Drug-Target Interaction Prediction for Drug Repurposing with Probabilistic Similarity Logic

Shobeir Fakhraei
University of Maryland
College Park, MD, USA
shobeir@cs.umd.edu

Louiqa Raschid
University of Maryland
College Park, MD, USA
louiqa@umiacs.umd.edu

Lise Getoor
University of Maryland
College Park, MD, USA
getoor@cs.umd.edu

ABSTRACT

The high development cost and low success rate of drug discovery from new compounds highlight the need for methods to discover alternate therapeutic effects for currently approved drugs. Computational methods can be effective in focusing efforts for such drug repurposing. In this paper, we propose a novel drug-target interaction prediction framework based on probabilistic similarity logic (PSL) [5]. Interaction prediction corresponds to link prediction in a bipartite network of drug-target interactions extended with a set of similarities between drugs and between targets. Using probabilistic first-order logic rules in PSL, we show how rules describing link predictions based on triads and tetrads can effectively make use of a variety of similarity measures. We learn weights for the rules based on training data, and report relative importance of each similarity for interaction prediction. We show that the learned rule weights significantly improve prediction precision. We evaluate our results on a dataset of drug-target interactions obtained from Drugbank [27] augmented with five drug-based and three target-based similarities. We integrate domain knowledge in drug-target interaction prediction and match the performance of the state-of-the-art drug-target interaction prediction systems [22] with our model using simple triad-based rules. Furthermore, we apply techniques that make link prediction in PSL more efficient for drug-target interaction prediction.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: biology and genetics

General Terms

Algorithms

Keywords

Link prediction, Drug repurposing, Drug discovery, Polypharmacology, System biology, Statistical relational learning, Machine learning

BIOKDD'13 Chicago, IL, USA

1. INTRODUCTION

The development of drugs based on novel chemistry is a time consuming and costly procedure. New drugs often take nearly a decade to reach market, and the development cost often reaches two billion US dollars. In addition, most novel drug candidates fail in or before the clinical trials and will never get approved. In fact, such failures are so common that the process is often referred to as the *valley of death*. The cost of these failures must be borne by the companies involved.

Drugs are organic small molecules that bind to the biomolecular targets in order to activate or inhibit their functions. However, drugs often effect not a single target, but multiple ones. *Polypharmacology*, the study of drugs acting on multiple targets, is an area of growing interest [4]. The interactions with multiple targets can potentially result in adversarial side effects or unintentional treatments.

A potentially effective method to new treatment discovery is finding new uses for drugs which have already been approved. Such *drug repositioning* or *repurposing* approaches bypass the need for many tests required for a new therapeutic compound, as they have already been pre-approved and their safety has already been established. One of the most famous examples of drug repositioning is *Sildenafil* which was originally developed as a treatment for pulmonary arterial hypertension. During clinical trials it was discovered to have a side effect potentially treating erectile dysfunction in men. The drug was repurposed and later renamed as *Viagra* [8].

The discovery of these new uses by chance highlights the need for a systematic approach for finding new treatment effects. However, experimental identification of drug-target associations is a labor intensive and costly procedure. Hence, computational prediction methods are promising approaches for focusing the experimental investigations. They can serve as a basis for biological investigations and experiments, thus, reducing the cost and time of new discoveries [13]. Recent computational approaches for addressing this problem make use of networks which describe the relationships among drugs and targets [29]. Such networks can be constructed via integration of data from multiple publicly accessible datasets [18].

In order to predict drug-target interactions, we can construct a bipartite graph between drugs and targets, where edges denote interactions. We can augment the bipartite graph with drug-drug and target-target similarities. In this graph, similar drugs tend to interact with the same targets, and similar targets tend to interact with the same drugs [7]. The similarities between drugs and between targets have different semantics. For example, drugs can have similarities based

on chemical structure, ligand[1], gene expression, side effect and, annotation [22]. Figure 1 shows a schematic overview of drug-target interaction network.

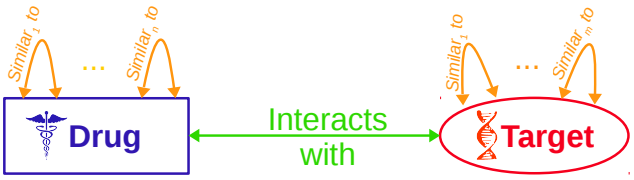

Figure 1: A schematic overview of drug-target interaction network, edges between drugs and between targets represent different similarities.

In this setting, links between drugs and targets indicate their interactions. Prediction of a new potential drug-target interaction can be cast as a link prediction problem. There are multiple standard link prediction methods established for networked data [10, 20]. However, traditional link prediction methods often fail to make use of the multi-relational (i.e. nodes and edges with different semantics) characteristics of this drug-target interaction network.

On the other hand, the structure of the network and the multi-relational aspects of it, makes it challenging to transform into a flat form for input to a standard link prediction algorithm. Because the links are not independent but depend on the similarities between their end points and other interactions, a more collective approach is appropriate. Attempts to convert the data to such flat format and apply traditional machine learning approaches rely on heuristics such as learning new combined features [22, 28]. While such methods may demonstrate good prediction performances, they suffer from low interpretability, and loss of information.

We propose a drug-target prediction framework based on Probabilistic Similarity Logic[2] (PSL) [5] that preforms on the original data format and captures the multi-relational nature of the network. We use probabilistic first-order logic rules in PSL to perform drug-target interaction prediction based on interpretable domain knowledge. We use rules based on triad and tetrad structures, and show that triad-based rules outperform tetrad-based rules in drug-target interaction prediction. We then use training data to learn weights for these rules with PSL, improving prediction performance. Weight learning can also provide insight into relative importance of each similarity for interaction prediction. Furthermore, we apply techniques that make link prediction in PSL more efficient in drug-target interaction domain. Finally we match the performance of the state-of-the-art drug-target interaction prediction approaches using simple triad-based rules.

2. RELATED WORK

There are multiple approaches for drug-target interaction prediction. In Similarity Ensemble Approach (SEA), Keiser et al. [13] predict drug-target interaction based on ligands. They represent targets with their ligands and consider chemical similarities between drugs and ligand sets as indicators for possible interactions. Lamb et al. [16, 17] in CMap represent diseases, genes and drugs with their mRNA expression profiles. They create gene-expression profiles from cultured human cells treated with bioactive small molecules, and compare these expressions only through up and down regulations, providing possibility of cross-platform comparisons. They propose drugs with opposite expression profiles, as potentially effective on diseases.

Networks of drugs and genetic products are of interest from multiple perspectives. Cockell et al. [8] describe how to build an integrated systems biology network for drug repurposing with elements such as drugs, targets, genes, proteins and pathways. In addition, they indicate that similar targets interact with same drugs and similar drugs tend to interact with same targets. Lee et al. [18] describe how a network can be used in tasks such as multi-agent drug development, drug repurposing, estimation of drug effects on target perturbations in the whole system. They summarize the information that should be integrated in networks for each task, and list resources that contain such data.

Yildirim et al. [29] provide an analysis on the drug-target networks and explain the trends in drug-discovery industry over time with emphasis on the effect of sequencing genome on such trends. They also highlight interesting characteristics of drug-target interactions from a network perspective, including preferential attachment and cluster formations.

There are methods that use one set of similarities for drugs and targets to predict interactions. Cheng et al. [7] build a bipartite graph and use similarities between drugs and between targets to predict new potential interactions. They use three link prediction approaches. They use drug-based similarity inference (DBSI), only considering similarities between drugs, and target-based similarity inference (TBSI), only considering target similarities. Furthermore, they propose network-based inference (NBI), combining both similarities. They calculated 2D chemical similarities with SIMCOMP for drugs, and Smith-Waterman score genomic sequence similarities for targets.

Yamanishi et al. [28] propose three interaction prediction methods based on using only the nearest neighbor drug or target, weighted k-nearest neighbors, and space integration. In the space integration method they describe genomic space (using Smith–Waterman score) for targets, and chemical space (using SIMCOMP) for drugs. They propose a method to integrate drugs and targets in a unified latent space called Pharmacological space, and predict interactions in that space based on proximity of the drugs and targets. They conduct their experiments on four categories of targets namely, Enzyme, Ion Channel, GPCR, Nuclear Receptor. They report that two compounds sharing high structure similarity tend to interact with similar target proteins. Likewise, two target proteins sharing high sequence similarity tend to interact with similar drugs.

Along this line of work, Bleakley and Yamanishi [3] build a bipartite drug-target graph inference method with local models to predict interactions. They model link prediction in the bipartite graph as two classifications for each drug-target interaction. Once the classifier treats the drug in each interaction as the label, and once the target is teated as such. The predictions acquired from each classifier are combined in later stages. They use Support Vector Machines (SVM) with similarity matrices as kernels.

Furthermore, some methods integrate multiples similarities for the prediction task. Chen et al. [6] develop an statistical

[1]A substance that binds with a biomolecule to serve a biological purpose

[2]Also referred to as Probabilistic Soft Logic.

model to assess the association of drug-target pairs based on their relation with other linked objects. They measure strength of a relation in the network according to distance, the number of shortest paths and other topological properties between the two nodes. They first find the paths between drugs and targets and assign a score to each path. Then for every drug-target pair, they integrate scores of all the paths between them as their final relation score.

Perlman et al. [22] also convert the link prediction problem into a classification setting, where instances are drug-target interactions and features are combination of drug-drug and target-target similarities. They use multiple similarities for drugs and targets (Five drug-drug and three target-target similarities). They train their classifiers using three online datasets and reserved another dataset for validation. They use cross validation in their experiments, and consider unobserved interactions as negative samples. They perform undersampling to deal with class-imbalanced issue imposed by the limited number of true interactions among a very large number of possibilities.

Gottlieb et al. [11] extend this approach to drug-disease relation prediction. In addition, they propose use of this method in personalized medicine by representing the disease via its genetic signature. In such setting, a drug can be recommended for an unknown disease with just a genetic signature representation.

3. OUR MODEL

We propose a drug-target prediction framework based on Probabilistic Similarity Logic (PSL) [5]. In the following sections, we provide details about PSL and describe the logical rules that we use for drug-target interaction prediction. We then describe techniques that help make link prediction in PSL more efficient.

3.1 Probabilistic Similarity Logic

PSL uses rules written in first-order logic as a template language for graphical models over random variables. For example, a typical PSL rule looks like the following:

$$w : P(A, B) \wedge Q(B, C) \rightarrow R(A, C) \tag{1}$$

where P , Q and R are *predicates* and A, B, and C are *variables*. e.g. $P(A, B)$ can be $Interacts(D, T)$ where D represents a drug and T is a target. Instantiation of predicates with data is called *grounding* (e.g. *Interacts(acetaminophen, cox2)*). Each ground predicate, often called ground atom has a soft-truth value in the range of $[0, 1]$. To build a PSL model in the drug-target interaction domain, we represent drugs and targets as variables and specify predicates to represent different similarities and interactions between them. Domain knowledge is captured by writing rules that govern the relationship between these predicates.

Since PSL uses soft-truth values for atoms, some relaxations from Boolean domain is required to use the first-order logical representation. PSL uses the *Lukasiewicz* t-norm and co-norm to provide a relaxation of the logical connectives, $AND(\wedge)$, $OR(\vee)$, and $NOT(\neg)$ as following:

$$\begin{aligned} p\tilde{\wedge}q &= max(0,\ p+q-1) \\ p\tilde{\vee}q &= min(1,\ p+q) \\ \tilde{\neg}p &= 1-p \end{aligned}$$

where $\tilde{}$ denotes relaxed form.

An assignment of soft-truth values to a set of ground atoms, is called an *interpretation* (I) of that set. In this setting, a ground instance of a rule r $(r_{body} \longrightarrow r_{head})$ is satisfied (i.e., $I(r) = 1$) when $I(r_{body}) \leq I(r_{head})$.

PSL converts these rules into an optimization problem for *Most Probable Explanation (MPE)* inference, which is finding the most probable interpretation given evidence (i.e., a given partial interpretation). To perform the MPE inference, PSL calculates a *distance to satisfaction* value for any grounded instance of a rule. The rule's distance to satisfaction under interpretation I is calculated via the following:

$$d_r(I) = max\{0, I(r_{body}) - I(r_{head})\} \tag{2}$$

PSL uses the model to define a probability distribution over interpretations I by combination of the weighted degree of satisfaction over all rules, as the following:

$$f(I) = \frac{1}{Z} exp\left[-\sum_{r \in R} w_r d_r(I)\right] \tag{3}$$

where R is the set of ground rules, w_r is the weight of rule r and Z is the continuous version of the normalization constant used in discrete Markov random fields:

$$Z = \int_I exp\left[-\sum_{r \in R} w_r d_r(I)\right] \tag{4}$$

MPE inference in PSL means maximizing the density function $f(I)$ in Equation (3), subject to both the evidence and the equality and inequality constraints. For example, given a drug-target interaction network and interactions between some drugs and some targets, MPE inference derives the most likely interactions between all other drugs and targets. Finding the most probable interpretation given a set of weighted rules reduces to solving a convex optimization problem and can be solved very efficiently [1].

Every rule in PSL is associated with a weight w. These weights indicate how much an assignment is penalized if a rule is not satisfied. They are a measure of importance for each rule. We can set the weights based on prior domain knowledge or, if we have training data, we can learn the weights. One way to learn the weights is via maximum-likelihood estimation [14]. The gradient of the log-likelihood with respect to a weight w_i is:

$$\frac{\partial}{\partial w_i} log f(I) = -\sum_{r \in R_i} (d_r(I))^p + \mathbb{E}\left[\sum_{r \in R_i} (d_r(I))^p\right]$$

where R_i is the set of ground rules parameterized with weight w_i. Bach et al. [2] discuss more advanced methods of weight learning.

3.2 PSL Model for Drug-Target Interaction

The rules in a PSL program capture domain knowledge about the task domain. For drug-target interaction prediction, established methods are based on triangles or triads between drugs and targets. Similar targets tend to interact with the same drug, and similar drugs tend to interact with the same target [7, 8, 28]. Figure 2 depicts the triad-based prediction of interaction for drugs and targets.

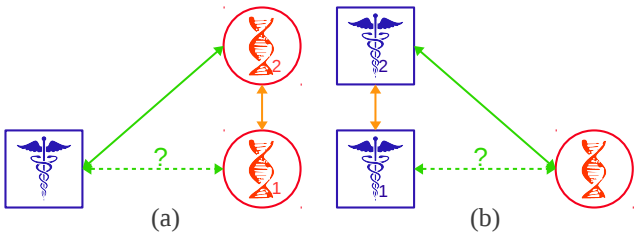

Figure 2: Similar targets tend to interact with the same drug (a), and similar drugs tend to interact with the same target (b).

The following rules capture the triads shown in Figure 2(a) and 2(b) respectively:

$$SimilarTarget_{\beta}(T_1, T_2) \wedge Interacts(D, T_2) \rightarrow Interacts(D, T_1) \quad (5)$$

$$SimilarDrug_{\alpha}(D_1, D_2) \wedge Interacts(D_2, T) \rightarrow Interacts(D_1, T) \quad (6)$$

where T denotes a target, D indicates a drug, $SimilarTarget_{\beta}$ represents a specific target-target similarity metric. For each similarity metric an instance of the rule (5) is added to the PSL model. As described in section 4, in our experiments we include three instances of this rule, such that $\beta \in \{$ *Sequence-based, PPI-network-based, Gene Ontology-based* $\}$. $SimilarDrug_{\alpha}$ is a specific drug-drug similarities. In our experiments we include five instances of rule (6), and $\alpha \in \{$ *Chemical-based, Ligand-based, Expression-based, Side-effect-based, Annotation-based* $\}$.

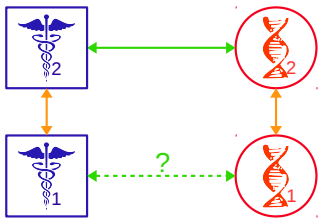

Figure 3: If two drugs and two targets are similar correspondingly, and one of the drugs interact with one of the targets, then other two could also interact with each other.

In addition to considering such triads, we can consider more complicated situations in which we reason about both drug and target similarities for predicting interactions. Figure 3 illustrates this setting. We refer to these rules as tetrad rules:

$$SimilarDrug_{\alpha}(D_1, D_2) \wedge SimilarTarget_{\beta}(T_1, T_2) \wedge Interacts(D_2, T_2) \rightarrow Interacts(D_1, T_1) \quad (7)$$

where α and β are from the same set of similarities described earlier. We include an instance of this rule for every combination of different target-target and drug-drug similarity metric.

We also include a negative prior indicating that *Interacts* predicate is most likely false. In other words, this rule promotes the possibility that drugs and targets do not interact. The prior rule is simply the following:

$$\neg Interacts(D, T) \quad (8)$$

3.3 Blocking

As PSL grounds every possible rule in the network for each link prediction, the number of grounded rules can be extremely large. If $|D|$ denote the number of drugs and $|\alpha|$ number of different similarities between them, and $|T|$ specify the number of targets and $|\beta|$ number of different similarities between targets, for each potential link, $O(|D| \times |\alpha|)$ instance of rule (6) and $O(|T| \times |\beta|)$ instances of rule (5) can be grounded. For the tetrad rules the situation is even worst. In addition, since there are $O(|D| \times |T|)$ potential interactions, the total number of ground rules is $O(|D| \times |T| \times (|D| \times |\alpha| + |T| \times |\beta|))$. Running inference on such huge number of ground rules is computationally very expensive.

To control such situation we limit some of the rules from being grounded, by reducing the number of tetrads that are considered for each potential link. To reduce this number we ignore some of the less significant similarities between drugs and between targets. More specifically, we reduce $|\beta|$ and $|\alpha|$. This is commonly referred to as *blocking* [21], and is a way of limiting the number of links considered. Typically, a fast method for computing the blocking criterion is used, to avoid the quadratic blow up.

There are several ways to achieve such goal; one is simply using a fixed threshold for all similarities and set the values below that threshold to zero. However, although the similarities are normalized to a $[0, 1]$ range, the distribution of the values could be highly different such that a fixed threshold either ignores most of the values in one similarity or includes most of the values from the other. Figure 4 which shows the distribution of similarities in our dataset, confirms this situation.

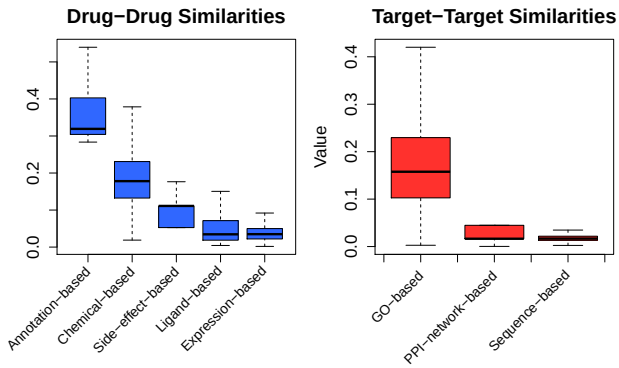

Figure 4: Distribution variation of different similarity values between drugs and between targets. Similarities with values of zero or one are omitted in this plot.

Another method of blocking is based on choosing a different threshold for each similarity measure. While better than the previous approach, similar problem could arise in a target or drug level similarities, i.e. similarities for each target or for each drug having highly variable values. Therefore, choosing a fixed threshold will include many similarities for some drugs or targets, and very few for others. Figure 5 which shows Annotation-based similarity for drugs, demonstrates

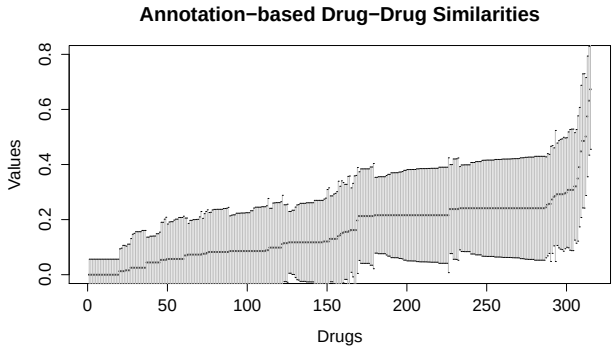

Figure 5: Distribution of Annotation-based similarity values for 315 drugs, where dots indicate mean similarity value between each drug and all others, and lines demonstrate standard deviation of the values. Similarities with values of zero or one are omitted in this plot. e.g., mean of all drug similarities with drug #200 is about 0.2 with stdev. of 0.15.

an instance of this situation.

Instead, here we propose an approach based on k-nearest-neighbors to ensure that for every drug and every target at least a few values from each similarity are not blocked. In this approach, we preserve the k-highest values in each similarity for each drug and each target and set the others to zero. However, depending on the method used for calculating the similarities, there are many cases that similarity values between multiple drugs or targets are the same. Hence, k-th nearest neighbor of a drugs or target is often not only a single instance but many drugs or targets with the same similarity values. To control this situation we consider the drugs or targets with similarities greater than the k-th nearest neighbor. In other words, we only include the similarities from k-1 drugs or targets. Formally, the *blocked* set of similarity predicates are as follows:

$$Similar_{\lambda}^{blocked} = \begin{cases} Similar_{\lambda}(x_i, x_j) & \text{if } Similar_{\lambda}(x_i, x_j) > Similar_{\lambda}(x_i, x_k); \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where λ is any drug-drug or target-target similarity and x_k is the k-th nearest neighbor of x_i.

3.4 Collective Classification

Most traditional machine learning approaches assume the data are *independent and identically distributed* (i.i.d.). In drug-target interaction prediction setting, the presence or absence of an interaction is often studied independently. However, interaction are highly interdependent as they are predicated based on each other (e.g., in triads).

It can be beneficial to consider interactions jointly. This approach often referred to as *collective classification* [24] and it results in global information propagation through the network. Since PSL performs MPE inference on the interpretation I over the whole network (eq. (3)), interaction predictions propagate and influence the prediction of other interactions. Therefore, even if a triad or tetrad structure does not initially include an observed interaction, such rules can reason about new interactions using other predicated interactions.

4. EXPERIMENTAL ANALYSIS

We performed an empirical evaluation studying:

- The effectiveness of triad ((5)&(6)) and tetrad rules (7).
- The effectiveness of our proposed blocking strategy.
- The results of weights learned for the different similarity functions.

4.1 Dataset

Our dataset is based on a network of drugs and genetic targets, where interactions between them are obtained from the DrugBank database [27]. The dataset includes 315 drugs, 250 targets and 1,306 interactions. We use 5 drug-drug and 3 target-target similarities, obtained from Perlman et al. [22]. A brief description of the methods used for similarity calculation are provided in this section. Drug-drug similarities include the following:

1. **Chemical-based:** Using the chemical development kit (CDK) [26], the hashed fingerprint of each drug based on the canonical SMILES[3] obtained from Drugbank, was computed. Considering each fingerprint as a set of elements, The Jaccard similarity of the fingerprints where calculated. The Jaccard similarity score between two sets X and Y is as follows:
$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$
2. **Ligand-based:** Drugs canonical SMILES obtained from Drugbank were compared against a collection of ligand sets using the similarity ensemble approach (SEA) search tool [13]. A list of relevant protein-receptor families were obtained for each drug, and Jaccard similarity was computed between the corresponding sets of receptor families for each drug pairs.
3. **Expression-based:** The Spearman rank correlation coefficient of gene expression responses to drugs retrieved from the Connectivity Map project [16, 17] was used as a similarity measure between drugs. Spearman rank correlation coefficient between two sets X and Y is calculated via:
$$Spearman(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$
where x_i and y_i are ranked elements of X and Y.
4. **Side-effect-based:** Drugs side-effects were obtained from SIDER [15] and the similarities between drugs were calculated using the Jaccard score between their common side-effects.

[3]Simplified Molecular Input Line Entry Specification

5. **Annotation-based:** Drug's ATC codes where obtained from DrugBank and matched against the World Health Organization ATC classification system [25], where drugs are categorized based on different characteristics. The semantic similarity algorithm of Resnik [23] was used to calculate the similarities.

Target-target similarities include the following:

1. **Sequence-based:** The Smith-Waterman sequence alignment scores, normalized via the method suggested in [3], which divides the pairwise score by the geometric mean of the alignment scores of each sequence against itself.
2. **Protein-protein interaction network-based:** Using an all-pairs shortest path algorithm, the distance between pairs of genes were calculated using their corresponding proteins in the human protein-protein interactions network.
3. **Gene Ontology-based:** Using the method of Resnik [23] the semantic similarity measure between Gene Ontology annotations, downloaded from UniProt [12] were calculated.

Perlman et al. [22] provide more detailed description of these similarities.

4.2 Evaluation Criteria

There are multiple evaluation methods for link prediction problems [19]. We use the Area Under the ROC Curve (AUC) [9] for our evaluations as this is the most common reported measure in the related publications, and it allows us to compare against the published results of other methods [3, 22, 28] on the same dataset.

ROC curves are created by plotting the true positive rate versus the false positive rate at various thresholds. The drug-target interaction network does not have information about the negative samples, i.e. lack of interaction. The missing links which we call unobserved interactions in the network, could represent interactions that are not yet studied. As described in the next section, the common practice is to assume the unobserved interactions as negative samples. In our inference we do not make such assumption. However, to be able to calculate the AUC in our evaluations, we treat the unobserved interaction as negative samples. This evaluation assumption does not effect the inference step.

We use 10-fold cross validation, where at each fold 10% of the positive interactions are left out. We perform inference, predicting interactions, and compare against the held out known interaction. We report the average values over all folds along with their standard deviations.

4.3 Results

We first conduct our experiments to study the effect of domain knowledge (or assumptions) on the predictions. We compare the rules based on triads ((5)&(6)), and the rules based on tetrads (7). We conduct the experiments using three different settings; once including only the triad-based rules, once including only the tetrad-based rules, and once including both set of rules in the model. We set the blocking parameter (k) to 5, in order to control the groundings in tetrad-based and combined settings. We learn the weights of the rules in each setting using separate hold-outs of interactions.

As Table 1 shows the rules inspired by triads are more predictive of the interactions comparing to the rules that are based on tetrads. This experiment not only provides insight into the prediction paradigm assumptions of triads and tetrads, but also demonstrates how we can easily test such domain assumptions.

Table 1: Comparision of triad-based and tetrad-based rules

Rules	AUC with k=5
Triad-based only	0.930 ± 0.016
Tetrad-based only	0.796 ± 0.025
Triad-based & tetrad-based	0.913 ± 0.017

In our next set of experiments, we study the effect of blocking and weight learning on performance using the triad rules ((5)&(6)). We conduct the blocking study by varying the number of neighbors (k). We study the effect of weight learning by running the experiments under two conditions. We once set all weights to 5 (arbitrary), and once learn the weights from a set of observed interactions. Table 2 shows performance of the model with different ks. The insignificant performance change suggests that even with limited number of similarities (i.e., k=5) PSL can provide valid predictions. Table 2 shows average computation time of a 10-fold cross validation experiment on a computer with a (2×4) 2.66 GHz Intel processor and 48GB of RAM. It should be noted that as we executed our experiments on machines with slightly different specifications and under different loads, these numbers are approximate representations. The results show that blocking causes significant improvement in processing time with no performance loss.

Table 2: Completion time and performance variations under the effect of blocking and weight learning with triad-based rules

Condition	Time to Complete and AUC		
	k=5	k=15	k=30
Exec. time	12mins	3h	9h
$\forall r : w_r = 5$	0.926 ± 0.016	0.929 ± 0.020	0.923 ± 0.021
Exec. time	1h	10h	28h
w_r Learned	0.930 ± 0.016	$\mathbf{0.931 \pm 0.018}$	0.924 ± 0.21

Table 2 also shows the effect of weight learning on performance. Although weight learning improves the results, it does not have a significant impact on AUC. Table 3 shows the average weights assigned by PSL to triad rule of each similarity. In this combined setting where rules are weighted against each other *Sequence-based* target similarities are slightly more important.

Perlman et al. [22] report experimental evaluation on the same dataset as the one we use for our experiments. They report that the method of Perlman et al. [22] achieves AUC of 0.935, method of Yamanishi et al. [28] achieves AUC of 0.884 and, Bleakley and Yamanishi [3] get AUC of 0.814.

In our experiments, although we use the same dataset as Perlman et al. [22], due to different sampling methods and

Table 3: Weights learned for each triad-based rule under variation of number of neighbors (k)

Similarity		Weights		
		k=5	k=15	k=30
Drugs	Annotation-based	1.21	1.28	1.46
	Chemical-based	1.29	1.40	1.42
	Ligand-based	1.92	2.06	2.02
	Expression-based	1.65	1.78	1.75
	Side-effect-based	1.74	1.75	1.67
Targets	PPI-network-based	1.55	1.24	1.20
	GO-based	1.81	1.90	1.84
	Sequence-based	2.72	2.45	2.18
Negative Prior		3.43	9.05	14.21

performance variations under the effect of sampling, our results are not exactly comparable. However, our results in Table 2 show that in addition to high interpretability, PSL with simple triad-based rules matches the prediction performance of the state-of-the-art methods.

Figure 6 shows the average precision of the top 100 interaction predictions for all 10 folds. As we did not sample the unobserved interactions for inference, our setting is highly class imbalanced. We have 1,306 real interactions (positive samples) and 78,750 total possible interactions. In such imbalanced setting, our predictions have high precision. In addition, Figure 6 shows that although weight learning did not have a high impact on the AUC, it significantly improves the precision of the predictions.

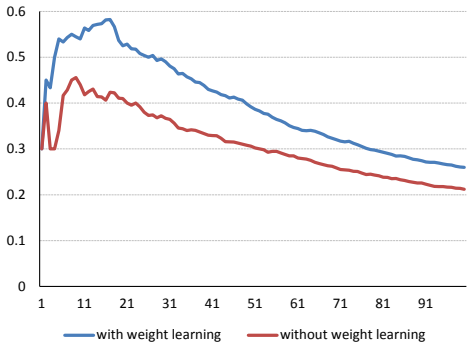

Figure 6: Average precision of the top 100 interaction predictions for all 10 folds with k=5.

5. DISCUSSION AND CONCLUSION

To use traditional machine learning algorithms we have to convert the data to a matrix of features and instances. However, the drug-target interaction network is not easily convertible to such format. The methods that transform the data to such suitable format [11, 22], rely mostly on heuristic intuitions, suffer from loss of information, and are hard to interpret. Since PSL is fundamentally designed based on multi-relational network environments there is no such requirement to change the data format.

In addition, modeling drug-target interaction prediction as classification, requires presence of both positive and negative samples to achieve optimal performance. However, it is not really known that the missing link between a drug and a target is due to a real lack of interaction, or lack of scientific investigation. A common technical assumption of considering all missing links as negative samples [3, 22, 28], is not valid. In contrast, we write PSL rules without the need for negative instances. In our rules, the missing links are treated as unobserved interactions. In the inference step a truth values will be assigned to all unobserved interactions by PSL.

In conclusion, we proposed a drug-target interaction prediction framework based on Probabilistic Similarity Logic. We showed how domain assumptions can be implemented and tested using rules in a bipartite network of drug-target interactions with a set of similarities between drugs and between targets. Furthermore, we studied rules based on triads and more advanced tetrad structures for link prediction. We described limitation of different blocking methods and proposed a highly efficient blocking methods in the network. Using weight learning, we provided insight into usefulness of each similarity with respect to interaction predication and improved the prediction precision. We also showed that in addition to higher interpretability, our model matches the performance of the state-of-the-art approach via experimental evaluations.

We can extend our method with study of clusters in the network and rules on that basis. We also plan to study performance of our model using other performance measures and evaluate new interaction predictions of the PSL with help of domain experts. Our proposed method can easily be generalized to all networks with similar structures.

6. ACKNOWLEDGMENT

The authors would like to thank Stephen Bach and Bert Huang for providing their invaluable insights on PSL. We also like to thank Alex Memory, Benjamin London and Jay Pujara for their technical comments. We are also grateful for the dataset that Assaf Gottlieb and Eytan Ruppin provided us with. This work is partially supported by the National Science Foundation (NSF) under contract numbers IIS0746930, CCF0937094 and IIS1218488.

References

[1] S. H. Bach, M. Broecheler, L. Getoor, and D. P. O'Leary. Scaling MPE Inference for Constrained Continuous Markov Random Fields with Consensus Optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2663–2671, 2012.

[2] S. H. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction. In *Uncertainty in Artificial Intelligence*, 2013.

[3] K. Bleakley and Y. Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, Sept. 2009.

[4] A. D. Boran and R. Iyengar. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297, 2010.

[5] M. Broecheler, L. Mihalkova, and L. Getoor. Probabilistic Similarity Logic. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

[6] B. Chen, Y. Ding, and D. J. Wild. Assessing Drug Target Association Using Semantic Linked Data. *PLoS Comput Biol*, 8(7):e1002574, July 2012.

[7] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *PLoS Comput Biol*, 8(5): e1002503, May 2012.

[8] S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson, M. Young, and A. Wipat. An integrated dataset for in silico drug discovery. *J Integr Bioinform*, 7(3):116, 2010.

[9] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31:1–38, 2004.

[10] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, Dec. 2005.

[11] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1), June 2011.

[12] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. E. Suzek, M. J. Martin, P. McGarvey, and E. Gasteiger. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics*, 10(1):136, 2009.

[13] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, Nov. 2009.

[14] A. Kimmig, S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. A Short Introduction to Probabilistic Soft Logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[15] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1), 2010.

[16] J. Lamb. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54–60, Jan. 2007.

[17] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313(5795):1929–1935, Sept. 2006.

[18] S. Lee, K. Park, and D. Kim. Building a drug–target network and its applications. *Expert Opinion on Drug Discovery*, 4(11):1177–1189, Nov. 2009.

[19] R. Lichtnwalter and N. V. Chawla. Link prediction: fair and effective evaluation. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 376–383. IEEE, 2012.

[20] L. Lu and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, Mar. 2011.

[21] A. E. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Proceedings of the second international Conference on Knowledge Discovery and Data Mining*, pages 267–270, 1996.

[22] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan. Combining Drug and Gene Similarity Measures for Drug-Target Elucidation. *Journal of Computational Biology*, 18(2):133–145, Feb. 2011.

[23] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[24] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective Classification in Network Data. *AI Magazine*, 29(3):93–106, 2008.

[25] A. Skrbo, B. Begović, and S. Skrbo. Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Medicinski arhiv*, 58(1 Suppl 2):138, 2004.

[26] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics. *Current pharmaceutical design*, 12(17):2111–2120, 2006.

[27] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

[28] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, July 2008.

[29] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, and M. Vidal. Drug–target network. *Nature biotechnology*, 25(10):1119–1126, Oct. 2007.