Supplementary Data

0.1 Supplementary tables for 10-fold cross-validation

Supplementary Tables 1–4 give results averaged over 10 trials of 10-fold cross-validation for each of the 4 benchmark data sets. Standard deviations are shown in parentheses. KRM means Kernel Regression Model (Yamanishi et al., 2008) and BLM means Bipartite Local Models. m is a given function that accepts several predictions for the same edge and outputs an aggregated prediction. Here, m outputs the largest score from the set of input scores, though other choices are possible. Each table is divided into four parts:

- The first gives AUC (Area Under ROC Curve) and AUPR (Area Under Precision-Recall) when performing leave-one-out on potential drugs (d).
- The second gives results when performing leave-one-out on potential target proteins (t).
- The third gives results when combining two or four leave-one-out predictions for the same edge.
- The fourth gives results using the nearest neighbor algorithm and leave-one-out on potential drugs (d), potential target proteins (t) and the result obtained when combining the two.

In each table, (*) indicates the original AUC results for the KRM method (Yamanishi et al., 2008) and (**) the best AUC results for methods introduced in this article. (***) is the best AUPR result across all methods.

Table 1: **Prediction performance for the enzyme data set.**

Method	AUC	AUPR
KRMd	85.9 (0.6)*	38.1 (2.7)
BLMd	83.3 (1.1)	38.9 (1.3)
m(KRMd,BLMd)	87.3 (0.9)	39.8 (0.9)
KRMt	96.4 (0.1)*	80.7 (0.3)
BLMt	93.5 (0.6)	81.0 (1.2)
m(KRMt,BLMt)	95.9 (0.0)	81.1 (0.2)
m(KRMd,KRMt)	97.2 (0.0)	82.9 (0.4)
m(BLMd,BLMt)	97.0 (0.2)	83.2 (0.6)
m(KRMd,KRMt,BLMd,BLMt)	97.5 (0.0)**	83.4 (0.2)***
NNd	68.3 (5.0)	34.4 (1.5)
NNt	89.5 (0.4)	77.2 (0.5)
m(NNd,NNt)	97.1 (0.1)	71.6 (0.6)

Table 2: **Prediction performance for the ion channel data set.**

Method	AUC	AUPR
KRMd	73.8 (0.9)*	31.4 (1.0)
BLMd	74.4 (1.1)	31.9 (1.4)
m(KRMd,BLMd)	73.8 (0.6)	32.5 (0.9)
KRMt	95.5 (0.3)*	80.8 (0.7)
BLMt	93.2 (0.1)	80.0 (0.5)
m(KRMt,BLMt)	96.0 (0.2)	82.7 (0.3)***
m(KRMd,KRMt)	96.7 (0.2)	76.3 (0.8)
m(BLMd,BLMt)	96.8 (0.1)	76.8 (0.8)
m(KRMd,KRMt,BLMd,BLMt)	97.4 (0.0)**	77.6 (0.6)
NNd	64.9 (0.7)	23.6 (1.7)
NNt	89.0 (0.4)	74.1 (0.9)
m(NNd,NNt)	91.9 (0.3)	55.1 (1.9)

Table 3: **Prediction performance for the GPCR data set.**

Method	AUC	AUPR
KRMd	88.3 (0.6)*	41.2 (1.3)
BLMd	81.7 (0.7)	38.1 (0.7)
m(KRMd,BLMd)	88.2 (0.4)	42.3 (1.4)
KRMt	92.8 (0.4)*	61.9 (2.4)
BLMt	85.7 (0.9)	54.7 (1.2)
m(KRMt,BLMt)	92.0 (0.3)	61.2 (1.7)
m(KRMd,KRMt)	95.2 (0.4)	67.5 (1.0)
m(BLMd,BLMt)	94.8 (0.3)	65.2 (0.6)
m(KRMd,KRMt,BLMd,BLMt)	95.8 (0.1)**	68.3 (0.8)***
NNd	69.3 (0.7)	33.5 (1.6)
NNt	82.4 (0.6)	60.1 (2.3)
m(NNd,NNt)	89.0 (0.4)	55.0 (2.5)

Table 4: **Prediction performance for the nuclear receptor data set.**

Method	AUC	AUPR
KRMd	87.9 (1.0)*	50.7 (3.3)
BLMd	80.7 (1.4)	40.2 (3.6)
m(KRMd,BLMd)	87.5 (1.0)	51.0 (3.9)
KRMt	90.1 (2.0)*	61.3 (8.1)
BLMt	53.1 (2.9)	34.8 (1.9)
m(KRMt,BLMt)	87.4 (1.7)	57.1 (6.9)
m(KRMd,KRMt)	94.0 (1.4)**	74.0 (3.2) ***
m(BLMd,BLMt)	85.0 (0.9)	58.1 (2.6)
m(KRMd,KRMt,BLMd,BLMt)	93.4 (1.6)	73.7 (4.2)
NNd	74.6 (3.4)	46.7 (6.1)
NNt	75.1 (4.4)	61.2 (7.1)
m(NNd,NNt)	88.6 (2.2)	68.8 (5.4)

0.2 Analysis of performance improvement

First, with respect to the the leave-one-out cross-validation results (see article), percentage improvements in AUC with respect to the (best of the two predictions furnished by the) previous method were 5.1%, 6.1%, 9.4% and 5.4% for the enzyme, ion channel, GPCR and nuclear receptor data sets, respectively. For the AUPR, the improvements were respectively 4.3%, 1.9%, 15.6% and 40.4%.

Second, to get some idea of the statistical significance of such improvements, we can consider Supplementary Tables 1-4 for the 10-fold cross-validation trials, each performed 10 times. The percentage improvements in AUC with respect to the previous method were 1.1%, 2.0%, 3.2% and 4.3% for the enzyme, ion channel, GPCR and nuclear receptor data sets, respectively. All of these improvements were statistically significant with respect to a one-sided Student t-test at 0.001 significance level. For the AUPR, the improvements were respectively 3.4%, 2.4%, 10.3% and 20.7%. Again, all of these improvements were statistically significant with respect to a one-sided Student t-test at 0.001 significance level.

0.3 A case study: the predicted edge between hsa:6095 and D00094 (see Fig. 2 of article)

A detailed explanation for the predicted pair hsa:6095 and D00094, and a comparison with the nearest neighbor method, is the following:

On the one hand, from a chemical viewpoint, D00094 does not share high structure similarity with D01441 (0.08) nor D00040 (0.19) which are known to interact with hsa:6095. We remark in passing that the nearest neighbor algorithm will predict the edge hsa:6095 to D00094 here with low probability.

On the other hand, from a genomic viewpoint, hsa:6095 shares high sequence similarity with hsa:6096 (0.57) and hsa:6097 (0.43), both of which have ligand D00094. Clearly in this case, the similarity captured in the sequence features would have been more pertinent than the similarity captured in the structural features. We remark that here, it is true that nearest neighbor will to some extent "notice" this high level of similarity, and will predict the edge hsa:6095 to D00094 with a larger probability than in the first case.

This does not imply that we too are using a nearest-neighbor algorithm. Indeed, let us examine the second case above in order to see what the SVM

algorithm is doing. It gives the set of proteins with known ligand D00094 a label +1. These include hsa:6096 and hsa:6097. It then gives all other proteins the label -1. It then takes the matrix of similarity scores (including the similarity score between proteins hsa:6096 and hsa:6097) and the assigned labels (+1,-1) and tries to project the proteins (excluding hsa:6095) into a high-dimensional space in which those with label +1 can be linearly separated (as well as possible) from those with label -1.

We then project protein hsa:6095 into this space, and give it a real-valued score which gets larger the further the protein is away from the hyperplane that is trying to separate the two classes. It gets larger "positive" if it is on the +1 side of the hyperplane, and larger "negative" if it is on the -1 side. This score is the score we use to estimate our "confidence" in the prediction of the edge between hsa:6095 and D00094. Indeed, to make the ROC and PR curves, we merely rank these scores. The answer to the question "where do the pink lines come from?" is that they are the highest positively ranked scores obtained in this way. More the score is positive and large, more we confidently predict an edge.

As protein hsa:6095 has a high similarity score with both hsa:6096 and hsa:6097, it will tend to get projected close to these two proteins in the new space. However, it is not this "closeness" in the nearest neighbor sense which gives our prediction score. In fact, all that matters is the distance of hsa:6095 to the hyperplane.

Thus, the relation with the NN algorithm can probably be characterized (in a very general sense) with the three following cases:

- (1) if hsa:6096 and hsa:6097 are very similar and they are also quite similar to the other neighbors of D00094 and this set of proteins are quite dissimilar to all the other proteins, then the hyperplane will separate with a large margin the +1 and -1 proteins. Then, hsa:6095 will be projected close to hsa:6096 and hsa:6097 and will have quite a "+" score as there is a large margin. Thus the SVM algorithm will predict the edge between hsa:6095 and D00094 with large probability. In this situation, it is also true that the nearest neighbor "strategy" would work, and have a similar predictive result to SVM;

- (2) if hsa:6096 and hsa:6097 are very similar to hsa:6095 but they are not very similar to the other neighbors of D00094 and/or they are also quite similar with some proteins that are not neighbors of D00094, then

the SVM algorithm will struggle to project hsa:6096 and hsa:6097 to a highly "+" location with respect to the hyperplane. In this case, the NN algorithm will still strongly predict an edge between hsa:6095 and D00094 (due to the proximity of hsa:6095 to hsa:6096 and hsa:6097) but the SVM algorithm will predict the edge with a lower probability;

- (3) if instead it had been the case that hsa:6095 was only to a medium extent similar to hsa:6096 and hsa:6097 and also to the other +1 proteins, and quite dissimilar to most or all of the -1 proteins, the NN algorithm would predict an edge between hsa:6095 and D00094 with a medium to low probability. However, the SVM algorithm may in this case still be able to project protein hsa:6095 to a relatively "+" position with respect to the hyperplane, even if it is not close to hsa:6096 and hsa:6097, thus predicting an edge between hsa:6095 and D00094 with a higher probability than the NN algorithm would have done.