

Guidelines for the supervised learning of species interactions

Timothée Poisot^{1,2}

¹ Université de Montréal ² Québec Centre for Biodiversity Sciences

Correspondance to:

Timothée Poisot — timothee.poisot@umontreal.ca

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



Last revision: *December 12, 2021*

1. The prediction of species interaction networks is gaining momentum as a way to circumvent limitations in data volume. Yet, ecological networks are challenging to predict because they are typically small and sparse. Dealing with extreme class imbalance is a challenge for most binary classifiers, and there are currently no guidelines as to how predictive models can be trained.
2. Using simple mathematical arguments and numerical experiments in which a variety of classifiers (for supervised learning) are trained on simulated networks, we develop a series of guidelines related to the choice of measures to use for model selection, and the degree of unbiasing to apply to the training dataset.
3. Classifier accuracy and the ROC-AUC are not informative measures for the performance of interaction prediction. PR-AUC is a fairer assessment of performance. In some cases, even standard measures can lead to selecting a more biased classifier because the effect of connectance is strong. The amount of correction to apply to the training dataset depends as a function of the classifier and the network connectance.
4. These results reveal that training machines to predict networks is a challenging task, and that in virtually all cases, the composition of the training set needs to be experimented on before performing the actual training. We discuss these consequences in the context of the low volume of data.

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

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

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

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

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

27 least amount of positive interaction data?

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

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

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

56 **Baseline values**

57 Intro

58 **Definition of the performance measures**

59 κ

60 F_β

61 informedness

62 MCC

63 ROC-AUC

64 PR-AUC

65 **Confusion matrix with skill and bias**

66 In this section, we will assume a network of connectance ρ , *i.e.* having ρS^2 interactions (where S is the
67 species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the
68 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
69 confusion matrix, which ends up expressing *relative* values).

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

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

76 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 77 to s , and off-diagonal elements equal to $(1 - s)$, and re-express the skill-adjusted confusion matrix as
 78 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

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

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

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

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

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

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

86 What are the baseline values of performance measures?

87 In this section, we will change the values of b , s , and ρ , and report how the main measures discussed in
 88 the introduction (MCC, F_1 , κ , and informedness) are responding to issues with the classifier. Before we do
 89 so, it is important to explain why we will not focus on accuracy too much. Accuracy is the number of
 90 correct predictions ($\text{Tr}(\mathbf{C})$) divided by the sum of the confusion matrix. For a no-skill, no-bias classifier,
 91 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
 92 other words, the values of accuracy are expected to be so high that they are not really informative (this is
 93 simply explained by the fact that for ρ small, $\rho^2 \ll (1 - \rho)^2$). More concerning is the fact that introducing

bias changes the response of accuracy in unexpected ways. Assuming a no-skill classifier, the numerator of accuracy becomes $b\rho^2 + (1 - b)(1 - \rho)^2$, which increases when b is low, which specifically means that at equal skill, a classifier that under-predicts interactions will have higher accuracy than an un-biased classifier. These issues are absent from balanced accuracy, but should nevertheless lead us to not report accuracy as the primary measure of network prediction success; moving forward, we will focus on other measures.

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

[Figure 1 about here.]

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

[Figure 2 about here.]

These two analyses point to the following recommendations: MCC is indeed more appropriate than κ , as

122 although sensitive to bias, it is sensitive in a consistent way. Informedness is appropriate at discriminating
123 between different skills, but confounded by bias. F_1 is increasing with bias, and should not be prioritized
124 to evaluate the performance of the model. The discussion of sensitivity to bias should come with a
125 domain-specific caveat: although it is likely that interactions documented in ecological networks are
126 correct, a lot of non-interactions are simply unobserved; as predictive models are used for data-inflation
127 (*i.e.* the prediction of new interactions), it is not necessarily a bad thing in practice to select models that
128 predict more interactions than the original dataset, because the original dataset misses some interactions.
129 Furthermore, the weight of positive interactions could be adjusted if some information about the extent of
130 undersampling exists (*e.g.* Branco et al., 2015).

131 **Illustration on empirical food webs**

132 In this section, we use a collection of 22 empirical networks (Thompson & Townsend, 2004; Thompson &
133 Townsend, 1999) from streams in New Zealand; these networks were previously analyzed by Pomeranz et
134 al. (2018), who found that a variety of rule-based models tended to increase the number of interactions.
135 Our goal here is not to infer which interactions are missing, but rather to estimate (i) how much precise
136 knowledge of the network influences the measures of classification performance, and (ii) whether an
137 increase in performance also results in more ecologically realistic networks. In order to do so, we use a
138 truncated Singular Value Decomposition (t-SVD) coupled to a Random Dot Product Graph (RDPG; *e.g.*
139 Dalla Riva & Stouffer, 2015) at increasingly high dimensions to reconstruct an approximation of the
140 network. Recent work by Poisot et al. (2021) and Strydom, Bouskila, et al. (2021) show that t-SVD and
141 RDPG are able to infer biotic interactions in, respectively, host-virus and predator-prey systems. The
142 t-SVD/RDPG approach is covered in detail by Strydom, Bouskila, et al. (2021); by approximating the
143 network at different ranks, where rank 1 is the coarsest, we can generate a matrix that can be thresholded
144 and compared to the original network. The variance explained at each rank is given by the relative sum of
145 the eigenvalues of the SVD up to the chosen rank, and measures how much information about the
146 network we can capture (Strydom, Dalla Riva, et al., 2021).

147 **Effect of classifier performance on metrics**

148 By changing the rank of approximation from 1 (variance explained is minimized) to 45 (variance
149 explained is close to 1), we show in fig. 3 that although all measures of classification performance increase
150 the more we know about the network, MCC is initially lower compared to informedness; similarly,
151 ROC-AUC is consistently higher compared to PR-AUC.

152 [Figure 3 about here.]

153 **Effect of classifier performance on network properties**

154 [Figure 4 about here.]

155 connectance, spectral radius (Staniczenko et al., 2013), entropy (Stock et al., 2021)

156 **Numerical experiments**

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

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

172 The dataset used for numerical experiments is composed of 64000 such (ξ, ν) pairs, on which four learners
173 are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random forest
174 regressor. All models were taken from the `MLJ.jl` package (Blaom et al., 2020; Blaom & Vollmer, 2020) in
175 Julia 1.7 (Bezanson et al., 2017). In order to pick the best adjacency matrix for a given learner, we
176 performed a thresholding approach using 500 steps on predictions from the testing set, and picking the
177 threshold that maximized Youden’s informedness, which is usually the optimized target for imbalanced
178 classification. During the thresholding step, we measured the area under the receiving-operator
179 characteristic (ROC-AUC) and precision-recall (PR-AUC) curves, as measures of overall performance over
180 the range of returned values. We report the ROC-AUC and PR-AUC, as well as a suite of other measures as
181 introduced in the next section, for the best threshold. The ensemble model was generated by summing the
182 predictions of all component models on the testing set (ranged in $[0, 1]$), then put through the same
183 thresholding process. The complete code to run the simulations is given as an appendix; running the final
184 simulation required 4.8 core days (approx. 117 hours).

185 After the simulations were completed, we removed all runs (*i.e.* pairs of ξ and ν) for which at least one of
186 the following conditions was met: the accuracy was 0, the true positive or true negative rates were 0, the
187 connectance was larger than 0.2. This removes both the obviously failed model runs, and the networks
188 that are more densely connected compared to the connectance of empirical food webs (and are therefore
189 less difficult to predict, being less imbalanced).

190 **Effect of training set bias on performance**

191 In fig. 5, we present the response of MCC and informedness to (i) four levels of network connectance and
192 (ii) a gradient of training set bias, for the four component models as well as the ensemble. All models
193 reached a higher performance on more connected networks, and using more biased training sets (with the
194 exception of ridge regression, whose informedness decreased in performance with training set bias). In all
195 cases, informedness was extremely high, which is an expected consequence of the fact that this is the
196 value we optimized to determine the cutoff. MCC increased with training set bias, although this increase
197 became less steep with increasing connectance. Interestingly, the ensemble almost always outclassed its
198 component models.

199 [Figure 5 about here.]

200 In fig. 6, we present the same information as fig. 5, this time using ROC-AUC and PR-AUC. ROC-AUC is
201 always high, and does not vary with training set bias. On the other hand, PR-AUC shows very strong
202 responses, increasing with training set bias. It is notable here that two classifiers that seemed to be
203 performing well (Decision Tree and Random Forest) based on their MCC are not able to reach a high
204 PR-AUC even at higher connectances. As in fig. 5, the ensemble outperforms its component models.

205 [Figure 6 about here.]

206 Based on the results presented in fig. 5 and fig. 6, it seems that informedness and ROC-AUC are not
207 necessarily able to discriminate between good and bad classifiers (although this result may be an artifact
208 for informedness, as it has been optimized when thresholding). On the other hand, MCC and PR-AUC
209 show a strong response to training set bias, and may therefore be more useful at model comparison.

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

211 The previous results revealed that the measure of classification performance responds both to the bias in
212 the training set *and* to the connectance of the network; from a practical point of view, assembling a
213 training set requires to withhold positive information, which in ecological networks are very scarce (and
214 typically more valuable than negatives, on which there is a doubt). For this reason, across all values of
215 connectance, we measured the training set bias that maximized a series of performance measures. When
216 this value is high, the training set needs to skew positive in order to get a good model; when this value is
217 about 0.5, the training set needs to be artificially balanced to optimize the model performance. These
218 results are presented in fig. 7.

219 [Figure 7 about here.]

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

227 Guidelines for the assesment of network predictive models

228 The results presented here highlight an interesting paradox: larger networks (with lower connectance)
 229 require more training set bias in order to maximize model performance fig. 7, but are also more difficult to
 230 predict according to MCC and PR-AUC fig. 8. This suggests that the task of network prediction will be
 231 difficult regardless of network size: by being limited by the *frequency* of interactions when the network is
 232 large, and by being limited by the *number* of interactions when the network is small. Nevertheless, based
 233 on the simulations and numerical experiments, it is possible to formulate a series of recommendations for
 234 the evaluation of network prediction models.

235 First, because we should have more trust in reported interactions than in reported absences of interactions,
 236 we can draw on previous literature to recommend informedness as a measure to decide on a threshold
 237 (Chicco et al., 2021); this being said, because informedness is insensitive to bias, the model performance is
 238 better evaluated through the use of MCC fig. 5. Because F_1 is monotonously sensitive to classifier bias
 239 fig. 1 and network connectance fig. 2, MCC should be prefered as a measure of model evaluation.

240 Second, because the PR-AUC responds more to network connectance fig. 8 and training set imbalance
 241 fig. 6, it should be used as a measure of model performance over the ROC-AUC. This is not to say that
 242 ROC-AUC should be discarded (in fact, a low ROC-AUC is a sign of an issue with the model), but that its
 243 interpretation should be guided by the PR-AUC value. This again echoes recommendations from other
 244 fields (Saito & Rehmsmeier, 2015).

245 Thirdly, regardless of network connectance *or* measure to evaluate the model performance, as long as the
 246 network connectance is larger than ≈ 0.1 , artificially balancing the training set to have equiprevalence will
 247 give the best possible results. This was true for all models.

248 Finally, it is noteworthy that the ensemble model was systematically better than the component models;
 249 even when poor models were included (Random Forest and Decision Tree), the ensemble was able to
 250 leverage the different biases expressed by the models to make an overall more accurate prediction. We do
 251 not expect that ensembles will *always* be better than single models. In a recent multi-model comparison,
 252 Becker et al. (2021) found that the ensemble was *not* the best model. There is no general conclusion to

draw from this besides reinforcing the need to be pragmatic about which models should be included in the ensemble, or whether to use an ensemble at all. In a sense, the surprising performance of the ensemble model should form the basis of the last recommendation: optimal training set bias and its interaction with connectance and binary classifier is, in a sense, an hyperparameter that should be assessed. The distribution of results in fig. 7 and fig. 8 show that there are variations around the trend; furthermore, networks with different structures than the one we simulated here may respond in different ways.

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. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computeCanada.ca) through the Narval general purpose cluster. TP is supported by a NSERC Discovery Grant and Discovery Acceleration Supplement, and by a grant from the Institut de Valorisation des Données (IVADO).

References

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>

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

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>

Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, 5(55), 2704. <https://doi.org/10.21105/joss.02704>

Blaom, A. D., & Vollmer, S. J. (2020, December 31). *Flexible model composition in machine learning and its implementation in MLJ*. <http://arxiv.org/abs/2012.15505>

280 Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews
 281 Correlation Coefficient metric. *PloS One*, 12(6), e0177678.
 282 <https://doi.org/10.1371/journal.pone.0177678>

283 Branco, P., Torgo, L., & Ribeiro, R. (2015, May 13). *A Survey of Predictive Modelling under Imbalanced*
 284 *Distributions*. <http://arxiv.org/abs/1505.01658>

285 Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1
 286 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6.
 287 <https://doi.org/10.1186/s12864-019-6413-7>

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

291 Dalla Riva, G. V., & Stouffer, D. B. (2015). Exploring the evolutionary signature of food webs' backbones
 292 using functional traits. *Oikos*, 125(4), 446–456. <https://doi.org/10.1111/oik.02305>

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

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

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

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

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

304 Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N.,
 305 & Carlson, C. J. (2021, May 31). *Imputing the mammalian virome with linear filtering and singular*
 306 *value decomposition*. <http://arxiv.org/abs/2105.14973>

307 Pomeranz, J. P., Thompson, R. M., Poisot, T., & Harding, J. S. (2018). Inferring predator-prey interactions
 308 in food webs. *Methods in Ecology and Evolution*, 0. <https://doi.org/10.1111/2041-210X.13125>

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

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

315 Staniczenko, P. P. A., Kopp, J. C., & Allesina, S. (2013). The ghost of nestedness in ecological networks.
 316 *Nature Communications*, 4(1), 1391. <https://doi.org/10.1038/ncomms2422>

317 Stock, M., Hoebeke, L., & De Baets, B. (2021). Disentangling the Information in Species Interaction
 318 Networks. *Entropy*, 23(6, 6), 703. <https://doi.org/10.3390/e23060703>

319 Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M. J., Fortin, M.-J., Hemming, V.,
 320 Mercier, B., Pollock, L., Runghen, R., Riva, G. V. D., & Poisot, T. (2021). *Food web reconstruction through*
 321 *phylogenetic transfer of low-rank network representation*. <https://doi.org/10.32942/osf.io/y7sdz>

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

327 Strydom, T., Dalla Riva, G. V., & Poisot, T. (2021). SVD Entropy Reveals the High Complexity of Ecological
 328 Networks. *Frontiers in Ecology and Evolution*, 9. <https://doi.org/10.3389/fevo.2021.623141>

329 Thompson, R. M., & Townsend, C. R. (2004). Land-use influences on New Zealand stream communities:
 330 Effects on species composition, functional organisation, and food-web structure. *New Zealand Journal*
 331 *of Marine and Freshwater Research*, 38(4), 595–608.
 332 <https://doi.org/10.1080/00288330.2004.9517265>

333 Thompson, R. M., & Townsend, C. R. (1999). The Effect of Seasonal Variation on the Community
 334 Structure and Food-Web Attributes of Two Streams: Implications for Food-Web Science. *Oikos*, 87(1),

335 75–88. <https://doi.org/10.2307/3546998>

336 Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine
337 learning in genomics. *Nature Reviews Genetics*, 1–13.

338 <https://doi.org/10.1038/s41576-021-00434-9>

339 Williams, R., & Martinez, N. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.

340 <http://userwww.sfsu.edu/>

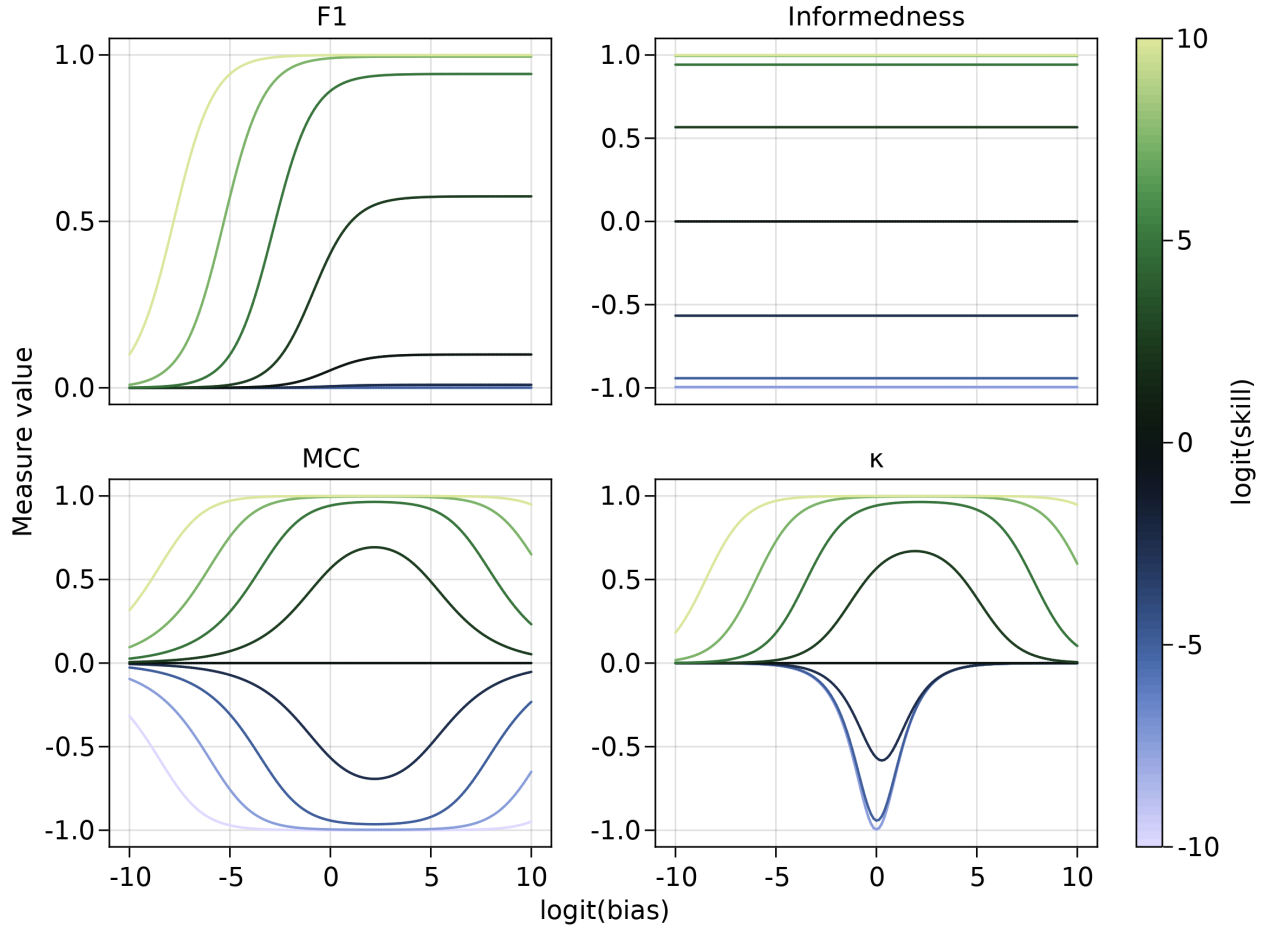


Figure 1: Consequences of changing the classifier skills (s) and bias (b) for a connectance $\rho = 0.15$, on accuracy, F_1 , positive predictive value, and κ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions. The F_1 score increases with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

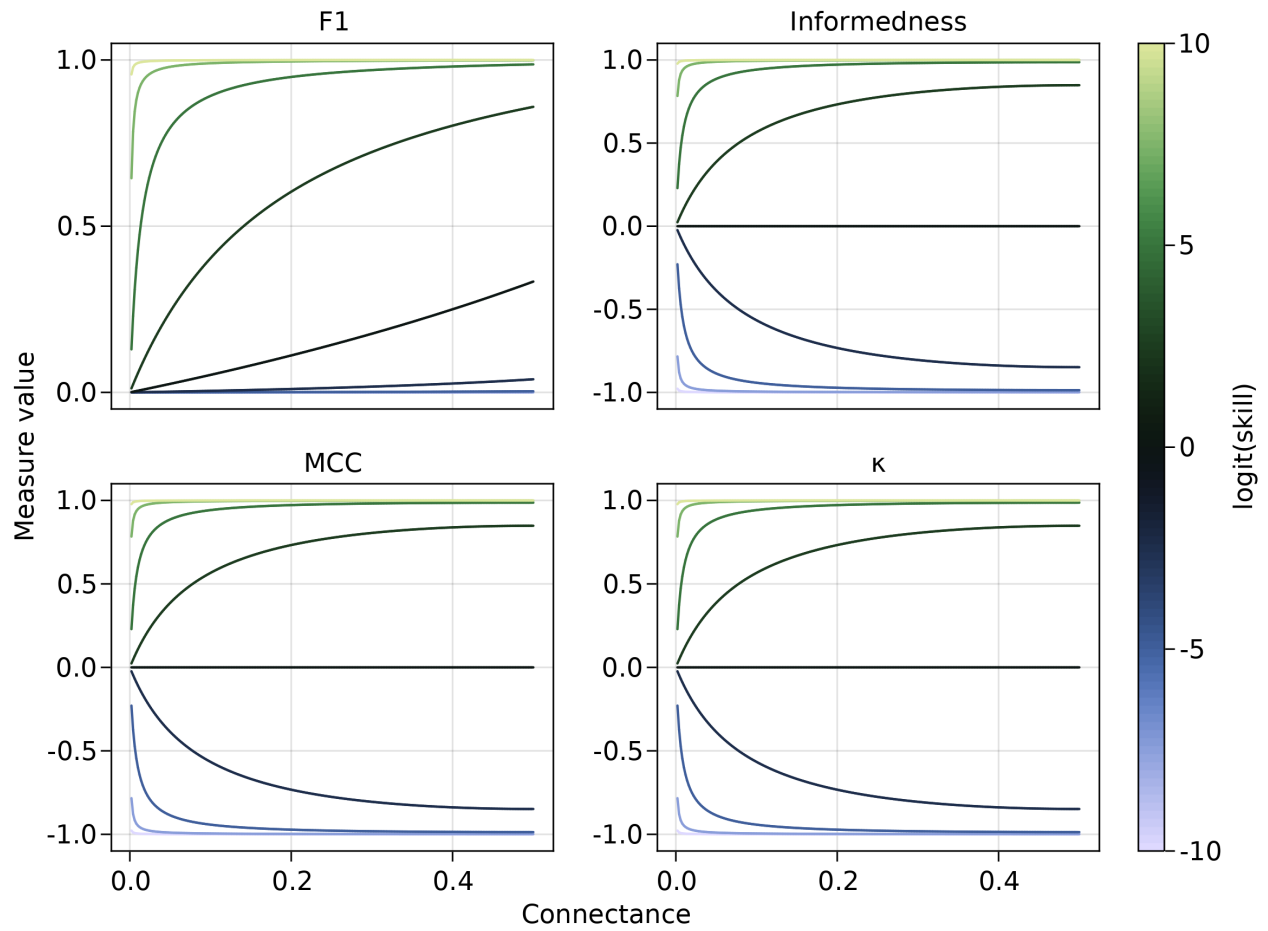


Figure 2: As in fig. 1, consequences of changing connectance for different levels of classifier skill, assuming no classifier bias. Informedness, κ , and MCC do increase with connectance, but only when the classifier is not no-skill; by way of contrast, a more connected network will give a higher F_1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

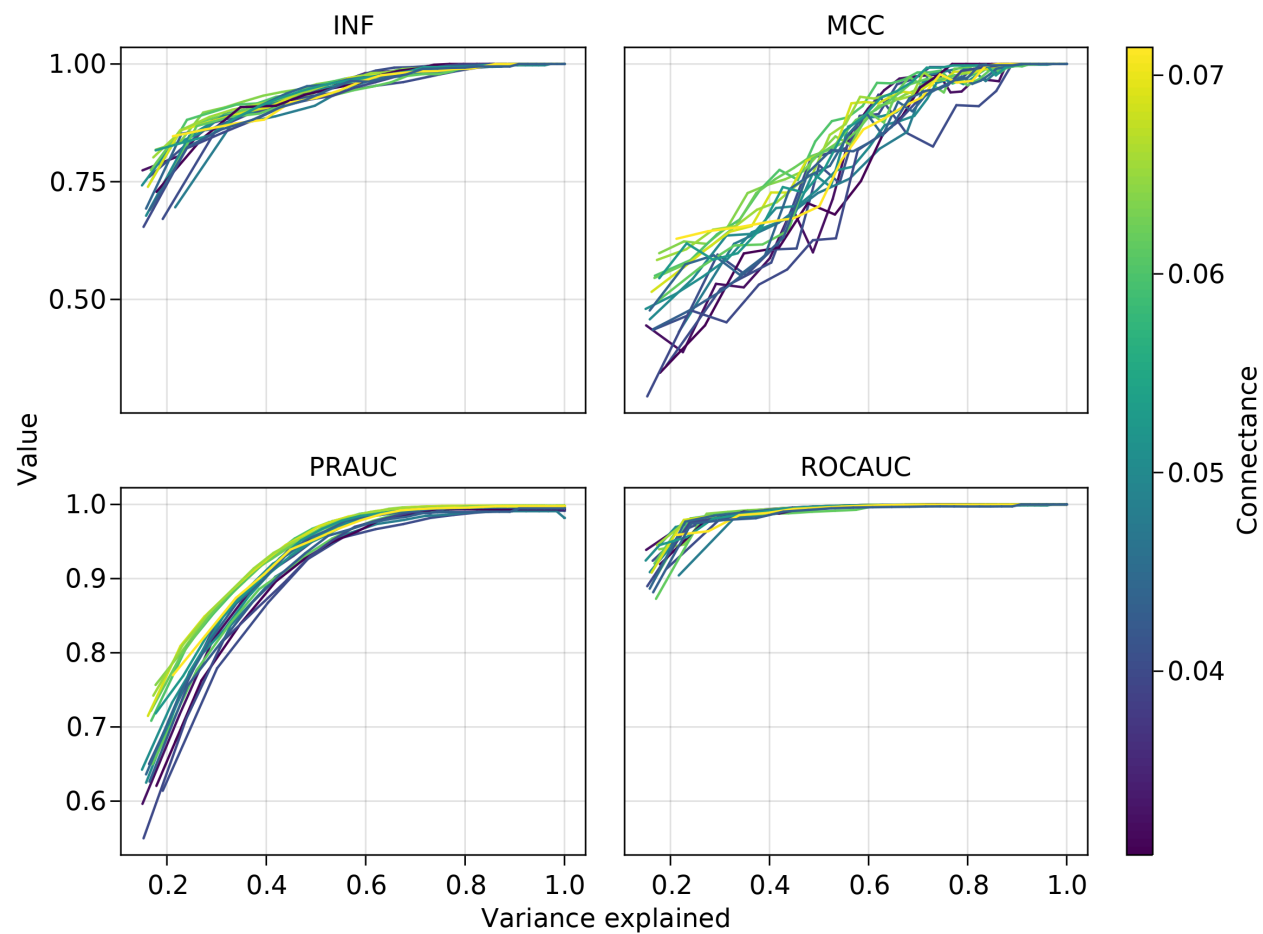


Figure 3: TODO

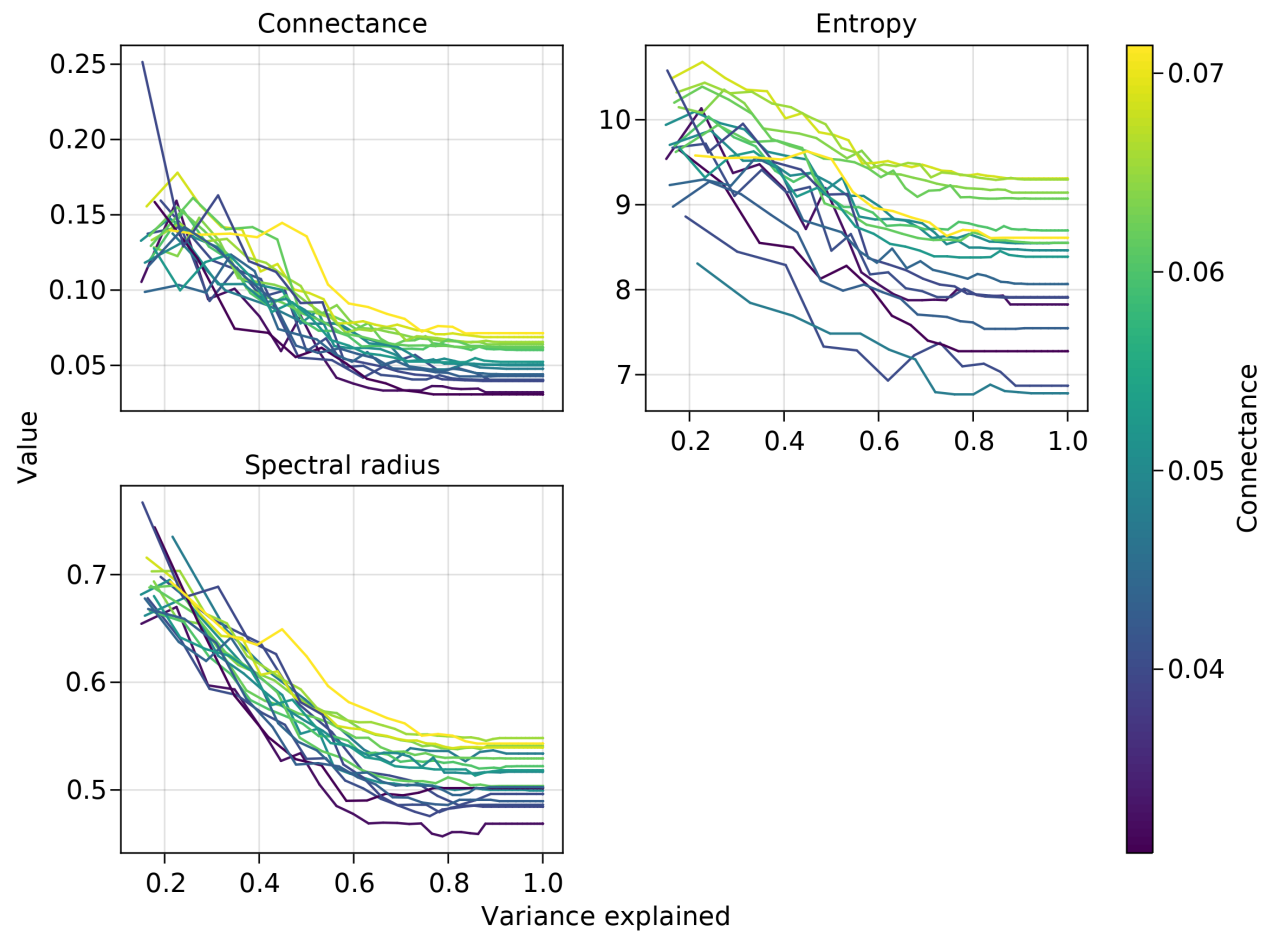


Figure 4: TODO

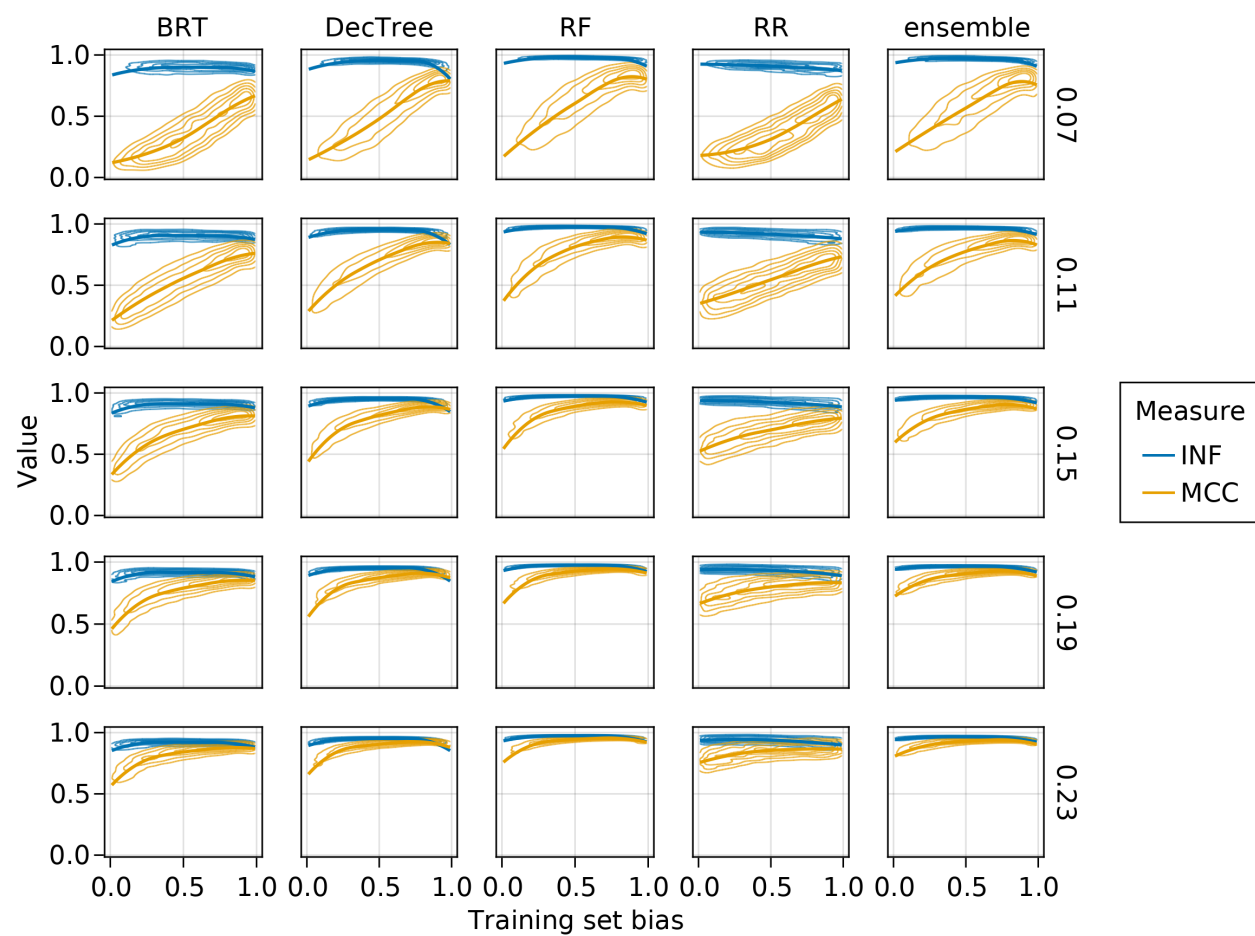


Figure 5: TODO

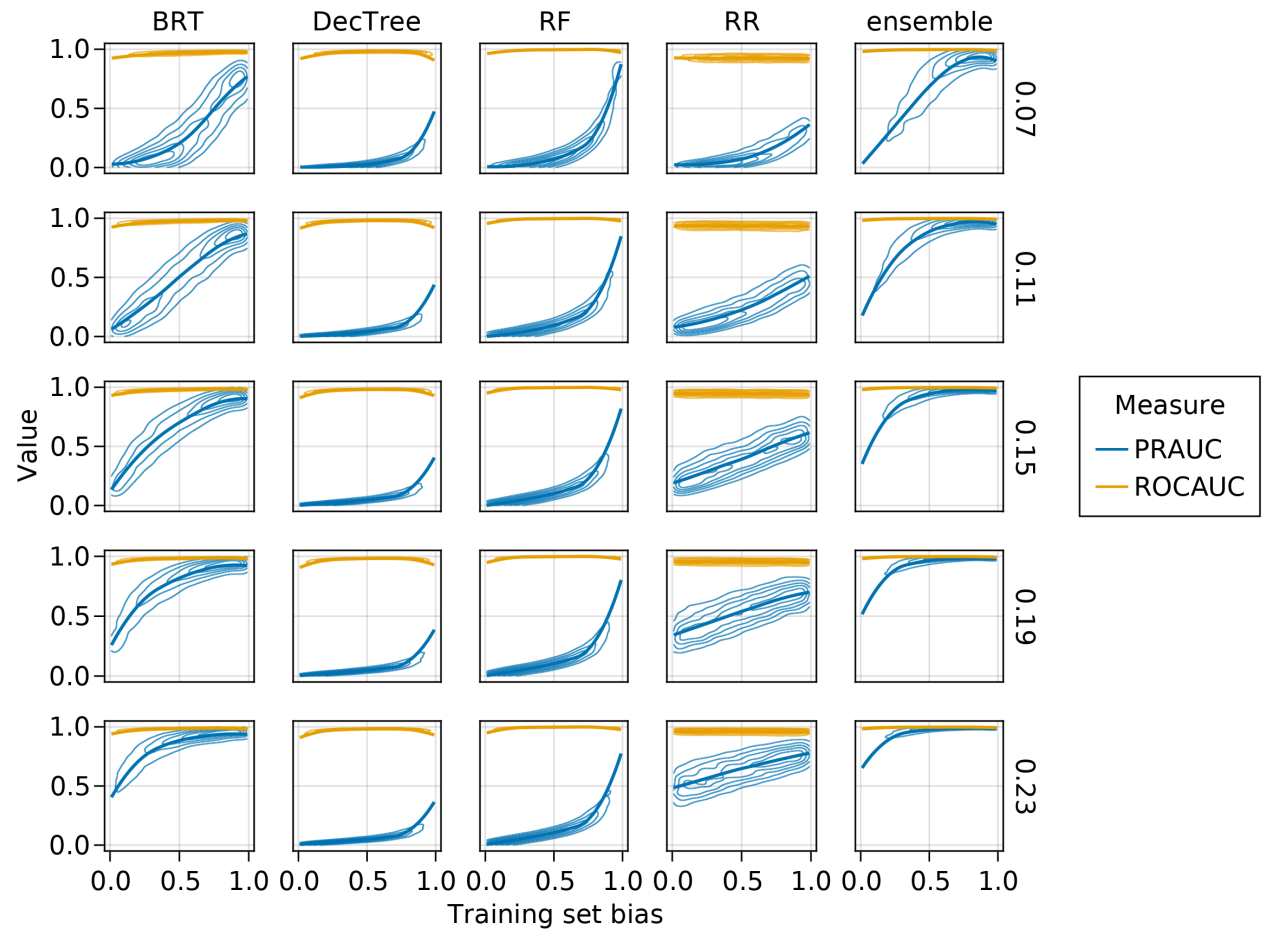


Figure 6: TODO

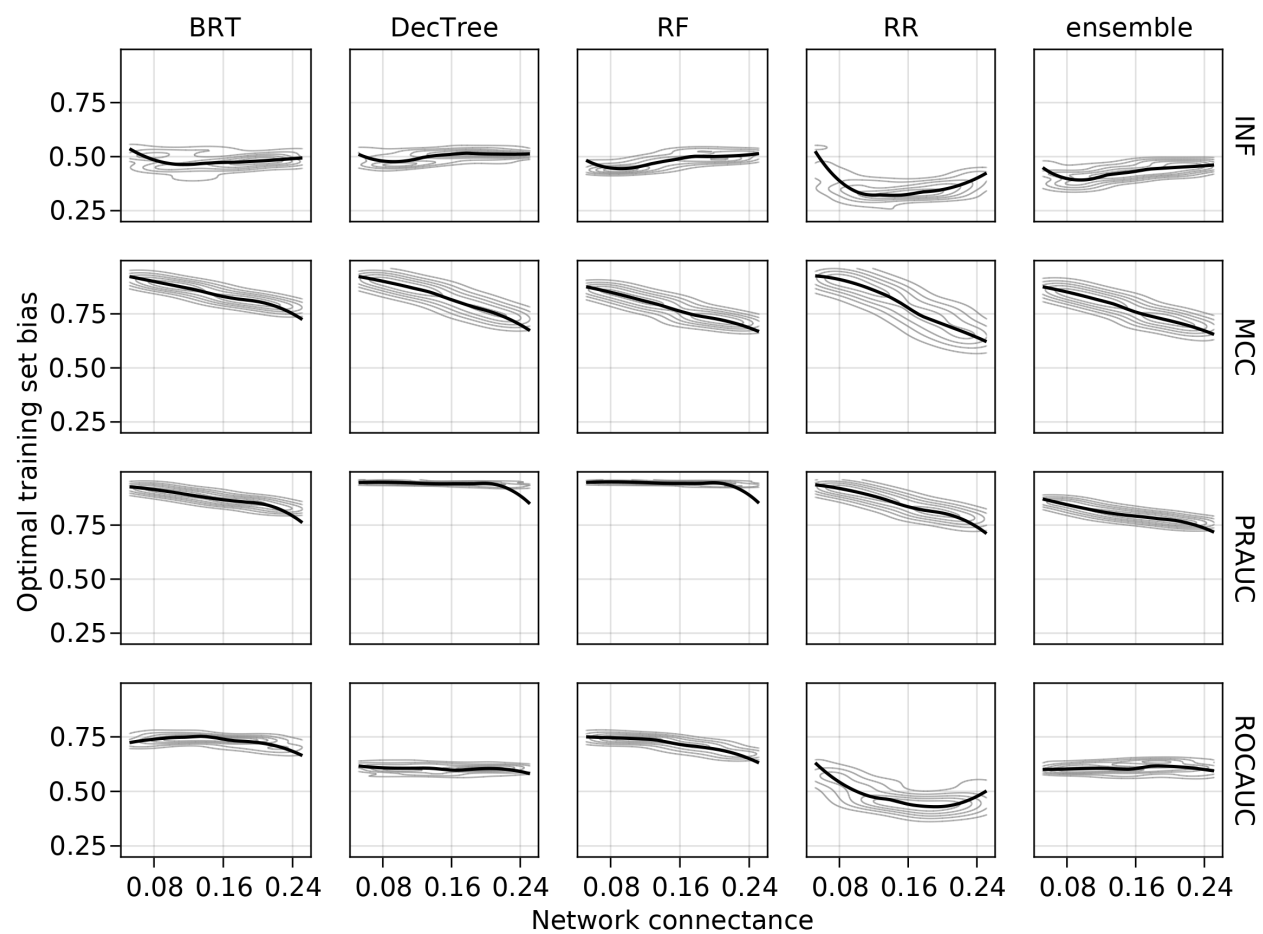


Figure 7: TODO

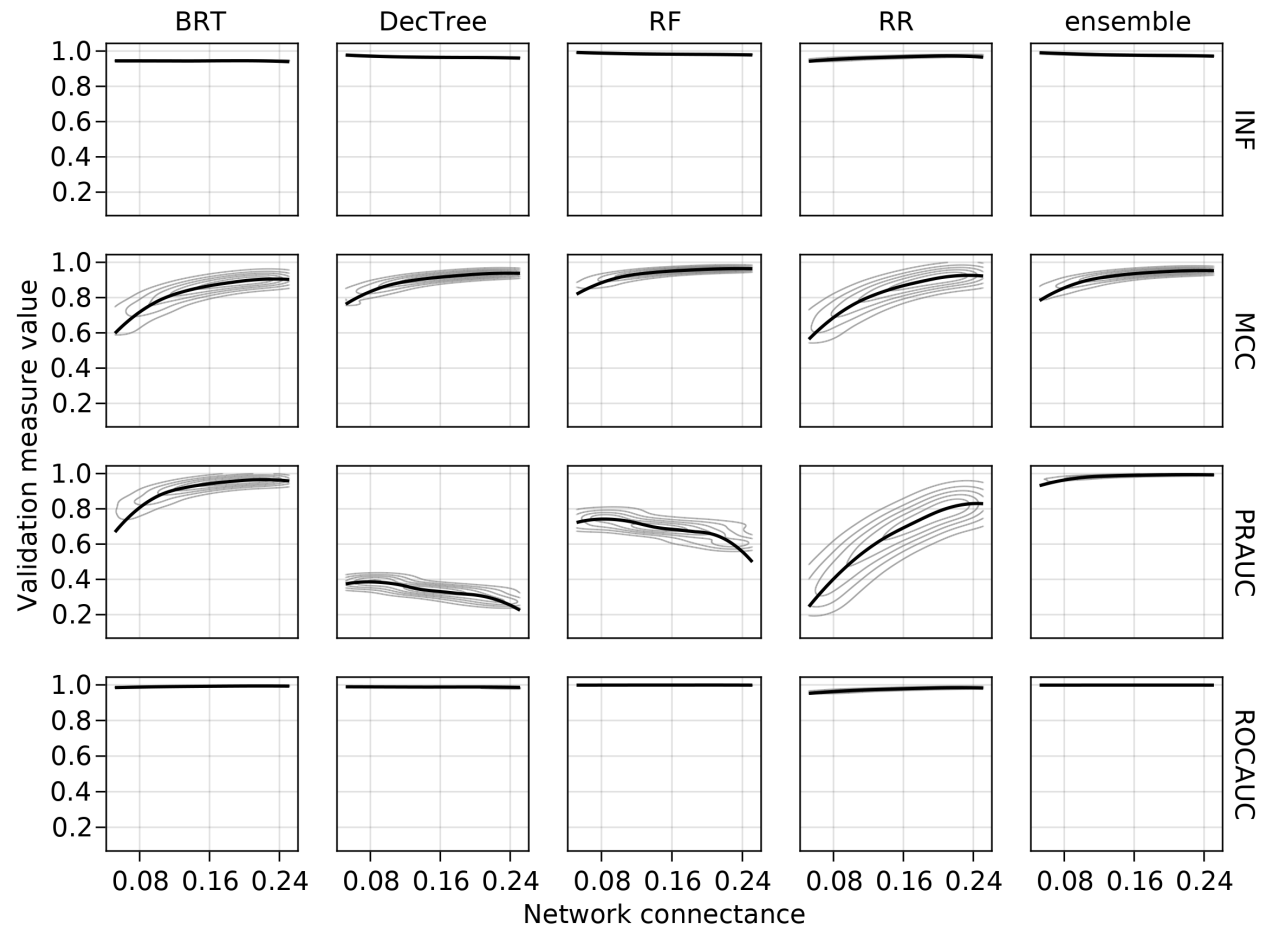


Figure 8: TODO