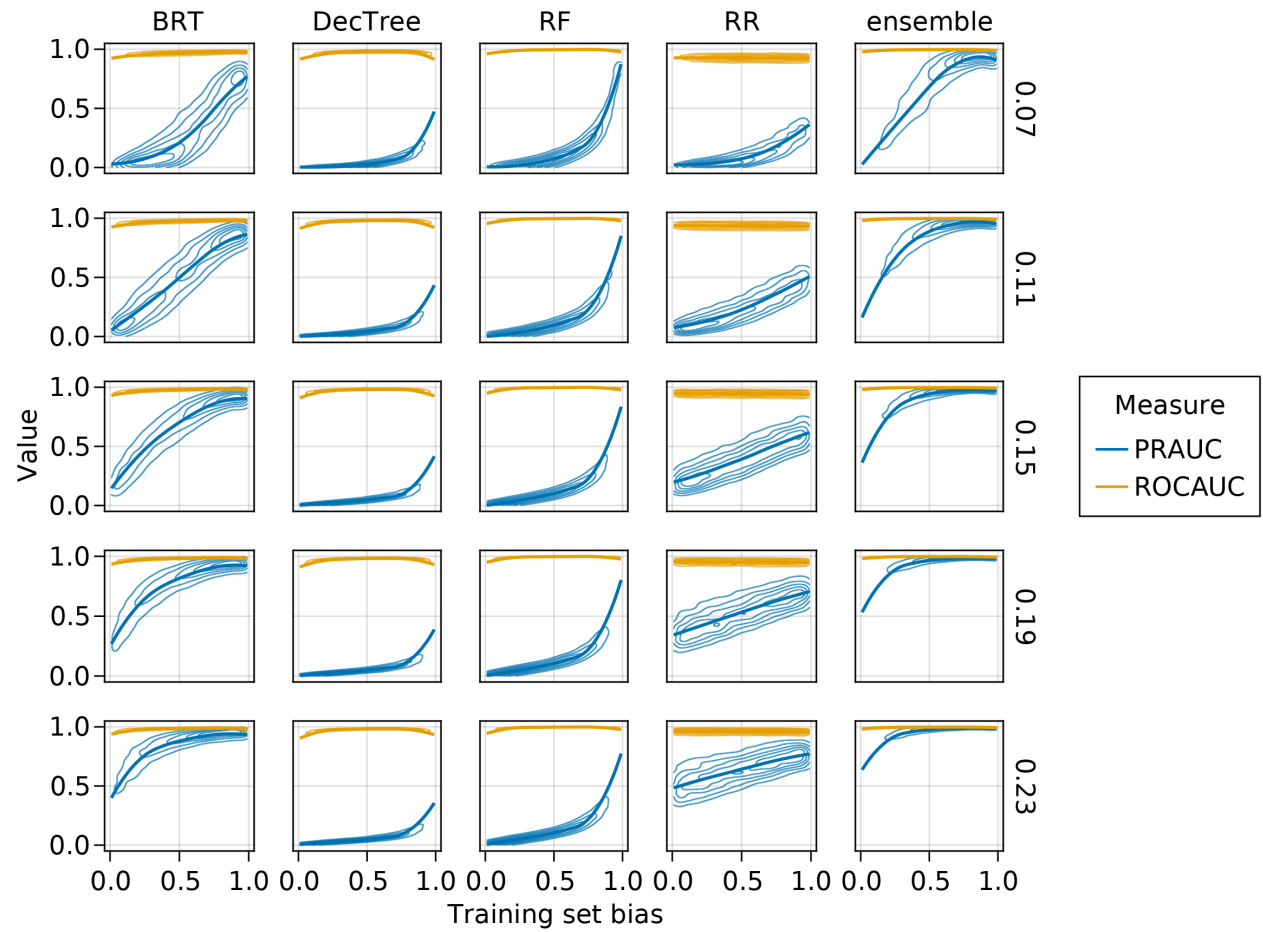


Figure 4: TODO

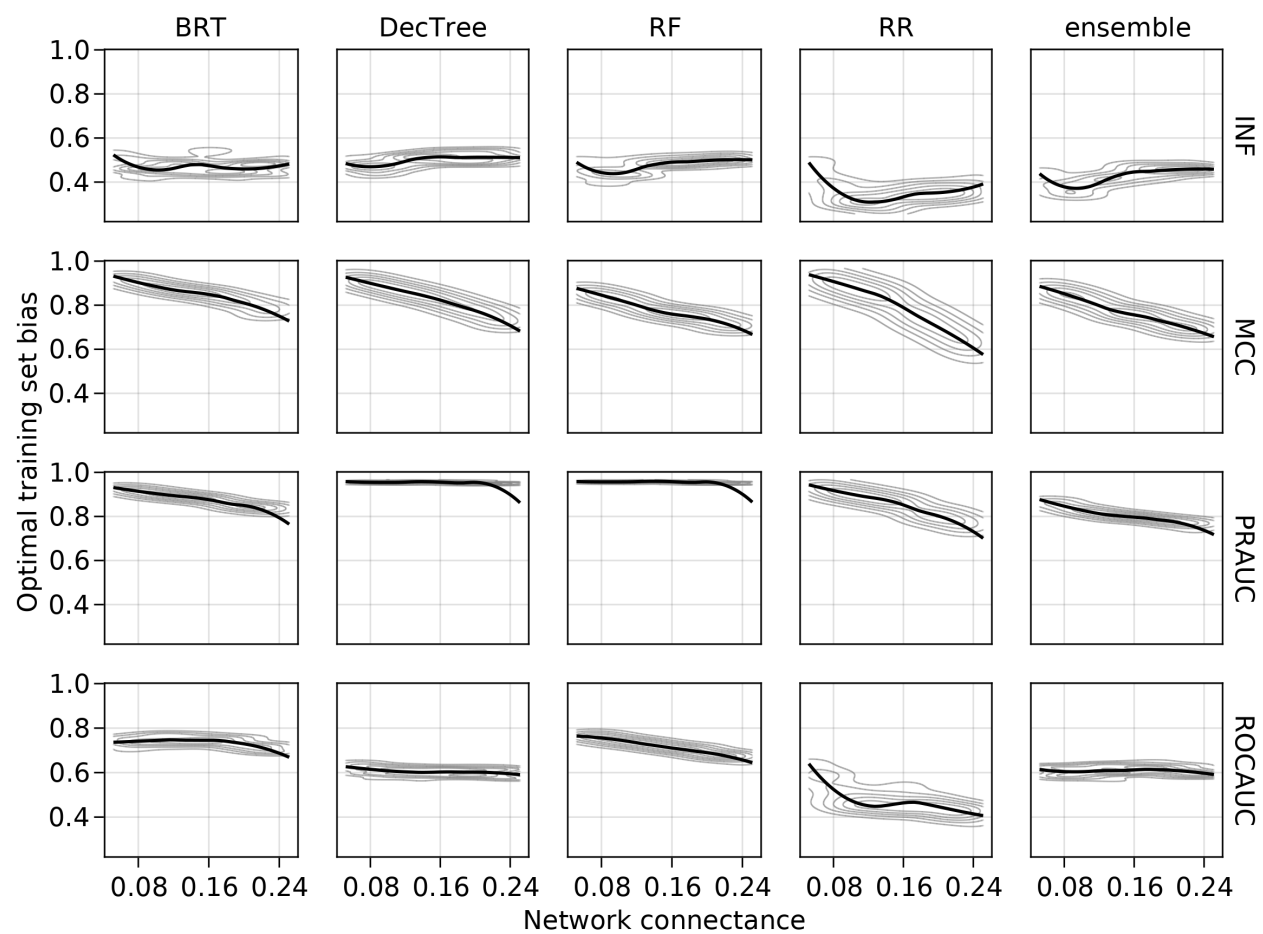


Figure 5: TODO

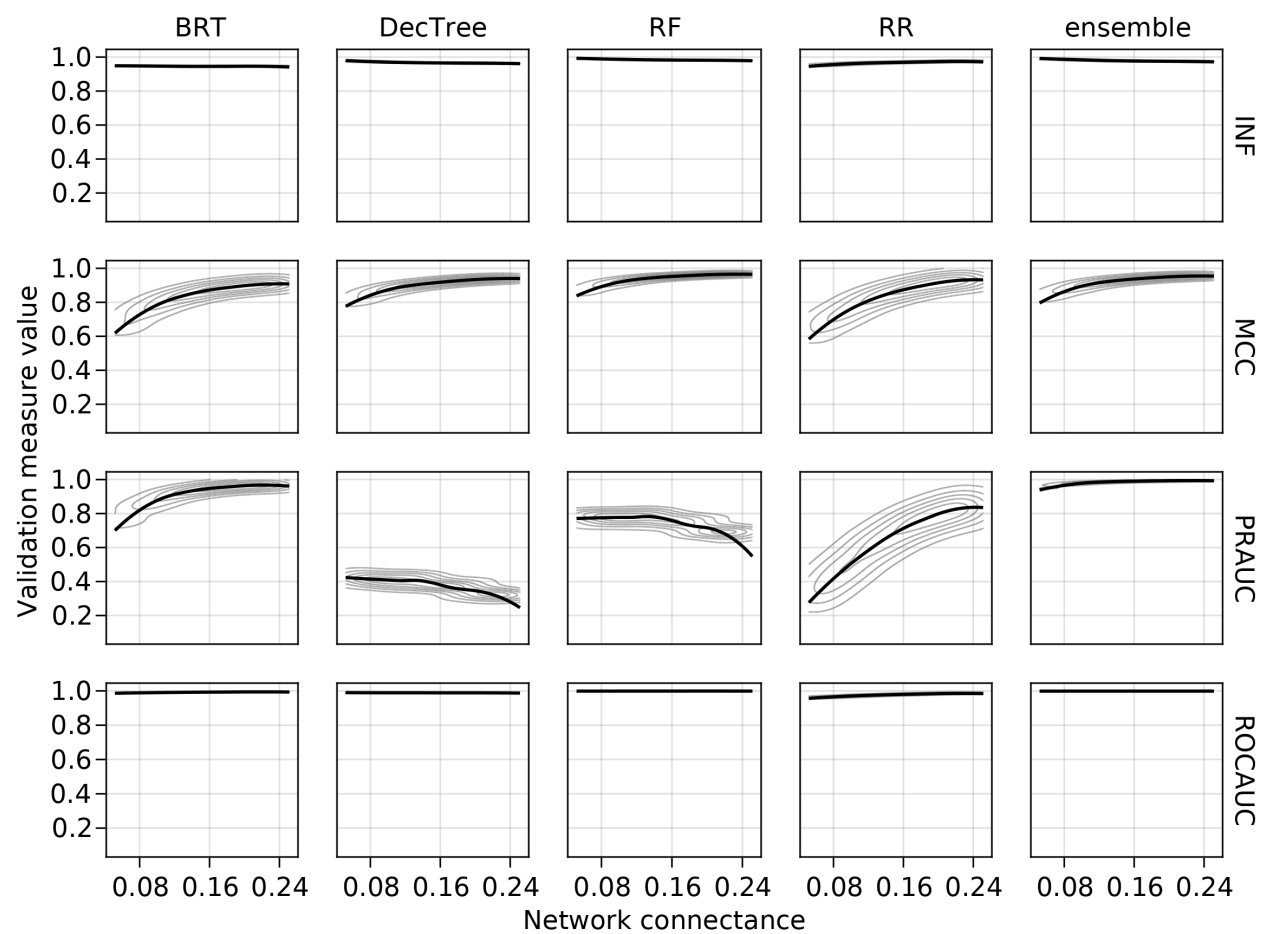


Figure 6: TODO

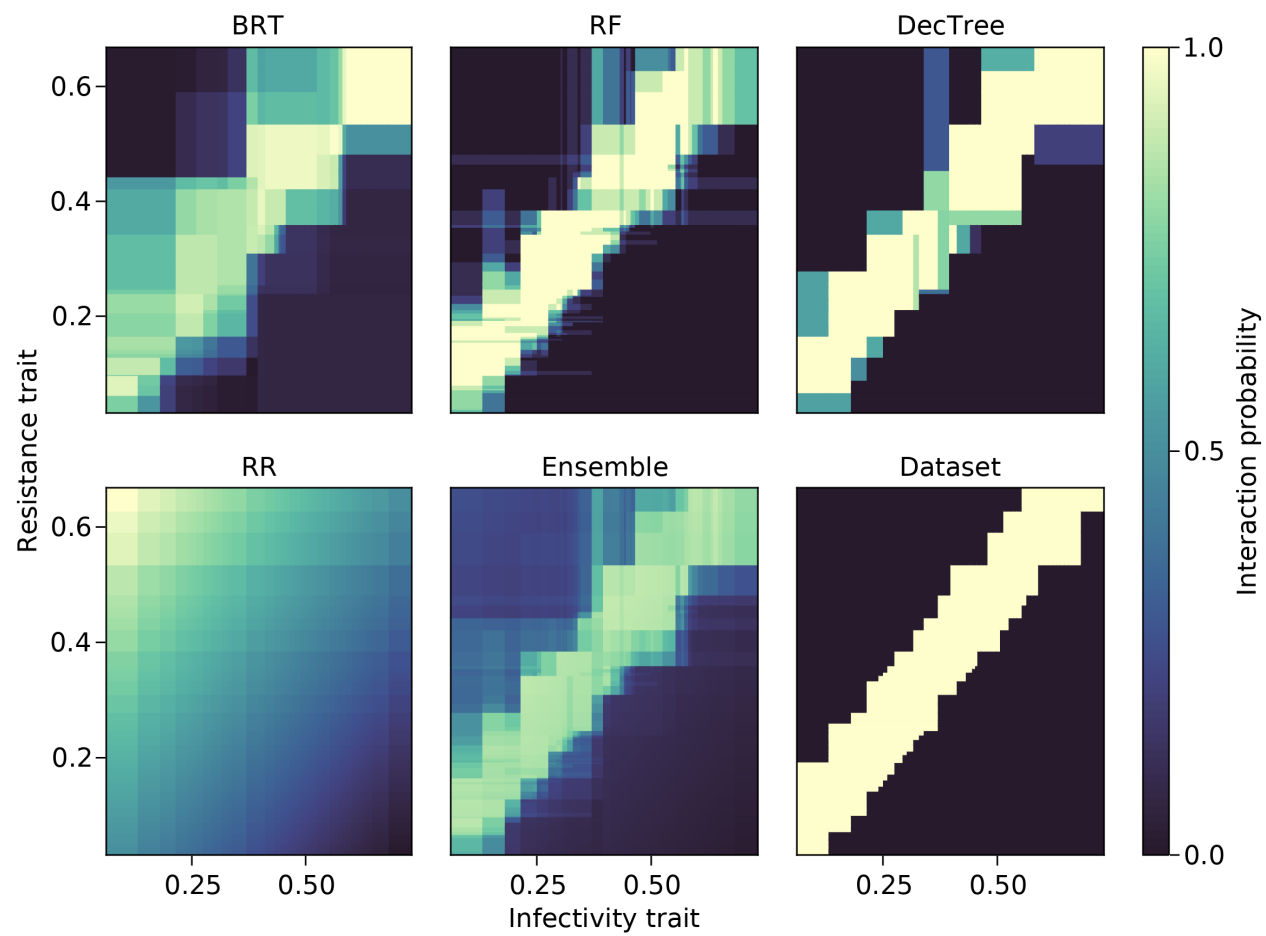


Figure 7: TODO