Hassanzadeh and Shabani-Mashcool
BMC Bioinformatics (2022) 23:278
https://doi.org/10.1186/s12859-022-04831-7

BMC Bioinformatics

RESEARCH

Open Access

Does adding the drug–drug similarity to drug–target interaction prediction methods make a noticeable improvement in their efficiency?

Reza Hassanzadeh[1*] and Soheila Shabani-Mashcool[2]

*Correspondence:
r.hassanzadeh@uma.ac.ir

[1] Department of Engineering Sciences, Faculty of Advanced Technologies, University of Mohaghegh Ardabili, Namin, Iran
[2] Laboratory of Bioinformatics and Drug Design, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Abstract

Predicting drug–target interactions (DTIs) has become an important bioinformatics issue because it is one of the critical and preliminary stages of drug repositioning. Therefore, scientists are trying to develop more accurate computational methods for predicting drug–target interactions. These methods are usually based on machine learning or recommender systems and use biological and chemical information to improve the accuracy of predictions. In the background of these methods, there is a hypothesis that drugs with similar chemical structures have similar targets. So, the similarity between drugs as chemical information is added to the computational methods to improve the prediction results. The question that arises here is whether this claim is actually true? If so, what method should be used to calculate drug–drug chemical structure similarities? Will we obtain the same improvement from any DTI prediction method we use? Here, we investigated the amount of improvement that can be achieved by adding the drug–drug chemical structure similarities to the problem. For this purpose, we considered different types of real chemical similarities, random drug–drug similarities, four gold standard datasets and four state-of-the-art methods. Our results show that the type and size of data, the method which is used to predict the interactions, and the algorithm used to calculate the chemical similarities between drugs are all important, and it cannot be easily stated that adding drug–drug similarities can significantly improve the results. Therefore, our results could suggest a checklist for scientists who want to improve their machine learning methods.

Keywords: Drug–target interaction, Drug repositioning, Machine learning, Drug–drug similarity

Introduction

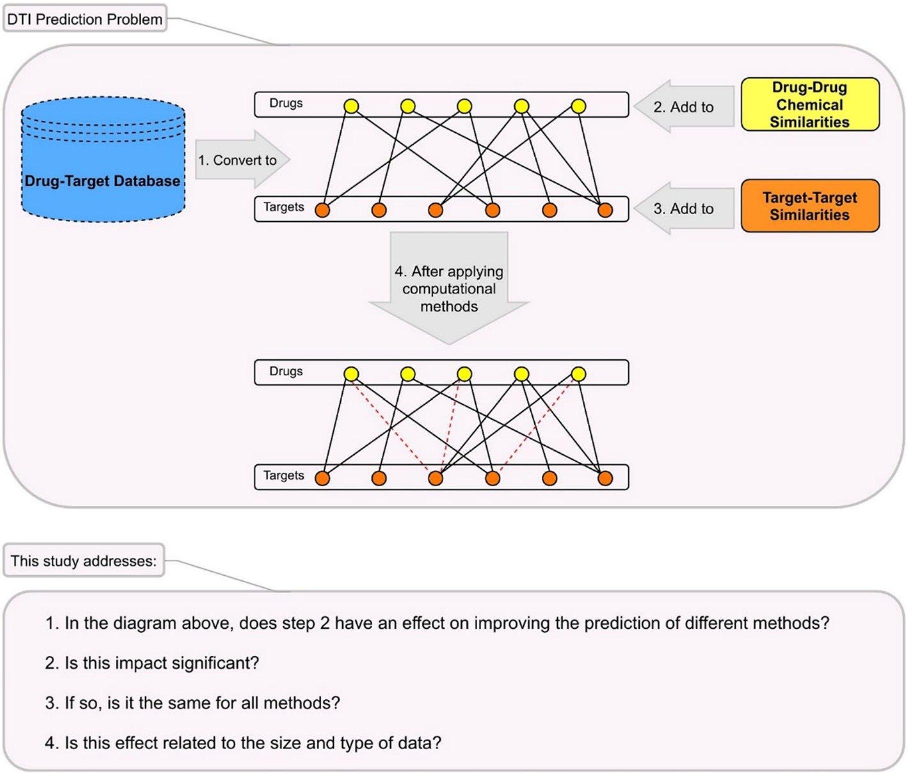
Most drugs fail in the early stages of a clinical trial and it takes a lot of time and cost for a drug to be successful in the market [1, 2]. These factors have led scientists to work on better and cheaper ways to find suitable drugs. One of the most effective and interesting solutions to solve these problems is drug repositioning (also called drug repurposing).

It is true that drug repositioning, by eliminating the early stages of drug design, can speed up research, but it also has drawbacks. For example, determining the dosage of a drug that is considered for a new disease using drug repositioning is one of the most important challenges of this viewpoint because the drug has already been considered for another disease with a specific dose. However, this viewpoint has found its place and we have to consider it today.

One of the most important steps of drug repositioning is identifying Drug–Target Interactions (DTI), which is a difficult task if laboratory and traditional methods are used. In contrast, computational methods can be more effective both in terms of time and cost. These methods can identify or predict DTI more quickly. The computational methods are usually based on machine learning or recommender systems. To predict interactions, these methods first consider a mathematical model for the information in the databases and then add biological and/or chemical information to the model, according to the guilt by association principle. For example, NRLMF [3], NetLapRLS [4], BLM-NII [5], WNN-GIP [6] and DT-Hybrid [7] are some of the well-known methods in this field. In addition to the DTI problem, computational methods are also widely used to predict drug–drug interactions and drug–disease associations [8–10].

NRLMF is a matrix factorization approach that predicts the probability that a drug would interact with a target. In this method, the properties of a drug and a target are represented by two latent vectors in the shared low dimensional latent space, respectively [3]. NetLapRLS is a semi-supervised learning method based on Laplacian regularized least square. NetLapRLS, by incorporating a new kernel established from the known drug-protein interaction network, is actually an improvement of the LapRLS [4, 11]. The bipartite local model (BLM) is a supervised learning approach introduced by Bleakley and Yamanishi in 2009 [12]. To improve the BLM, Mei et al. presented a simple procedure called neighbor-based interaction-profile inferring (NII) and integrated it into the existing BLM method and called it BLM-NII [5]. WNN-GIP is actually a combination of a simple weighted nearest neighbor algorithm and the GIP method [6, 13]. An example of recommender systems method introduced for DTI prediction problem is DT-Hybrid. It is a network-based interface method that extends a well-established recommendation technique by domain-based knowledge including drug and target similarity [7]. Many other algorithms have been introduced for this problem, but the algorithms mentioned are the most popular and can be considered as the state-of-the-art methods in this field.

As mentioned, the methods first model the information in databases. There are some public databases, for example, KEGG [14], PubChem [15], DrugBank [16], and ChEMBL [17] that contain information about drugs, targets, and interactions between them. Usually, all methods introduced for predicting DTI interactions use DrugBank to evaluate their results or compare them to other methods. Regardless of what algorithm each of these methods uses, they all add similarity between targets and chemical structure similarity between drugs to improve the prediction (Fig. 1). The similarities between targets (proteins) are always calculated by the Smith-Waterman method [18] and the chemical structure similarities between drugs are usually computed with SIMCOMP [19] which has been implemented in the KEGG system for searching similar chemical structures in the chemical structure databases. SIMCOMP is a graph-based method and uses a graph alignment algorithm to get a global similarity score based on the size of the common



1. In the diagram above, does step 2 have an effect on improving the prediction of different methods?
2. Is this impact significant?
3. If so, is it the same for all methods?
4. Is this effect related to the size and type of data?

Fig. 1 Schematic illustration of the DTI prediction problem and the questions addressed in this study

substructures between two compounds [5]. Of course, other information such as molecular fingerprints can be used to calculate similarities between drugs, but it is not usually used. There are several types of molecular fingerprints (e.g., MACCS [20], PubChem fingerprint [21], BCI fingerprints [22] and TGD [23]). PubChem fingerprints are 2D fingerprints that make a drug to be expressed by a vector and used to discover similar conformers by the PubChem database. These fingerprints are very popular and easily calculated for every drug.

The purpose of this study is not to identify the best method. Here, we want to discuss the following questions specifically for the DTI prediction problem (Fig. 1):

- Does considering the similarity of drugs indeed improve the results of computational methods?
- Is SIMCOMP the best way to calculate drug–drug similarities in any computational method?
- Do the type and size of the dataset affect the improvement that occurs with adding drug–drug similarities?

In this paper, the similarity between drugs refers to the chemical structure similarity.

Materials and methods

Datasets

Yamanishi et al. have provided four benchmark drug–target interaction datasets including Nuclear Receptors, G-Protein Coupled Receptors (GPCR), Ion Channels, and Enzymes [24]. The datasets are publicly available at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. The interactions were retrieved from databases KEGG BRITE [25], BRENDA [26], SuperTarget [27], and DrugBank [16]. These datasets of known DTIs are commonly considered as the gold standard for evaluating the performance of any new method introduced for DTI prediction problem. Each dataset contains three types of information in the form of matrices:

1. The drug–target interaction matrix, where the presence or absence of an interaction is indicated by 1 or 0, respectively.
2. The drug–drug similarity matrix calculated by SIMCOMP [19].
3. The target-target similarity matrix obtained by Smith-Waterman method [18].

To obtain the matrix mentioned in the second case, the chemical structure of a drug is treated as a 2D graph consisting of atoms as vertices and covalent bonds as edges. SIMCOMP provides the atom alignments between two chemical compound graphs, then it can also calculate the similarity of two chemical compounds by counting the number of matched atoms in those atom alignments. The calculation of similarity is based on the algorithm to solve the maximal common subgraphs of two graphs as the maximum vertex induced common subgraph or as the maximum edge induced common subgraph. The maximal common subgraphs of two graphs can be found by searching for maximal cliques in the association graph [19].

Some properties of datasets are shown in Table 1. The abbreviations in Table 1 are as follows:

- N_D: Number of drugs.
- N_T: Number of targets.
- N_I: Number of interactions.
- $$Density = N_I/(N_D \times N_T)$$
- AD_T: Average number of drugs per target.

Table 1 The properties of the benchmark datasets

Dataset	Nuclear Receptors	GPCR	Ion Channels	Enzymes
N_D	54	223	210	445
N_T	26	95	204	664
N_I	90	635	1476	2926
Density	0.0641	0.0299	0.0344	0.0099
AD_T	3.46	6.68	7.24	4.41
AT_D	1.67	2.85	7.03	6.58
D_{1T}	72.22%	47.53%	38.57%	39.78%
T_{1D}	30.77%	35.79%	11.27%	43.37%

- AT_D: Average number of targets per drug.
- D_{1T}: Percentage of drugs with only one target.
- T_{1D}: Percentage of targets with only one drug.

Evaluation

For each data set, in addition to the default drug similarity matrix obtained by SIMCOMP, we calculated 104 other matrices including one hundred random similarity matrices, one matrix where every element is equal to one, and three matrices calculated from PubChem 2D fingerprint using Tanimoto coefficient, Dice coefficient and Cosine similarity. The software PaDEL was used to obtain PubChem 2D fingerprints of all drugs [28]. For a drug, PubChem 2D fingerprints is a binary vector of length 881 that encodes the presence or absence of specific molecular substructures. Then, for the fingerprints of two drugs A and B, the Tanimoto, Dice and Cosine similarity can be calculated as follows:

- $Tanimoto(A, B) = \frac{c}{a+b-c}$,
- $Cosine(A, B) = \frac{c}{\sqrt{ab}}$,
- $Dice(A, b) = \frac{2c}{a+b}$,

  where *a* equals the amount of bit set to 1 in A, *b* equals the amount of bits set to 1 in B and *c* equals the amount of bits set to 1 in both A and B.

We considered a matrix where every element is equal to one to find out what happens to the results of the algorithms if the similarity of the drugs is not affected. To show how far the effect of adding drug–drug similarity to DTI problem is from the random effect that may occur, we generated 100 random similarity matrices between drugs. To make the comparison fair, we consider four state-of-the-art methods NRLMF, NetLapRLS, BLM-NII and WNN-GIP. Therefore, in short, we executed every algorithm on every dataset using every drug–drug similarity matrix. To do this, we slightly modified the PyDTI package [3] to perform the evaluation. Like most studies in this field, results are assessed using the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPR). Similar to [3, 4, 6, 13], we performed tenfold CV for five times to evaluate the performance of the methods on datasets. Then, we calculated the average AUC and AUPR over the five repetitions. In the next section, we will illustrate the results of the evaluations.

Results and discussion

Before discussing the results, it is necessary to state some of the abbreviations given in the tables and figures as follows:

- All-onesSim: The value obtained for the matrix where every element is equal to one.
- MeanRandoms: The average value obtained for random matrices.
- BestRandom: The best value obtained for the random matrices.
- WorstRandom: The worst value obtained for the random matrices.

- CosinePF: The value obtained for the matrix calculated by Cosine similarity for the PubChem fingerprint.
- DicePF: The value obtained for the matrix calculated by Dice similarity for the PubChem fingerprint.
- TanimotoPF: The value obtained for the matrix calculated by Tanimoto similarity for the PubChem fingerprint.

The evaluation results on Enzyme, GPCR, Ion Channel and Nuclear Receptors datasets are shown in Tables 2, 3, 4 and 5, respectively. In these tables, higher value cells have a green color, middle value cells have a yellow color, and lower value cells have a red color. It is worth noting that the best parameters for each algorithm are obtained in [3], and we have used these parameters here as well.

In each row of tables, the best similarity matrix for each algorithm is bolded. The best AUC and AUPR are also marked with underlines. The first point about these tables is that the use of random matrices has degraded the efficiency of the methods. In fact, what the first four columns of the tables show is that ignoring the drug–drug similarities yields far better results than using inaccurate drug–drug similarities. It should be noted that the NRLMF and NetLapRLS have less tolerance than other methods in this case.

Table 2 Comparing different drug–drug similarities on Enzyme dataset

	Method	All-onesSim	MeanRandoms	BestRandom	WorstRandom	CosinePF	DicePF	TanimotoPF	SIMCOMP
AUC	NRLMF	0.971239	0.968016	0.969477	0.966555	**0.976691**	0.97665	0.975809	0.97632
	BLM-NII	0.977584	0.7542	0.80648	0.718349	0.977368	**0.977764**	0.976215	0.969431
	NetLapRLS	0.959789	0.96367	0.964636	0.962813	0.966335	0.966613	0.968903	**0.972169**
	WNN-GIP	0.938578	0.515036	0.524567	0.507309	0.914265	0.897733	0.875283	**0.964062**
AUPR	NRLMF	0.84053	0.841717	0.845043	0.839261	0.870242	0.870329	0.870117	**0.875611**
	BLM-NII	0.592729	0.023396	0.034385	0.019237	0.605514	0.60798	0.535238	**0.703746**
	NetLapRLS	0.784019	0.787323	0.787526	0.787082	0.789326	0.789748	0.791864	**0.794216**
	WNN-GIP	0.476497	0.011065	0.01174	0.010693	0.256565	0.281493	0.243454	**0.69719**

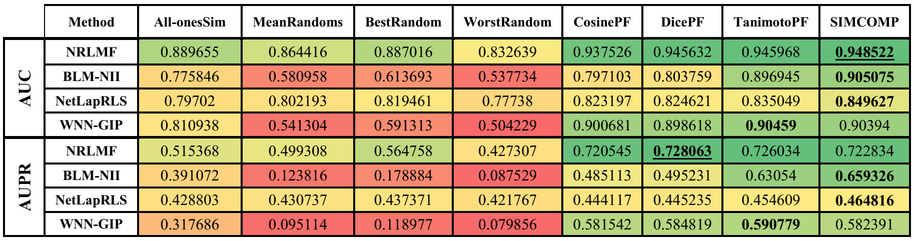
Table 3 Comparing different drug–drug similarities on GPCR dataset

	Method	All-onesSim	MeanRandoms	BestRandom	WorstRandom	CosinePF	DicePF	TanimotoPF	SIMCOMP
AUC	NRLMF	0.932221	0.922694	0.929836	0.917277	0.956879	0.957188	0.95682	**0.960355**
	BLM-NII	**0.94386**	0.671366	0.692179	0.647643	0.934594	0.928518	0.879454	0.943664
	NetLapRLS	0.902196	0.90289	0.905388	0.896996	0.910593	0.910846	0.91363	**0.914909**
	WNN-GIP	0.872255	0.528443	0.540304	0.517898	0.804141	0.787323	0.901193	**0.933079**
AUPR	NRLMF	0.570196	0.62361	0.642159	0.60265	0.69302	0.689631	0.688301	**0.702622**
	BLM-NII	0.373081	0.054418	0.062578	0.046693	0.342311	0.33531	0.324491	**0.514827**
	NetLapRLS	0.606391	0.611795	0.612422	0.611115	0.613065	0.613264	**0.615776**	0.615446
	WNN-GIP	0.278136	0.033394	0.035729	0.031326	0.2326	0.230504	0.428247	**0.466361**

Table 4 Comparing different drug–drug similarities on Ion Channels dataset

	Method	All-onesSim	MeanRandoms	BestRandom	WorstRandom	CosinePF	DicePF	TanimotoPF	SIMCOMP
AUC	NRLMF	0.979234	0.975785	0.977846	0.973896	0.981475	0.980925	0.980701	**0.983564**
	BLM-NII	0.974675	0.702874	0.744834	0.672745	0.96044	0.958388	0.944077	**0.981287**
	NetLapRLS	0.958158	0.95734	0.957955	0.956605	0.959433	0.959498	0.959527	**0.959882**
	WNN-GIP	0.861103	0.525954	0.535912	0.516046	0.930855	0.919196	0.944477	**0.956789**
AUPR	NRLMF	**0.865326**	0.856477	0.863683	0.847016	0.864683	0.85956	0.858608	0.863386
	BLM-NII	0.521158	0.058516	0.068181	0.051707	0.484567	0.482101	0.636176	**0.821476**
	NetLapRLS	0.81846	0.820111	0.820284	0.819911	0.821028	0.821095	0.821819	**0.823003**
	WNN-GIP	0.34916	0.038653	0.04019	0.037466	0.53947	0.524961	0.594643	**0.667893**

Table 5 Comparing different drug–drug similarities on Nuclear Receptors dataset



	Method	All-onesSim	MeanRandoms	BestRandom	WorstRandom	CosinePF	DicePF	TanimotoPF	SIMCOMP
AUC	NRLMF	0.889655	0.864416	0.887016	0.832639	0.937526	0.945632	0.945968	**0.948522**
	BLM-NII	0.775846	0.580958	0.613693	0.537734	0.797103	0.803759	0.896945	**0.905075**
	NetLapRLS	0.79702	0.802193	0.819461	0.77738	0.823197	0.824621	0.835049	**0.849627**
	WNN-GIP	0.810938	0.541304	0.591313	0.504229	0.900681	0.898618	**0.90459**	0.90394
AUPR	NRLMF	0.515368	0.499308	0.564758	0.427307	0.720545	**0.728063**	0.726034	0.722834
	BLM-NII	0.391072	0.123816	0.178884	0.087529	0.485113	0.495231	0.63054	**0.659326**
	NetLapRLS	0.428803	0.430737	0.437371	0.421767	0.444117	0.445235	0.454609	**0.464816**
	WNN-GIP	0.317686	0.095114	0.118977	0.079856	0.581542	0.584819	**0.590779**	0.582391

Although the purpose of this study is not to identify a better method, the performance of NRLMF is better than other methods in most cases. In the Enzyme dataset (Table 2), the AUPR value for all methods and the AUC value for NetLapRLS and WNN-GIP methods are the best values when SIMCOMP similarity is considered. The NRLMF and BLM-NII methods obtain the best AUC value if they use the CosinePF and DicePF similarities, respectively. In the GPCR dataset (Table 3), the AUC for BLM-NII and the AUPR for NetLapRLS are the best values if they use the All-onesSim and TanimotoPF similarities, respectively. Except for these two cases, according to Table 3, the use of SIMCOMP has given the best results in all cases. Table 4 shows that, in the Ion Channels dataset, using All-onesSim for the NRLMF method leads to a better AUPR. In all other cases, it is clear that SIMCOMP is the best.

In the Nuclear Receptors dataset (Table 5), the SIMCOMP gives both the best AUC and AUPR for NetLapRLS and BLM-NII methods. The same thing happens with TanimotoPF and WNN-GIP. The AUC and AUPR values for NRLMF are the best if it uses the SIMCOMP and DicePF similarities, respectively. In summary, these tables show that in almost 94% of experiments, the use of drug–drug chemical structure similarities has led to better results.

So far we have seen that drug–drug similarities can increase the accuracy of DTI predictions. But which method of calculating chemical structure similarity between drugs is more appropriate for the DTI predictions problem? The answer shown in Tables 2, 3, 4 and 5 is clearly SIMCOMP. But the results shown in these tables are obtained by parameters tuned for SIMCOMP [3]. Therefore, we randomly selected a dataset for each method and tuned the parameters of that method for all drug–drug similarities except random similarities. Nuclear Receptors, GPCR, Ion Channel and Enzyme datasets were considered for NRLMF, NetLapRLS, WNN-GIP and BLM-NII methods respectively. The results of these experiments are illustrated in Fig. 2. The use of SIMCOMP for NetLapRLS and WNN-GIP methods gives the best AUC in GPCR and Ion Channel datasets, respectively. The AUCs and AUPRs calculated in the rest of the experiments, i.e., 75% of them, show that TanimotoPF gave better results than the rest of the similarities. In general, it can be concluded that for these datasets and these methods, TanimotoPF and SIMCOMP are more appropriate than other similarities in the DTI prediction problem.

To investigate the effect of the type and size of the datasets on the values obtained in the experiments, we check the values in Tables 2, 3, 4 and 5 in a different way. Figures 3, 4, 5 and 6 are given for this purpose. In each figure, we considered a method and illustrated the values of AUC and AUPR obtained for that method across all datasets.

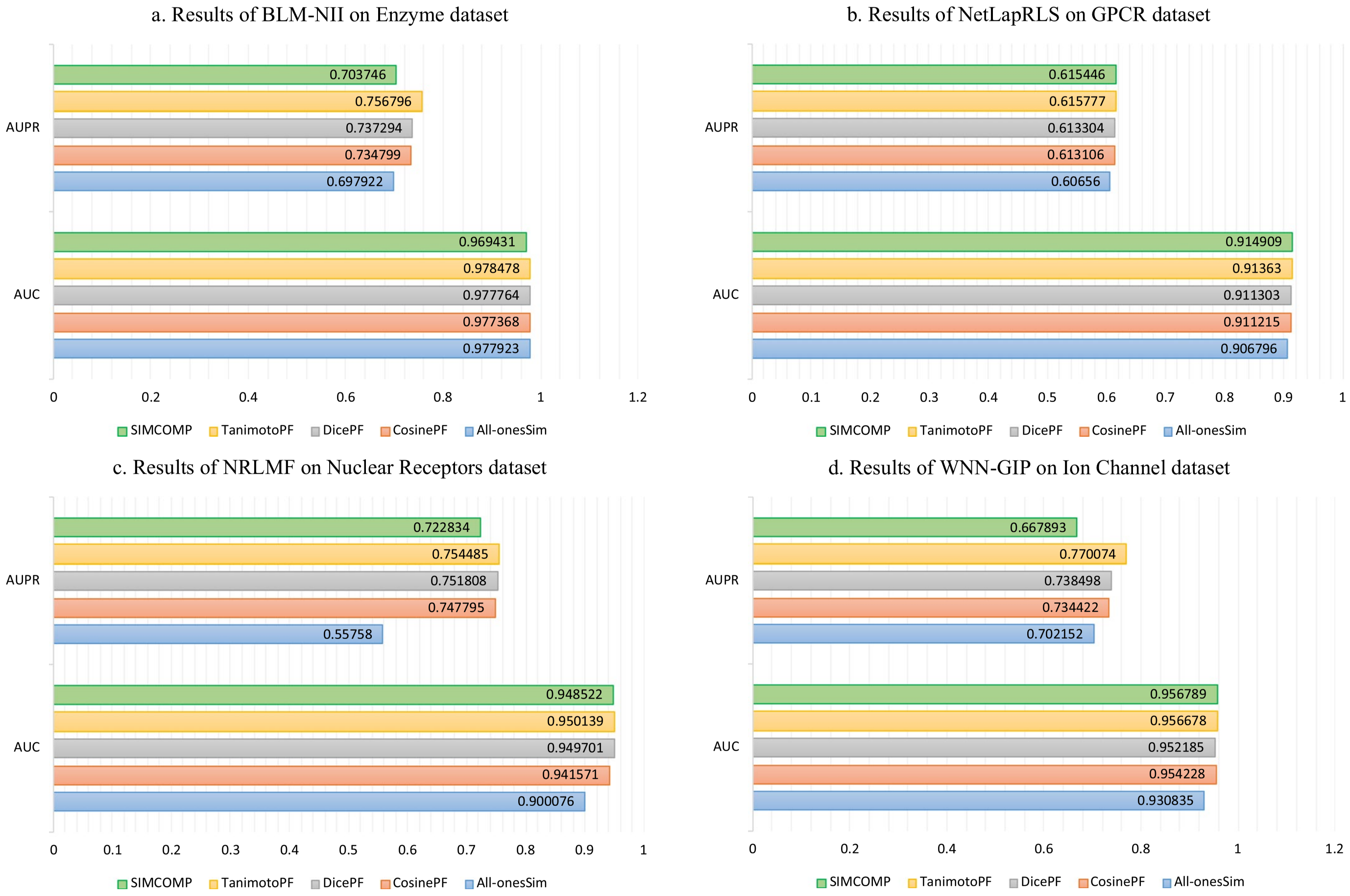
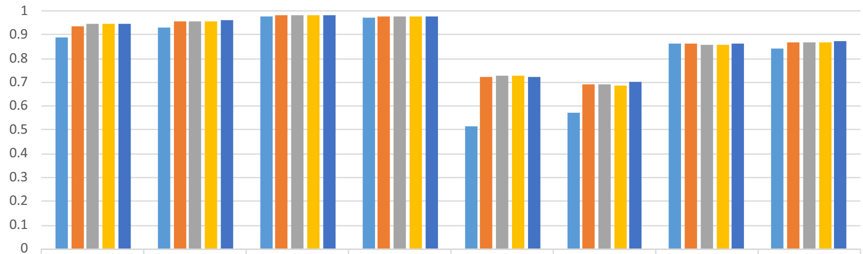
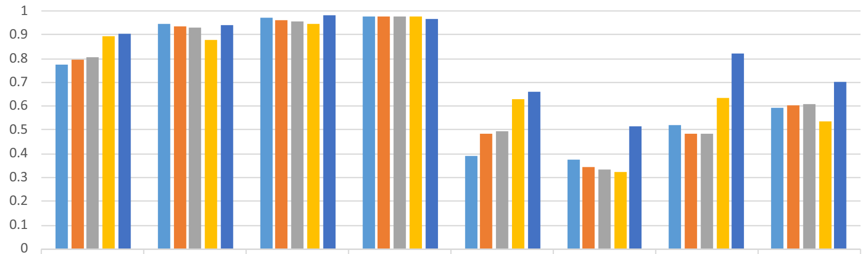


Fig. 2 The AUC and AUPR values obtained by tuning the parameters of each method for different drug–drug similarities: **a** BLM-NII and Enzyme dataset, **b** NetLapRLS and GPCR dataset, **c** NRLMF and Nuclear Receptors dataset, **d** WNN-GIP and Ion Channel dataset.



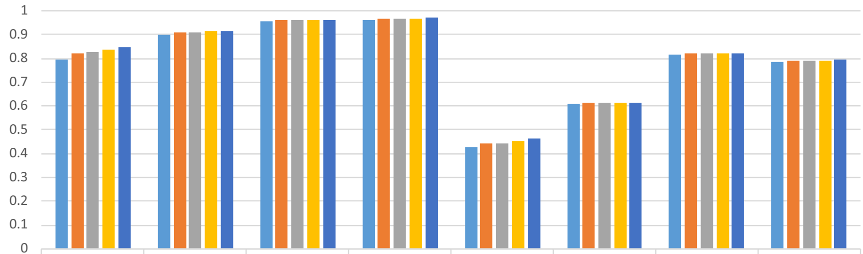
	Nuclear Receptor (AUC)	GPCR (AUC)	Ion Channel (AUC)	Enzyme (AUC)	Nuclear Receptor (AUPR)	GPCR (AUPR)	Ion Channel (AUPR)	Enzyme (AUPR)
All-onesSim	0.889655	0.932221	0.979234	0.971239	0.515368	0.570196	0.865326	0.84053
CosinePF	0.937526	0.956879	0.981475	0.976691	0.720545	0.69302	0.864683	0.870242
DicePF	0.945632	0.957188	0.980925	0.97665	0.728063	0.689631	0.85956	0.870329
TanimotoPF	0.945968	0.95682	0.980701	0.975809	0.726034	0.688301	0.858608	0.870117
SIMCOMP	0.948522	0.960355	0.983564	0.97632	0.722834	0.702622	0.863386	0.875611

Fig. 3 Investigating the effect of data type on the use of different drug similarities for NRLMF method



	Nuclear Receptor (AUC)	GPCR (AUC)	Ion Channel (AUC)	Enzyme (AUC)	Nuclear Receptor (AUPR)	GPCR (AUPR)	Ion Channel (AUPR)	Enzyme (AUPR)
All-onesSim	0.775846	0.94386	0.974675	0.977584	0.391072	0.373081	0.521158	0.592729
CosinePF	0.797103	0.934594	0.96044	0.977368	0.485113	0.342311	0.484567	0.605514
DicePF	0.803759	0.928518	0.958388	0.977764	0.495231	0.33531	0.482101	0.60798
TanimotoPF	0.896945	0.879454	0.944077	0.976215	0.63054	0.324491	0.636176	0.535238
SIMCOMP	0.905075	0.943664	0.981287	0.969431	0.659326	0.514827	0.821476	0.703746

Fig. 4 Investigating the effect of data type on the use of different drug similarities for BLM-NII method

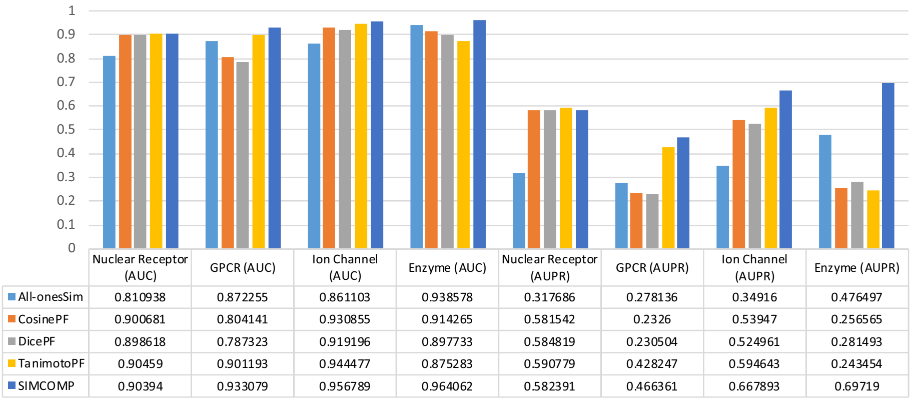


	Nuclear Receptor (AUC)	GPCR (AUC)	Ion Channel (AUC)	Enzyme (AUC)	Nuclear Receptor (AUPR)	GPCR (AUPR)	Ion Channel (AUPR)	Enzyme (AUPR)
All-onesSim	0.79702	0.902196	0.958158	0.959789	0.428803	0.606391	0.81846	0.784019
CosinePF	0.823197	0.910593	0.959433	0.966335	0.444117	0.613065	0.821028	0.789326
DicePF	0.824621	0.910846	0.959498	0.966613	0.445235	0.613264	0.821095	0.789748
TanimotoPF	0.835049	0.91363	0.959527	0.968903	0.454609	0.615776	0.821819	0.791864
SIMCOMP	0.849627	0.914909	0.959882	0.972169	0.464816	0.615446	0.823003	0.794216

Fig. 5 Investigating the effect of data type on the use of different drug similarities for NetLapRLS method

The results for the NRLMF, BLM-NII, NetLapRLS and WNN-GIP methods are shown in Figs. 3, 4, 5 and 6, respectively.

The results of Figs. 3, 4, 5 and 6 can be summarized as follows:



	Nuclear Receptor (AUC)	GPCR (AUC)	Ion Channel (AUC)	Enzyme (AUC)	Nuclear Receptor (AUPR)	GPCR (AUPR)	Ion Channel (AUPR)	Enzyme (AUPR)
🟦 All-onesSim	0.810938	0.872255	0.861103	0.938578	0.317686	0.278136	0.34916	0.476497
🟧 CosinePF	0.900681	0.804141	0.930855	0.914265	0.581542	0.2326	0.53947	0.256565
⬜ DicePF	0.898618	0.787323	0.919196	0.897733	0.584819	0.230504	0.524961	0.281493
🟨 TanimotoPF	0.90459	0.901193	0.944477	0.875283	0.590779	0.428247	0.594643	0.243454
🟦 SIMCOMP	0.90394	0.933079	0.956789	0.964062	0.582391	0.466361	0.667893	0.69719

Fig. 6 Investigating the effect of data type on the use of different drug similarities for WNN-GIP method

Table 6 Percentage of AUC improvement after considering drug–drug similarity

	Percentage improvement (enzyme) (%)	Percentage improvement (ion channels) (%)	Percentage improvement (GPCR) (%)	Percentage improvement (nuclear receptors) (%)
NRLMF	0.56	0.44	3.02	6.62
BLM-NII	0.02	0.68	0	16.66
NetLapRLS	1.29	0.18	1.41	6.6
WNN-GIP	2.72	11.11	6.97	11.55

- By replacing the similarities, the change in the value of AUPR is greater than that of AUC.
- Ion Channel and Enzyme datasets seem to be less dependent on similarity matrices replacement.
- In almost all figures, when the similarity matrix is replaced, the amount of AUC and AUPR changes for the Nuclear Receptors dataset is greater than what happens for other datasets. This has sometimes happened with less tolerance for the GPCR dataset.
- Compared to other methods, the NRLMF and NetLapRLS methods are less dependent on similarities and by replacing the matrices, their AUC and AUPR values change slightly.

In addition to the more changes that occur in the results on Nuclear Receptors and GPCR datasets, all methods perform worse on these two data, compared to other data. If we review Table 1 again, we find that these two datasets are smaller than the Ion Channel and Enzyme datasets, and the difference between the AD_T and AT_D criteria in these two data is a larger number. Also, the D_{1T} criterion has a larger value for these two data, especially for the Nuclear Receptors dataset. Probably, these factors have caused that the different methods cannot have better performance and less tolerance on these two datasets.

We did not settle for these results and did more analysis to make sure that the impact of adding chemical structure similarities between drugs is completely related to the type and size of the data. For this purpose, for each dataset and each method, we compared

Table 7 Percentage of AUPR improvement after considering drug–drug similarity

	Percentage improvement (enzyme) (%)	Percentage improvement (ion channels) (%)	Percentage improvement (GPCR) (%)	Percentage improvement (nuclear receptors) (%)
NRLMF	4.17	0	23.22	41.27
BLM-NII	18.73	57.63	37.99	68.59
NetLapRLS	1.3	0.56	1.55	8.4
WNN-GIP	46.32	91.29	67.67	85.96

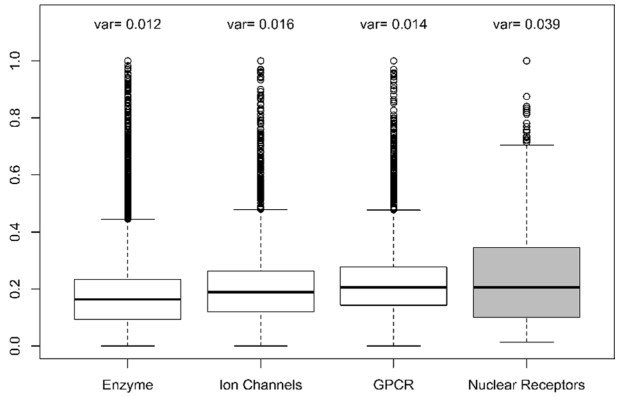


Fig. 7 Variance and Boxplot of chemical structure similarities between drugs obtained by SIMCONP for all datasets

the value obtained by the All-onesSim similarity matrix with its best value from Tables 2, 3, 4 and 5 and calculated the percentage of improvement. Tables 6 and 7 show these values for AUC and AUPR, respectively. What can be deduced from these tables is that, in general, the value of AUPR has improved more than that of AUC. Our datasets are all imbalances (Table 1), so it is appropriate to use the AUPR criterion for evaluation [29]. Since AUPR focuses mainly on the positive interactions, Tables 6 and 7 show that adding similarities between drugs has made the methods work better in predicting positive interactions. This improvement is quite evident in methods BLM-NII and especially WNN-GIP. The results of method WNN-GIP have improved by 46% in the lowest case and 91% in the highest case. The nature of the NRLMF and NetLapRLS methods is apparently such that the adding drug–drug similarities does not have much effect on them. As mentioned before, NRLMF works great compared to other methods. So, if its developers can make changes to the algorithm to get more impact from drug–drug similarities, then the results will be even better.

Another important case that can be deduced from Tables 6 and 7 is that the improvement of both the AUC and AUPR criteria for all methods in the case of Nuclear Receptors dataset is large compared to the other datasets. The size of this dataset may have caused this to happen because it is smaller than other datasets, but certainly not the only possible reason. Hence, we performed an analysis on the drug–drug similarity matrices of drugs for all datasets. Since the SIMCOMP similarities performed better in almost all Tables 2, 3, 4 and 5, we calculated the variance and drew boxplots only on these similarities. The results of this analysis are shown in Fig. 7. In this figure, the variance is denoted by *var*. It is clear that the dispersion of drug–drug similarities in Nuclear Receptors dataset, both variance and interquartile range, is greater than in other datasets. In other words, there is more information in the drug–drug similarity matrix for this dataset. Therefore, it can have a greater impact on the performance of methods, even in the case of the NRLMF and NetLapRLS methods. In fact, if the dispersion of similarities within the drug–drug matrix is low, it means that the chemical structures of the drugs are very similar, and this is equivalent to the fact that the similarities between the drugs are not considered.

Conclusions

This paper presents a meta-analysis of adding drug–drug chemical structure similarities to DTI prediction problem. Four state-of-the-art methods were selected and implemented on four benchmark datasets. The results show that using a meaningful drug–drug similarity can improve the performance of all methods. Tables 2, 3, 4 and 5 indicated that chemical structure similarity between drugs obtained by SIMCOMP has acceptable results for almost all computational methods and all datasets. It is worth noting that these methods have some parameters which can be optimized for the different similarities.

The other important conclusion is that the improvement that occurs by adding drug–drug similarities is not the same for every dataset and every method. It strongly depends on the nature of the DTI predictor method, data type and data size. The results of a method may be greatly improved, but this improvement for another method may be negligible. Perhaps, the nature of these methods is such that the effect of adding drug–drug similarities in the processes of the various stages of their algorithm is lost and wasted. For example, the WNN-GIP method is strongly influenced by the addition of drug–drug similarities, and the results are sometimes even improved by up to 90%. But for method NRLMF, which works better than all other methods, considering drug–drug similarities has little effect on the accuracy of its predictions.

Finally, we analyzed the relationship between the datasets and the improvement discussed. We showed that if the dispersion of similarities between drugs is low then adding drug–drug similarities to the DTI problem will have little effect on improving the results. That is why the improvement of both the AUC and AUPR criteria for all methods in the case of Nuclear Receptors dataset is large compared to other datasets.

Briefly, we should mention that using drug–drug chemical structure similarity can improve the prediction results in the DTI problem. However, this improvement depends on the nature of computational predictor method, the size and type of dataset, and the type of the method used to obtain the similarities between drugs. This means that this improvement may be very small for one method and very desirable for another. It may work well on some datasets and not so much on another. If a method wants to improve the results by using the drug–drug similarities, it must increase the effect of the drug–drug similarities in some steps of its algorithm. Otherwise, it may not achieve the desired results. One direction for future work is that all the experiments performed here can be done on the target-target similarities.

Acknowledgements
The authors would like to express thanks to Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao and Xiao-Li Li for making the PyDTI package available.

Author contributions
All authors implements the codes, analyzed the experiments, wrote the main manuscript text, and reviewed it. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets and codes can be freely accessed via https://github.com/Reza-HZ/DSComparisons.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 April 2022 Accepted: 12 July 2022
Published online: 14 July 2022

References

1. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really $802 million? Health Aff. 2006;25(2):420–8.
2. Krantz A. Diversification of the drug discovery process. Nat Biotechnol. 1998;16(13):1294–1294.
3. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol. 2016;12(2):e1004760.
4. Xia Z, Wu LY, Zhou X, Wong ST. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010;4(2):1–16.
5. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J. Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45.
6. Van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS ONE. 2013;8(6):e66952.
7. Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug–target interaction prediction through domain-tuned network-based inference. Bioinformatics. 2013;29(16):2004–8.
8. Zhao BW, Hu L, You ZH, Wang L, Su XR. Hingrl: predicting drug–disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform. 2022;23(1):bbab15.
9. Su X, You ZH, Huang DS, Wang L, Wong L, Ji B, Zhao B (2022) Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. IEEE Trans Knowl Data Eng
10. Su X, Hu L, You Z, Hu P, Zhao B. Attention-based knowledge graph representation learning for predicting drug-drug interactions. Brief Bioinform. 2022;23(3):140.
11. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7(11):2399–434.
12. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–403.
13. Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics. 2011;27(21):3036–43.
14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
15. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202–13.
16. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(suppl 1):D668–72.
17. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(D1):D1100–7.
18. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
19. Hattori M, Okuno Y, Goto S, Kanehisa M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc. 2003;125(39):11853–65.
20. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42(6):1273–80.
21. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Annu Rep Comput Chem. 2008;4:217–41.

22. Barnard JM, Downs GM. Chemical fragment generation and clustering software. J Chem Inf Comput Sci. 1997;37(1):141–2.
23. Sheridan RP, Miller MD, Underwood DJ, Kearsley SK. Chemical similarity using geometric atom pair descriptors. J Chem Inf Comput Sci. 1996;36(1):128–36.
24. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–40.
25. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34(suppl_1):D354–7.
26. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 2004;32(suppl_1):D431–3.
27. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R. SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 2007;36(suppl_1):D919–22.
28. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
29. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.