What can connectivity characteristics of networks tell us about the quality of link predictions?

Maksim Koptelov, Albrecht Zimmermann

What can connectivity characteristics of networks tell us about the quality of link predictions?

Maksim Koptelov and Albrecht Zimmermann

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

Abstract. Link prediction in networks works better when those networks are connected and not sparse. But can we use common connectivity characteristics to decide once a network is well enough connected to allow a random walk process to predict links best? Recent results in our work on link prediction lead us to ask this question and we attempt to shed some light on it. We do this by combining networks stemming from different data sources into networks combining different numbers of layers, and connecting their connectivity characteristics to the AUC that can be achieved by a random walk algorithm for link prediction. What we find is that it seems to be very important to reduce the radius and diameter of the network as much as possible, and get close to having a single connected component in the network. We also argue that the five benchmark data sets that have been used in the literature on drug-target activity prediction might be too easy to allow meaningful evaluations.

1 Introduction

As a rule of thumb, having more data in a learning or mining setting is better. There are a number of qualifiers to this statement, however: unreliable data, noisy data, the curse of dimensionality, strongly correlated descriptors can all turn the addition of data from a boon to an impediment.

The motivation for the work described in this paper can be found in work on drug-target activity prediction that we published recently [7], which started from the rule of thumb. In that problem setting, the goal is to predict the likelihood that a drug and a target interact, based on information of their other interactions (and the interactions of other drugs and targets). The network induced by introducing edges between interacting entities (drugs or targets) is often spare and not fully connected, limiting how well additional interactions can be predicted. This is typically alleviated in the literature by adding additional information, e.g. about entities' similarities [1] or interactions of entities of the same type, e.g. drug-drug ones. When this is done, choices are made about which similarity measure to use or which additional networks to use or discard. The typical configuration in the literature consists of a three-layer network: the interaction layer, as well as one each connecting drugs to drugs and targets to targets, respectively.

The hypothesis of our work was a simple one: instead of making those choices, we would use all networks at our disposal and let the prediction algorithm sort things out. The random walk algorithm we proposed, NEWERMINE, did exactly that on a large and sparse interaction network, IUPHAR: as we showed experimentally, running NEWERMINE on a six-layer network from IUPHAR and the five additional networks which are available gave superior results to using *any* three-layer network we could construct. We tentatively explained this result with the fact that only the six-layer network formed a fully connected component while having the same sparsity as some of the better-performing three-layer networks.

Further results, however, call this assumption into question. Performing the same kind of experimental evaluation in a number of networks that have become the de facto benchmark standard in the field [12, 3], we found that using only the three layers selected in the original publications, NEWERMINE consistently performed better than six- or eight-layer networks.

There are at least two possible explanations for this: 1) some of the networks do not provide reliable information, or 2) once the network has a certain structure that is sufficient for the success of the random walk, adding additional layers does not help anymore. To shed some light on this issue, we present a systematic exploration of how combining different networks into multi-layer networks affects connectivity characteristics of the resulting network and whether we can use those characteristics to tentatively explain the quality of the derived link predictions.

2 Definitions

A *graph* is a tuple $G =< V, E >$, where $V = \{v_1, v_2, ..., v_n\}$ denotes a set of vertices or nodes, and $E \subseteq V \times V$ a set of edges defined by distinct vertex pairs $(u, v) \in V \times V$ with $u \neq v$ (without self-loops). We also use the notion of a *bipartite graph*, which we define as a graph the vertices of which can be divided into two classes V_1 and V_2 such that there is no edge between vertices of the same class: $G =< V_1 \cup V_2, E >$, $E \subset V_1 \times V_2$.

We address weighted and unweighted graphs in the same manner. We define a *weighted graph* as one with a labeling function for edges $E \mapsto A_e$ with $A_e \in [0, 1]$, where 0 means no interaction between vertices, 1 confirmed interaction, and an intermediate value represents interaction probability. An *unweighted graph* is one where every edge is labeled by 1.

To exploit different sources of information in one single structure, we employ multi-layer networks. We define a *multi-layer network* as a weighted graph where more than one edge (u, v) can exist for a pair of vertices u, v. Multi-layer networks can be decomposed into disjunct set of graphs G_l that contain at most a single edge for each pair of vertices, called *layers* or just *networks*. As we wrote above, our original setting is a bipartite one. To combine it with the multi-layer framework, we define a *bipartite multi-layer graph* as a multi-layer network the vertices of which can be divided into two classes, and where exactly one of the

layers is a bipartite graph. Note, according to the classification of Kivelä *et al.* [6], the multi-layer networks used in this work are **not** *node-aligned*[1], **not** *layer-disjoint*[2], have *diagonal couplings*[3] which are *categorical*[4], and the number of layers can be any.

3 Link prediction via random walk

A long-established method for exploring a network is the random walk [**?**], which proceeds roughly as follows: starting from a randomly selected node, it performs walks along edges of the graph at random. In every step, the edge to follow is chosen uniformly from all outgoing links (in the case of an unweighted graph) or proportional to link weights (in the case of a weighted graph). Node importance is based on how frequently the walker visits the node: a node with higher frequency is considered more important than a node with a low value.

This idea can be modified in a number of ways to improve network exploration: the walker can be constrained to perform at most *max_steps* steps, to not visit any of the last c vertices it encountered, or with small probability $1-\beta$ the process can be restarted at any time to avoid getting trapped by those vertices it mustn't visit. The product of the probabilities of edges the walker traversed gives the cumulative probability of a path between two nodes and can be used to *predict* a link between a starting node and an end node: if the path probability is greater than a given threshold, a new edge is predicted.

To extend this approach to multi-layer graphs, one needs to add how to choose the layer to walk in. We propose to select a network uniformly at random from the set of networks, and multiply the path strength by $\frac{1}{|\{G_l^i\}|+|\{G_p^j\}|+|\{G_{lp}^k\}|}$. Repeat the process until a user-defined target vertex is reached or the maximum number of steps have been performed. Due to the randomized nature, random walks are usually repeated several times to derive more robust estimates.

In our prior work, we have adapted a PageRank algorithm to the setting using an arbitrary number of layers.

4 The data

As mentioned in the introduction, there are several drug-target interaction data sets, five in total, that have been used in prior work on drug-target interaction prediction [2, 13, 8, 1], and can be considered benchmarks. We downloaded the original data of four of those datasets (Enzyme, GPCR, IC, NR) from the supplementary information of Yamanishi *et al.*'s work [12][5], and the fifth (Kinase) from the supplementary information of Davis *et al.*'s work [3][6].

[1] All nodes are shared between all layers
[2] Each node is present only in a single layer
[3] Inter-layer edges, that cross layers, are only between nodes and their counterparts
[4] Diagonal couplings for which all possible inter-layer edges are present
[5] `http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget`
[6] `http://staff.cs.utu.fi/aatapa/data/DrugTarget`

The original representation of these data consists of 3 networks: a drug similarity network obtained by the use of the SimComp score that we will refer to as DS(sc), a target similarity network obtained by the use of the Smith-Waterman score (TS(sw)), and a drug-target interaction network (DT) constructed from the KEGG BRITE [5], BRENDA [9], SuperTarget [4], and DrugBank [11] databases. We have augmented this representation by constructing additional networks:

- a drug-drug interaction network based on DrugBank (DB),
- a target-target interaction network based on BioGrid [10] (BG),
- a drug similarity network based on similarities calculated using the Tanimoto coefficient on binary vectors constructed from the presence/absence of frequent subgraphs (DS(sg)), and
- two target similarity networks calculated using the Tanimoto coefficient on feature vectors constructed from the presence/absence *frequent substrings* (TS(ss)) and *Prosite motifs*(TS(mot)) [7].

The IUPHAR multi-layer network[8] we constructed and used in our earlier work is not a standard benchmark for drug-target activity prediction in the literature, however.

1. IUPHAR – an open-access database of drugs, biological targets and their interactions. We used version 2017.5 (released on 22/08/2017). The full dataset has 8978 drugs, 2987 proteins, and 17198 interactions (edges) between them[9]. In order to satisfy the designed setting conditions, we removed duplicate interactions (based on different affinity measures), leaving 12456 interactions in total. For existing interactions, we label an edge with 1 if the negative logarithm of the affinity measure is ≥ 5, non-interacting otherwise.[10] We treat all affinity measures available in the data (pKi, pIC50, pEC50, pKd, pA2, pKB) as equivalent.
2. DrugBank (DB) – an open-access database of drug-drug interactions. We used version 5.0.11 (released 20-12-2017). It has 658079 interactions of 3138 distinct drugs. 242922 of these interactions involve 1254 distinct drugs that are present in IUPHAR. The database was also used as a source of 2D representations of drugs to compute drug similarities.
3. BioGrid (BG) – an open-access database of protein-protein interactions mined from a corpus of biomedical literature. We used version 3.4.154 (25/10/2017). It has 1482649 interactions of 67372 distinct proteins. Only 15410 of these interactions involve proteins present in IUPHAR (1925 distinct proteins).
4. NCBI Protein database – The National Center for Biotechnology Information proteins database[11] was used to obtain amino acids sequences to represent targets. The data was parsed from the website of NCBI and mapped

[7] An open-access database is available at `http://prosite.expasy.org`
[8] https://zimmermanna.users.greyc.fr/supplementary-material.html
[9] in drugs.csv, interactions.csv, and targets_and_families.csv, respectively
[10] Cutoff proposed by researchers from CERMN (`http://cermn.unicaen.fr`)
[11] `https://www.ncbi.nlm.nih.gov/protein/`

to IUPHAR using the RefSeq attribute (human protein sequence identifier) available in IUPHAR. The database was accessed 20/12/2017.

drugs were mapped between networks by numerical identifiers provided by IUPHAR as well as by INN (International Non-proprietary Name) and Common name attributes. Proteins were mapped by IUPHAR identifiers as well as by Human Entrez Gene attribute.[12]

As the preceding listing shows, not all drugs and proteins contained in IUPHAR are available in DB and BG. In addition, not all proteins and drugs are annotated with molecular information so that not all entities will be connected in the similarity networks.

5 Experimental evaluation

As we wrote above, our earlier work indicated that using eight networks leads to better results than using three networks [7]. In this section, we first explore whether the same holds for the benchmark data sets. We then explore how many and which combinations of networks allow for best performance before discussing the connectivity characteristics of those best-performing combinations. Finally, we evaluate our newly derived insights on the IUPHAR data.

5.1 Link prediction in benchmark data sets

In this section, we report the results of NEWERMINE, the random-walk based method we have proposed in [7], on the benchmark data sets. We use optimal parameter setting ($\eta = 0.2$, $\beta = 0.7$) defined for this method. Unlike the evaluation framework used in [7], we performed a *5x5-fold cross-validation*, with each fold containing 20% of all drug-target interactions, acting as test set for link prediction once, while the method is run on the other 80%. The process is repeated 5 times, the results are averaged among all runs. We switched to a *5x5-fold cross-validation* to allow comparison to state-of-the-art methods in order to be able to compare the results in the future work.

Table 1 reports both area under the ROC curve (AUC) and area under the precision-recall curve (AUPR), both for using the three layers proposed in the literature, and the eight layers that result from augmenting those networks with the additional information we outlined above.

As we can see, using all layers clearly underperforms compared to using only the three that have been proposed in the literature.

5.2 How many/which networks are useful for link prediction?

The question to ask in light of these results is: "Why?" We can formulate at least two plausible working hypotheses:

[12] Global Query Cross-Database Search System gene identifiers: `https://www.ncbi.nlm.nih.gov/gene`

Data set	3 layers		8 layers	
	AUC	AUPR	AUC	AUPR
Enzyme	0.844	0.15	0.655	0.15
GPCR	0.81	0.22	0.674	0.112
IC	0.76	0.26	0.61	0.09
NR	0.635	0.26	0.489	0.15
Kinase	0.61	0.13	0.553	0.08

Table 1. Three layer results, eight-layer results of NEWERMINE link predictions on the benchmark data

1. **Certain networks are less reliable/informative than others.** DrugBank information, for instance, is arguably anecdotal since interactions are based on published results in the literature, whereas similarity edges are reliable: the molecular information about a compound exists and has been translated into a similarity value in a transparent manner. But not all similarities are equal: the SimComp score takes chemical aspects into account that the use of frequent subgraphs might miss.
2. **There are diminishing returns to adding additional layers.** To arrive at a network that can be exploited for link predictions by a random walk process, it is necessary to create as few connected components as possible, and a network that is as dense as possible. But once the entire network consists of a single connected component, and paths between any two vertices are short, it might be impossible to improve the state of the network.

Either of these effects, or, even worse, a combination of the two, can lead a random walk process astray. Similar to how adding noisy attributes or ones that are strongly correlated with attributes that are already present can degrade the performance of a predictor for vectorial data, adding non-reliable or redundant edges can undermine link prediction.

5.3 Connectivity characteristics and redundancy

We start with exploring the second hypothesis. Table 2 illustrates the main characteristics of the different data sets and we can see that IUPHAR is not like the others: significantly less dense, with a much higher number of distinct connected components. Sparsity refers to the number of edges in the network, divided by the maximal number of edges possible. Notably, the Kinase network is also different since it is the only one that is already connected.

And as Table 3 shows, adding two additional networks boosts the density of the benchmark data sets and turns them into a single connected component, whereas even the best-case situation for IUPHAR still results in a sparse, fragmented network. The reason for this is, as mentioned before, that even in the case of the rich similarity networks there are missing connections. The reader will also notice that the three-layer IUPHAR combination indicates fewer vertices since we removed isolated vertices, i.e. those not connected to any other vertex in any layer.

Table 2. Basic properties of Benchmark and IUPHAR data sets

Data set	Drugs	Targets	$\|V\|$	$\|E\|$	Sparsity	CC
Enzyme	445	664	1109	2926	0.005	44
GPCR	223	95	318	635	0.013	19
IC	210	204	414	1476	0.017	3
NR	54	26	80	90	0.028	10
Kinase	68	442	510	1527	0.016	1
IUPHAR	8137	2502	10639	12456	0.00017	443

Table 3. Basic properties of three-layer networks

Data set	Drugs	Targets	Layers	$\|V\|$	$\|E\|$	Sparsity	CC
Enzyme	445	664	3	1109	321832	0.524	1
GPCR	223	95	3	318	29853	0.592	1
IC	210	204	3	414	44127	0.516	1
NR	54	26	3	80	1846	0.584	1
Kinase	442	1527	3	510	101266	0.780	1
IUPHAR	7025	2010	3	9126	1786917	0.0215	103

This is probably a first indication for why using three layers for the IUPHAR network is not enough – it does not explain, however, why adding additional layers degrades link prediction performance on the benchmark data sets.

Table 4. Connectivity characteristics of different network combinations for the Enzyme data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC	AUPR
				Radius	Diameter	$\|E\|$	C_{coef}	Tr		
BG	2	0.007	12	6	10	4153	0.1599	0.0562	0.24	0.01
DB	2	0.018	14	5	10	10856	0.2127	0.435	0.66	0.1
DS(sg)	2	0.166	1	2	3	101716	0.7298	0.971	0.82	0.14
DS(sc)	2	0.166	1	2	3	101716	0.7298	0.971	0.82	0.14
DS(sg),TS(sw)	**3**	**0.524**	**1**	**2**	**2**	**321832**	**0.9825**	**0.9838**	**0.84**	**0.15**
DS(sc),TS(sw)	***3***	***0.524***	***1***	***2***	***2***	***321832***	***0.9825***	***0.9838***	***0.84***	***0.15***
DS(sg),DB,TS(sw),TS(ss)	5	0.895	1	2	2	321832	0.9825	0.9838	0.74	0.09
DS(sg),DB,TS(sw),TS(mot)	5	0.895	1	2	2	321832	0.9825	0.9838	0.74	0.09
DS(sc),DB,TS(sw),TS(ss)	5	0.895	1	2	2	321832	0.9825	0.9838	0.74	0.10
DS(sc),DB,TS(sw),TS(mot)	5	0.895	1	2	2	321832	0.9825	0.9838	0.74	0.09

To gain more insight into this question, we present a number of connectivity characteristics of derived multi-layer networks in Tables 4 – 8. For reasons of space constraints and readability, we do not show the complete list but only representative results. The first column lists which additional networks have been added to the LT network (which is always present and therefore omitted). The tables then list, in order,

– the number of layers that the network has ($|L|$),

– its sparsity, i.e. the number of existing edges divided by the maximum possible number of edges,
– the number of connected components (CC), where by *connected component* we assume a graph any two vertices of which are connected to each other by a path, and which is connected to no additional vertices in the supergraph,
– the radius of the network, i.e. the *minimum* shortest path between any two different vertices in the network,
– its diameter, the *maximum* shortest path between any two different vertices,
– the number of edges of the flattened graph ($|E|$),
– the clustering coefficient (C_{coef}), the number of existing closed triangles divided by the maximum possible number of open and close triangles, and
– the transitivity (Tr), the number of existing closed triangles divided by the number of connected triples.

Finally, the right-most two columns list the AUC and AUPR scores for link prediction via NEWERMINE. It should be noted that connectivity characteristics such as clustering coefficient and transitivity are typically not defined for multilayer networks – we therefore report these measures for a flattened network: when there is an edge available between any two nodes, they are treated as connected, no matter the network the edge occurs in. It should also be mentioned also that in networks having more than one connected component, we compute and report radius and diameter values for the largest such component in terms of the number of vertices. For a network with the number of $CC = 1$ that giant component is equal to the network itself.

Table 5. connectivity characteristics of different network combinations for the GPCR data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC	AUPR
				Radius	Diameter	$\|E\|$	C_{coef}	Tr		
BG	2	0.013	14	6	12	650	0.0251	0.0078	0.28	0.02
DB	2	0.087	7	5	9	4369	0.3706	0.5779	0.65	0.19
DS(sg)	2	0.504	1	2	3	25388	0.8771	0.9764	0.77	0.22
DS(sc)	2	0.504	1	2	3	25388	0.8771	0.9764	0.77	0.22
DS(sg),TS(sw)	**3**	**0.592**	**1**	**2**	**2**	**29853**	**0.9546**	**0.9694**	**0.80**	**0.21**
DS(sc),TS(sw)	***3***	***0.592***	***1***	***2***	***2***	***29853***	***0.9546***	***0.9694***	***0.80***	***0.22***
DS(sg),DB,TS(sw),TS(ss)	5	0.755	1	2	2	29853	0.9546	0.9694	0.73	0.17
DS(sg),DB,TS(sw),TS(mot)	5	0.755	1	2	2	29853	0.9546	0.9694	0.74	0.17
DS(sc),DB,TS(sw),TS(ss)	5	0.755	1	2	2	29853	0.9546	0.9694	0.74	0.18
DS(sc),DB,TS(sw),TS(mot)	5	0.755	1	2	2	29853	0.9546	0.9694	0.74	0.18

What we see for every data set except Kinase is that reducing the number of connected components improves link prediction, especially once only a single connected component remains. For all of these benchmark data sets, the latter is achieved at the latest once the network contains three layers.

There seems to be a secondary effect of decreasing the diameter to two: for Enzyme, GPCR, and IC, the final bump in AUC comes once the network moves from a diameter of three to one of two. This decrease in the diameter goes together with a clear increase in the value of the clustering coefficient. Notably,

Table 6. connectivity characteristics of different network combinations for the IC data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC	AUPR
				Radius	Diameter	$\|E\|$	C_{coef}	Tr		
BG	2	0.019	3	5	9	1583	0.0958	0.034	0.32	0.02
DB	2	0.107	3	4	7	6666	0.3639	0.6033	0.63	0.14
DS(sg)	2	0.481	1	2	3	23421	0.9169	0.9367	0.71	0.16
DS(sc)	2	0.481	1	2	3	23421	0.9169	0.9367	0.72	0.17
DS(sg),TS(sw)	**3**	**0.516**	**1**	**2**	**2**	**44127**	**0.9469**	**0.9414**	**0.76**	**0.26**
DS(sc),TS(sw)	***3***	***0.516***	***1***	***2***	***2***	***44127***	***0.9469***	***0.9414***	***0.76***	***0.27***
DS(sg),DB,TS(sw),TS(ss)	5	0.819	1	2	2	44127	0.9469	0.9414	0.67	0.15
DS(sg),DB,TS(sw),TS(mot)	5	0.819	1	2	2	44127	0.9469	0.9414	0.67	0.15
DS(sc),DB,TS(sw),TS(ss)	5	0.819	1	2	2	44127	0.9469	0.9414	0.68	0.15
DS(sc),DB,TS(sw),TS(mot)	5	0.819	1	2	2	44127	0.9469	0.9414	0.67	0.16

however, we achieve better results on NR for a diameter of three than one of two.

Table 7. connectivity characteristics of different network combinations for the NR data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC	AUPR
				Radius	Diameter	$\|E\|$	C_{coef}	Tr		
BG	2	0.043	2	4	6	137	0.2241	0.2318	0.26	0.04
DB	2	0.107	3	3	5	339	0.2717	0.3026	0.5	0.10
DS(sg)	2	0.481	1	2	3	1521	0.8648	0.9452	0.61	0.22
DS(sc)	2	0.481	1	2	3	1521	0.8648	0.9452	0.63	0.28
DS(sg),TS(sw)	3	0.584	1	2	2	1846	0.9096	0.9295	0.60	0.20
DS(sc),TS(sw)	*3*	*0.584*	*1*	*2*	*2*	*1846*	*0.9096*	*0.9295*	*0.63*	*0.26*
DS(sg),BG	3	0.496	1	2	3	1568	0.7892	0.9413	0.66	0.19
DS(sc),BG	**3**	**0.496**	**1**	**2**	**3**	**1568**	**0.7892**	**0.9413**	**0.69**	**0.22**
DS(sg),DB,TS(sw),TS(ss)	5	0.766	1	2	2	1846	0.9096	0.9295	0.51	0.13
DS(sg),DB,TS(sw),TS(mot)	5	0.766	1	2	2	1846	0.9096	0.9295	0.48	0.12
DS(sc),DB,TS(sw),TS(ss)	5	0.766	1	2	2	1846	0.9096	0.9295	0.53	0.15
DS(sc),DB,TS(sw),TS(mot)	5	0.766	1	2	2	1846	0.9096	0.9295	0.52	0.16

Once those two aspects – the number of connected components and the diameter – have been addressed, there does not seem to be anything to be gained from adding additional networks. Notably, making the multi-layer network significantly denser *does not* result in additional gains but instead reduces quality.

5.4 Reliability/informativeness of networks

We would suspect that this degradation of link prediction quality is related to the first hypothesis that we formulated above: that certain networks simply provide better information.

Tables 4 – 8 also show that not all multi-layer networks that lead to similar or even the same connectivity characteristics allow for the same performance of NEWERMINE. In each table, we have indicated the best-performing layer combination (or combinations) in **bold** and the layer combination found in the literature – using SimComp for drug similarity and Smith-Waterman for target similarity – in *italics*.

Table 8. connectivity characteristics of different network combinations for the Kinase data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC	AUPR
				Radius	Diameter	$\|E\|$	C_{coef}	Tr		
BG	2	0.019	1	4	6	2433	0.1447	0.0459	0.36	0.04
DB	2	0.017	1	5	10	1615	0.0363	0.0285	0.56	0.13
DS(sg)	**2**	**0.038**	**1**	**2**	**4**	**3700**	**0.7136**	**0.5108**	**0.66**	**0.17**
DS(sc)	2	0.039	1	2	3	3805	0.7319	0.5451	0.65	0.16
DS(sg),DB	3	0.039	1	2	4	3701	0.7137	0.5111	0.65	0.16
DS(sc),DB	3	0.040	1	2	3	3805	0.7319	0.5451	0.64	0.14
DS(sg),BG	3	0.039	1	3	6	4606	0.5234	0.4938	0.65	0.12
DS(sc),BG	3	0.040	1	3	6	4711	0.532	0.5272	0.64	0.11
DS(sg),TS(sw)	3	0.779	1	2	3	101161	0.954	0.9847	0.61	0.14
DS(sc),TS(sw)	*3*	*0.780*	*1*	*2*	*3*	*101266*	*0.9578*	*0.9848*	*0.61*	*0.13*
DS(sg),DB,TS(sw),TS(ss)	5	1.516	1	2	3	101162	0.9541	0.9847	0.58	0.12
DS(sg),DB,TS(sw),TS(mot)	5	1.516	1	2	3	101162	0.9541	0.9847	0.58	0.12
DS(sc),DB,TS(sw),TS(ss)	5	1.517	1	2	3	101266	0.9578	0.9848	0.58	0.11
DS(sc),DB,TS(sw),TS(mot)	5	1.517	1	2	3	101266	0.9578	0.9848	0.58	0.11

As we can see, for Enzyme, GPCR, and IC, the combination used in the literature does indeed give best results. But while Smith-Waterman seems to encode vital information about target similarity, the SimComp score can be replaced by a very simple similarity measure, the Tanimoto coefficient over a vector of frequent subgraphs, i.e. subgraphs mined in an unsupervised manner!

The results are even more interesting for the other two data sets: for NR, using the SimComp/Smith-Water does *not* give best results – instead combining BioGrid information about target-target interactions with SimComp drug similarity results in the highest AUC. The reason for that could be in the fact that NR is too small in terms of the number of vertices and the number of edges for the evaluation framework we use (when we remove 20% of edges the network might become destructed too much what results in such unpredictable behavior). And there is an explanation for Kinase also. As we already wrote above, it is different, particularly because it already consists of a single connected component. Adding subgraph-based drug similarity to the interaction network is enough to achieve the highest AUC, with the SimComp/Smith-Waterman combination not only behind that two-layer network but also behind a number of others that do not use the TS(sw) network.

5.5 How does this transfer to IUPHAR?

After having explored how connectivity characteristics on benchmark data sets align with the quality of link prediction, we obviously want to come back to our original problem settings – link prediction on the IUPHAR data set – and see whether we can observe similar behavior.

In Table 9, we list the connectivity characteristics of the *best-performing* network combinations for IUPHAR, one for each number of layers from two to six. There is an exception for four layers where the AUC was virtually indistinguishable. We had neither DS(sc) nor TS(sw) networks for the IUPHAR set.

There are a number of interesting observations here.

Table 9. connectivity characteristics of different network combinations for the IUPHAR data set

Networks	$\|L\|$	Sparsity	CC	Giant component		Flattened graph			AUC
				Radius	Diameter	$\|E\|$	C_{coef}	Tr	
DS(sg)	2	0.5147	69	5	9	23272066	0.7992	0.9996	0.5781
DB, BG	3	0.0033	87	7	14	143922	0.1285	0.5118	0.5255
DS(sg), DB, TS(ss)	4	0.4780	14	5	9	24929849	0.9003	0.9992	0.5805
DS(sg), DB, TS(mot)	4	0.4780	14	5	9	24929849	0.9003	0.9992	0.5800
DS(sg), DB, TS(ss), TS(mot)	5	0.5095	14	5	9	24929849	0.9003	0.9992	0.5812
all	6	0.4719	1	5	9	24932609	0.8802	0.9992	0.5783

1. The first is that we were wrong in our NEWERMINE publication – the best layer combination does *not* contain all six layers but only five, the BioGrid layer seems to reduce the quality of the information in the multi-layer network.
2. The second is that the importance of reducing radius and diameter hold – five and nine are the two lowest values we can achieve on IUPHAR.
3. The third observation is that the exception to this rule is the three-layer setting – this setting, so common in the literature, seems to be the single worst setting for IUPHAR and the improvement on it therefore easiest.
4. Fourth, once radius and diameter have been driven down, the number of connected components seems to matter less, a phenomenon we could not observe on the smaller benchmark data sets.
5. Finally, there were several other combinations that achieved a radius of five and a diameter of nine, without leading to best AUC. The difference is that the "winning" combinations have lower average distance between vertices (not shown).

6 Conclusion

This is very consciously a workshop paper, intended to provoke discussion and additional work to explore this issue further. As such, we do not have clear conclusions to draw. What seems to emerge from our experimental data, however, is that the most important issue to pay attention to when augmenting networks for link prediction is to add data that reduces the radius and diameter of the final network as much as possible. We have to admit, however, that those results are also preliminary in the sense that, for running time reasons, we did not in fact test all layer combinations for all networks. To base our insights on firmer ground, we will do this for a future publication.6

We would add that the five benchmark data sets that are being widely used in the literature on drug-target activity prediction might not be the best choice. They are all small, can easily augmented to form only a single connected component and achieve minimal diameter. This is doubly true for the Kinase data set that forms a single connected component to begin with.

Similarly to how data sets in other fields have become outdated and are rarely used nowadays – e.g. the Iris data set for machine learning and clustering, or

the Mutagenicity data set for ILP and substructure mining – it might be time to retire these networks and curate new benchmarks.

References

1. Buza, K., Peska, L.: Aladin: A new approach for drug–target interaction prediction. In: ECML/PKDD. pp. 322–337. Springer (2017)
2. Chen, X., Liu, M.X., Yan, G.Y.: Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems **8**(7), 1970–1978 (2012)
3. Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Zarrinkar, P.P.: Comprehensive analysis of kinase inhibitor selectivity. Nature biotechnology **29**(11), 1046 (2011)
4. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G., Gewiess, A., Jensen, L.J., et al.: Supertarget and matador: resources for exploring drug-target relationships. Nucleic acids research **36**(suppl_1), D919–D922 (2007)
5. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al.: Kegg for linking genomes to life and the environment. Nucleic acids research **36**(suppl_1), D480–D484 (2007)
6. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. Complex Networks **2**(3), 203–271 (2014)
7. Koptelov, M., Zimmermann, A., Crémilleux, B.: Link prediction in multi-layer networks and its application to drug design. In: IDA. pp. 175–187. Springer (2018)
8. Lim, H., Gray, P., Xie, L., Poleksic, A.: Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Scientific reports **6**, 38860 (2016)
9. Schomburg, I., Chang, A., Schomburg, D.: Brenda, enzyme data and metabolic information. Nucleic acids research **30**(1), 47–49 (2002)
10. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic acids research **34**(suppl_1), D535–D539 (2006)
11. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research **34**(suppl_1), D668–D672 (2006)
12. 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 **24**(13), i232–i240 (2008)
13. Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: KDD. pp. 1025–1033. ACM (2013)