

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 11, 2021*

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

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

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

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

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

The same approach is used to evaluate *e.g.* species distribution models (SDMs). Indeed, the training and evaluation of SDMs as binary classifiers suffers from the same issue of low prevalence. In a previous work, Allouche et al. (2006) suggested that 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 Confusion matrix with skill and bias

58 In this section, we will assume a network of connectance ρ , *i.e.* having ρS^2 interactions (where S is the
 59 species richness), and $(1 - \rho)S^2$ non-interactions. Therefore, the vector describing the *true* state of the
 60 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
 61 confusion matrix, which ends up expressing *relative* values).

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

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

68 In order to regulate the skill of this classifier, we can define a skill matrix \mathbf{S} with diagonal elements equal
 69 to s , and off-diagonal elements equal to $(1 - s)$, and re-express the skill-adjusted confusion matrix as
 70 $\mathbf{M} \odot \mathbf{S}$, *i.e.*

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

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

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

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

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

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

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

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

79 In this section, we will change the values of b , s , and ρ , and report how the main measures discussed in
80 the introduction (MCC, F_1 , κ , and informedness) are responding to issues with the classifier. Before we do
81 so, it is important to explain why we will not focus on accuracy too much. Accuracy is the number of
82 correct predictions ($\text{Tr}(\mathbf{C})$) divided by the sum of the confusion matrix. For a no-skill, no-bias classifier,
83 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
84 other words, the values of accuracy are expected to be so high that they are not really informative (this is
85 simply explained by the fact that for ρ small, $\rho^2 \ll (1-\rho)^2$). More concerning is the fact that introducing
86 bias changes the response of accuracy in unexpected ways. Assuming a no-skill classifier, the numerator
87 of accuracy becomes $b\rho^2 + (1-b)(1-\rho)^2$, which increases when b is low, which specifically means that at
88 equal skill, a classifier that under-predicts interactions will have higher accuracy than an un-biased
89 classifier. These issues are absent from balanced accuracy, but should nevertheless lead us to not report
90 accuracy as the primary measure of network prediction success; moving forward, we will focus on other
91 measures.

92 In order to examine how MCC, F_1 , κ , and informedness change w.r.t. the imbalance, skill, and bias, we
93 performed a grid exploration of the values of $\text{logit}(s)$ and $\text{logit}(b)$ linearly from -10 to 10 , of ρ linearly in
94 $[0, 0.5]$, which is within the range of usually observed connectance values for empirical food webs. Note
95 that at this point, there is no food web model to speak of; rather, the confusion matrix we discuss can be
96 obtained for any classification task.

[Figure 1 about here.]

Are the measures affected by connectance?

[Figure 2 about here.]

Numerical experiments

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

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

The dataset used for numerical experiments is composed of 20000 such (ξ, ν) pairs, on which four learners are trained: a decision tree regressor, a boosted regression tree, a ridge regressor, and a random forest regressor. All models were taken from the `MLJ.jl` package (Blaom et al., 2020; Blaom & Vollmer, 2020) in Julia 1.7 (Bezanson et al., 2017). In order to pick the best adjacency matrix for a given learner, we performed a thresholding approach using 500 steps on predictions from the testing set, and picking the threshold that maximized Youden's informedness, which is usually the optimized target for imbalanced classification. During the thresholding step, we measured the area under the receiving-operator

characteristic (ROC-AUC) and precision-recall (PR-AUC) curves, as measures of overall performance over the range of returned values. We report the ROC-AUC and PR-AUC, as well as a suite of other measures as introduced in the next section, for the best threshold. The ensemble model was generated by summing the predictions of all component models on the testing set (ranged in $[0, 1]$), then put through the same thresholding process. The complete code to run the simulations is given as an appendix.

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

Effect of training set bias on performance

Required amount of positives to get the best performance

Guidelines for prediction

References

- Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>
- Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, 5(55), 2704. <https://doi.org/10.21105/joss.02704>
- Blaom, A. D., & Vollmer, S. J. (2020, December 31). *Flexible model composition in machine learning and its implementation in MLJ*. <http://arxiv.org/abs/2012.15505>

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

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

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

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

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

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

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

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

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

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

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

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

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

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

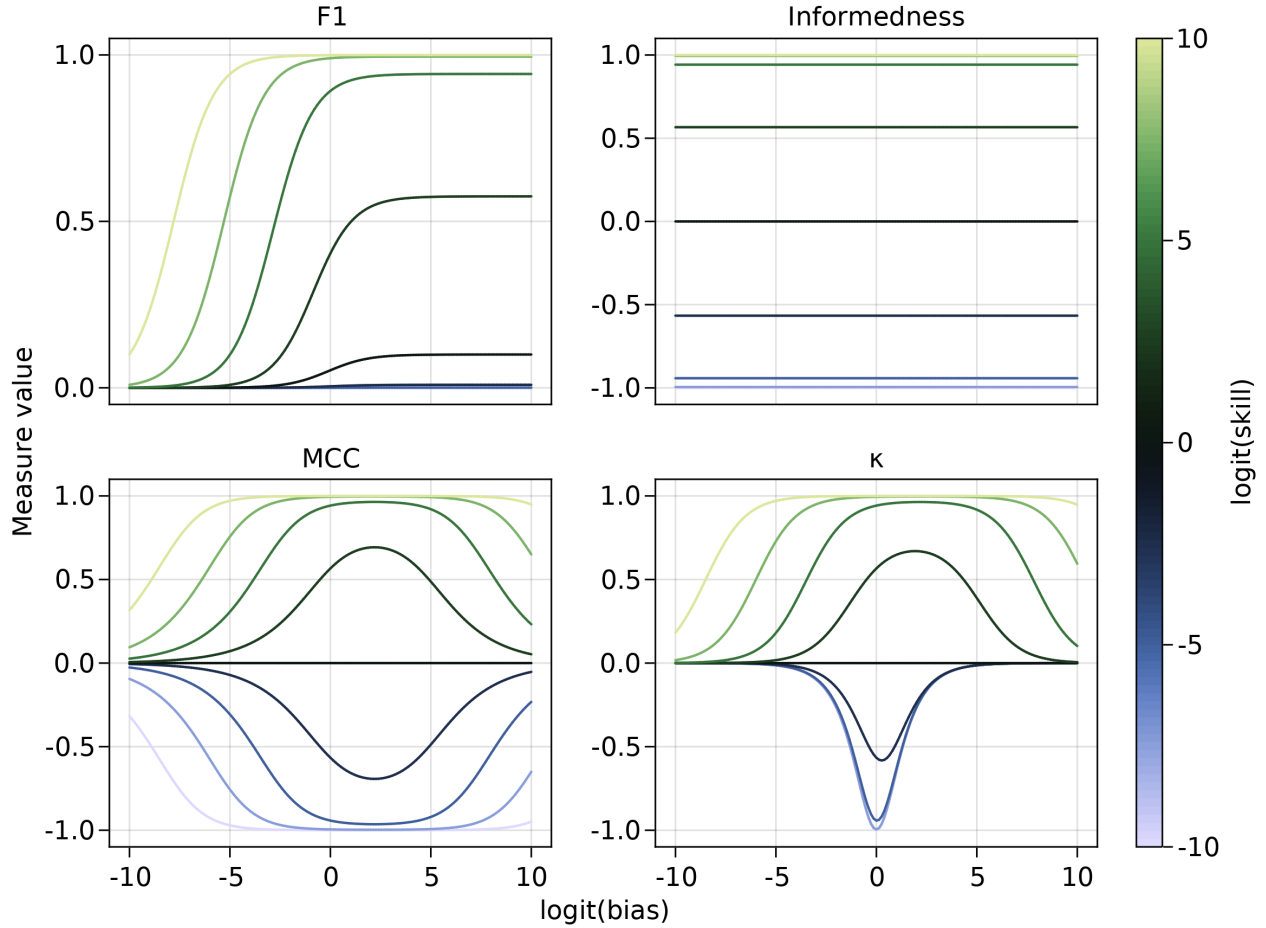


Figure 1: Consequences of changing the classifier skills (s) and bias (b) for a connectance $\rho = 0.15$, on accuracy, F_1 , positive predictive value, and κ . Accuracy increases with skill, but also increases when the bias tends towards estimating *fewer* interactions. The F_1 score increases with 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.

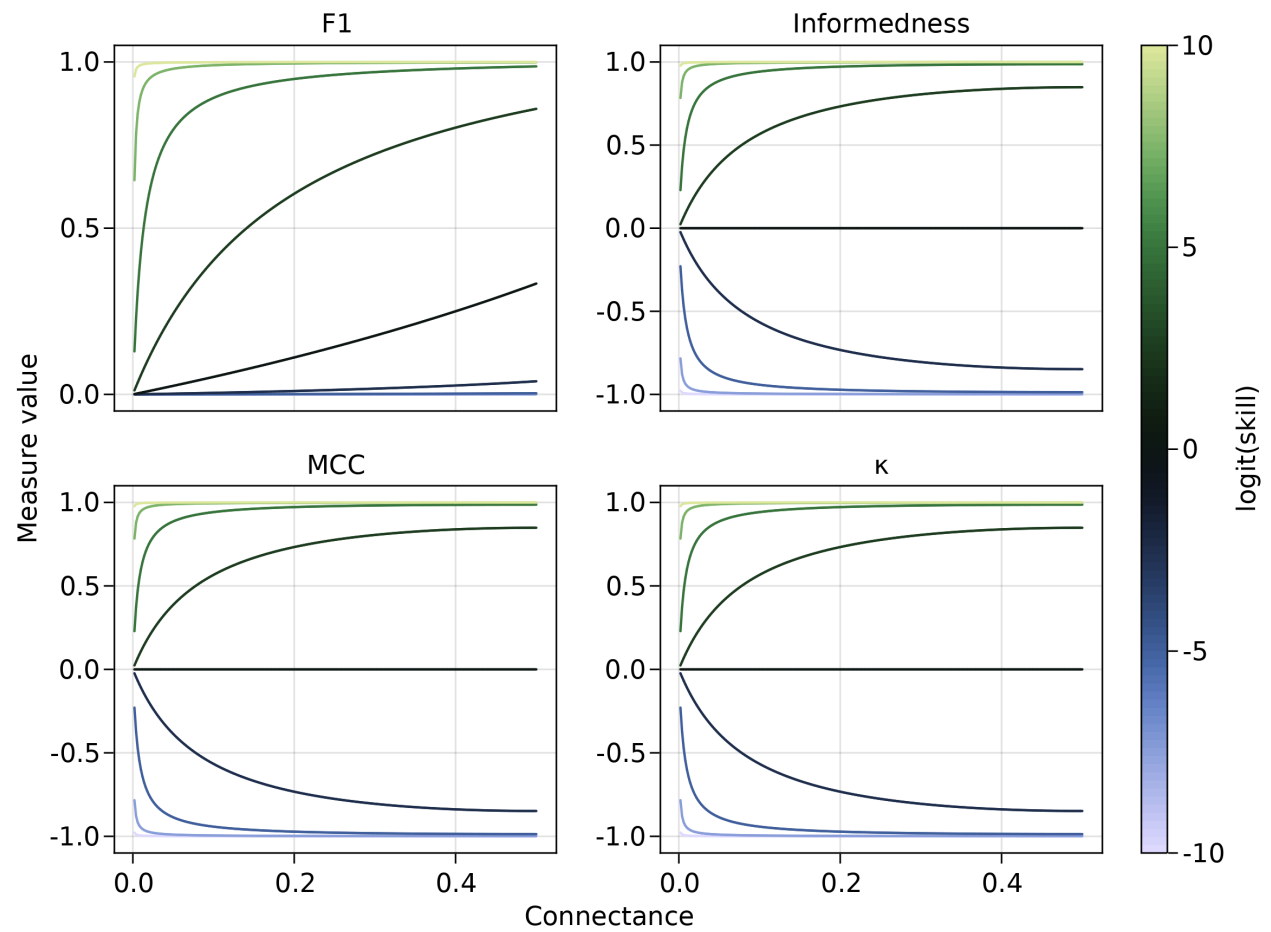


Figure 2: TODO