

Guidelines for the prediction of species interactions through binary classification

Timothée Poisot ^{1,2}

¹ Université de Montréal ² Québec Centre for Biodiversity Sciences

Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

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. 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 ways to assemble the training dataset. classifier accuracy nor the area under the receiver operating characteristic curve (ROC-AUC) are informative measures for the performance of interaction prediction. The area under the precision-recall curve (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. 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 fine-tuned before performing the actual training. We discuss these consequences in the context of the low volume of data.

1 Species interactions, forming ecological networks, are a backbone for key ecological and evolutionary
2 processes; yet enumerating all of the interactions between S species is a daunting task, as it scales with S^2 ,
3 *i.e.* the squared species richness (Martinez, 1992). Recent contributions to the field of ecological network
4 prediction (Becker et al., 2022; Pichler et al., 2020; Strydom et al., 2021) highlight that although
5 interactions can be predicted by adding ecologically relevant information (in the form of, *e.g.* traits), we do
6 not have robust guidelines as to how the predictive ability of models recommending species interactions
7 should be evaluated, nor about how these models should be trained. Here, by relying on simple
8 derivations and a series of simulations, we formulate a number of such guidelines, specifically for the case
9 of binary classifiers derived from thresholded values. Specifically, we conduct an investigation of the
10 models in terms of their skill (ability to make the right prediction), bias (trends towards systematically
11 over-predicting one class), class imbalance (the relative number of cases representing interactions), and
12 show how these effects interact. We conclude on the fact that models with the best interaction-scale
13 predictive score do not necessarily result in the most accurate representation of the true network.

14 The prediction of ecological interactions shares conceptual and methodological issues with two fields in
15 biology: species distribution modelling (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). An important challenge lies in the fact that
18 the best measure to quantify the performance of a model is not necessarily a point of consensus (these
19 methods, their interpretation, and the way they are measured, are covered in depth in the next section). In
20 previous work, Allouche et al. (2006) suggested that Cohen’s κ agreement score (κ thereafter) was a better
21 test of model performance than the True Skill Statistic (TSS; which we refer to as Youden’s informedness
22 thereafter); these conclusions were later criticized by Somodi et al. (2017), who emphasized that
23 informedness is affected both by prevalence and bias. Although this work offers recommendations about
24 the comparison of models, it doesn’t establishes baselines or good practices for training on imbalanced
25 ecological data, or ways to remedy the imbalance. Steen et al. (2021) show that, when applying spatial
26 thinning (artificially re-balancing observation data in space to avoid artifacts due to auto-correlation), the
27 best approach to train ML-based SDMs varies according to the balancing of the dataset, and the evaluation
28 measures used; there is no single “recipe” that is guaranteed to give the best model. By contrast to
29 networks, SDMs have the advantage of being able to both thin datasets to remove some of the sampling
30 bias (*e.g.* Inman et al., 2021), but also to create pseudo-absences to inflate the number of supposed

negatives in the dataset (e.g. Iturbide et al., 2015). These powerful ways to remove data bias often have no analogue in networks, removing one potential tool from our methodological toolkit, and making the task of network prediction through classification potentially more demanding, and more prone to underlying data biases.

An immense body of research on machine learning application to life sciences is focused on genomics (which has very specific challenges, see a recent discussion by Whalen et al., 2021); this sub-field has generated recommendations that do not necessarily match the current best-practices for SDMs, and therefore hint at the importance of domain-specific guidelines. Chicco & Jurman (2020) suggest using Matthews correlation coefficient (MCC) over F_1 , as a protection against over-inflation of predicted results; Delgado & Tibau (2019) advocate against the use of Cohen's κ , again in favor of MCC, as the relative nature of κ means that a worse classifier can be picked over a better one; similarly, Boughorbel et al. (2017) recommend MCC over other measures of performance for imbalanced data, as it has more desirable statistical properties. More recently, Chicco et al. (2021) temper the apparent supremacy of the MCC, by suggesting it should be replaced by Youden's informedness (also known as J , bookmaker's accuracy, and the True-Skill Statistic) when the imbalance in the dataset may not be representative of the actual imbalance. In a way, the measures themselves need not be a strong focus for network prediction, as they are routinely used in other field; the discipline-specific question we seek to address is: 'which metric should be employed when predicting networks, and how to optimize it?'

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

61 imbalance correction methods (see *e.g.* Branco et al., 2015) hard to apply, which measures of model
62 performance can be optimized by sacrificing least amount of positive interaction data?

63 A preliminary question is to examin the baseline performance of these measures, *i.e.* the values they
64 would take on hypothetical networks based on a classifier that has no-skill. It may sound counter-intuitive
65 to care so deeply about how good a classifier with no-skill is, as by definition, is has no skill. The necessity
66 of this exercise has its roots in the paradox of accuracy: when the desired class (“two species interact”) is
67 rare, a model that gets less ecologically performant by only predicting the opposite class (“these two
68 species do not interact”) sees its accuracy increase; because most of the guesses have “these two species do
69 not interact” as a correct answer, a model that never predicts interactions would be right an overwhelming
70 majority of the time; it would also be utterly useless. Herein lies the core challenge of predicting species
71 interactions: the extreme imbalance between classes makes the training of predictive models difficult, and
72 their validation even more so as we do not reliably know which negatives are true. The connectance (the
73 proportion of realized interactions, usually the number of interactions divided by the number of species
74 pairs) of empirical networks is usually well under 20%, with larger networks having a lower connectance
75 (MacDonald et al., 2020), and therefore being increasingly difficult to predict.

76 **A primer on binary classifier evaluation**

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

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

80 In this matrix, tp is the number of times the model predicts an interaction that exists in the network (true
81 positive), fp is the number of times the model predicts an interaction that does not exist in the network
82 (false positive), fn is the number of times the model fails to predict an interaction that actually exists in the
83 network (false negatives), and tn is the number of times the model correctly predicts that an interaction
84 does not exist (true negatives). From these values, we can derive a number of measures of model

85 performance (see Strydom et al., 2021 for a review of their interpretation in the context of networks). At a
86 coarse scale, a classifier is *accurate* when the trace of the matrix divided by the sum of the matrix is close
87 to 1, with other measures informing us on how the predictions fail.

88 A lot of binary classifiers are built by using a regressor (whose task is to guess the value of the interaction,
89 and can therefore return a value considered to be a pseudo-probability); in this case, the optimal value
90 below which predictions are assumed to be negative (*i.e.* the interaction does not exist) can be determined
91 by picking a threshold maximizing some value on the ROC or the PR curve. The area under these curves
92 (ROC-AUC and PR-AUC henceforth) give ideas on the overall goodness of the classifier, and the ideal
93 threshold is the point on these curves that minimizes the tradeoff represented in these curves. Saito &
94 Rehmsmeier (2015) established that the ROC-AUC is biased towards over-estimating performance for
95 imbalanced data; on the contrary, the PR-AUC is able to identify classifiers that are less able to detect
96 positive interactions correctly, with the additional advantage of having a baseline value equal to
97 prevalence. Therefore, it is important to assess whether these two measures return different results when
98 applied to ecological network prediction. The ROC curve is defined by the false positive rate on the x axis,
99 and the true positive rate on the y axis, and the PR curve is defined by the true positive rate on the x axis,
100 and the positive predictive value on the y axis.

101 There is an immense diversity of measures to evaluate the performance of classification tasks (Ferri et al.,
102 2009). Here we will focus on five of them with high relevance for imbalanced learning (He & Ma, 2013).
103 The choice of metrics with relevance to class-imbalanced problems is fundamental, because as Japkowicz
104 (2013) unambiguously concluded, “relatively robust procedures used for unskewed data can break down
105 miserably when the data is skewed”. Following Japkowicz (2013), we focus on two ranking metrics (the
106 areas under the Receiver Operating Characteristic and Precision Recall curves), and three threshold
107 metrics (κ , informedness, and MCC; we will briefly discuss F_1 but show early on that it has undesirable
108 properties).

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

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

111 Informedness (Youden, 1950) (also known as bookmaker informedness or the True Skill Statistic) is

112 TPR + TNR = 1, where $\text{TPR} = tp/(tp + fn)$ and $\text{TNR} = tn/(tn + fp)$. Informedness can be used to find
 113 the optimal cutpoint in thresholding analyses (Schisterman et al., 2005); indeed, the maximal
 114 informedness corresponds to the point on the ROC curve that is closest to the perfect classifier point. The
 115 formula for informedness is

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

116 The MCC is defined as

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

117 Finally, F_1 is the harmonic mean of precision (the chance that interaction was correctly detected as such)
 118 and sensitivity (the ability to correctly classify interactions), and is defined as

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

119 One noteworthy fact is that F_1 and MCC have ties to the PR curve (being close to the expected PR-AUC),
 120 and that informedness has ties to the ROC curve (whereby the threshold maximizing informedness is also
 121 the point of maximal inflection on the ROC curve). One important difference between ROC and PR is that
 122 the later does not prominently account for the size of the true negative compartments: in short, it is more
 123 sensitive to the correct positive predictions. In a context of strong imbalance, PR-AUC is therefore a more
 124 stringent test of model performance.

125 **Baseline values for the threshold metrics**

126 In this section, we will assume a network with connectance equal to a scalar ρ , *i.e.* having ρS^2 interactions
 127 (where S is the species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true*
 128 state of the network (assumed to be an unweighted, directed network) is a column vector $\mathbf{o}^T = [\rho, (1 - \rho)]$
 129 (we can safely drop the S^2 terms, as we will work on the confusion matrix, which ends up expressing
 130 *relative* values). We will apply skill and bias to this matrix, and measure how a selection of performance
 131 metrics respond to changes in these values, in order to assess their suitability for model evaluation.

132 Confusion matrix with skill and bias

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

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

139 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 140 to s , and off-diagonal elements equal to $(1 - s)$, which allows to regulate how many predictions are wrong,
 141 under the assumption that the bias is the same (*i.e.* the classifier is as likely to make a false positive or a
 142 false negative). The skill-adjusted confusion matrix is $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

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

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

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

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

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

149 In all further simulations, the confusion matrix \mathbf{C} is transformed so that it sums to unity, *i.e.* the entries
150 are the *proportions* of guesses.

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

152 In this section, we will change the values of b , s , and ρ , and report how the main measures discussed in
153 the introduction (MCC, F_1 , κ , and informedness) respond. Before we do so, it is important to explain why
154 we will not focus on accuracy too much. Accuracy is the number of correct predictions ($\text{Tr}(\mathbf{C})$) divided by
155 the sum of the confusion matrix. For a no-skill, no-bias classifier, accuracy is equal to $\rho^2 + (1-\rho)^2$; for
156 $\rho = 0.05$, this is ≈ 0.90 , and for $\rho = 0.01$, this is equal to ≈ 0.98 . In other words, the values of accuracy are
157 high enough to be uninformative (for ρ small, $\rho^2 \ll (1-\rho)^2$). More concerning is the fact that introducing
158 bias changes the response of accuracy in unexpected ways. Assuming a no-skill classifier, the numerator
159 of accuracy becomes $b\rho^2 + (1-b)(1-\rho)^2$, which increases when b is low, which specifically means that at
160 equal skill, a classifier that under-predicts interactions will have higher accuracy than an un-biased
161 classifier (because the value of accuracy is dominated by the size of tn, which will increase). These issues
162 are absent from balanced accuracy, but should nevertheless lead us to not report accuracy as the primary
163 measure of network prediction success; moving forward, we will focus on other measures.

164 In order to examine how MCC, F_1 , κ , and informedness change w.r.t. the imbalance, skill, and bias, we
165 performed a grid exploration of the values of $\text{logit}(s)$ and $\text{logit}(b)$ linearly from -10 to 10 ; $\text{logit}(x) = -10$
166 means that x is essentially 0, and $\text{logit}(x) = 10$ means it is essentially 1 – this choice was motivated by the
167 fact that most responses are non-linear with regards to bias and skill. The values of ρ were taken linearly
168 in $]0, 0.5]$, which is within the range of connectance for species interaction networks. Note that at this
169 point, there is no network model to speak of; the confusion matrix we discuss can be obtained for any
170 classification task. Based on the previous discussion, the desirable properties for a measure of classifier
171 success should be: an increase with classifier skill, especially at low bias; a hump-shaped response to bias,
172 especially at high skill, and ideally centered around $\text{logit}(b) = 0$; an increase with prevalence up until
173 equiprevalence is reached.

[Figure 1 about here.]

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

[Figure 2 about here.]

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

200 Numerical experiments on training strategy

201 In the following section, we will generate random bipartite networks, and train four binary classifiers (as
 202 well as an ensemble model using the sum of ranged outputs from the component models) on 50% of the
 203 interaction data. In practice, testing usually uses 70% of the total data; for ecological networks, where
 204 interactions are sparse *and* the number of species is low, this may not be the best solution, as the testing
 205 set becomes constrained not by the *proportion* of interactions, but by their *number*. Preliminary
 206 experiments using different splits revealed no qualitative change in the results. Networks are generated by
 207 picking a random infectiousness trait v_i for 100 species (from a beta distribution $B(\alpha = 6, \beta = 8)$
 208 distribution), and a resistance trait h_j for 100 species (from $B(\alpha = 2, \beta = 8)$ distribution). There is an
 209 interaction between i and j when $v_i - \xi/2 \leq h_j \leq v_i + \xi/2$, where ξ is a constant regulating the
 210 connectance of the network (visual exploration of the parameters show that there is an almost 1:1
 211 relationship between ξ and connectance), and varies uniformly in $[0.05, 0.35]$. This model gives fully
 212 interval networks that are close analogues to the bacteria–phage model of Weitz et al. (2005), with both a
 213 modular structure and a non-uniform degree distribution. This dataset is easy for almost any algorithm to
 214 learn: when trained with features $[v_i, h_j, \text{abs}(v_i, h_j)]^T$ to predict the interactions between i and j , all four
 215 models presented below were able to reach almost perfect predictions all the time (data not presented
 216 here) – this is in part because the rule (there is maximum value of the distance between traits for which
 217 there is an interaction) is fixed for all interactions, and any method able to learn non-linear relationships
 218 should infer it without issues. In order to make the problem more difficult to solve, we use $[v_i, h_j]$ as a
 219 feature vector (*i.e.* the traits on which the models are trained), and therefore the models will have to
 220 uncover that the rule for interaction is $\text{abs}(v_i, h_j) \leq \xi$. The models therefore all have the following form,
 221 where $i_{i,j}$ is an interaction from species i to species j :

$$\begin{bmatrix} i_{1,1} \\ i_{1,2} \\ \vdots \\ i_{m,n-1} \\ i_{m,n} \end{bmatrix} \propto \begin{bmatrix} v_1 & h_1 \\ v_1 & h_2 \\ \vdots & \vdots \\ v_m & h_{n-1} \\ v_m & h_n \end{bmatrix}$$

222 The training sample is composed of a random pick of up to 50% of the 10^4 possible entries in the network,

223 *i.e.* $n = 5000$. Out of these interactions, we pick a proportion ν (the training set balance) to be positive, so
224 that the training set has νn interactions, and $(1 - \nu)n$ non-interactions. We vary ν uniformly in $]0, 1[$. This
225 allows to evaluate how the measures of binary classification performance respond to artificially
226 rebalanced dataset for a given network connectance. The rest of the dataset is used as a testing set, on
227 which all further measures are calculated. Note that although the training set is balanced arbitrarily, the
228 testing set is assembled so that it has the exact connectance of the entire network; this ensures that the
229 model is evaluated under the class imbalance where the predictions will be made, which represents a
230 more meaningful evaluation. Furthermore, to avoid artifacts due to different sizes of the training and
231 testing set within a single network, the number of entries in both sets are equal. Note also that although
232 the simulated networks are bipartite, the algorithms have no “knowledge” of the network structure, and
233 simply look at pairs of species; therefore, the approach outlined here would also work for unipartite
234 networks.

235 The dataset used for numerical experiments is composed of a grid of 35 values of connectance (from 0.011
236 to 0.5) and 35 values of ν (from 0.02 to 0.98); for each pair of values, 500 networks are generated and
237 predicted. For each network, we train four machines: a trait-based k-NN (*e.g.* Desjardins-Proulx et al.,
238 2017), a regression tree, a regression random forest, and a boosted regression tree; the later three methods
239 are turned into classifiers using thresholding, which oftentimes provides better results than classification
240 when faced with class imbalance (Hong et al., 2016). Following results from Pichler et al. (2020), linear
241 models have not been considered (in any way, the relationship in the simulated networks is non-linear).
242 The point of these numerical experiments is *not* to recommend the best model (this is likely
243 problem-specific), but to highlight a series of recommendations that would work for supervised learning
244 tasks. All models were taken from the MLJ .jl package (Blaom et al., 2020; Blaom & Vollmer, 2020) in Julia
245 1.7 (Bezanson et al., 2017). All machines use the default parameterization; this is an obvious deviation
246 from best practices, as the hyperparameters of any machine require training before its application on a real
247 dataset. As we use 612500 such datasets, this would require over 2 millions unique instances of tweaking
248 the hyperparameters, which is prohibitive from a computing time point of view. An important thing to
249 keep in mind is that the problem we simulate has been designed to be simple to solve: we expect all
250 machines with sensible default parameters to fare well — the results presented in the later sections show
251 that this assumption is warranted, and we further checked that the models do not overfit by ensuring that
252 there is never more than 5% of difference between the accuracy on the training and testing sets. All

253 machines return a quantitative prediction, usually (but not necessarily) in $[0, 1]$, which is proportional
254 (but not necessarily linearly) to the probability of an interaction between i and j . The ROC-AUC and
255 PR-AUC (and therefore the thresholds) can be measured by integrating over the domain of the values
256 return by each machine, but in order to make the average-based ensemble model more meaningful, all
257 predictions are expressed in $[0, 1]$.

258 In order to pick the best confusion matrix for a given trained machine, we performed a thresholding
259 approach using 500 steps on predictions from the testing set, and picking the threshold that maximized
260 Youden's informedness. During the thresholding step, we measured the area under the receiver operating
261 characteristic (ROC-AUC) and precision-recall (PR-AUC) curves, as measures of overall performance over
262 the range of returned values. We report the ROC-AUC and PR-AUC, as well as a suite of other measures as
263 introduced in the next section, for the best threshold. The ensemble model was generated by summing the
264 predictions of all component models on the testing set (ranged in $[0, 1]$), then put through the same
265 thresholding process. The complete code to run the simulations is available at [10.17605/OSF.IO/JKEWD](https://doi.org/10.17605/OSF.IO/JKEWD).

266 After the simulations were completed, we removed all runs (*i.e.* triples of model, ξ , and ν) for which at
267 least one of the following conditions was met: the accuracy was 0, the true positive or true negative rates
268 were 0, the connectance was larger than 0.25. This removes both the obviously failed model runs, and the
269 networks that are more densely connected compared to the connectance of empirical food webs (and are
270 therefore less difficult to predict, being less imbalanced; preliminary analyses of data with a connectance
271 larger than 0.3 revealed that all machines reached consistently high performance).

272 **Effect of training set balance on performance**

273 In fig. ??, we present the response of two thresholding measures (PR-AUC and ROC-AUC) and two
274 ranking measures (Informedness and MCC) to a grid of 35 values of training set balance, and 35 values of
275 connectance, for the four component models as well as the ensemble. ROC-AUC is always high, and does
276 not vary with training set balance. On the other hand, PR-AUC shows very strong responses, increasing
277 with training set balance. It is notable here that two classifiers that seemed to be performing well (Decision
278 Tree and Random Forest) based on their MCC are not able to reach a high PR-AUC even at higher
279 connectances. All models reached a higher performance on more connected networks, and using more
280 balanced training sets. 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 balance, although this increase became less steep with increasing connectance. Three of the models (kNN, decision tree, and random forest) only increased their PR-AUC sharply when the training set was heavily imbalanced towards more interactions. Interestingly, the ensemble almost always outclassed its component models. For larger connectances (less difficult networks to predict, as they are more balanced), MCC and informedness started decreasing when the training set bias got too close to one, suggesting that a training set balance of 0.5 may often be appropriate if these measures are the one to optimize.

[Figure 3 about here.]

Based on the results presented in fig. ??, 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 balance, 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 one 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 balance 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. ??.

[Figure 4 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 balance 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

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

312 [Figure 5 about here.]

313 When trained at their optimal training set balance, connectance still had a significant impact on the
314 performance of some machines (fig. ??). Notably, Decision Tree, and k-NN, as well as Random forest to a
315 lower extent, had low values of PR-AUC. In all cases, the Boosted Regression Tree was reaching very good
316 predictions (especially for connectances larger than 0.1), and the ensemble was almost always scoring
317 perfectly. This suggests that all the models are biased in different ways, and that the averaging in the
318 ensemble is able to correct these biases. We do not expect this last result to have any generality, and
319 provide a discussion of a recent example in which the ensemble was performing worse than its
320 components models.

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

322 In this last section, we generate a network using the same model as before, with $S_1, S_2 = 50, 80$ species, a
323 connectance of ≈ 0.16 ($\xi = 0.19$), and a training set balance of 0.5, as fig. ?? suggests this is the optimal
324 training set balance for this range of connectance. The prediction made on the complete dataset is
325 presented in fig. ??.

326 [Figure 6 about here.]

327 The trained models were then thresholded (again by optimising informedness), and their predictions
328 transformed back into networks for analysis; specifically, we measured the connectance, nestedness [η ;
329 Bastolla et al. (2009)], modularity [Q ; Barber (2007)], asymmetry [A ; Delmas et al. (2018)], and Jaccard
330 network dissimilarity (Canard et al., 2014). This process was repeated 250 times, and the results are
331 presented in tbl. ?. The k-NN model is an interesting instance here: it produces the network that looks

the most like the original dataset, despite having the lowest PR-AUC, suggesting it hits high recall at the cost of low precision. The ensemble was able to reach a very high PR-AUC (and a very high ROC-AUC), which translated into more accurate reconstructions of the structure of the network (with the exception of modularity, which is underestimated by 0.03). 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). In this specific example, the combination of individually “adequate” models resulted in an extremely strong ensemble, suggesting that the correct prediction of interactions (as measured by MCC, Inf., ROC-AUC, and PR-AUC) and network properties is indeed a feasible task under appropriately hyper-parameterized models.

Table 1: Values of four performance metrics, and five network structure metrics, for 500 independent predictions similar to the ones presented in fig. ???. 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	A	Jaccard
Decision tree	0.59	0.94	0.97	0.04	0.17	0.64	0.37	0.42	0.1
BRT	0.46	0.91	0.97	0.36	0.2	0.78	0.29	0.41	0.19
Random Forest	0.72	0.98	0.99	0.1	0.16	0.61	0.38	0.42	0.06
k-NN	0.71	0.98	0.99	0.02	0.16	0.61	0.39	0.42	0.06
<i>Ensemble</i>	0.74	0.98	1.0	0.79	0.16	0.61	0.38	0.42	0.06
<i>Data</i>					0.16	0.56	0.41	0.42	0.0

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

It is noteworthy that the ensemble model was systematically better than the component models. We do not expect that ensembles will *always* be better than single models. Networks with different structures than the one we simulated here may respond in different ways, especially if the rules are fuzzier than the simple rule we used here. In a recent multi-model comparison involving supervised and unsupervised learning, Becker et al. (2022) found that the ensemble was *not* the best model, and was specifically under-performing compared to models using biological traits. This may be because the dataset of Becker et al. (2022) was known to be under-sampled, and so the network alone contained less information than the combination of the network and species traits. There is no general conclusion to draw from either these results or ours, besides reinforcing the need to be pragmatic about which models should be included in the ensemble, and whether to use an ensemble at all. In a sense, the surprising performance of the ensemble model should form the basis of the first broad recommendation: optimal training set balance and its interaction with connectance and the specific binary classifier used is, in a sense, an hyperparameter that should be assessed following the approach outlined in this manuscript. The distribution of results in fig. ?? and fig. ?? show that there are variations around the trend, and multiple models should probably be trained on their “optimal” training/testing set, as opposed to the same ones.

The results presented here highlight an interesting paradox: although the k-NN model was ultimately able to get a correct estimate of network structure (see tbl. ?? and fig. ??), it ultimately remains a poor classifier, as evidenced by its low PR-AUC. This suggests that the goal of predicting *interactions* and predicting

377 *networks* may not always be solvable in the same way – of course a perfect classifier of interactions would
378 make a perfect network prediction; indeed, the best scoring predictor of interactions (the ensemble model)
379 had the best prediction of network structure. The tasks of predicting networks structure and of predicting
380 interactions within networks are essentially two different ones. For some applications (e.g. comparison of
381 network structure across gradients), one may care more about a robust estimate of the structure, at the cost
382 at putting some interactions at the wrong place. For other applications (e.g. identifying pairs of interacting
383 species), one may conversely care more about getting as many pairs right, even though the mistakes
384 accumulate in the form of a slightly worse estimate of network structure. How these two approaches can
385 be reconciled is something to evaluate on a case-by-case basis, especially since there is no guarantee that
386 an ensemble model will always be the most precise one. Despite this apparent tension at the heart of the
387 predictive exercise, we can use the results presented here to suggest a number of guidelines.

388 First, because we have more trust in reported interactions than in reported absences of interactions (which
389 are overwhelmingly *pseudo*-absences), we can draw on previous literature to recommend informedness as
390 a measure to decide on a threshold for binary classification (Chicco et al., 2021); this being said, because
391 informedness is insensitive to bias (although it is a linear measure of skill), the overall model performance
392 is better evaluated through the use of MCC (figs. ??, ??). Because F_1 is monotonously sensitive to classifier
393 bias (fig. ??) and network connectance (fig. ??), MCC should be preferred as a measure of model evaluation
394 and comparison. When dealing with multiple models, we therefore suggest to find the optimal threshold
395 using informedness, and to pick the best model using MCC (assuming one does not want to use an
396 ensemble model).

397 Second, accuracy alone should not be the main measure of model performance, but rather an expectation
398 of how well the model should behave given the class balance in the set on which predictions are made;
399 this is because, as derived earlier, the expected accuracy for a no-skill no-bias classifier is $\rho^2 + (1 - \rho)^2$
400 (where ρ is the class balance), which will most often be large. This pitfall is notably illustrated in a recent
401 food-web model (Caron et al., 2022) wherein the authors, using a training set of $n = 10^4$ with only 100
402 positive interactions (representing 0.1% of the total interactions), reached a good accuracy. Reporting a
403 good accuracy is not informative, especially when accuracy isn't (i) compared to the baseline expected
404 value under the given class balance, and (ii) interpreted in the context of a measure that is not sensitive to
405 the chance prediction of many negatives (like MCC).

406 Third, because the PR-AUC responds more to network connectance (fig. ??) and training set imbalance

(fig. ??) than ROC-AUC, it should be used as a measure of model performance over the ROC-AUC. This is not to say that ROC-AUC should be discarded (in fact, a low ROC-AUC is undoubtedly a sign of an issue with the model), but that its interpretation should be guided by the PR-AUC value. Specifically, a high ROC-AUC is not informative, as it can be associated to a low PR-AUC (see e.g. Random Forest in tbl. ??). This again echoes recommendations from other fields (Jeni et al., 2013; Saito & Rehmsmeier, 2015). We therefore expect to see high ROC-AUC values, and then to pick the model that maximizes the PR-AUC value. Taken together with the previous two guidelines, we strongly encourage to (i) ensure that accuracy and ROC-AUC are high (in the case of accuracy, higher than expected under no-skill no-bias situation), and (ii) to discuss the performance of the model in terms of the most discriminant measures, *i.e.* PR-AUC and MCC.

Finally, network connectance (*i.e.* the empirical class imbalance) should inform the composition of the training and testing set, because it is an ecologically relevant value. In the approach outlined here, we treat the class imbalance of the training set as an hyper-parameter, but *test* the model on a set that has the same class imbalance as the actual dataset. This is an important distinction, as it ensure that the prediction environment matches the testing environment (as we cannot manipulate the connectance of the empirical dataset on which the predictions will be made), and so the values measured on the testing set (or validation set if the data volume allows one to exists) can be directly compared to the values for the actual prediction. A striking result from fig. ?? is that Informedness was almost always maximal at 50/50 balance (regardless of connectance), whereas MCC required *more* positives to be maximized when connectance *increases*, matching the idea that it is a more stringent measure of performance. This has an important consequence in ecological networks, for which the pool of positive cases (interactions) to draw from is typically small: the most parsimonious measure (*i.e.* the one requiring to discard the least amount of interactions to train the model) will give the best validation potential, and in this light is very likely informedness [maximizing informedness is, in fact, the generally accepted default for imbalanced classification regardless of the problem domain; Schisterman et al. (2005)]. This last result further strengthens the assumption that the amount of bias *is* an hyper-parameter that must be fine-tuned, as using the wrong bias can lead to models with lower performance; for this reason, it makes sense to not train all models on the same training/testing set, but rather to optimize the set composition for each of them.

One key element for real-life data that can make the prediction exercise more tractable is that some interactions can safely be assumed to be impossible; indeed, a lot of networks can be reasonably well

described using a stochastic block model (e.g. Xie et al., 2017). In ecological networks, this can be due to spatial constraints (Valdovinos, 2019), or to the long-standing knowledge that some links are “forbidden” due to traits (Olesen et al., 2011) or abundances (Canard et al., 2014). The matching rules (Olito & Fox, 2015; Strona & Veech, 2017) can be incorporated in the model either by adding compatibility traits, or by *only* training the model on pairs of species that are not likely to be forbidden links. Knowledge of true negative interactions could be propagated in training/testing sets that have true negatives, and in this situation, it may be possible to use the more usual 70/30 split for training/testing folds as the need to protect against potential unbalance is lowered. Besides forbidden links, a real-life case that may arise is multi-interaction or multi-layer networks (Pilosof et al., 2017). These can be studied using the same general approach outlined here, either by assuming that pairs of species can interact in more than one way (wherein one would train a model for each type of interaction, based on the relevant predictors), or by assuming that pairs of species can only have one type of interaction (wherein this becomes a multi-label classification problem).

Acknowledgements: We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and Omàmiwininiwak nations. We thank Colin J. Carlson, Michael D. Catchen, Giulio Valentino Dalla Riva, and Tanya Strydom for inputs on earlier versions of this manuscript. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) through the Narval general purpose cluster. TP is supported by the Fondation Courtois, a NSERC Discovery Grant and Discovery Acceleration Supplement, by funding to the Viral Emergence Research Initiative (VERENA) consortium including NSF BII 2021909, 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>
- Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6), 066102. <https://doi.org/10.1103/PhysRevE.76.066102>
- Bastolla, U., Fortuna, M. A., Pascual-García, A., Ferrera, A., Luque, B., & Bascompte, J. (2009). The

architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, 458(7241), 1018–1020. <https://doi.org/10.1038/nature07950>

Beauchesne, D., Desjardins-Proulx, Archambault, P., & Gravel, D. (2016). Thinking Outside the Box—predicting Biotic Interactions in Data-poor Environments. *Vie Et Milieu-Life and enVironment*, 66(3-4), 333–342.

Becker, D. J., Albery, G. F., Sjödin, A. R., Poisot, T., Bergner, L. M., Chen, B., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A. C., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., Stock, M., Teeling, E. C., & Carlson, C. J. (2022). Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *The Lancet Microbe*. [https://doi.org/10.1016/S2666-5247\(21\)00245-7](https://doi.org/10.1016/S2666-5247(21)00245-7)

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*. <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*. <http://arxiv.org/abs/1505.01658>

Canard, E. F., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D., & Gravel, D. (2014). Empirical evaluation of neutral interactions in host-parasite networks. *The American Naturalist*, 183(4), 468–479. <https://doi.org/10.1086/675363>

Caron, D., Maiorano, L., Thuiller, W., & Pollock, L. J. (2022). Addressing the Eltonian shortfall with trait-based interaction models. *Ecology Letters*, 25(4), 889–899. <https://doi.org/10.1111/ele.13966>

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.

493 <https://doi.org/10.1186/s12864-019-6413-7>

494 Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable
 495 than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix
 496 evaluation. *BioData Mining*, 14, 13. <https://doi.org/10.1186/s13040-021-00244-z>

497 de Aguiar, M. A. M., Newman, E. A., Pires, M. M., Yeakel, J. D., Boettiger, C., Burkle, L. A., Gravel, D.,
 498 Guimarães, P. R., O'Donnell, J. L., Poisot, T., Fortin, M.-J., & Hembry, D. H. (2019). Revealing biases in
 499 the sampling of ecological interaction networks. *PeerJ*, 7, e7566.
 500 <https://doi.org/10.7717/peerj.7566>

501 Delgado, R., & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in
 502 classification. *PloS One*, 14(9), e0222916. <https://doi.org/10.1371/journal.pone.0222916>

503 Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães,
 504 P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D., & Poisot, T. (2018).
 505 Analysing ecological networks of species interactions. *Biological Reviews*, 112540.
 506 <https://doi.org/10.1111/brv.12433>

507 Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix
 508 problem. *PeerJ*, 5(e3644). <https://doi.org/10.7717/peerj.3644>

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

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

514 Hong, C., Ghosh, R., & Srinivasan, S. (2016). *Dealing with Class Imbalance using Thresholding*.
 515 <https://doi.org/10.48550/arXiv.1607.02705>

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

519 Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., & Gutiérrez, J. M. (2015). A framework for
 520 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>

Martinez, N. D. (1992). Constant Connectance in Community Food Webs. *The American Naturalist*, 139(6), 1208–1218. <https://www.jstor.org/stable/2462337>

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

Olesen, J. M., Bascompte, J., Dupont, Y. L., Elberling, H., Rasmussen, C., & Jordano, P. (2011). Missing and forbidden links in mutualistic networks. *Proc. R. Soc. B*, 278(1706), 725–732. <https://doi.org/10.1098/rspb.2010.1371>

Olito, C., & Fox, J. W. (2015). Species traits and abundances predict metrics of plant–pollinator network structure, but not pairwise interactions. *Oikos*, 124, 428–436.

Pichler, M., Boreux, V., Klein, A., Schleuning, M., & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, 11(2), 281–293. <https://doi.org/10.1111/2041-210X.13329>

548 Pilosof, S., Porter, M. A., Pascual, M., & Kéfi, S. (2017). The multilayer nature of ecological networks.
 549 *Nature Ecology & Evolution*, 1, 0101. <https://doi.org/10.1038/s41559-017-0101>

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

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

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

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

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

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

568 Strona, G., & Veech, J. A. (2017). Forbidden versus permitted interactions: Disentangling processes from
 569 patterns in ecological network analysis. *Ecology and Evolution*, 7(14), 5476–5481.
 570 <https://doi.org/10.1002/ece3.3102>

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

- 576 Valdovinos, F. S. (2019). Mutualistic networks: Moving closer to a predictive theory. *Ecology Letters*, 0(0).
577 <https://doi.org/10.1111/ele.13279>
- 578 Weitz, J. S., Hartman, H., & Levin, S. A. (2005). Coevolutionary arms races between bacteria and
579 bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27),
580 9535–9540. <https://doi.org/10.1073/pnas.0504062102>
- 581 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine
582 learning in genomics. *Nature Reviews Genetics*, 1–13.
583 <https://doi.org/10.1038/s41576-021-00434-9>
- 584 Wood, S. A., Russell, R., Hanson, D., Williams, R. J., & Dunne, J. A. (2015). Effects of spatial scale of
585 sampling on food web structure. *Ecology and Evolution*, 5(17), 3769–3782.
586 <https://doi.org/10.1002/ece3.1640>
- 587 Xie, J.-R., Zhang, P., Zhang, H.-F., & Wang, B.-H. (2017). Completeness of Community Structure in
588 Networks. *Scientific Reports*, 7(1), 5269. <https://doi.org/10.1038/s41598-017-05585-6>
- 589 Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
590 [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)

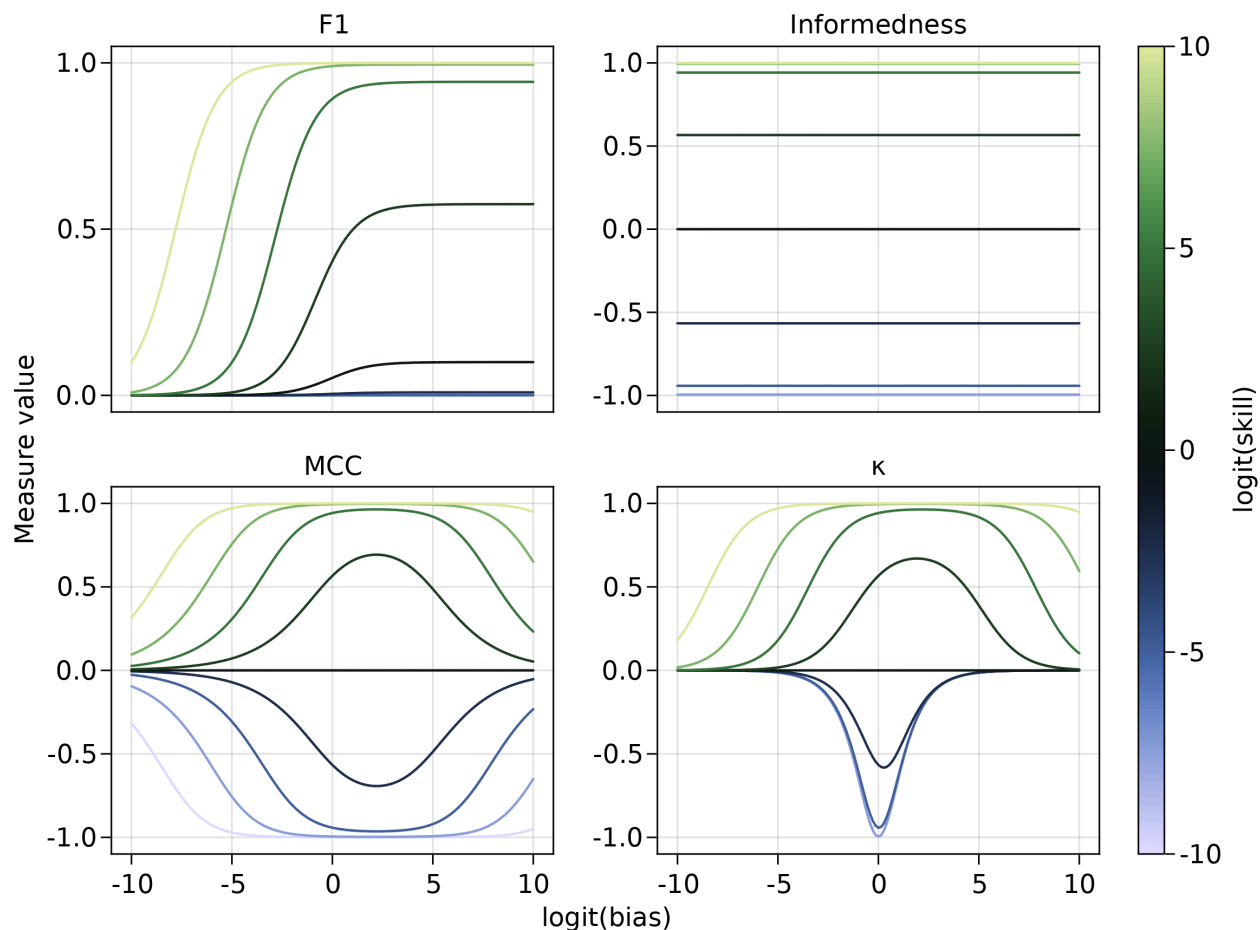


Figure 1: Consequences of changing the classifier skills (s) and bias (s) for a connectance $\rho = 0.15$, on F_1 , informedness, MCC, and κ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions (this follows from the derivations in the text, not shown in the figure). 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.

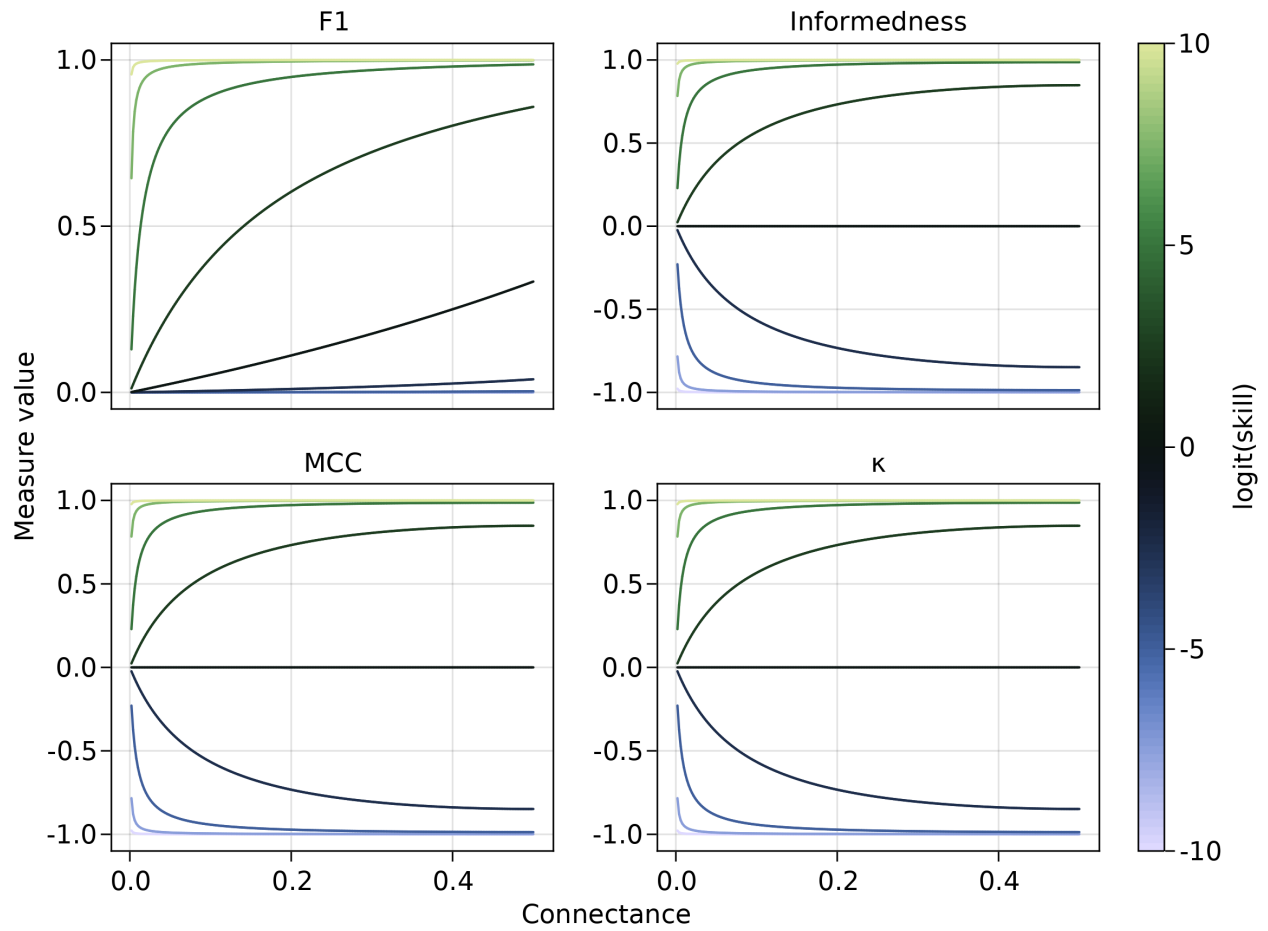


Figure 2: As in fig. ??, consequences of changing connectance for different levels of classifier skill, assuming no classifier bias. Informedness, κ , and MCC do increase with connectance, but only when the classifier is not no-skill; by way of contrast, a more connected network will give a higher F_1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

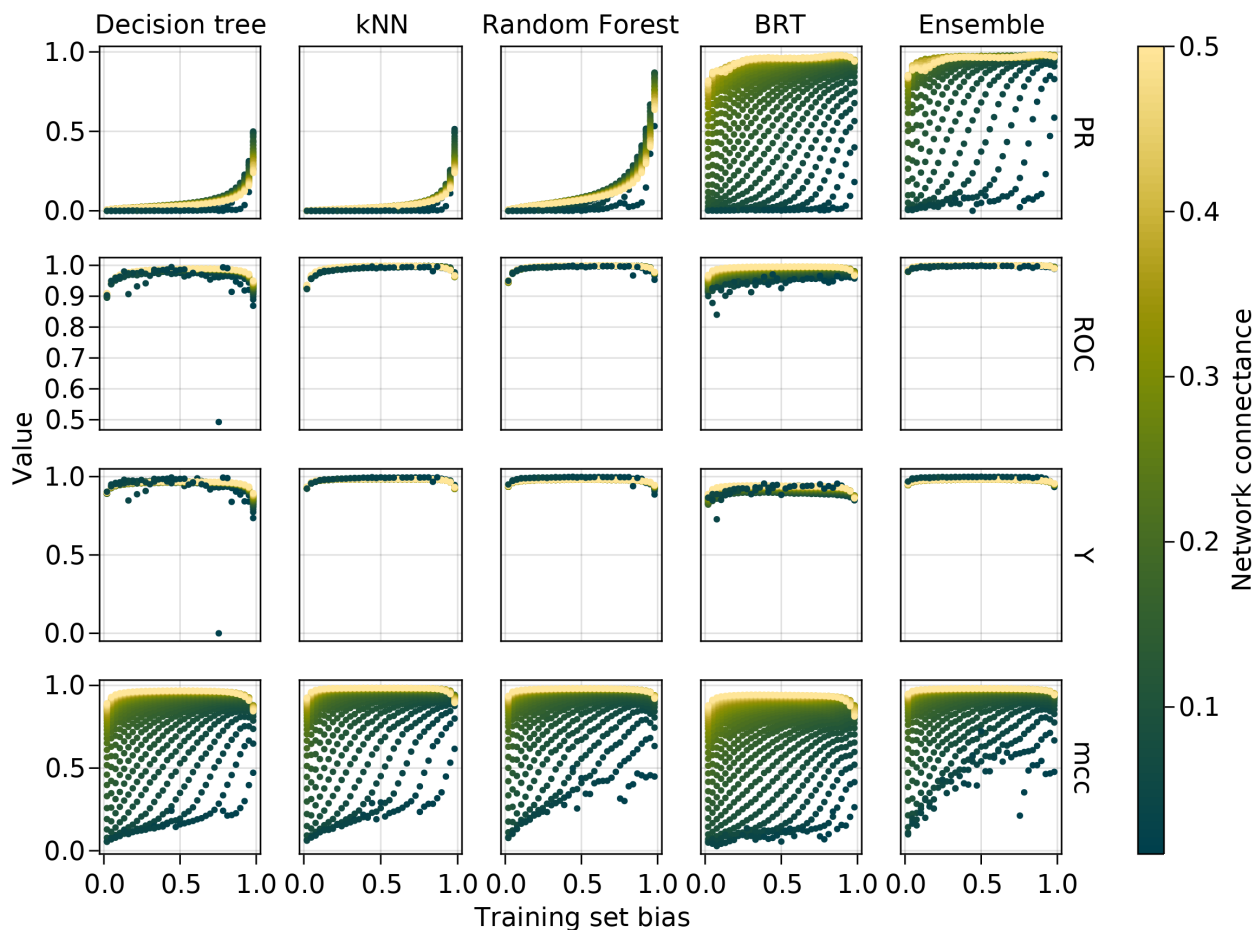


Figure 3: Response of MCC, Informedness, ROC-AUC, and PR-AUC to changes in the training set balance (on the x axis) for a series of increasing connectances (color). All of these values approach 1 for a good model, but should be lower when the prediction is more difficult. Informedness is consistently high, and by contrast, MCC increases with additional training set balance. Across all models, training on a more connected network is easier. 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.

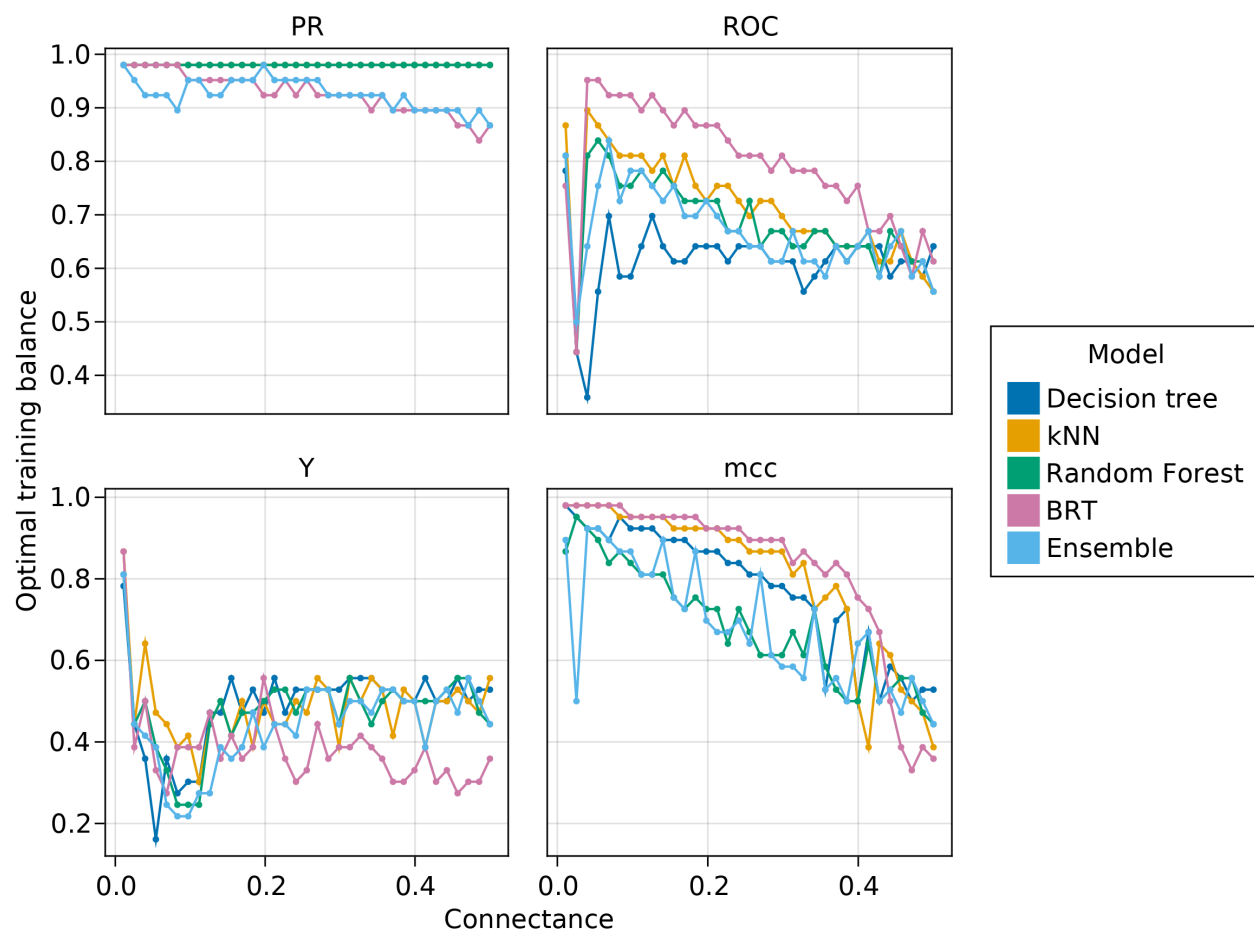


Figure 4: Value of the optimal training set balance for the different models and measures evaluated here, over a range of connectances. Informedness was reliably maximized for balanced training sets, and kept this behavior across models. For other measures, larger connectances in the true network allowed 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.

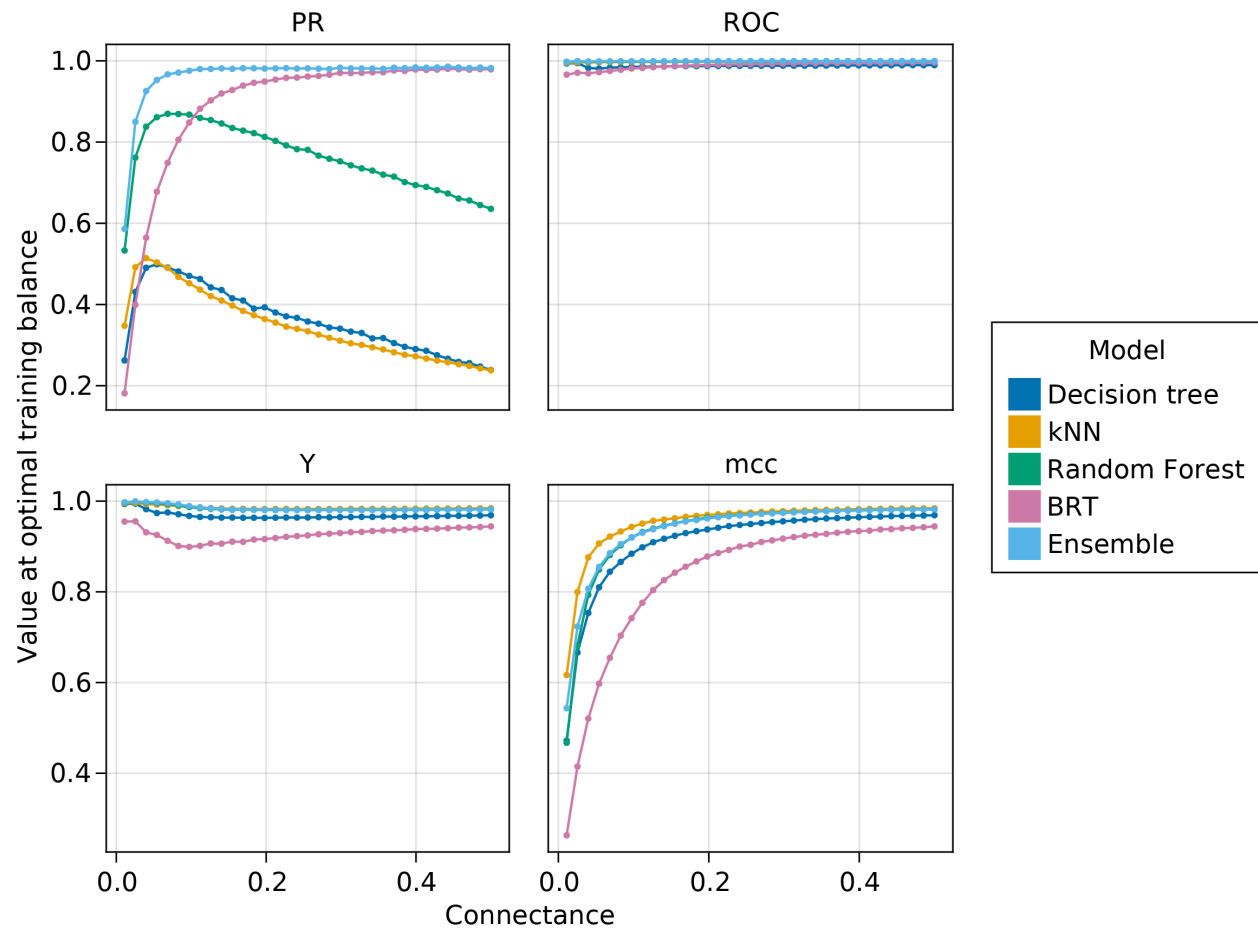


Figure 5: When trained on their optimally biased training set, most models were able to maximize their performance; this is not true when measuring PR-AUC for decision tree, k-NN, and to a lower extent RF. The ensemble had a consistently high performance despite incorporating low-performing models.

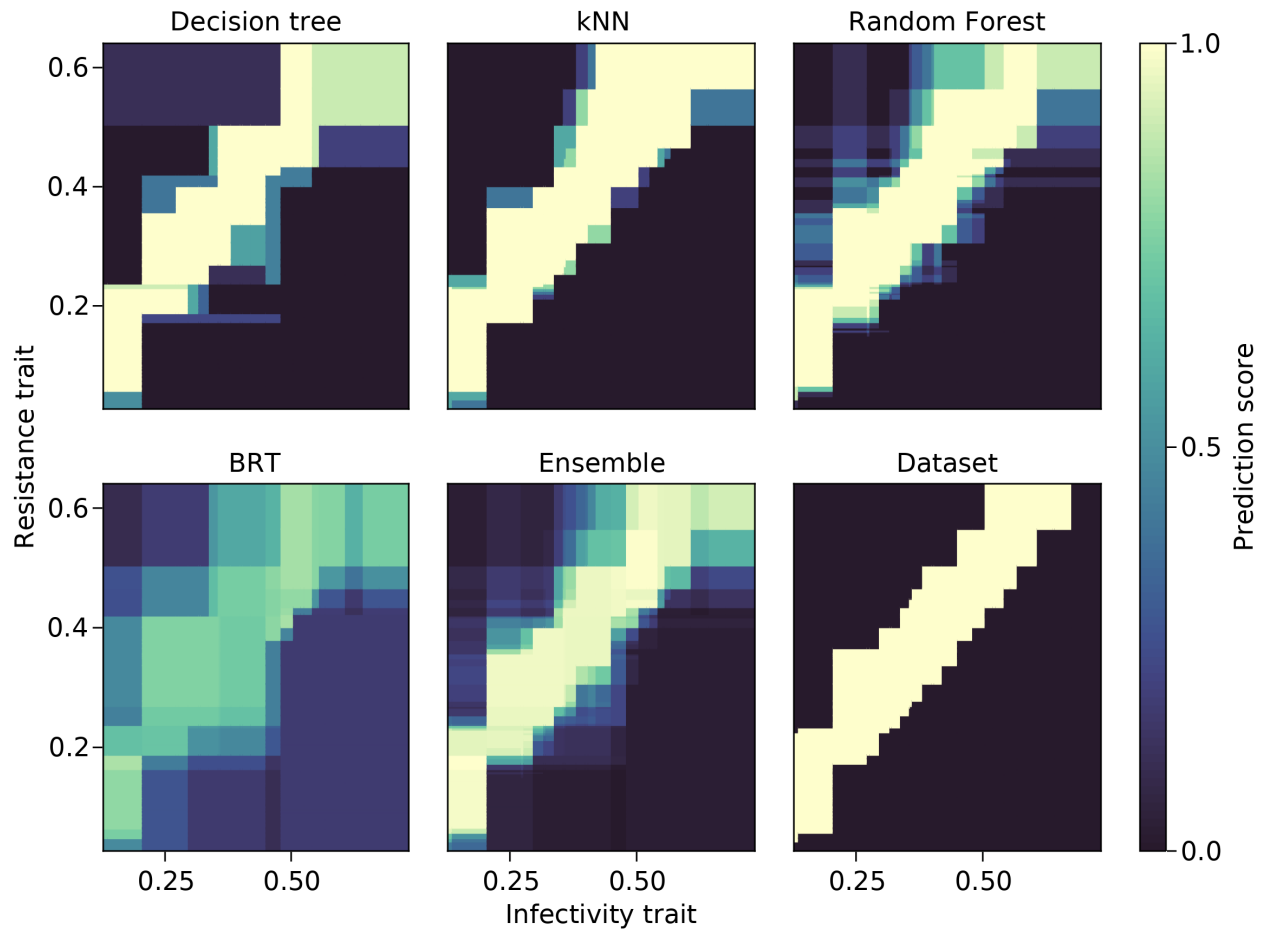


Figure 6: Visualisation of the raw (un-thresholded) models predictions for one instance of a network prediction problem (shown in the “Dataset” panel). Increasing the value of the ξ parameter would make the diagonal structure “broader”, leading to more interactions. A visual inspection of the results is important, as it highlights how some models can “miss” parts of the network; by combining them in an ensemble, these gaps compensate one another, and lead (in this case) to a better prediction.