

Guidelines for the validation of 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: *February 26, 2022*

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. Neither classifier accuracy nor the ROC-AUC are 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 Ecological networks are a backbone for key ecological and evolutionary processes; yet enumerating all of
2 the interactions they contain is a daunting task, as it scales with S^2 , *i.e.* the squared species richness
3 TODO. Recent contributions to the field of ecological network prediction (Strydom et al., 2021;
4 **Becker2021OptPre?**) highlight that although interactions can be predicted by adding ecologically
5 relevant information (in the form of, *e.g.* traits), we do not have robust guidelines as to how the predictive
6 ability of these models should be evaluated, nor about how the models should be trained. Here, by relying
7 on simple derivations and a series of simulations, we formulate a number of such guidelines, specifically
8 for the case of binary classifiers derived from thresholded values. Specifically, we conduct an investigation
9 of the models in terms of their skill (ability to make the right prediction), bias (trends towards
10 systematically over-predicting one class), class imbalance (the relative number of cases representing
11 interactions), and show how these effects interact. We conclude on the fact that models with the best
12 interaction-scale predictive score do not necessarily result in the most accurate representation of the
13 network.

14 The prediction of ecological interactions shares conceptual and methodological issues with two fields in
15 biology: species distribution models (SDMs), and genomics. SDMs suffers from issues affecting
16 interactions prediction, namely low prevalence (due to sparsity of observations/interactions) and data
17 aggregation (due to bias in sampling some locations/species). In previous work, Allouche et al. (2006)
18 suggested that κ was a better test of model performance than the True Skill Statistic (TSS; which we refer
19 to as Youden's informedness); these conclusions were later criticized by Somodi et al. (2017), who
20 emphasized that informedness' is affected both by prevalence and bias. Although this work offers
21 recommendations about the comparison of models, it doesn't establishes baselines or good practices for
22 training on imbalanced ecological data, or ways to remedy the imbalance. Steen et al. (2021) show that,
23 when applying spatial thinning (a process that has no analogue in networks), the best approach to train
24 ML-based SDMs varies according to the balancing of the dataset, and the evaluation measures used. This
25 suggests that there is no single "recipe" that is guaranteed to give the best model. By contrast to networks,
26 SDMs have the advantage of being able to both thin datasets to remove some of the sampling bias (*e.g.*
27 Inman et al., 2021), but also to create pseudo-absences to inflate the number of supposed negatives in the
28 dataset (*e.g.* Iturbide et al., 2015).

29 An immense body of research on machine learning application to life sciences is focused on genomics
30 (which has very specific challenges, see a recent discussion by Whalen et al., 2021); this sub-field has

31 generated recommendations that do not necessarily match the current best-practices for SDMs, and
32 therefore hint at the importance of domain-specific guidelines. Chicco & Jurman (2020) suggest using
33 Matthews correlation coefficient (MCC) over F_1 , as a protection against over-inflation of predicted results;
34 Delgado & Tibau (2019) advocate against the use of Cohen's κ , again in favor of MCC, as the relative
35 nature of κ means that a worse classifier can be picked over a better one; similarly, Boughorbel et al.
36 (2017) recommend MCC over other measures of performance for imbalanced data, as it has more
37 desirable statistical properties. More recently, Chicco et al. (2021) temper the apparent supremacy of the
38 MCC, by suggesting it should be replaced by Youden's informedness (also known as J , bookmaker's
39 accuracy, and the True-Skill Statistic) when the imbalance in the dataset may not be representative of the
40 actual imbalance.

41 Species interaction networks are often under-sampled (Jordano, 2016b, 2016a), and this under-sampling is
42 structured taxonomically (Beauchesne et al., 2016), structurally (de Aguiar et al., 2019) and spatially
43 (Poisot, Bergeron, et al., 2021; Wood et al., 2015). As a consequence, networks suffer from data
44 deficiencies both within and between datasets. This implies that the comparison of classifiers across
45 space, when undersampling varies locally (see *e.g.* McLeod et al., 2021) is non-trivial. Furthermore, the
46 baseline value of classifiers performance measures under various conditions of skill, bias, and prevalence,
47 has to be identified to allow researchers to evaluate whether their interaction prediction model is indeed
48 learning. Taken together, these considerations highlight three specific issues for ecological networks.
49 First, what values of performance measures are indicative of a classifier with no skill? This is particularly
50 important as it can evaluate whether low prevalence can lull us into a false sense of predictive accuracy.
51 Second, independently of the question of model evaluation, is low prevalence an issue for *training* or
52 *testing*, and can we remedy it? Finally, because the low amount of data on interaction makes a lot of
53 imbalance correction methods (see *e.g.* Branco et al., 2015) hard to apply, which indicators can be
54 optimized by sacrificing least amount of positive interaction data?

55 It may sound counter-intuitive to care so deeply about how good a classifier with no-skill is, as by
56 definition, it has no skill. The necessity of this exercise has its roots in the paradox of accuracy: when the
57 desired class ("two species interact") is rare, a model that gets less ecologically performant by only
58 predicting the opposite class ("these two species do not interact") sees its accuracy increase; because most
59 of the guesses have "these two species do not interact" as a correct answer, a model that never predicts
60 interactions would be right an overwhelming majority of the time; it would also be utterly useless. Herein

61 lies the core challenge of predicting species interactions: the extreme imbalance between classes makes
62 the training of predictive models difficult, and their validation even more so as we do not reliably know
63 which negatives are true. The connectance (the proportion of realized interactions, usually the number of
64 interactions divided by the number of species pairs) of empirical networks is usually well under 20%, with
65 larger networks having a lower connectance (MacDonald et al., 2020), and therefore being increasingly
66 difficult to predict.

67 **A primer on binary classifier evaluation**

68 Binary classifiers, which it to say, machine learning algorithms whose answer is a categorical value, are
69 usually assessed by measuring properties of their confusion matrix, *i.e.* the contingency table reporting
70 true/false positive/negative hits. A confusion matrix is laid out as

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

71 In this matrix, tp is the number of times the model predicts an interaction that exists in the network (true
72 positive), fp is the number of times the model predicts an interaction that does not exist in the network
73 (false positive), fn is the number of times the model fails to predict an interaction that actually exists in the
74 network (false negatives), and tn is the number of times the model correctly predicts that an interaction
75 does not exist (true negatives). From these values, we can derive a number of measures of model
76 performance (see Strydom et al., 2021 for a review of their interpretation in the context of networks). At a
77 coarse scale, a classifier is *accurate* when the trace of the matrix divided by the sum of the matrix is close
78 to 1, with other measures informing us on how the predictions fail.

79 There is an immense diversity of measures to evaluate the performance of classification tasks (Ferri et al.,
80 2009). Here we will focus on five of them with high relevance for imbalanced learning (He & Ma, 2013).
81 The choice of metrics with relevance to class-imbalanced problems is fundamental, because as Japkowicz
82 (2013) unambiguously concluded, “relatively robust procedures used for unskewed data can break down
83 miserably when the data is skewed.” Following Japkowicz (2013), we focus on two ranking metrics (the
84 areas under the Receiving Operator Characteristic and Precision Recall curves), and three threshold

85 metrics (κ , informedness, and MCC; we will briefly discuss F_1 but show early on that it has undesirable
86 properties).

87 The κ measure (Landis & Koch, 1977) establishes the extent to which two observers (the network and the
88 prediction) agree, and is measured as

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

89 Informedness (Youden, 1950) (also known as bookmaker informedness or the True Skill Statistic) is
90 $TPR + TNR - 1$, where $TPR = tp/(tp + fn)$ and $TNR = tn/(tn + fp)$; informedness can be used to find
91 the optimal cutpoint in thresholding analyses (Schisterman et al., 2005). The formula for informedness is

$$\frac{tp}{tp + fn} + \frac{tn}{tn + fp} - 1.$$

92 The MCC is defined as

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

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

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

95 A lot of binary classifiers are built by using a regressor (whose task is to guess the value of the interaction,
96 and can therefore return a value considered to be a pseudo-probability); in this case, the optimal value
97 below which predictions are assumed to be negative (*i.e.* the interaction does not exist) can be determined
98 by picking a threshold maximizing some value on the ROC or the PR curve. The area under these curves
99 (ROC-AUC and PR-AUC henceforth) give ideas on the overall goodness of the classifier, and the ideal
100 threshold is the point on these curves that minimizes the tradeoff represented in these curves. Saito &
101 Rehmsmeier (2015) established that the ROC-AUC is biased towards over-estimating performance for
102 imbalanced data; on the contrary, the PR-AUC is able to identify classifiers that are less able to detect

positive interactions correctly, with the additional advantage of having a baseline value equal to prevalence. Therefore, it is important to assess whether these two measures return different results when applied to ecological network prediction. The ROC curve is defined by the false positive rate on the x axis, and the true positive rate on the y axis, and the PR curve is defined by the true positive rate on the x axis, and the positive predictive value on the y axis. By comparison with the previous paragraph, it is obvious that F_1 and MCC have ties to the PR curve (being close to the expected PR-AUC), and that informedness has ties to the ROC curve (whereby the threshold maximizing informedness is also the point of maximal inflection on the ROC curve). One important difference between ROC and PR is that the later does not prominently account for the size of the true negative compartments: in short, it is more sensitive to the correct positive predictions. In a context of strong imbalance, PR-AUC is therefore a more stringent test of model performance.

Baseline values for the threshold metrics

In this section, we will assume a network of connectance ρ , *i.e.* having ρS^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 (assumed to be an unweighted, directed 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 confusion matrix, which ends up expressing *relative* values). We will apply skill and bias to this matrix, and measure how a selection of performance metrics respond to changes in these values, in order to assess their suitability for model evaluation.

Confusion matrix with skill and bias

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

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

128 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 129 to s , and off-diagonal elements equal to $(1-s)$, and re-express the skill-adjusted confusion matrix as
 130 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

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

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

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

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

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

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

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

139 In this section, we will change the values of b , s , and ρ , and report how the main measures discussed in
 140 the introduction (MCC, F_1 , κ , and informedness) are responding to issues with the classifier. Before we do
 141 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 ; $\text{logit}(x) = -10$ means that x is essentially 0, and $\text{logit}(x) = 10$ means it is essentially 1 – this choice was motivated by the fact that most responses are non-linear with regards to bias and skill. The values of ρ were taken linearly in $]0, 0.5]$, which is within the range of usually observed connectance values for empirical species interaction networks. 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 contrast, 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

171 a value of 0 for a no-skill classifier, and are forced towards their correct maximal value when skill changes
172 (*i.e.* a more connected networks will have higher values for a skilled classifier, and lower values for a
173 classifier making mostly mistakes).

174 [Figure 2 about here.]

175 These two analyses point to the following recommendations: MCC is indeed more appropriate than κ , as
176 although sensitive to bias, it is sensitive in a consistent way. Informedness is appropriate at discriminating
177 between different skills, but confounded by bias. As both of these measures bring valuable information on
178 the model behavior, we will retain them for future analyses. F_1 is increasing with bias, and should not be
179 prioritized to evaluate the performance of the model. The discussion of sensitivity to bias should come with
180 a domain-specific caveat: although it is likely that interactions documented in ecological networks are
181 correct, a lot of non-interactions are simply unobserved; as predictive models are used for data-inflation
182 (*i.e.* the prediction of new interactions), it is not necessarily a bad thing in practice to select models that
183 predict more interactions than the original dataset, because the original dataset misses some interactions.
184 Furthermore, the weight of positive interactions could be adjusted if some information about the extent of
185 undersampling exists (*e.g.* Branco et al., 2015). In a recent large-scale imputation of interactions in the
186 mammal-virus networks, Poisot, Ouellet, et al. (2021) for example estimated that 93% of interactions are
187 yet to be documented.

188 **Numerical experiments on training strategy**

189 In the following section, we will generate random bipartite networks (this works without loss of generality
190 on unipartite networks), and train four binary classifiers (as well as an ensemble model using the sum of
191 ranged outputs from the component models) on 30% of the interaction data. In practice, testing usually
192 uses 70% of the total data; for ecological networks, where interactions are sparse *and* the number of species
193 is low, this may not be the best solution, as the testing set becomes constrained not by the *proportion* of
194 interactions, but by their *number*. Preliminary experiments using different splits revealed no qualitative
195 change in the results. Networks are generated by picking a random infectiousness trait v_i for 100 species
196 (from a beta distribution $B(\alpha = 6, \beta = 8)$ distribution), and a resistance trait h_j for 100 species (from
197 $B(\alpha = 2, \beta = 8)$ distribution). There is an interaction between i and j when $v_i - \xi/2 \leq h_j \leq v_i + \xi/2$,

198 where ξ is a constant regulating the connectance of the network (there is an almost 1:1 relationship
 199 between ξ and connectance), and varies uniformly in $[0.05, 0.35]$. This model gives fully interval networks
 200 that are close analogues to the bacteria–phage model of Weitz et al. (2005), with both a modular structure
 201 and a non-uniform degree distribution. This dataset is easy for almost any algorithm to learn: when
 202 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
 203 presented below were able to reach almost perfect predictions all the time (data not presented here) – this
 204 is in part because the rule (there is maximum value of the distance between traits for which there is an
 205 interaction) is fixed for all interactions. In order to make the problem more difficult to solve, we use
 206 $[v_i, h_j]$ as a feature vector (*i.e.* the traits on which the models are trained), and therefore the models will
 207 have to uncover that the rule for interaction is $\text{abs}(v_i, h_j) \leq \xi$. The models therefore all have the form

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

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

215 The dataset used for numerical experiments is composed of 64000 such (ξ, ν) pairs, on which four
 216 machines are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random
 217 forest regressor. All models were taken from the MLJ.jl package (Blaom et al., 2020; Blaom & Vollmer,
 218 2020) in Julia 1.7 (Bezanson et al., 2017). All machines use the default parameterization; this is an obvious
 219 deviation from best practices, as the hyperparameters of any machine require training before its
 220 application on a real dataset. As we use 64000 such datasets, this would require 256000 unique instances
 221 of tweaking the hyperparameters, which is not realistic. Therefore, we assume that the default
 222 parameterizations are comparable across networks. All machines return a quantitative prediction, usually

(but not necessarily) in $[0, 1]$, which is proportional (but not necessarily linearly) to the probability of an interaction between i and j .

In order to pick the best adjacency matrix for a given trained machine, 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 predictions of all component models on the testing set (ranged in $[0, 1]$), then put through the same thresholding process. The complete code to run the simulations is given as an appendix; running the final simulation required 4.8 core days (approx. 117 hours).

After the simulations were completed, we removed all runs (*i.e.* pairs of ξ and ν) for which at least one of the following conditions was met: the accuracy was 0, the true positive or true negative rates were 0, the 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.

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 more positive in order to get a performant 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.]

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

276 stringent tests of model performance. These results suggest that learning from a dataset with very low
277 connectance can be a different task than for more connected networks: it becomes increasingly important
278 to capture the mechanisms that make an interaction *exist*, and therefore having a slightly more biased
279 training dataset might be beneficial. As connectance increases, the need for biased training sets is less
280 prominent, as learning the rules for which interactions *do not* exist starts gaining importance.

281 [Figure 6 about here.]

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

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

290 In this last section, we generate a network using the same model as before, with $S_1, S_2 = 50, 80$ species, a
291 connectance of ≈ 0.16 ($\xi = 0.19$), and a training set bias of 0.7. The prediction made on the complete
292 dataset is presented in fig. 7. Visualizing the results this way highlights the importance of exploratory data
293 analysis: whereas all models return a network with interactions laying mostly on the diagonal (as
294 expected), the Ridge Regression is quite obviously biased. Despite this, we can see that the ensemble is
295 close to the initial dataset.

296 [Figure 7 about here.]

297 The trained models were then thresholded (again by optimising informedness), and their predictions
298 transformed back into networks for analysis; specifically, we measured the connectance, nestedness
299 (REF), and modularity (REF). This process was repeated 250 times, and the results are presented in tbl. 1.
300 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 able 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 |
| 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 |

Guidelines for the assesment of network predictive models

We establish that due to the low prevalence of interactions, even poor classifiers applied to food web data will reach a high accuracy; this is because the measure is dominated by the accidentally correct predictions of negatives. On simulated confusion matrices with ranges of imbalance that are credible for ecological networks, MCC had the most desirable behavior, and informedness is a linear measure of

318 classifier skill. By performing simulations with four models and an ensemble, we show that informedness
319 and ROC-AUC are consistently high on network data, and that MCC and PR-AUC are more accurate
320 measures of the effective performance of the classifier. Finally, by measuring the structure of predicted
321 networks, we highlight an interesting paradox: the models with the best performance measures are not
322 the models with the closest reconstructed network structure. We discuss these results in the context of
323 establishing guidelines for the prediction of ecological interactions.

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

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

343 Second, because the PR-AUC responds more to network connectance fig. 6 and training set imbalance
344 fig. 4, it should be used as a measure of model performance over the ROC-AUC. This is not to say that
345 ROC-AUC should be discarded (in fact, a low ROC-AUC is a sign of an issue with the model), but that its
346 interpretation should be guided by the PR-AUC value. Specifically, a high ROC-AUC is not informative, as
347 it can be associated to a low PR-AUC (see *e.g.* Random Forest in tbl. 1) This again echoes

348 recommendations from other fields (Jeni et al., 2013; Saito & Rehmsmeier, 2015).

349 Thirdly, regardless of network connectance, maximizing informedness required a training set bias of about
350 0.5, and maximizing the MCC required a training set bias of 0.7 and more. This has an important
351 consequence in ecological networks, for which the pool of positive cases (interactions) to draw from is
352 typically small: the most parsimonious measure (*i.e.* the one requiring to discard the least amount of
353 information to train the model) will give the best validation potential, and is probably the informedness
354 (maximizing informedness is the generally accepted default for imbalanced classification; Schisterman et
355 al., 2005).

356 Finally, it is noteworthy that the ensemble model was systematically better than the component models;
357 even when the models were individually far from perfect, the ensemble was able to leverage the different
358 biases expressed by the models to make an overall more accurate prediction. We do not expect that
359 ensembles will *always* be better than single models. In a recent multi-model comparison,
360 (Becker2021OptPre?) found that the ensemble was *not* the best model. There is no general conclusion to
361 draw from this besides reinforcing the need to be pragmatic about which models should be included in the
362 ensemble, or whether to use an ensemble at all. In a sense, the surprising performance of the ensemble
363 model should form the basis of the last recommendation: optimal training set bias and its interaction with
364 connectance and binary classifier is, in a sense, an hyperparameter that should be assessed. The
365 distribution of results in fig. 5 and fig. 6 show that there are variations around the trend; furthermore,
366 networks with different structures than the one we simulated here may respond in different ways.

367 **Acknowledgements:** We acknowledge that this study was conducted on land within the traditional
368 unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and
369 Omàmiwininiwak nations. We thank Colin J. Carlson, Michael D. Catchen, Giulio Valentino Dalla Riva,
370 and Tanya Strydom for inputs on earlier versions of this manuscript. This research was enabled in part by
371 support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada
372 (www.computeCanada.ca) through the Narval general purpose cluster. TP is supported by a NSERC
373 Discovery Grant and Discovery Acceleration Supplement, by funding to the Viral Emergence Research
374 Initiative (VERENA) consortium including NSF BII 2021909, and by a grant from the Institut de
375 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>

Beauchesne, D., Desjardins-Proulx, Archambault, P., & Gravel, D. (2016). Thinking Outside the

Boxpredicting Biotic Interactions in Data-poor Environments. *Vie Et Milieu-Life and enVironment*,

66(3-4), 333–342.

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). Flexible model composition in machine learning and its

implementation in MLJ. *arXiv:2012.15505 [cs]*. <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). A Survey of Predictive Modelling under Imbalanced

Distributions. *arXiv:1505.01658 [cs]*. <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>

de Aguiar, M. A. M., Newman, E. A., Pires, M. M., Yeakel, J. D., Boettiger, C., Burkle, L. A., Gravel, D.,

Guimarães, P. R., O'Donnell, J. L., Poisot, T., Fortin, M.-J., & Hembry, D. H. (2019). Revealing biases in

the sampling of ecological interaction networks. *PeerJ*, 7, e7566.
<https://doi.org/10.7717/peerj.7566>

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>

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

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

Inman, R., Franklin, J., Esque, T., & Nussear, K. (2021). Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere*, 12(3), e03422.
<https://doi.org/10.1002/ecs2.3422>

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., & Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312, 166–174. <https://doi.org/10.1016/j.ecolmodel.2015.05.018>

Japkowicz, N. (2013). Assessment Metrics for Imbalanced Learning. In *Imbalanced Learning* (pp. 187–206). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118646106.ch8>

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

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

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

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

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

434 Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., MacDonald, A., Mercier, B., Violet, C., &
 435 Vissault, S. (2021). Global knowledge gaps in species interaction networks data. *Journal of*
 436 *Biogeography*, jbi.14127. <https://doi.org/10.1111/jbi.14127>

437 Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N.,
 438 & Carlson, C. J. (2021). Imputing the mammalian virome with linear filtering and singular value
 439 decomposition. *arXiv:2105.14973 [q-Bio]*. <http://arxiv.org/abs/2105.14973>

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

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

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

449 Steen, V. A., Tingley, M. W., Paton, P. W. C., & Elphick, C. S. (2021). Spatial thinning and class balancing:
 450 Key choices lead to variation in the performance of species distribution models with citizen science
 451 data. *Methods in Ecology and Evolution*, 12(2), 216–226. <https://doi.org/10.1111/2041-210X.13525>

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

457 Weitz, J. S., Hartman, H., & Levin, S. A. (2005). Coevolutionary arms races between bacteria and
 458 bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27),
 459 9535–9540. <https://doi.org/10.1073/pnas.0504062102>

- 460 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine
461 learning in genomics. *Nature Reviews Genetics*, 1–13.
462 <https://doi.org/10.1038/s41576-021-00434-9>
- 463 Wood, S. A., Russell, R., Hanson, D., Williams, R. J., & Dunne, J. A. (2015). Effects of spatial scale of
464 sampling on food web structure. *Ecology and Evolution*, 5(17), 3769–3782.
465 <https://doi.org/10.1002/ece3.1640>
- 466 Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
467 [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.

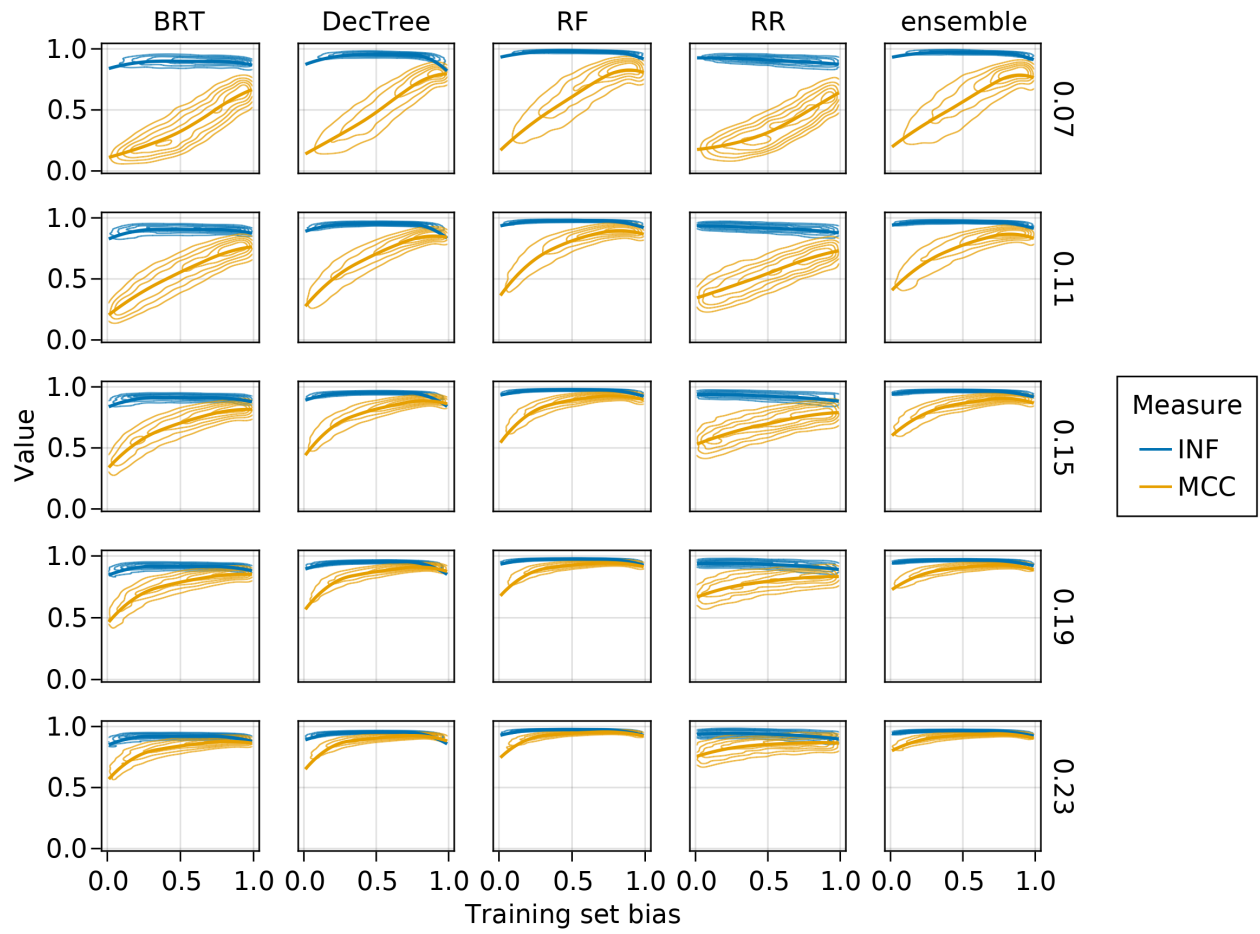


Figure 3: Response of MCC and Informedness to changes in the training set bias for a fixed connectance (rows). Both of these values approach 1 for a good model. Informedness is consistently high, and by contrast, MCC increases with additional training set bias. Across all models, training on a more connected network is easier.

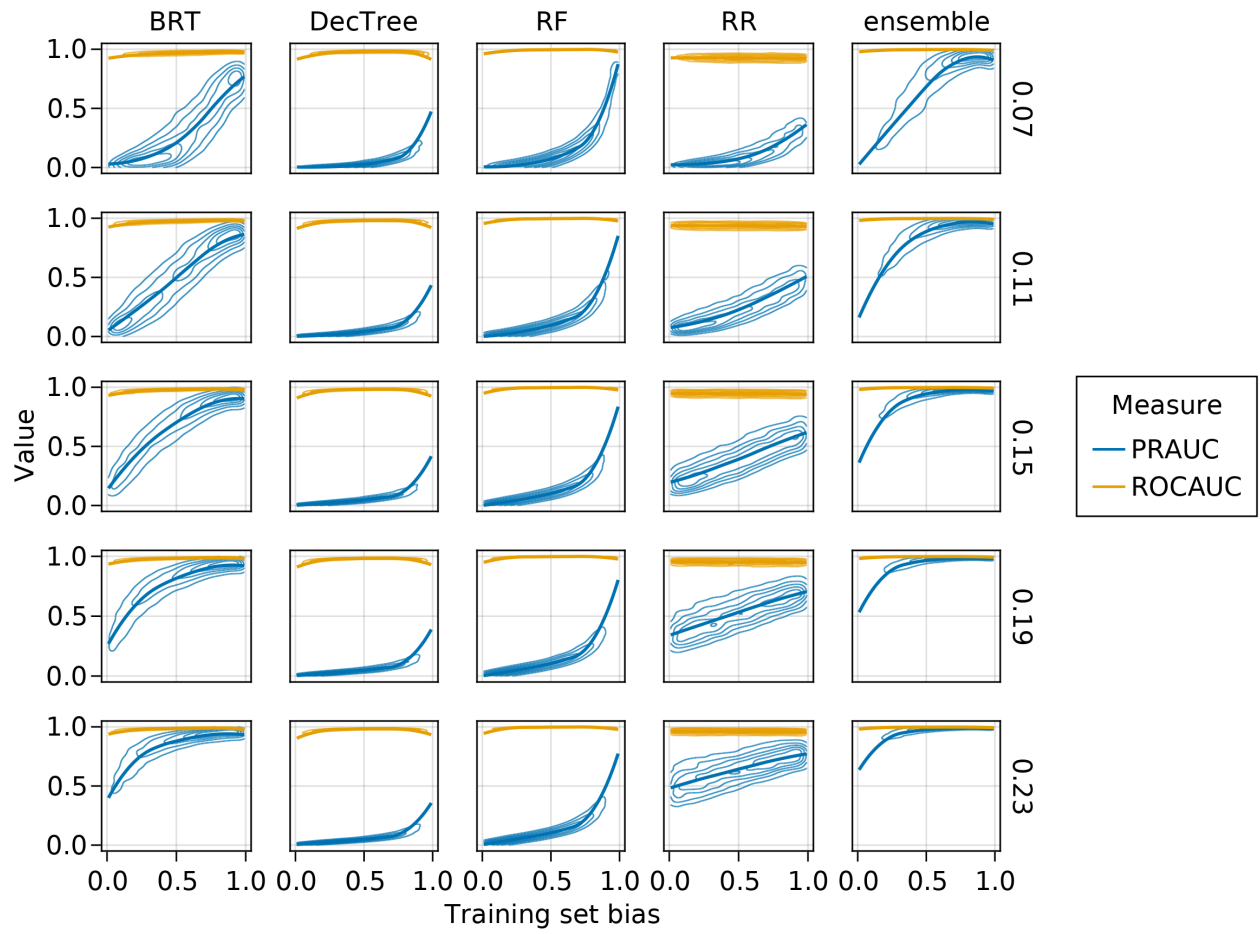


Figure 4: Response of ROC-AUC and PR-AUC to changes in the training set bias for a fixed connectance (rows). ROC-AUC is consistently high, and therefore not properly able to separate good from poor classifiers. On the other hand, PR-AUC responds to changes in the training set. As in fig. 3, training on more connected networks is easier.

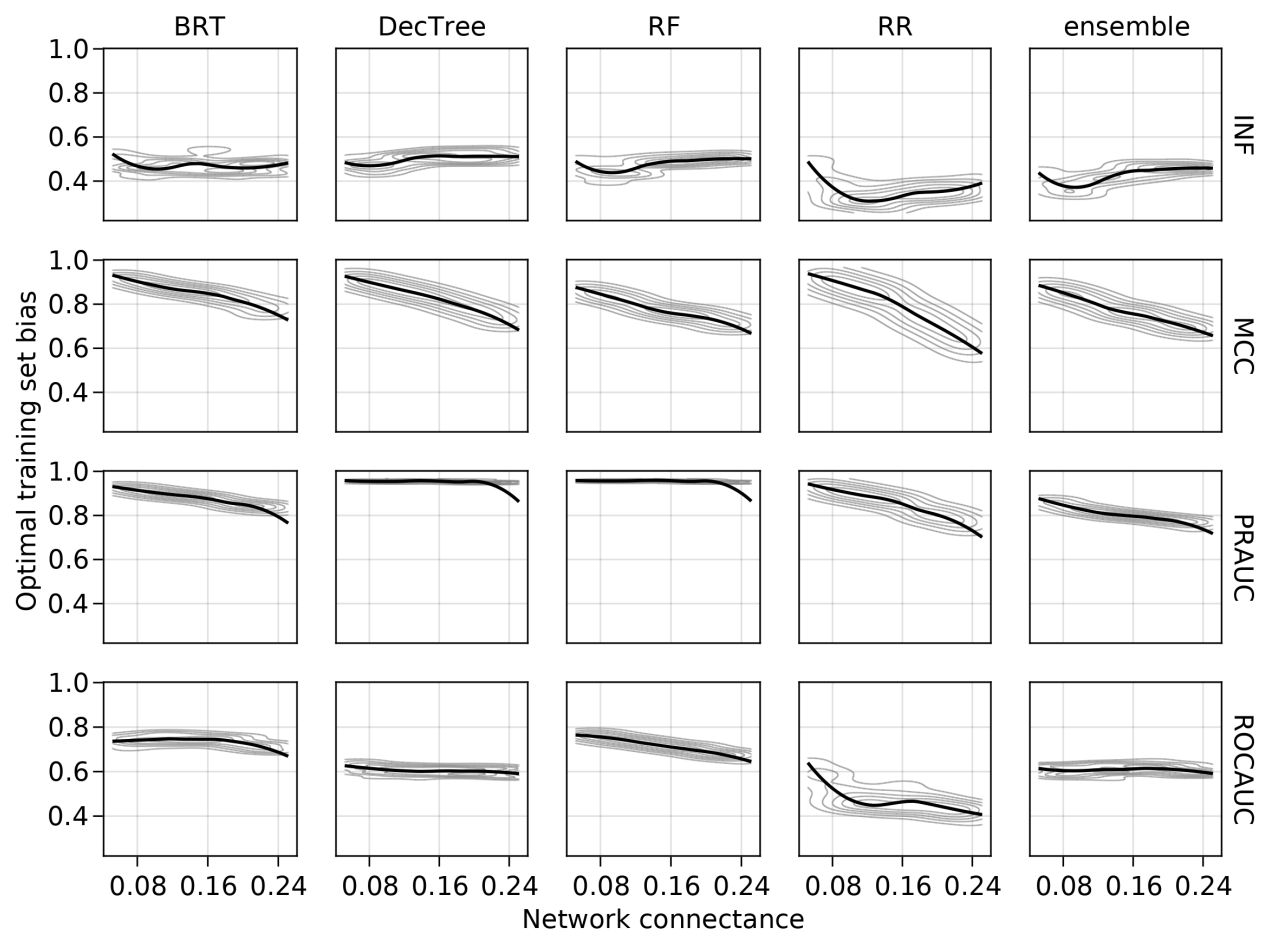


Figure 5: Value of the optimal training set bias for the different models and measures evaluated here, over a range of connectances. Informedness was reliably maximized for balanced training sets, and kept this behavior across models. For other measures, larger connectances in the true network allowed lower biases in the training set. In a large number of cases, “over-correcting” by having training sets with more than half instances representing interactions would maximize the values of the model performance measures.

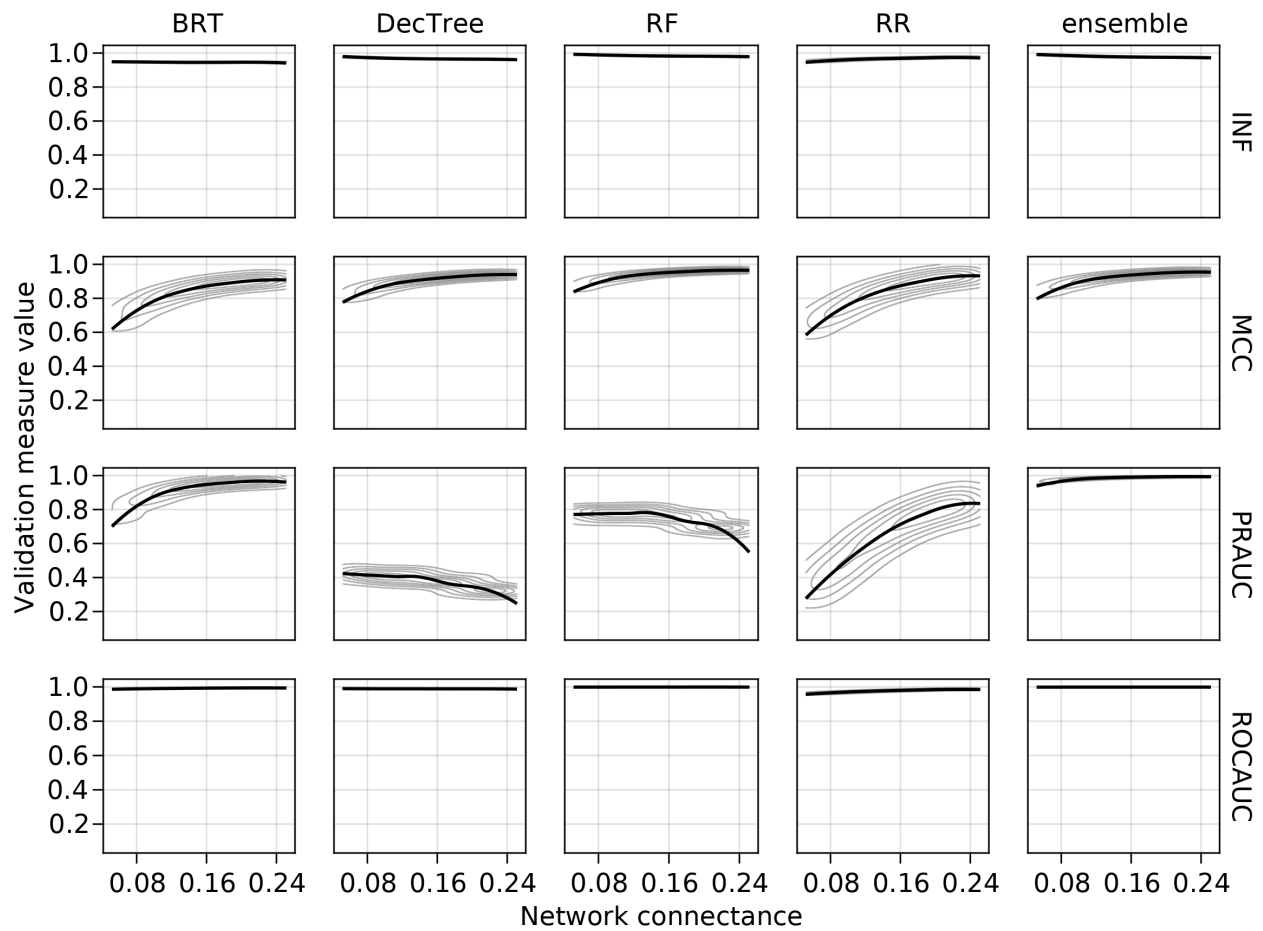


Figure 6: When trained on their optimally biased training set, most models were able to maximize their performance; this is not true for decision tree, which had a very low PR-AUC, and to some extent for ridge regression who had a slow increase with network connectance. The ensemble had a consistently high performance despite incorporating poor models.

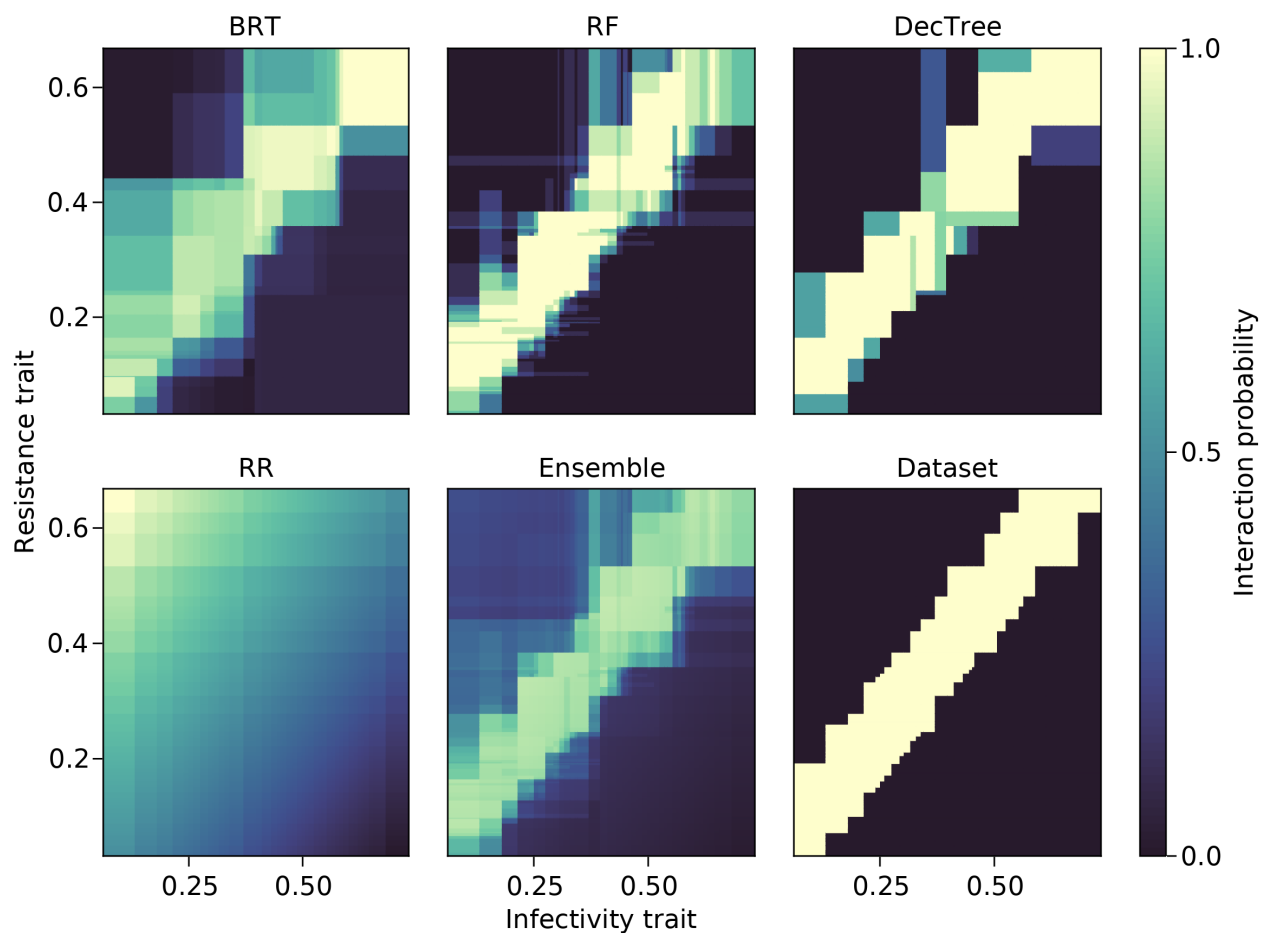


Figure 7: Visualisation of the models predictions for one instance of a network prediction problem (shown in the “Dataset” panel). This figure reveals how inspecting the details of the prediction is important: indeed, although the performance measures hint at the fact that ridge regression is mediocre, this figure reveals that it is making predictions that correspond to a network with an entirely different topology (namely, nested as opposed to diagonal).