# Collective Inference and Multi-Relational Learning for Drug–Target Interaction Prediction

**Shobeir Fakhraei, Bert Huang and Lise Getoor**
Computer Science Department, University of Maryland, College Park, MD 20742
{shobeir,bert,getoor}@cs.umd.edu

## Abstract

State-of-the-art methods for drug-target interaction prediction make use of interaction networks, drug similarities, and target similarities. In this paper we study the importance of *multi-relational* and *collective* prediction in these domains. We implement different models with *probabilistic soft logic* (PSL) to empirically show the effect of each assumption on prediction performance and demonstrate that a model using collective inference and combination of similarities significantly outperforms other models. In other words, we show the superiority of the models that combine multiple heterogeneous evidence and take advantage of the relational structure of the data.

## 1 Introduction

Drugs are microscopic organic molecules that bind to bio-molecular targets to activate or inhibit their functions. They often affect multiple targets, causing unexpected therapeutic or adverse side effects. Due to these adverse side effects, many novel therapeutic compounds fail during clinical trials. Predicting side effects during the drug developmental phase reduces 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 extremely interested in drug repositioning or repurposing, which involves finding new therapeutic effects of pre-approved drugs.

Drug-target interaction studies are important to predict both unexpected therapeutic or adverse drug side-effects. Experimental identification of all drug-target associations is labor intensive and costly. Thus, computational predictions of potential interactions are highly valuable to focus biological experiments. Several frameworks have been proposed for such discoveries including recent network-based approaches [1, 2].

Drug-target interaction networks are bipartite graphs between drugs and targets, where edges denote interactions. These graphs can be augmented with different drug-drug similarities such as chemical-structure-based similarity, and target-target similarities such as sequence-based similarity [3]. Drug-target interaction prediction corresponds to link prediction in this augmented network. More specifically, given the partially observed network (interactions and similarities) we want to predict the probability of the unobserved links (interactions).

In this paper, we study two aspects of this task using our previous prediction framework built via *probabilistic soft logic (PSL)* [4]. First, we study the *multi-relational* nature of the network with different drug-drug and target-target similarities. We compare performance of models using only a single similarity with models using multiple similarities. We empirically show that PSL [5] efficiently combines similarities to improve prediction performance.

Next, we study the *independent and identically distributed (i.i.d.)* assumption on data points (i.e., drug-target interactions). We argue that the fundamental *i.i.d.* assumption of most machine learning techniques does not hold in this domain, and empirically show that models that account for this property gain performance advantage over the ones that do not.

## 2 Our Model

We use the recently proposed PSL framework for drug-target interaction prediction [4] in our experiments. PSL uses rules written in first-order logic-like syntax as a templating language for *hinge-loss Markov random fields* over random

variables [6]. A typical PSL rule looks like the following:

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

where $P$ , $Q$ and $R$ are *predicates*, $A$, $B$, and $C$ are *variables*, and $\omega$ is a weight associated with each rule. We represent drugs and targets as variables and specify predicates to represent different similarities and interactions between them (e.g., *Interacts(D, T)*). Each predicate will be grounded with data to make a ground atom, or fact. A soft truth value will be assigned to each observed ground atom (e.g., *Similar(Acetaminophen, metformin)=0.64*).

An assignment of soft-truth values to a set of ground atoms is called an *interpretation* ($\mathcal{I}$) of that set. In this setting, a ground instance of a rule $r$ ($r_{body} \rightarrow r_{head}$) is satisfied (i.e., $\mathcal{I}(r) = 1$) when $\mathcal{I}(r_{body}) \leq \mathcal{I}(r_{head})$. Using this definition, PSL defines a distance from satisfaction for each rule ($\delta_r$). To assign truth values to unobserved atoms, PSL performs *most probable explanation (MPE)* inference give the partial interpretation. PSL defines a probability distribution over interpretations $\mathcal{I}$ by combining the weighted degree of satisfaction over all rules, as the following:

$$f(\mathcal{I}) = \frac{1}{\mathcal{Z}} \exp\left[-\sum_{r \in \mathcal{R}} \omega_r \delta_r(\mathcal{I})\right] \tag{2}$$

where $\mathcal{R}$ is a set of ground rules, $\omega_r$ is the weight of rule $r$, and $\mathcal{Z}$ is a normalization constant.

We use triad-based rules for our experimental studies [4]. The rules are motivated by the observation that similar targets tend to interact with the same drugs, and similar drugs tend to interact with the same targets [7]. Figure 1 depicts the triad-based prediction of interaction for drugs and targets.

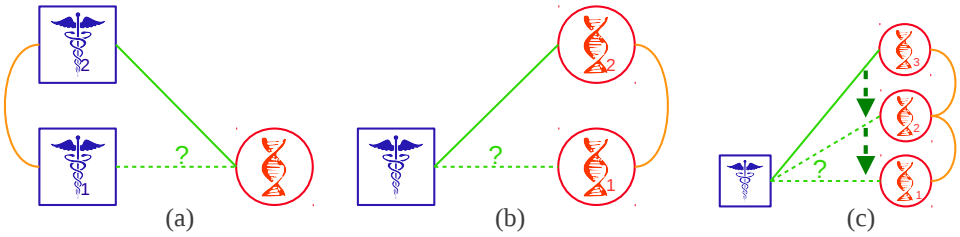


Figure 1: Similar drugs tend to interact with the same target (a), and similar targets tend to interact with the same drug (b). In a collective model information can propagate through the network (c).

We proposed the following rules to capture the triad relations of Figure 1(a) and 1(b) respectively [4]:

$$SimilarDrug_{\alpha}(D_1, D_2) \wedge Interacts(D_2, T) \rightarrow Interacts(D_1, T) \tag{3}$$

$$SimilarTarget_{\beta}(T_1, T_2) \wedge Interacts(D, T_2) \rightarrow Interacts(D, T_1) \tag{4}$$

where we use $T$ for a target, $D$ for a drug, $SimilarTarget_{\beta}$ for a specific target-target similarity, and $SimilarDrug_{\alpha}$ for a specific drug-drug similarity.

We design the following two studies based on these rule structures:

**Combining Similarities** We first study the effect of incorporating multi-relational heterogeneous information, and combining multiple drug-drug and target-target similarities. Different similarities can replace $SimilarTarget_{\beta}$ and $SimilarDrug_{\alpha}$ in the rules above as described in [4]. For each similarity metric, we add an instance of the rule (3) and (4) 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.

**Collective Inference** We then study the effect of making the simplifying *i.i.d.* assumption in this setting. The presence or absence of a drug-target interaction is often studied independently [3]. However, interactions are highly interdependent and can be predicted based on each other (e.g., in triads). Inferred probability of one interaction should affect the probability of other interactions. Rules (3) and (4) 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 1(c) shows a situation where a predicted interaction is used in predicting other interactions.

In this study we want to show the effect of such assumption by writing analogous rules that do not allow collective inference. We perform predictions using a non-collective model with the following rules:

$$SimilarDrug_{\alpha}(D_1, D_2) \wedge ObservedInteracts(D_2, T) \rightarrow Interacts(D_1, T) \tag{5}$$

$$SimilarTarget_{\beta}(T_1, T_2) \wedge ObservedInteracts(D, T_2) \rightarrow Interacts(D, T_1) \tag{6}$$

We will ground *ObservedInteracts* with the observed interactions from the dataset, and use predicate *Interacts* for predictions. In contrast to rules (3) and (4), the predictions cannot be used in the body of the rule to activate new instances of the rules. Hence, predictions are only made based on observed evidence. In other words, rules (3) and (4) make collective inference where rules (5) and (6) do not.

## 3 Experiments

Our dataset is based on a network of drugs and genetic targets, where interactions between them are obtained from the DrugBank database [4]. The dataset includes 315 drugs, 250 targets, and 1,306 interactions. We use five drug-drug and three target-target similarities. Drug-drug similarities include *Chemical-based, Ligand-based, Expression-based, Side-effect-based*, and *Annotation-based* and target-target similarities include *Sequence-based, PPI-network-based*, and *Gene Ontology-based*. Each of these similarities will have a corresponding predicate and a corresponding rule in the experiments, i.e., there are five instances of the rules (3) and (5) and three instances of rules (4) and (6).

As PSL grounds every possible rule to predict each link, the number of grounded instances of rules can be extremely large. To control such situations, we limit some of the rules from being grounded by ignoring some of the less significant similarities. We use our previous *K*-nearest-neighbors-based (with K=15) approach [4]. This is commonly referred to as *blocking* in the entity-resolution domain [8] and is a method to avoid the quadratic cost of the full problem.

We evaluate the model performances via ten-fold cross validation over the observed interactions. We measure area under the ROC curve (AUC), area under the precision-recall curve (AUPR) of the positive class, and precision of the top 130 predictions (i.e., the number held out in each fold). Due to high class imbalance (130 positive examples to 77576 negative examples), AUC changes are subtle and AUPR performance is relatively low, thus precision of the top 130 predictions highlights the importance of each model modification more clearly. In reality, only the top portion of the predicted interactions are interesting for domain experts for further evaluations.

Table 1: Prediction based on one similarity and all similarities combined.

| | SIMILARITY | AUC | AUPR | prec@130 |
|---|---|---|---|---|
| Drugs | Annotation-based | $0.788 \pm 0.022$ | $0.122 \pm 0.016$ | $0.198 \pm 0.019$ |
| | Chemical-based | $0.755 \pm 0.023$ | $0.064 \pm 0.015$ | $0.155 \pm 0.025$ |
| | Ligand-based | $0.774 \pm 0.025$ | $0.069 \pm 0.014$ | $0.151 \pm 0.027$ |
| | Expression-based | $0.606 \pm 0.024$ | $0.005 \pm 0.001$ | $0.020 \pm 0.009$ |
| | Side-effect-based | $0.726 \pm 0.015$ | $0.068 \pm 0.015$ | $0.151 \pm 0.032$ |
| Targets | PPI-network-based | $0.851 \pm 0.021$ | $0.167 \pm 0.046$ | $0.225 \pm 0.045$ |
| | GO-based | $0.678 \pm 0.029$ | $0.025 \pm 0.006$ | $0.080 \pm 0.022$ |
| | Sequence-based | $0.826 \pm 0.027$ | $0.129 \pm 0.034$ | $0.213 \pm 0.045$ |
| All Similarities Combined | | $0.931 \pm 0.018$ | $0.190 \pm 0.032$ | $0.249 \pm 0.041$ |

Table 1 shows the results from the first study, which compares using single similarities with using all similarities. Annotation-based drug-drug similarity, and PPI-network-based target-target similarities generate the best performance among models using a single similarity. There is a significant difference between the single similarity setting (AUC=$0.851 \pm 0.021$) and the all-similarities-combined setting (AUC=$0.931 \pm 0.018$). This study clearly shows that considering the multi-relational nature of the problem increases accuracy.

Figure 2 shows the average (over ten folds) precision of the top 130 predictions for the second study, which compares collective against non-collective modeling. It is clear that collective modeling significantly improves the performance in this setting, as it generates a much higher precision at various thresholds in the top 130 links.

Table 2 shows the variation of performance under collective and non-collective conditions using different measures. The significant superiority of collective inference is clear. Due to high class-imbalance, AUC does not reflect the change in performance as well as the other measures.

Table 2: Effect of collective inference

| CONDITION | AUC | AUPR | 130-PRECISION |
|---|---|---|---|
| Non-collective inference | $0.926 \pm 0.016$ | $0.058 \pm 0.009$ | $0.090 \pm 0.024$ |
| Collective inference | $0.931 \pm 0.018$ | $0.190 \pm 0.032$ | $0.249 \pm 0.041$ |

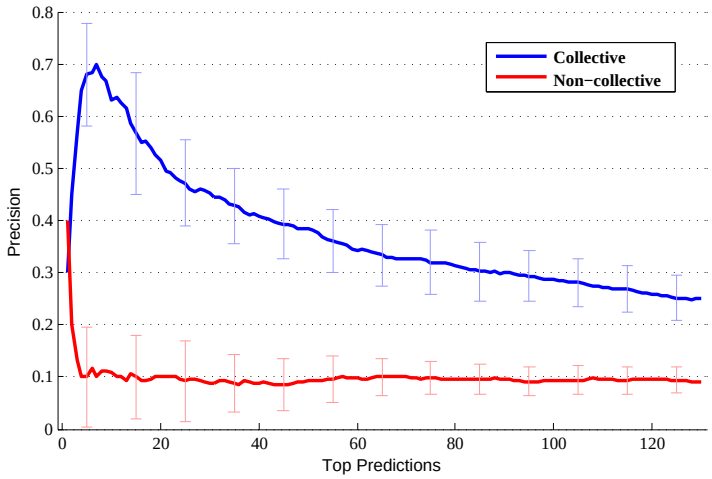


Figure 2: Collective vs. non-collective average precision of the top 130 prediction.

## 4 Conclusion

We empirically demonstrate that considering the multi-relational nature of drug-target interaction prediction improves performance. We used probabilistic soft logic (PSL) to effectively combine different similarities from heterogeneous sources to achieve better performance. We used rules (3) and (4) to study the combination of different similarities. Because these rules imply collective reasoning, predictions from one similarity can be used as evidence in rules with other similarities, which boosts the performance of the combined-similarities setting. We also showed that collective inference significantly improves prediction performance. This strongly suggests that the common *i.i.d.* assumption of most machine learning techniques is too strong for this domain, and frameworks such as PSL that enable collective inference are necessary for more accurate predictions.

### Acknowledgements

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

### References

[1] 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, Jrme 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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*, 2013.

[7] 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.

[8] Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 2007.