Network-Based Drug-Target Interaction Prediction with Probabilistic Soft Logic

Shobeir Fakhraei, Bert Huang, Louiqa Raschid, and Lise Getoor

Abstract—Drug-target interaction studies are important because they can predict drugs' unexpected therapeutic or adverse side effects. *In silico* predictions of potential interactions are valuable and can focus effort on *in vitro* experiments. We propose a prediction framework that represents the problem using a bipartite graph of drug-target interactions augmented with drug-drug and target-target similarity measures and makes predictions using probabilistic soft logic (PSL). Using probabilistic rules in PSL, we predict interactions with models based on triad and tetrad structures. We apply (blocking) techniques that make link prediction in PSL more efficient for drug-target interaction prediction. We then perform extensive experimental studies to highlight different aspects of the model and the domain, first comparing the models with different structures and then measuring the effect of the proposed blocking on the prediction performance and efficiency. We demonstrate the importance of rule weight learning in the proposed PSL model and then show that PSL can effectively make use of a variety of similarity measures. We perform an experiment to validate the importance of collective inference and using multiple similarity measures for accurate predictions in contrast to non-collective and single similarity assumptions. Finally, we illustrate that our PSL model achieves state-of-the-art performance with simple, interpretable rules and evaluate our novel predictions using online datasets.

Index Terms—Link prediction, Collective inference, Heterogeneous similarities, Drug target prediction, Drug target interaction prediction, Drug repurposing, Drug discovery, Polypharmacology, Drug adverse effect prediction, Statistical relational learning, Hinge-loss Markov random fields, Machine learning, Bipartite networks, Systems biology

1 Introduction

The cost of successful novel chemistry-based drug development often reaches billions of dollars, and the time to introduce the drug to market often comes close to a decade. Most new compounds fail during clinical trials or show adverse side effects. Because of the high failure rate of drugs during this process, the trial phase is often referred to as the "valley of death" [1].

Most drugs[1] affect multiple targets, and *Polypharmacology*, the study of such interactions, is an area of growing interest [2]. These multi-target interactions potentially result in both unintentional therapeutic and adverse side effects. Predicting side effects during the drug developmental phase can reduce the high cost of clinical trials and is crucial for the commercial success of new drugs. Moreover, due to the high cost and low success rate of novel drug development, pharmaceutical companies are particularly interested in *drug repositioning* or *repurposing*, which involves finding new therapeutic effects of pre-approved drugs.

Sildenafil—originally developed for pulmonary arterial hypertension treatment—is a famous drug repurposing example. In clinical trials, it was discovered by chance to have a side effect of treating erectile dysfunction in men, and it was eventually re-branded as *Viagra* [3].

Drug-target interaction identification is an essential step of drug repurposing and drug adverse effect prediction. *In vitro* identification of drug-target associations is a labor-intensive and costly procedure. Hence, *in silico* prediction methods are promising approaches for focusing *in vitro* investigations [4].

There are several methods to model the drug-target interaction prediction task [5], many of which use a network representation [6]. We can construct a bipartite interaction network where nodes represent drugs and targets, and edges denote interactions. Drug-drug and target-target similarities can augment this network on each side. Data from multiple publicly accessible datasets can be integrated toward building these networks [7]. The similarities between drugs and between targets have different semantics. For example, targets can have similarity measures based on their sequences and their ontology annotations [8]. Figure 1 shows a schematic overview of a drug-target interaction network.

A link prediction method can predict new potential drug-target interactions in this setting [9, 10]. However, traditional link prediction methods often ignore the multi-relational characteristics of this drug-target interaction network (i.e., nodes and edges with different semantics) or make oversimplifying assumptions that neglect key, interdependent phenomena during prediction.

The structure of the network and the multi-relational aspects make it challenging to convert such knowledge into the (flat) data formats that are typically used with standard prediction algorithms. At-

• *S. Fakhraei, B. Huang, L. Raschid and L. Getoor are with the Computer Science Department, University of Maryland, College Park, MD, 20740. L. Getoor is also affiliated with Computer Science Department, University of California, Santa Cruz, CA, 95064.*
E-mail: shobeir@cs.umd.edu

1. Organic molecules that bind to bio-molecular targets and inhibit or activate their functions.

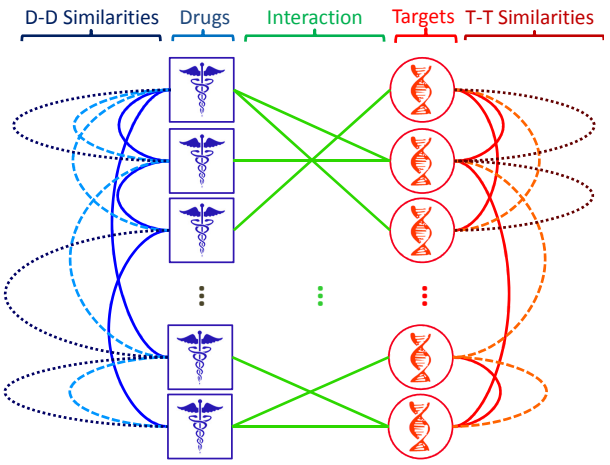


Fig. 1: A schematic overview of a drug-target interaction network. Edges between drugs and between targets represent different similarities.

tempts to make such conversions often rely on potentially ad-hoc feature engineering approaches [8, 11]. Such methods may sometimes yield good prediction performance, but they suffer from low interpretability and loss of information. Our approach is based on the premise that links depend on the similarities between their endpoints and on other interactions. Hence, a collective approach is more appropriate than standard machine learning models that make simplifying independence assumptions. As described in Section 5, in a collective setting, the presence or absence of interactions are studied interdependently.

In this paper we present a drug target prediction framework based on *probabilistic soft logic* (PSL) [12, 13]. We reason collectively over the unknown interactions using a structured representation that captures the multi-relational nature of the network[2]. We design a PSL model for drug-target interaction prediction that reasons over structured rules. We consider two types of structured rules, triad rules and tetrad rules, and consider a variety of similarity metrics. We propose a blocking method to manage the large computational cost of inference in this task and perform experimental studies on different aspects of link prediction in the drug-target interaction domain with this model. We compare the relative improvement provided by each type of structural rule and find that triad-based rules enable more accurate predictions. We also experiment with the effect of using different similarity metrics and show that combining all similarity metrics in a single probabilistic model produces the most effective model. We additionally test the importance of collective inference in such models by comparing against an analogous model that makes independent predictions. Our PSL based solution outperforms the state-of-the-art drug-target interaction prediction method proposed by Perlman et al. [8]. We further validate that our PSL models can outperform Perlman et al. [8] on a set of new interactions that were not considered in the original evaluation of Perlman et al. [8].

2. Our previous workshop papers [14, 15] contain a preliminary version of part of this research.

2 Related Work

In the *similarity ensemble approach* (SEA), Keiser et al. [4] use ligands to predict interaction. They use ligands for target representation and chemical similarities between drugs and ligand sets as potential interaction indicators. In CMap, by Lamb et al. [16, 17], mRNA expressions are used to represent diseases, genes, and drugs. They compare up- and down-regulations of the gene-expression profiles from cultured human cells treated with bioactive molecules and provide cross-platform comparisons. They predict new potential interactions based on opposite-expression profiles of drugs and diseases. Chang et al. [18, 19] proposed a method for predicting drug targets for a given phenotype predicting phenotypes given specific genetic perturbations.

A number of methods reason about network structures to predict interactions. Cockell et al. [3] describe how to integrate drugs, targets, genes, proteins, and pathways into a network for different tasks. They present a hypothesis that similar targets interact with the same drugs, and similar drugs tend to interact with the same targets. Lee et al. [7] describe drug repurposing, multi-agent drug development, and estimation of drug effects on target perturbations via network-based solutions.

Yildirim et al. [6] explain trends in the drug-discovery industry over time using a network-based analysis and show the effect of sequencing the genome on drug development. They also discuss different structural aspects of this network including preferential attachment and cluster formation.

Network-based approaches integrate drug-drug and target-target similarities extracted via different methods (e.g. SEA and CMap) with the drug-target interactions network. The following methods use a single similarity measure for drugs and targets to predict interactions: Cheng et al. [20] predict potential interactions using drug-drug and target-target similarities and a bipartite interaction graph. Using SIMCOMP [21], they compute the 2D chemical drug similarities and sequence similarities for targets via the Smith-Waterman score. They use the following three link-prediction methods: drug-based similarity inference (DBSI)—only considering similarities between drugs; target-based similarity inference (TBSI)—only considering target similarities; network-based inference (NBI)—combining both similarities. Alaimo et al. [22] extend this approach by proposing a DT-hybrid method that integrates prior domain-dependent knowledge.

Yamanishi et al. [11] propose the following three methods for interaction prediction: a nearest neighbor

approach; weighted *k*-nearest neighbors; and space integration. For the space integration method, they describe a genomic space, using the Smith-Waterman score for targets, and the SIMCOMP score for drugs. They propose a method to integrate drugs and targets in a unified latent *pharmacological space*, and they predict interactions in that space based on the proximity of drugs and targets. They separate out four categories of targets, namely enzymes, ion channels, GPCR, and nuclear receptors for their experiments. They also report that similar drugs tend to interact with similar targets and vise versa.

Bleakley and Yamanishi [23] extend this method and construct local models for graph inference. They classify each interaction twice and combine the results to provide a prediction. First, they build a classifier based on drugs and then based on targets. They use the similarities as the *support vector machine* (SVM) kernels. Extending this method, Mei et al. [24] propose to infer training data from neighbors' interaction profiles to make predictions for new drug or target candidate that do not have any interactions in the network. Wang and Zeng [25] propose a method based on restricted Boltzmann machines for drug-target interaction prediction.

More advanced methods predict interactions based on multiple similarities. Chen et al. [26] reason about the possibility of a drug-target interaction in relation with other linked objects. They use distance, shortest paths, and other topological properties in the network to assess the strength of a relation. They assign scores to paths between drugs and targets and combine path scores for each drug-target pair.

Perlman et al. [8] propose a feature-engineering method based on combinations of drug-drug and target-target similarities and use classification to predict interactions. They build their method based on five drug-drug and three target-target similarities. They evaluate their model using cross validation over three online datasets and validate their predication on a fourth dataset. They show significant performance improvement over Bleakley and Yamanishi [23] and Yamanishi et al. [11] that only use one type of drug-drug and target-target similarities for prediction. To the best of our knowledge Perlman et al. [8] method is the state-of-the-art mutli-similarity based approach for drug-target interaction prediction.

Gottlieb et al. [27] extend their method in [8] to drug-disease domain and propose a personalized medicine approach, representing diseases via their genetic signatures. This method can predict the most effective compound for a genetic signature of an unknown disease.

3 Our Model

Our proposed drug-target prediction framework uses *probabilistic soft logic* (PSL) [12, 13]. In this section, we review PSL and present the logical rules that we use for drug-target interaction prediction.

3.1 Probabilistic Soft Logic

PSL uses syntax based on first-order logic as a templating language for probabilistic models over continuous variables. For example, a PSL rule can be of the following form:

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

where P, Q, and R are *predicates*, and A, B, and C are *variables*. For instance, $P(A, B)$ can be *Interacts*(D, T) where D represents a drug and T is a target. Predicates become instantiated with data, generating *groundings* (e.g., *Interacts(acetaminophen, cox2)*). Each grounding forms a ground atom, or logical fact, that has a soft-truth value in the range $[0, 1]$. In our PSL model for drug-target interaction, we represent drugs and targets as variables and specify predicates to represent different similarities and interactions between them. Then the rules encode domain knowledge about dependencies between these predicates.

Since PSL uses continuous variables to represent the soft truth of atoms, its semantics are based on relaxations from Boolean logic. PSL uses the *Lukasiewicz* t-norm and co-norm to provide relaxations of the logical connectives AND ($\wedge$), OR ($\vee$), and NOT ($\neg$) as follows:

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

A full assignment of soft-truth values to a set of ground atoms is called an *interpretation* (I) of that set. Using the above relaxations and the logical identity $p \rightarrow q \equiv \neg \ p \vee q$, a ground instance of a rule r ($r_{\text{body}} \longrightarrow r_{\text{head}}$) is satisfied (i.e., $I(r) = 1$) when $I(r_{\text{body}}) \leq I(r_{\text{head}})$.

PSL defines a probability distribution by quantifying a *distance to satisfaction* for each grounded instance of a rule. A rule's distance to satisfaction under interpretation I is calculated as follows:

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

The distance to satisfaction for all rules are used as features in a log-linear distribution, where the weights are nonnegative. The density function for the distribution is

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

where R is the set of ground rules, w_r is the weight of rule r, p is a modeling parameter in $\{1, 2\}$, and Z is the normalization constant

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

Because each factor in this density function uses a hinge function (2) reminiscent of hinge-losses used

in classification, PSL's probability distributions are instances of *hinge-loss Markov random fields* (HL-MRFs). For HL-MRFs, inference of a *most probable explanation* (MPE), which finds the most probable interpretation given evidence (i.e., a given partial interpretation) is computed by 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, the goal of MPE inference is to output 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 [28, 29].

The PSL rule 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 using a number of different training objective and learning algorithms [12, 13, 29]. In particular, the primary methods for weight learning are voted-perceptron approximate maximum likelihood, maximum pseudo-likelihood, and large-margin estimation.

In this work, we use approximate maximum likelihood, which we review here. We seek to maximize the log-likelihood of the full data, including both the observed data and the training labels. We can do so using gradient ascent, where 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. This gradient is intractable to compute exactly because the expectation term enumerates all possible interpretations, so we approximate the expectation by the values at the MPE solution. This approximation—whose resulting algorithm can be interpreted as a form of structured perceptron—is effective in practice and has been used on various other structured models [30, 31].

3.2 PSL Model for Drug-Target Interaction

We design a PSL program using rules that capture domain knowledge about the drug-target interaction problem. Our rules model the idea that similarity among drugs may imply similar interactions with targets, and similarity among targets may imply similar interactions with drugs. We incorporate many types of similarities into a single joint probabilistic model, simultaneously reasoning about the various possible interactions.

Triad-based rules: For drug-target interaction prediction, many established methods are based on triangles or triads between drugs and targets. These triads occur between two similar targets and a drug that interacts with both of them, or two similar drugs and a target that both drugs interacts with. The hypothesis is that similar targets tend to interact with the same drug and that similar drugs tend to interact with the same target [3, 11, 20]. Figure 2 depicts the triad-based prediction of interactions for drugs and targets.

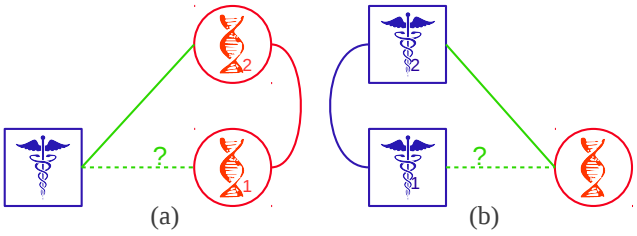


Fig. 2: Propagation via triads. Similar targets tend to interact with the same drug (a), and similar drugs tend to interact with the same target (b).

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

$$\begin{aligned} &\mathit{SimilarTarget}_\beta(T_1, T_2) \wedge \mathit{Interacts}(D, T_2) \\ &\qquad \rightarrow \mathit{Interacts}(D, T_1), \end{aligned} \tag{5}$$

$$\begin{aligned} &\mathit{SimilarDrug}_\alpha(D_1, D_2) \wedge \mathit{Interacts}(D_2, T) \\ &\qquad \rightarrow \mathit{Interacts}(D_1, T), \end{aligned} \tag{6}$$

where T denotes a target, D indicates a drug, predicate $\mathit{SimilarTarget}_\beta$ represents a specific target-target similarity metric. For each similarity metric, we add an instance of rule (5) to the PSL model. Our model is capable of integrating any set of similarities with these rules. As described in Section 6, we include three instances of this rule, where β is *sequence-based*, *PPI-network-based*, or *gene ontology-based*. Predicate $\mathit{SimilarDrug}_\alpha$ represents a specific drug-drug similarity measure. We consider five instances of rule (6), where α is *chemical-based*, *ligand-based*, *expression-based*, *side-effect-based*, or *annotation-based*.

Tetrad-based rules: In addition to triads, we also consider more complex templates for reasoning about both drug and target similarities to predict interactions. Specifically, when a drug interacts with a target, we may expect another similar drug to interact with another similar target. Figure 3 illustrates this hypothesis. We encode this hypothesis using the following *tetrad rules*:

$$\begin{aligned} &\mathit{SimilarDrug}_\alpha(D_1, D_2) \wedge \mathit{SimilarTarget}_\beta(T_1, T_2) \\ &\qquad \wedge \mathit{Interacts}(D_2, T_2) \rightarrow \mathit{Interacts}(D_1, T_1), \end{aligned} \tag{7}$$

where α and β are drug-drug or target-target similarity measures as discussed earlier. We include multiple instances of triad and tetrad rules corresponding to the three drug-drug and five target-target similarity measures.

To further enhance our model, we also experiment with an extension of the tetrad-based rules which we call *exclusive tetrad* rules. The idea behind this extension is to exclusively ground rules for the tetrad structures that do not include any triads inside them:

$$\begin{aligned} &\neg \mathit{Interacts}(D_1, T_2) \wedge \neg \mathit{Interacts}(D_2, T_1) \\ \wedge &\mathit{SimilarDrug}_\alpha(D_1, D_2) \wedge \mathit{SimilarTarget}_\beta(T_1, T_2) \\ &\wedge \mathit{Interacts}(D_2, T_2) \rightarrow \mathit{Interacts}(D_1, T_1). \end{aligned} \quad (8)$$

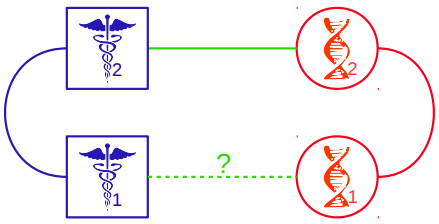


Fig. 3: Propagation via tetrads. Consider a pair of similar drugs and a pair of similar targets. If one of the drugs interacts with one of the targets, then the other drug may interact with the other target.

Negative prior: We also include a negative prior indicating that the *Interacts* predicate is most likely false, accounting for the natural sparsity in the drug-interaction network. The negative prior rule is as follows:

$$\neg \mathit{Interacts}(D, T). \quad (9)$$

The similarity predicates $\mathit{SimilarDrug}_\alpha$ and $\mathit{SimilarTarget}_\beta$ represent observed values, and the interaction predicate *Interacts* represents values that are partially observed. These rules all combine to form a complex, structured model that captures a large number of dependencies between unknown *Interacts* values that we aim to predict. In the next section, we discuss techniques to manage the high complexity of this rich model.

4 Blocking

Because PSL inference considers all possible substitutions for the rules, the number of ground rules can be extremely large. Let $|D|$ denote the number of drugs, $|\alpha|$ the number of different similarities between them, $|T|$ the number of targets, and $|\beta|$ the number of different similarities between targets. Then each potential link can be involved in $O(|D| \times |\alpha|)$ instances of rule (6), $O(|T| \times |\beta|)$ instances of rule (5). For tetrad-based rules, the situation is even worse because the number of possible substitutions is even greater. In addition, since there are $O(|D| \times |T|)$ potential interactions, the total number of ground rules is $O(|D||T|(|D||\alpha| + |T||\beta|))$. Running inference on such a massive number of ground rules is too computationally expensive for many practical settings.

To limit the number of ground rules, we prevent some of the rules from being grounded by reducing the number of triads and tetrads that are considered for each potential link. To reduce this number, we essentially ignore some of the less similar drugs and targets pairs. This strategy is reminiscent of *blocking* [32, 33, 34], which is a term that refers to the process of limiting the number of links considered. Typically, blocking decisions are done using a fast computation to fully avoid the quadratic costs inherent in link prediction settings.

There are several ways to approach blocking in our problem; the most basic strategy simply uses a fixed threshold for all similarities and sets the values below that threshold to zero. However, although the similarities are normalized to $[0, 1]$, the distribution of the values tends to be highly varied such that a fixed-threshold approach can ignore most of the values in some similarity measures or include most of the values from another. Figure 4 plots the distribution of similarities in our dataset, illustrating the diversity in the shapes of distributions these similarity measures generate.

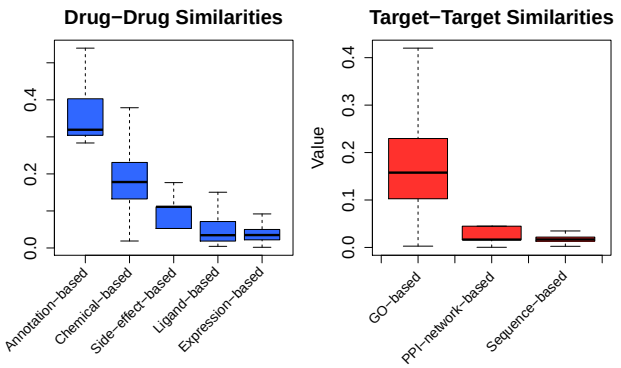


Fig. 4: Distribution variation of different similarity values between drugs and between targets. Similarities with values of zero or one are omitted in this plot.

Another method of blocking chooses a different threshold for each similarity measure. While potentially better than the previous approach, similar issues occur due to variability in individual target and drug similarity distributions. Similarities for each target or for each drug can have highly variable values, and choosing a fixed threshold will include too many similarities for some particular drugs or targets and very few for others. Figure 5 shows the annotation-based similarity for drugs and demonstrates an instance of this situation.

Instead, our proposed approach uses k-nearest-neighbors to ensure that every drug and every target considers at least a few values from each similarity. 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, the k-th nearest

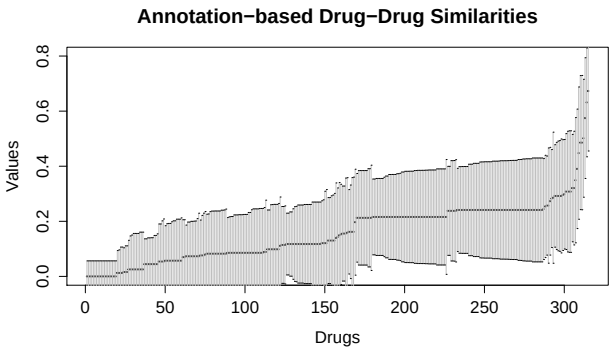


Fig. 5: Distribution of annotation-based similarity values for 315 drugs, where dots indicate mean similarity value between each drug and all others, and lines demonstrate standard deviation of the values. Similarities with values of zero or one are omitted in this plot. E.g., the mean of all annotation-based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

neighbor of a drug or target can have the same similarity as a large set of other drugs or targets. To address the possibility of ties, 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:

$$Similar_{\lambda}^{\text{blocked}} = \begin{cases} Similar_{\lambda}(x_i, x_j) & \text{if } Similar_{\lambda}(x_i, x_j) > Similar_{\lambda}(x_i, x_k); \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

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

5 Collective Inference

Traditional machine learning approaches often predict outputs independently, separating examples into distinct, unrelated instances. For example, in the drug-target interaction prediction setting, the presence or absence of each interaction is determined based on the evidence and independent of the other interactions. However, actual interactions may be interdependent.

Classification problems that consider interdependencies are known as *collective classification* problems [35]. Algorithms that perform collective classification exploit global information propagation through networks defined over the data. Since PSL performs MPE inference on the interpretation I over the whole network, interaction predictions propagate and influence the prediction of other interactions. Thus, PSL models perform collective classification and can reason about new interactions using other predicated interactions.

Specifically, rules (5) and (6) adopt a *collective inference* approach, using inferred links to imply the existence of other links, that results in global information propagation through the network. Figure 6 shows

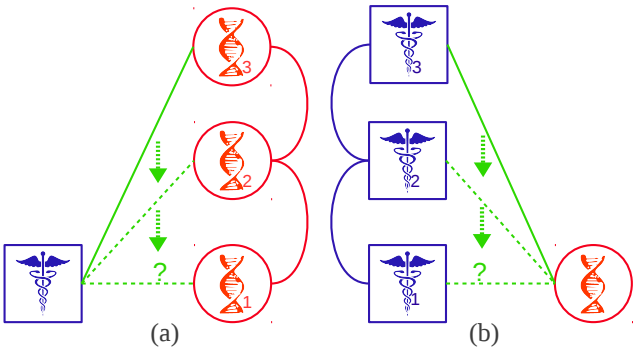


Fig. 6: Collective inference. Predicted interactions can be used for other inferences using target (a) or drug similarities (b), or both.

a situation where a predicted interaction is used in predicting other interactions.

We designed a model to experiment with the potential detrimental effect of making independent predictions. We use rules analogous to (5) and (6) that do not allow collective inference:

$$SimilarTarget_{\beta}(T_1, T_2) \wedge ObservedInteracts(D, T_2) \rightarrow Interacts(D, T_1), \tag{11}$$

$$SimilarDrug_{\alpha}(D_1, D_2) \wedge ObservedInteracts(D_2, T) \rightarrow Interacts(D_1, T). \tag{12}$$

We ground *ObservedInteracts* with the observed interactions from the dataset and use predicate *Interacts* for predictions. In contrast to rules (5) and (6), here only observed interactions imply the presence of new interactions, whereas in our full joint model, inferred interactions can imply other interactions. Hence, predictions using these new rules are only made based on observed evidence. We report the comparative results of the collective and non-collective models in the experimental analysis section.

6 Experimental Analysis

We perform an extensive evaluation of our PSL based method on a dataset that was obtained from Perlman et al. [8]. We first report on the behavior of the PSL based method[3] for different configurations as follows:

- **Rule structure:** We first compare the effectiveness of triad-based (5 & 6) and tetrad-based rules (7). This study serves as an example to test different domain assumptions for this task.
- **Blocking:** We show the effectiveness of our proposed blocking strategy by showing the speedup and performance stability of our blocking method.
- **Weight learning:** We measure the effect of weight learning on performance by comparing models with and without weight learning.

3. Our code and data we used for our experiments along with our implementation of the approach from Perlman et al. [8] can be obtained from: https://github.com/shobeir/fakhraei_tcbb2014

- **Collective inference:** We show the strength of models with collective inference, by comparing our collective model versus the non-collective version of our model.
- **Combining similarities:** Finally, we measure the effectiveness of PSL for combining information from different similarities. In this study, we compare models with all similarities against models using a single similarity.

We then compare the (best) performance of our method against the method of Perlman et al. [8].

6.1 Dataset

We obtained our dataset from Perlman et al. [8]. The interactions between drugs and targets are obtained from DrugBank [36], KEGG Drug [37], DCDB (Drug Combination database) [38], and Matador [39]. We also use the same five drug-drug and three target-target similarities used by Perlman et al. [8].

We filtered the dataset to remove the drugs and targets that do not have any computed similarities. The final dataset includes 315 drugs, 250 targets, and 1,306 interactions.

Using NodeXL [40], we calculate graph statistics and visualize the graph. The graph contains 16 connected components, and the largest component includes 518 vertices and 1280 edges. The average geodesic distance[4] in the graph is 5.31 with a maximum of 15. The vertices' degrees range from 1 to 37, with average of 4.6. Figure 7 shows an overall visualization of the drug-target interactions in the dataset, where drugs are drawn as blue squares and targets as red circles.

This section includes a brief description of the methods used for similarity calculation and how they were computed by Perlman et al. [8]. Drug-drug similarities include the following:

Chemical-based: Using the chemical development kit (CDK) [41], Perlman et al. [8] computed the hashed fingerprint of each drug based on the canonical SMILES[5] obtained from Drugbank. Considering each fingerprint as a set of elements, they computed the Jaccard similarity of the fingerprints. The Jaccard similarity score between two sets X and Y is

$$\text{Jaccard}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Ligand-based: Drugs' canonical SMILES obtained from Drugbank are compared against a collection of ligand[6] sets using the similarity ensemble approach (SEA) search tool [4]. A list of relevant protein-receptor families are obtained for each drug, and they computed Jaccard similarity between the corresponding sets of receptor families for each drug pair.

Expression-based: The Spearman rank correlation coefficient of gene expression responses to drugs retrieved from the Connectivity Map Project [16, 17] are used as a similarity measure between drugs. The Spearman rank correlation coefficient between two sets X and Y is calculated as

$$\text{Spearman}(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where x_i and y_i are ranked elements of X and Y.

Side-effect-based: Similarities between drugs are calculated using the Jaccard score between their common side-effects obtained from SIDER [42].

Annotation-based: Drugs' ATC codes are obtained from DrugBank and matched against the World Health Organization ATC classification system [43], where drugs are categorized based on different characteristics. They calculated the similarities using the semantic similarity algorithm of Resnik [44].

Target-target similarities include the following:

Sequence-based: Perlman et al. [8] compute sequence-based similarities using the Smith-Waterman sequence alignment scores, normalized via the method suggested in [23]—which divides the pairwise score by the geometric mean of the alignment scores of each sequence against itself.

Protein-protein interaction network-based: Using an all-pairs shortest path algorithm, they calculated the distance between pairs of genes using their corresponding proteins in the human protein-protein interactions network.

Gene Ontology-based: Using the method of Resnik [44], they calculated the semantic similarity measure between Gene Ontology annotations, downloaded from UniProt [45].

Perlman et al. [8] provide more detailed descriptions of these similarities.

6.2 Evaluation Criteria

We use ten-fold cross validation, where each fold randomly leaves out 10% of the positive and negative (unknown) interactions for testing. We infer interactions and compare against the held-out interactions, measuring performance using the area under the ROC curve (*AUC*), area under the precision-recall curve (*AUPR*) of the positive class, and the precision of the top n predictions (*P@n*) where $n = 130$ (i.e., the number of positive links held out in each fold) for our evaluations.

AUC is the most commonly reported measure in our related publications and it allows us to compare against the published results of other methods [23, 8, 11] on the same dataset. Lichtnwalter and Chawla [46] discuss different link prediction evaluation methods.

ROC curves are created by plotting the true positive rate versus the false positive rate at various thresholds [47]. Precision-recall (PR) curves are created by plotting the precision (or positive predictive value) versus

4. The number of edges in a shortest path between two vertices.
5. Simplified Molecular Input Line Entry Specification
6. A substance that binds with a biomolecule to serve a biological purpose.

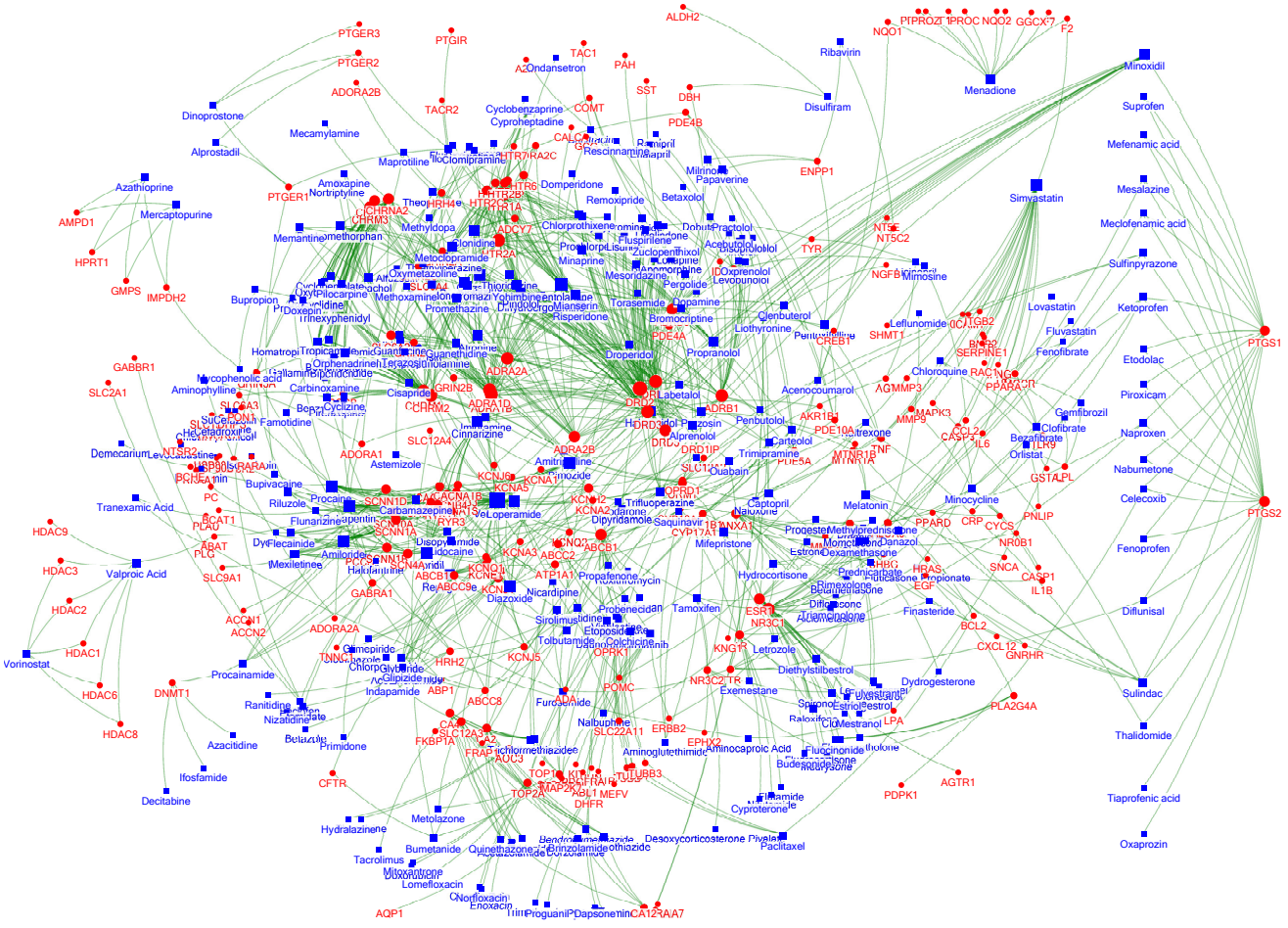


Fig. 7: Network of drug-target interactions in the dataset. Drugs are shown with blue squares and targets with red circles, where size of the node represent their degree. Similarities are not shown to simplify the graph.

the recall (or true positive rate) at various thresholds. ROC and PR curves are visually different but they are highly correlated [48], and PR curves are more informative in settings with heavy class imbalance, such as link prediction [46].

Due to this high class imbalance (130 positive to 7,744 negative examples), AUC changes are subtle. We also report AUPR performance and precision of the top 130 predictions which can highlight the importance of each model modification more clearly. This metric is of importance in practice as well, since only the top portion of the predicted interactions are typically actionable for domain experts to further evaluate.

To evaluate our model's performance with cross-validation we used the common method of random sampling of the interactions for hold-outs in each fold. However, we have to assign a value to all the possible grounding to perform weight learning on our PSL model. In order to avoid assigning an arbitrary value to the held-out interactions in each fold we add a dummy predicate (*IgnoredInteracts*) to the rules only for weight learning. We can avoid grounding of the rules for the held-out interactions using this predicate. For example, we change rule (5) to the following form:

$$\begin{aligned} &\neg \mathit{IgnoredInteracts}(D, T_2) \\ &\wedge \neg \mathit{IgnoredInteracts}(D, T_1) \\ &\wedge \mathit{SimilarTarget}_\beta(T_1, T_2) \wedge \mathit{Interacts}(D, T_2) \\ &\rightarrow \mathit{Interacts}(D, T_1). \end{aligned} \tag{13}$$

Although this change negatively affects the performance of our model in the cross-validation setting, we believe it provides an unbiased evaluation and avoids an arbitrary assignment to the held-out variables. This is only an artifact of the cross validation evaluation setting.

6.3 Analysis Results

We report the results of our five experimental analysis in this section.

Rule Structure: We first study the effectiveness of each assumption for predictions. We compare the rules based on triads (5 & 6) and the rules based on tetrads (7). We compare four different settings: the model with only the triad-based rules, the model with

only the tetrad-based rules, model with both set of rules, and model with triads and exclusive tetrads. We set the blocking parameter (k) to 5 in all models to control the growth of tetrad-based rules, and we learn the weights using separate held-out set of interactions (equal to the size of cross-validation hold-outs) in each fold.

As Table 1 shows, the rules inspired by triads are more predictive of the interactions compared to the rules that are based on tetrads. It may be the case that, in a collective setting, triad-based rules capture the effect of tetrad-based rules and perform the best. This experiment not only provides insight into the behavior of prediction using triads and tetrads in this domain, but it also demonstrates how we can easily test such assumptions using PSL's flexibility. One can easily generalize this to quickly evaluate different hypothesis about interactions.

TABLE 1: Comparison of triad-based and tetrad-based rules with $k = 5$

Rules	AUC	AUPR	P@130
Triad-based only	0.920±0.016	**0.617±0.048**	**0.616±0.035**
Tetrad-based only	0.775±0.023	0.188±0.029	0.250±0.033
Triad & tetrad	0.909±0.015	0.416±0.047	0.443±0.025
Triad & excl. tetrad	**0.924±0.013**	0.560±0.048	0.588±0.036

Blocking speedup: Next, we study the effect of blocking on performance by varying the number of neighbors (k) and measuring the effect on the PSL model based on the triad rules (5 & 6). We measure the completion time in two settings: first, a setup where we only perform inference, and, second, a setting where we run weight learning and inference. Table 2 lists the average computation time of ten-fold cross-validation experiments on computers with a (2×4) 2.66 GHz Intel processor and 48GB of RAM.[7] The results show that blocking causes significant improvement in processing time.

Our proposed blocking achieves this speed-up with no significant performance loss. The columns of Table 3 lists the performance as we change the blocking aggressiveness. The insignificant performance change as we block more aggressively suggests that, even with limited number of similarities (i.e., $k = 5$), PSL can produce accurate predictions. AUPR results in Table 3 (Inference + W. learning) show that blocking sometimes even helps performance and suggests that the similarities with higher values are most predictive of interactions. These results suggest that, in the drug-target interaction domain, models that rely on sparse similarities with high value are often more predictive than the ones that include many similarities with low values. Perlman et al. [8] report relatively similar findings with their own model.

7. We used machines with slightly different specifications and under different loads, so the reported times are approximate.

TABLE 2: Blocking speedup with triad-based rules

Condition	Time to Complete		
	k=5	k=15	k=30
Inference only	14mins	3h	9.5h
Inference + Weight learning	30mins	6h	22h

Weight learning: We study the effect of weight learning by running experiments under two conditions: with all weights set to 5 (arbitrarily hand-tuned) and with weights being learned from a set of observed interactions. Table 3 lists the performance improvement of the models with weight learning with different k values for blocking.

TABLE 3: Performance variations under the effect of weight learning with triad-based rules

Condition	AUC		
	k=5	k=15	k=30
Inference only	0.917±0.017	**0.933±0.014**	0.928±0.016
Inference + W. learning	0.920±0.016	0.931±0.016	0.924±0.019
Condition	**AUPR**		
	k=5	k=15	k=30
Inference only	0.563±0.047	0.578±0.067	0.504±0.061
Inference + W. learning	**0.617±0.048**	0.579±0.062	0.486±0.063
Condition	**P@130**		
	k=5	k=15	k=30
Inference only	0.580±0.042	0.585±0.045	0.532±0.051
Inference + W. learning	**0.616±0.035**	0.594±0.039	0.515±0.037

It is also notable that the performance improvement caused by weight learning is more significant than increasing the number of unblocked similarities used as evidence. Although weight learning improves the results, AUC changes are subtle and AUPR and the precision at the top predictions show the improvement more clearly.

Weight learning performance improvement in AUPR and AUC (for $k = 5$) is statistically significant ($p < 0.005$).[8] Figure 8 plots the average precision of the top 130 interaction predictions (i.e., P@n) over all ten folds with and without weight learning. It demonstrates how weight learning improves the precision of the predictions, providing steady improvement for the top 130 predictions.

Figure 9 illustrates the average relative weights assigned by PSL to triad-based rules for each similarity. We normalize the weights by dividing their value by the learned prior weight, so the resulting

8. We performed paired one-tailed t-test on the corresponding values of the ten folds.

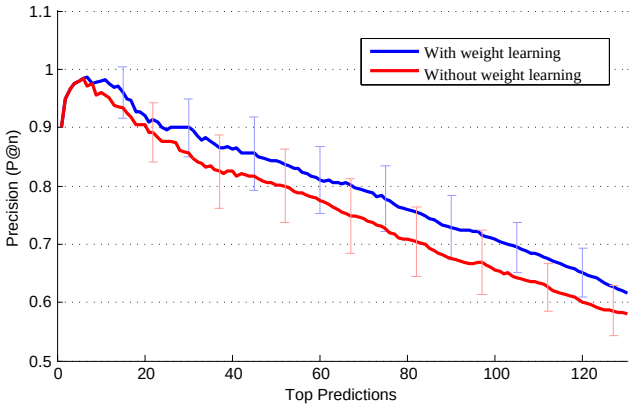


Fig. 8: Average precision of the top 130 interaction predictions for all 10 folds with $k = 5$.

quantity represents how much more heavily the rule is weighted than the prior. It is important to note that neither the absolute nor the relative value of the weights provide precise insight into the predictive power of the rules, since the features and predictions are dependent. Nevertheless, they provide some hints as to how the PSL model makes its joint prediction.

An example of low rule weight and high prediction performance is PPI-network-based similarity (Figure 9 and Table 4), which produces high accuracy for a single-similarity-based model, but has a low normalized weight. A more accurate method of measuring the effectiveness of each rule (and similarity) for prediction is building models with single rules (as described in the next section) and directly measuring their prediction performance. Even in models with single similarities, rule weight does not correlate with prediction power. Figure 9 also shows the rule weights of models with single similarities, which follows the previous trend.

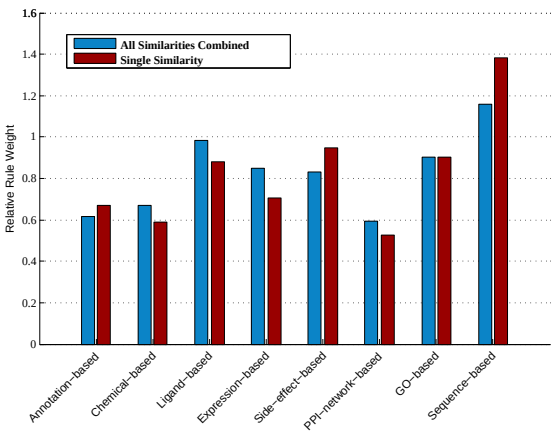


Fig. 9: Relative triad-based rule weights in models with all similarities included and models with only one similarity.

Combining similarities: We study the effect of incorporating multiple heterogeneous drug-drug and target-target similarities in PSL models. Different similarities can replace $SimilarTarget_\beta$ and $SimilarDrug_\alpha$ in rules (5) and (6) as described in previous sections. For each similarity metric, we add an instance of the rule (5) and (6) to the PSL model correspondingly. We study the situation where PSL models predict new interactions using only one drug-drug or target-target similarity versus when they are all combined and the results are shown in Table 4.

TABLE 4: Prediction based on one similarity and all similarities combined.

	Similarity	AUC	AUPR	P@130
Drugs	Annotation-based	0.660±0.017	0.224±0.026	0.319±0.026
	Chemical-based	0.670±0.023	0.234±0.042	0.289±0.032
	Ligand-based	0.713±0.023	0.270±0.035	0.337±0.037
	Expression-based	0.540±0.025	0.031±0.009	0.069±0.026
	Side-effect-based	0.631±0.016	0.209±0.032	0.271±0.023
Targets	PPI-network-based	0.781±0.021	0.389±0.047	0.480±0.041
	GO-based	0.611±0.023	0.103±0.027	0.213±0.039
	Sequence-based	0.811±0.026	0.516±0.062	0.574±0.055
All Similarities		**0.920±0.016**	**0.617±0.048**	**0.616±0.035**

Ligand-based drug-drug similarity and Sequence-based target-target similarity generate the best performance among models using a single similarity. Nevertheless, there is a significant difference between the best single similarity setting (AUC=0.811 ± 0.026 and AUPR=0.516 ± 0.062) and the all-similarities-combined setting (AUC=0.920 ± 0.016 and AUPR=0.617 ± 0.048). This study clearly shows that considering multiple similarities is critical for optimal prediction accuracy and that PSL can efficiently consider the multi-similarity nature of the problem.

Collective inference: We list the results from comparing the collective versus non-collective PSL models in Table 5. Due to high class-imbalance, AUC does not reflect the change in performance as well as the other measures.

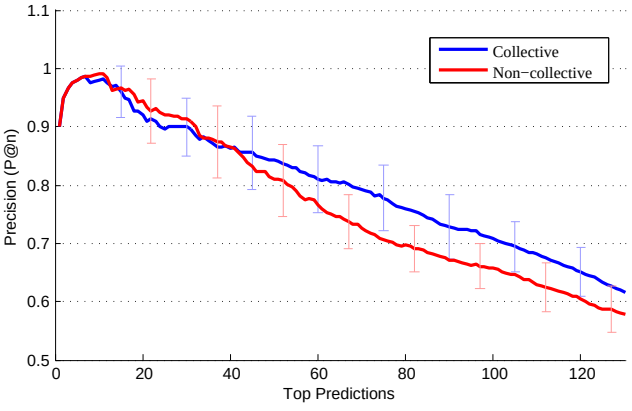
TABLE 5: Effect of collective inference

Condition	AUC	AUPR	P@130
Non-collective inference	0.916±0.016	0.556±0.039	0.577±0.039
Collective inference	**0.920±0.016**	**0.617±0.048**	**0.616±0.035**

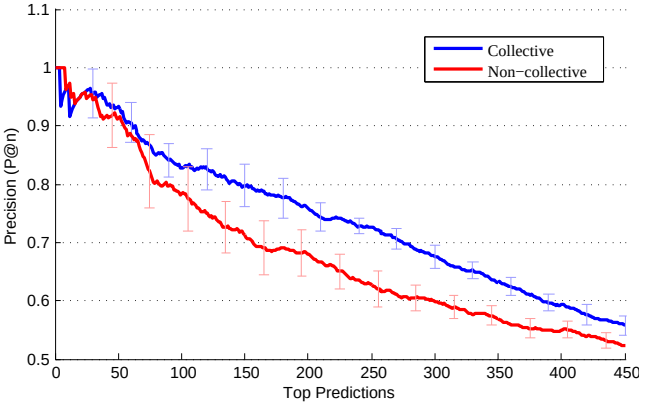
Collective inference performance improvement in AUPR and AUC is statistically significant ($p < 0.005$).[9] Figure 10(a) highlights the effect of collective model by showing the average (over ten folds) precision of

9. We performed paired one-tailed t-test on the corresponding values of the ten folds.

the top 130 predictions. Collective modeling improves the performance in this setting, as it generates a higher overall precision in the top 130 predictions. It is notable that non-collective model is only more effective for the first few predictions. This may be the results of those interactions being predicted based on direct observed evidence. However, collective inference outperforms non-collective setting for higher number of predictions. In addition, collective setting may be more effective when there are more missing links to predict. The significance of collective inference improvement is more clear in Figure 10(b) which shows the same experiment with three-fold cross validation and 450 top prediction.



(a) Top 130 prediction via ten-fold cross validation.



(b) Top 450 prediction via three-fold cross validation.

Fig. 10: Collective vs. non-collective average precision of the top predictions.

6.4 Predictions Evaluation

Finally, we can compare our results with the reported results of other state-of-the-art methods. Perlman et al. [8] report experimental evaluation on the same dataset. They report state-of-the-art performance and show that their method significantly outperforms those of Yamanishi et al. [11] and Bleakley and Yamanishi [23].

Thus, we compare our model's performance with Perlman et al. [8][10] using the same experimental setting. We report the results based on our folds that contain 130 positives and 7,744 negative links.[11] Table 6 lists the results of different PSL models comparing to the Perlman et al. [8].

TABLE 6: Comparison with Perlman's method using ten-fold cross validation

Methods	AUC	AUPR	P@130
Perlman et al. [8]	**0.937**±**0.018**	0.564±0.050	0.594±0.040
PSL triads $k = 5$	0.920±0.016	**0.617**±**0.048**	**0.616**±**0.035**
PSL triads $k = 15$ & excl. tetrads $k = 5$	**0.937**±**0.012**	0.585±0.056	**0.616**±**0.039**

PSL models with triads and blocking parameter $k = 5$ score a higher AUPR and P@n. This shows that our predictions have higher precision. Although we argue that AUC in such highly imbalanced settings is not as important, we can match the AUC of the previous state-of-the-art by using a more complicated PSL model with triads and exclusive tetrads and using two different blocking parameters of k for each set of rules. Since tetrads generate more groundings, we set k to 5 for tetrads and set k to 15 for triads. Figure 11 plots the precision of the top 130 predictions of the PSL model with triad rules and k set to 5 in comparison the predictions of Perlman et al. [8]. The results show that the PSL model with simple, triad-based rules improves the AUPR and P@n prediction performance of the state-of-the-art methods. We achieve statistically significant improvements in AUPR using the PSL triad model with $k = 5$ over the Perlman's method ($p < 0.005$), and we match the AUC performance of their method using our second PSL model with no significant difference ($p > 0.49$).[12]

New interaction predictions: Additionally, we aimed to evaluate our new interaction predictions by comparing them with new interactions that were not in our initial dataset. Using NodeXL's [40] motif clustering tool, we find targets (or drugs) that share several drugs (or targets). One example is *PTGS1* and *PTGS2*, which is shown on the right side of Figure 7.

Three unobserved interactions in this structure are *Minoxidil-PTGS2*, *Tiaprofenic acid-PTGS1*, and *Oxaprozin-PTGS1*. Our model ranks *Oxaprozin-PTGS1* and *Tiaprofenic acid-PTGS1* as the 46th and 158th most probable interactions out of the 77,444 total interactions, placing them in the top 0.2 percentile of all possible interactions. Since the time our dataset was

10. We implemented the method of Perlman et al. [8] in correspondence with the authors and reproduced their results. Our implementation of their method is also available for download.

11. The difference between our AUPR results and the ones reported in [8] is due to the down-sampling of the unobserved interactions for testing in their paper. We choose our testing folds based on the real ratio of positives to negatives

12. We performed paired one-tailed t-test on the corresponding values of the ten folds.

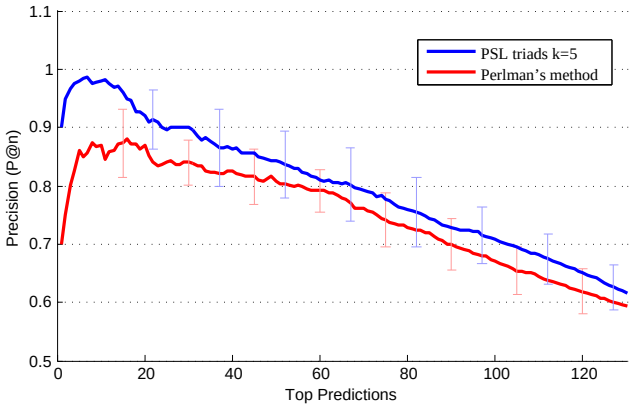


Fig. 11: Comparing Perlman's method with PSL's top 130 predictions using ten-fold cross validation.

collected by Perlman et al. [8], the online databases have been updated. Examining the latest version of Drugbank, the database now contains these two interactions.[13] Our model ranked *Minoxidil-PTGS2* significantly lower than the other two at 1,807, and we could not find any indication in the drug-target interaction dataset that this interaction exists.[14]

The two targets are shared between all three drugs in this structure, hence, only the target-target similarities were discriminative. The results show that PSL effectively uses different similarities between *Oxaprozin* and *Tiaprofenic acid*, and the other targets in the structure to rank their interactions higher than the one involving *Minoxidil*.

Furthermore, there are 197 interactions that were added to the Drugbank database since our dataset was collected. We used these newly reported interactions to further evaluate the performance of our models. We generate ten folds consisting of these new interactions and samples of the unobserved interactions, to rank these new interactions against all the other possible predictions. Table 7 lists the performance scores of our models and Perlman et al. [8] on the newly reported interactions.

TABLE 7: Comparison with Perlman's method using new interactions

Methods	AUC	AUPR	P@130
Perlman et al. [8]	0.921±0.016	0.309±0.014	0.393±0.018
PSL triads $k = 5$	0.881±0.001	0.324±0.008	0.456±0.017
PSL triads $k = 15$ & excl. tetrads $k = 5$	**0.926±0.001**	**0.344±0.018**	**0.460±0.010**

Although our more complex model demonstrates superior numbers on all performance measures, the predictions made by the PSL model with triads and $k = 5$ are more actionable due to higher precision at the top of the predictions list. That portion of the predictions are the most critical for domain experts to further evaluate. Our more complex PSL model with triads and tetrads achieves higher AUPR due to better recall. Figure 12 plots the precision of the top 150 prediction of the PSL model with triads with $k = 5$, the PSL model with triads with $k = 15$ and exclusive tetrads with $k = 5$, and Perlman et al. [8]. The simpler PSL model with triads and $k = 5$ ranks the newly reported interactions higher and performs significantly better than Perlman's method, especially beyond the top 40 predictions. Since there is potential bias in which interactions have been explored *in vitro*, these results, while encouraging, should be interpreted with discretion.

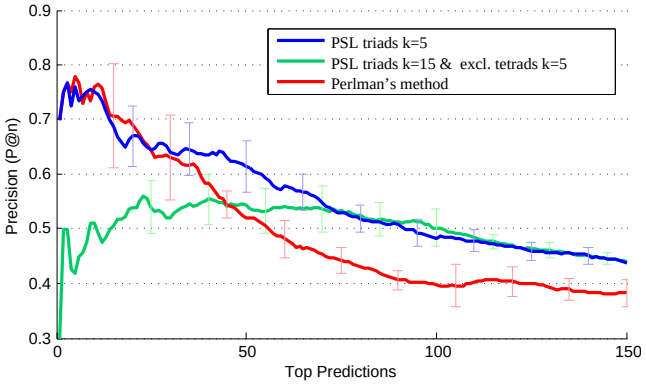


Fig. 12: Comparing Perlman's method with PSL's top 150 predictions using new interactions.

7 Discussion and Conclusion

In this paper, we propose a model using probabilistic soft logic (PSL) for drug target interaction prediction. We propose a blocking method and demonstrate how PSL enables rich, large-scale analysis of drug-target networks, combining similarities and collective inference to produce state-of-the-art prediction accuracy using an interpretable model. In our experimental evaluation, we isolate the contributions of collective inference, blocking, the combination of similarities, and weight learning to prediction quality. Our results indicate that each of these components plays a positive role, and the high accuracy and efficiency of the full PSL model results from their combination.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive reviews which helped improve the quality of this paper. We are also grateful to Stephen Bach for providing his invaluable insights on PSL. We also like to thank Alex Memory, Ben London, and Jay Pujara for their technical comments. We appreciate Eytan Ruppin's support, and Assaf Gottlieb's help with providing us with the dataset and going through our implementation of their method.

13. http://www.drugbank.ca/drugs/DB00991
http://www.drugbank.ca/drugs/DB01600
14. http://www.drugbank.ca/drugs/DB00350

This work is partially supported by the National Science Foundation (NSF) under contract numbers IIS0746930, CCF0937094, IIS1218488, and DBI1147144.

REFERENCES

[1] David J. Adams. The valley of death in anticancer drug development a reassessment. *Trends in Pharmacological Sciences*, 33(4):173 – 180, 2012.

[2] Aislyn D.W. Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3): 297, 2010.

[3] 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.

[4] Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462 (7270):175–181, November 2009.

[5] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in bioinformatics*, page bbt056, 2013.

[6] Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drug–target network. *Nature biotechnology*, 25(10):1119–1126, October 2007.

[7] Soyoung Lee, Keunwan Park, and Dongsup Kim. Building a drug–target network and its applications. *Expert Opinion on Drug Discovery*, 4(11):1177–1189, November 2009.

[8] Liat Perlman, Assaf Gottlieb, Nir Atias, Eytan Ruppin, and Roded Sharan. Combining drug and gene similarity measures for drug-target elucidation. *Journal of Computational Biology*, 18(2):133–145, February 2011.

[9] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, December 2005.

[10] Linyuan Lu and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, March 2011.

[11] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, July 2008.

[12] Matthias Broecheler, Lilyana Mihalkova, and Lise Getoor. Probabilistic similarity logic. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

[13] Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[14] Shobeir Fakhraei, Louiqa Raschid, and Lise Getoor. Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In *ACM SIGKDD 12th International Workshop on Data Mining in Bioinformatics (BIOKDD)*. ACM, 2013.

[15] Shobeir Fakhraei, Bert Huang, and Lise Getoor. Collective inference and multi-relational learning for drug-target interaction prediction. In *NIPS Workshop on Machine Learning in Computational Biology (MLCB)*, 2013.

[16] Justin Lamb. The connectivity map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54–60, January 2007.

[17] Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.

[18] Rui Chang, Robert Shoemaker, and Wei Wang. A novel knowledge-driven systems biology approach for phenotype prediction upon genetic intervention. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(5):1170–1182, 2011.

[19] Rui Chang, Robert Shoemaker, and Wei Wang. Systematic search for recipes to generate induced pluripotent stem cells. *PLoS computational biology*, 7(12):e1002300, 2011.

[20] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, May 2012.

[21] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Heuristics for chemical compound matching. *GENOME INFORMATICS SERIES*, pages 144–153, 2003.

[22] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. Drug-target interaction prediction through domain-tuned network based inference. *Bioinformatics*, 2013.

[23] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, September 2009.

[24] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2):238–245, 2013.

[25] Yuhao Wang and Jianyang Zeng. Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics*, 29(13):i126–i134, 2013.

[26] Bin Chen, Ying Ding, and David J. Wild. Assessing drug target association using semantic linked data. *PLoS Comput Biol*, 8(7):e1002574, July 2012.

[27] Assaf Gottlieb, Gideon Y. Stein, Eytan Ruppin, and Roded Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1), June 2011.

[28] Stephen H. Bach, Matthias Broecheler, Lise Getoor, and Dianne 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.

[29] Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

[30] Michael Collins. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 2002.

[31] P. Singla and P. Domingos. Entity resolution with Markov logic. In *International Conference on Data Mining*, 2006.

[32] Alvaro E. Monge and Charles Elkan. The field matching problem: Algorithms and applications. In *Proceed-*

ings of the Second International Conference on Knowledge Discovery and Data Mining, pages 267–270, 1996.

[33] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endow.*, 3(1-2):484–493, September 2010.

[34] Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, SIGMOD '09, pages 219–232. ACM, 2009.

[35] Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[36] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

[37] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl 1):D355–D360, 2010.

[38] Yanbin Liu, Bin Hu, Chengxin Fu, and Xin Chen. Dcdb: drug combination database. *Bioinformatics*, 26(4):587–588, 2010.

[39] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids research*, 36(suppl 1):D919–D922, 2008.

[40] Derek Hansen, Ben Shneiderman, and Marc A Smith. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.

[41] Christoph Steinbeck, Christian Hoppe, Stefan Kuhn, Matteo Floris, Rajarshi Guha, and Egon 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.

[42] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1), 2010.

[43] 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.

[44] Philip 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.

[45] Eric Jain, Amos Bairoch, Severine Duvaud, Isabelle Phan, Nicole Redaschi, Baris E. Suzek, Maria J. Martin, Peter McGarvey, and Elisabeth Gasteiger. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC bioinformatics*, 10(1):136, 2009.

[46] Ryan Lichtnwalter and Nitesh 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.

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

[48] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Shobeir Fakhraei is a Ph.D. student at Computer Science Department of University of Maryland College Park. He has received his M.Sc. degree in data mining and bioinformatics. He has recieved multiple awards and recognitions including outstanding graduate research assistant recognition award and General Motors academic scholarship award at Wayne State University. His research interests include machine learning, networked data, link prediction, and biomedical and health informatics.

Bert Huang Bert Huang is a postdoctoral research associate in the Department of Computer Science at the University of Maryland. He earned his Ph.D. and M.Sc. from Columbia University in 2011 for his work on efficient learning and inference using probabilistic models of graph structure, and his B.S. from Brandeis University. Bert's research investigates topics including structured prediction, statistical relational learning, and computational social science, and his papers have been published at conferences including NIPS, ICML, UAI, and AISTATS.

Louiqa Raschid Louiqa Raschid is a professor at the University of Maryland where she holds appointments in the Institute of Advanced Computer Studies, the Smith School of Business and Computer Science. She has over two decades of experience in data science, a computational paradigm to exploit data driven decision making to support a broad range of human activities. Raschid has made pioneering contributions towards meeting data integration, data management and data mining challenges in multiple non-traditional domains. Her research in data science ranges from the life sciences and health sciences to financial BIGDATA to behavior modeling in social streams to humanitarian disaster relief.

Lise Getoor Lise Getoor is a Full Professor in the Computer Science Department at the University of California, Santa Cruz. Her primary research interests are in machine learning and reasoning with uncertainty, applied to graphs and structured data. She has eight best paper awards, an NSF Career Award, was PC co-chair for the International Machine Learning Conference in 2011, and is an Association for the Advancement of Artificial Intelligence (AAAI) Fellow. She was a Professor in the Computer Science Department at the University of Maryland, College Park (2001-2013). She received her Ph.D. from Stanford University in 2001, her M.S. from UC Berkeley, and her B.S. from UC Santa Barbara.