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

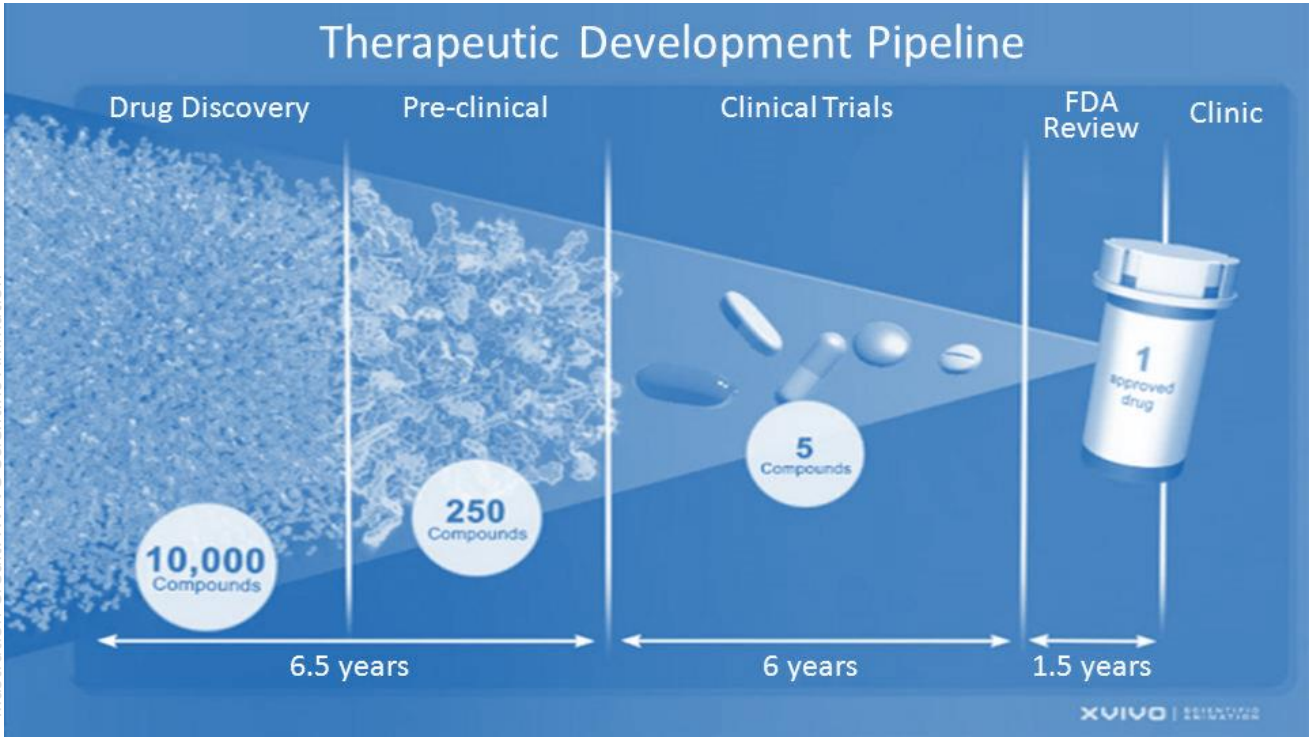
SHOBEIR FAKHRAEI*
LOUIQA RASCHID
LISE GETOOR

University of Maryland, College Park, MD, USA

# Outline

- Drug Repurposing
- Drug-Target Interaction Network
- Probabilistic Similarity Logic (PSL)
- Drug-Target Interaction Prediction with PSL
- Experimental Results

# New Drug Development

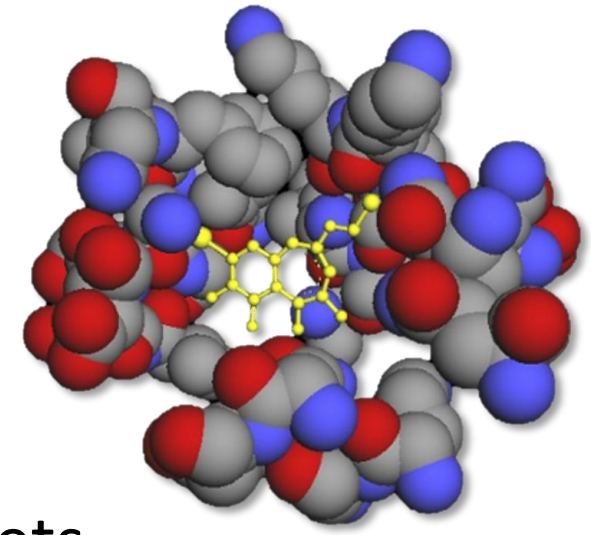



- Time Consuming: New drugs take a decade to reach market.
- Costly: Development cost reaches 2 billion US dollars.

**Valley of death:**
**Most novel drug candidates never get approved!**

# Drugs

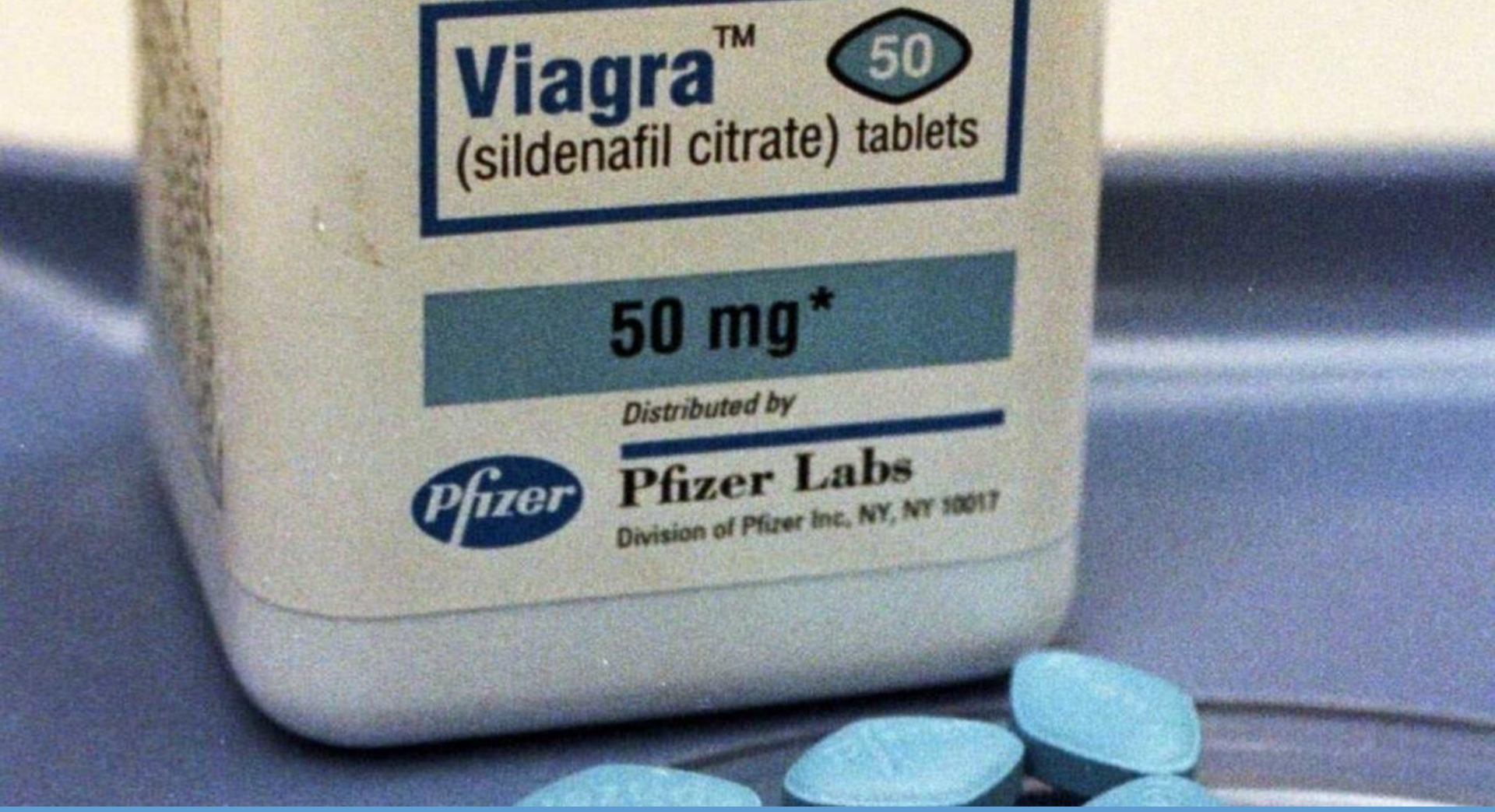
- Drugs:
  Organic small molecules that bind to bio-molecular targets to activate/inhibit their functions



- Drug often affect multiple targets.
- Poly-pharmacology is an area of growing interest

# Drug Repurposing

- Drug Affecting Multiple Targets:
  - Adverse side-effects
  - Unexpected therapeutic effect
- Drug Repurposing/Repositioning: Finding new uses for approved drugs.
- No need for tests required for a new therapeutic compound (Already approved)




**Sildenafil was originally developed for pulmonary arterial hypertension**

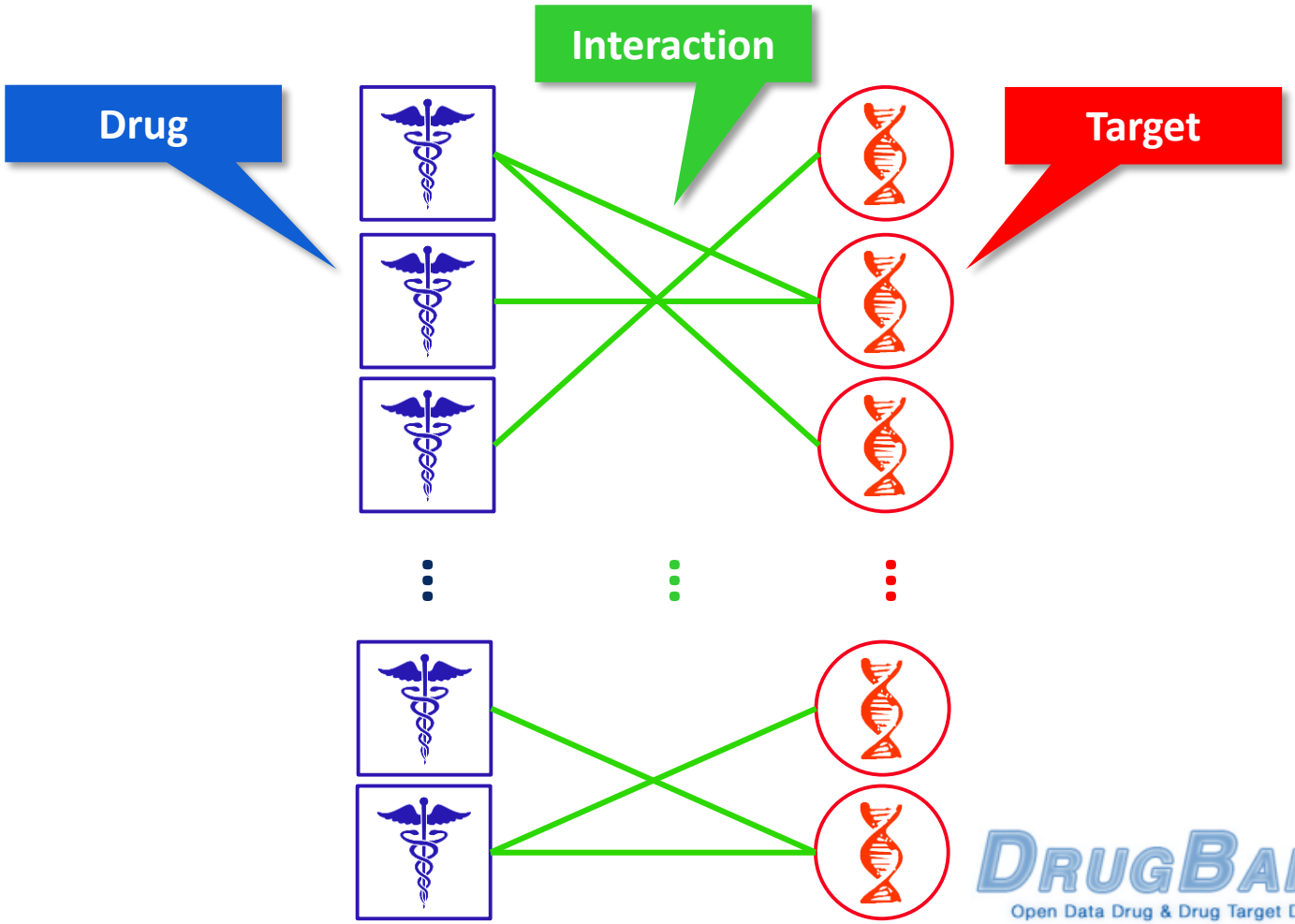
# Need for Systematic Search

- Most new treatment are discovered by chance during clinical trials.
- There is a need for a better systematic approach.
- Experimental identification of drug-target associations is labor intensive and costly

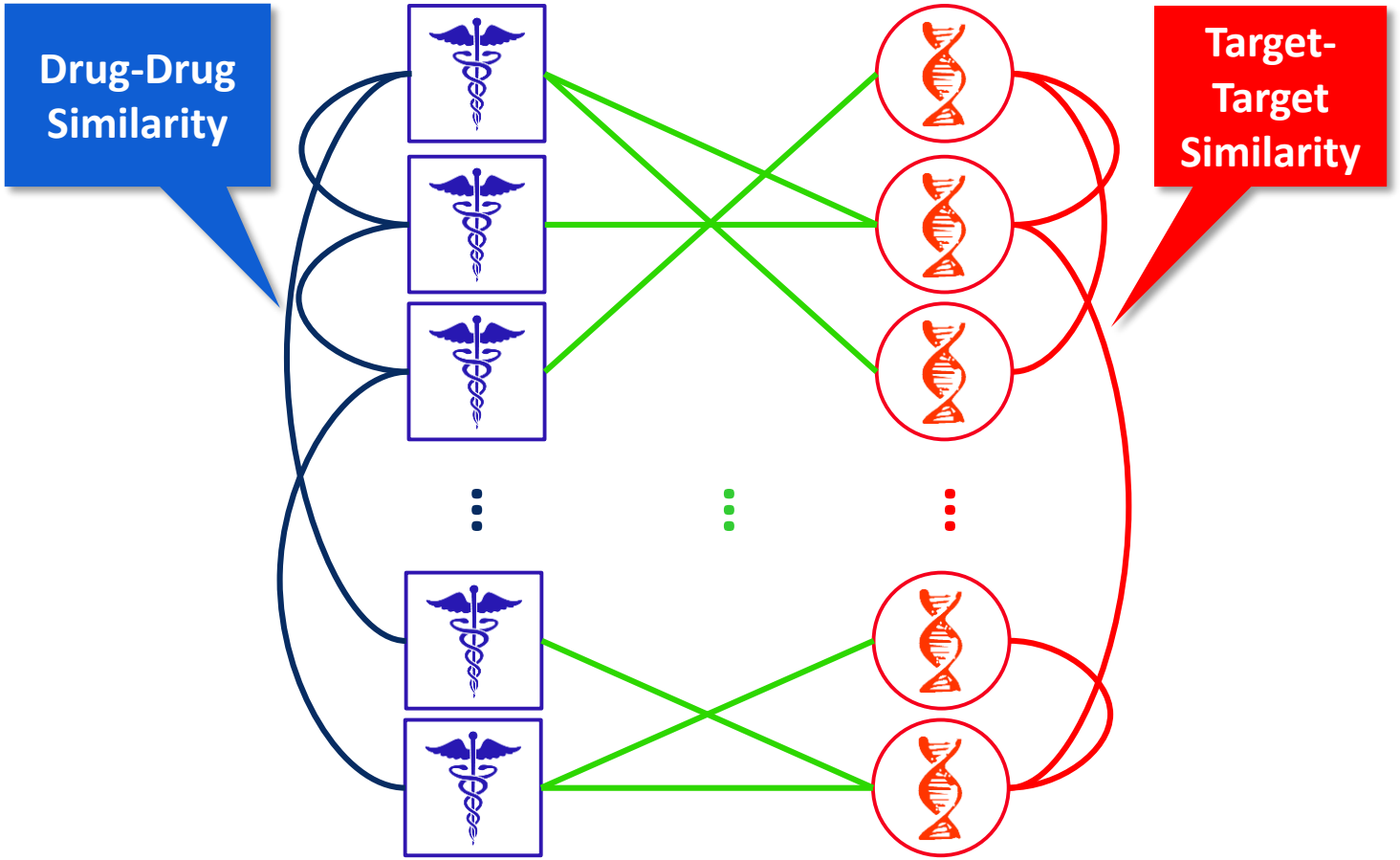
- A better solution?

# Using computational predictions to focus biological search

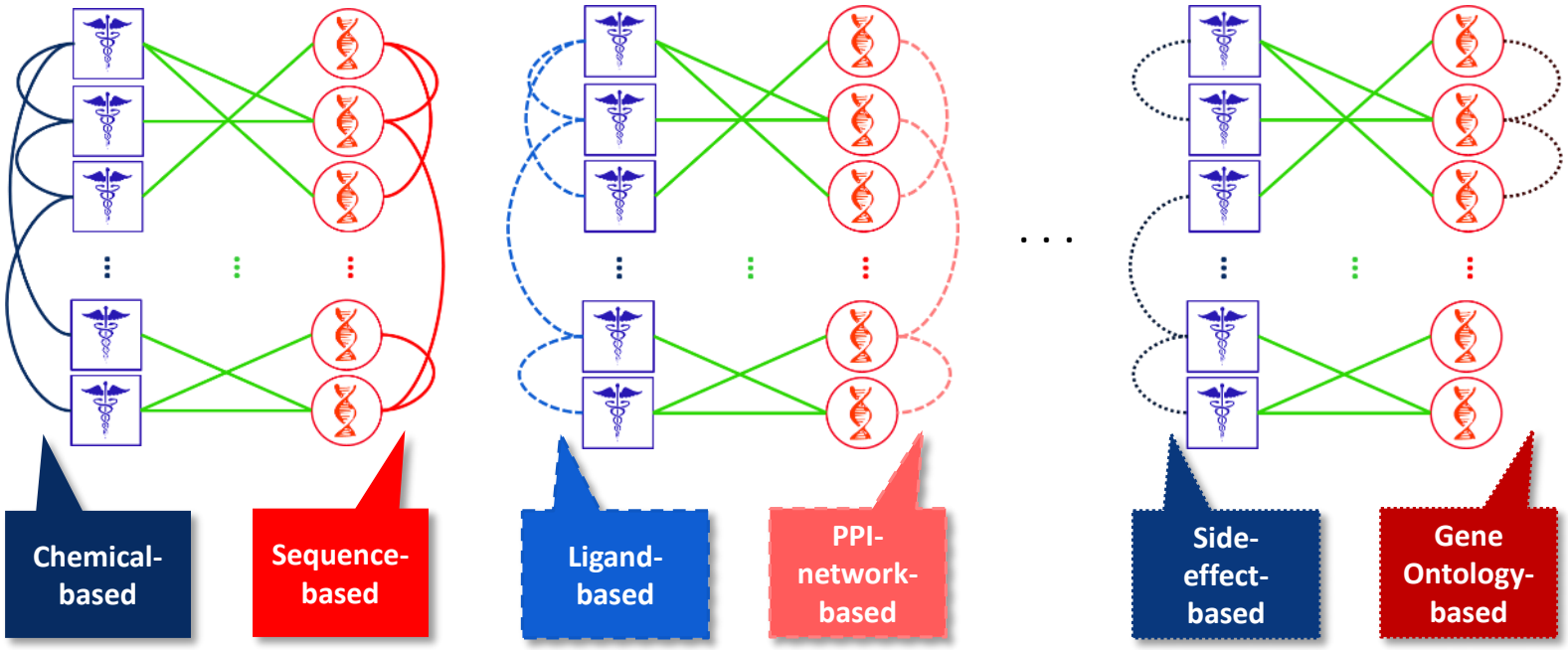
# Drug-Target Interaction Network



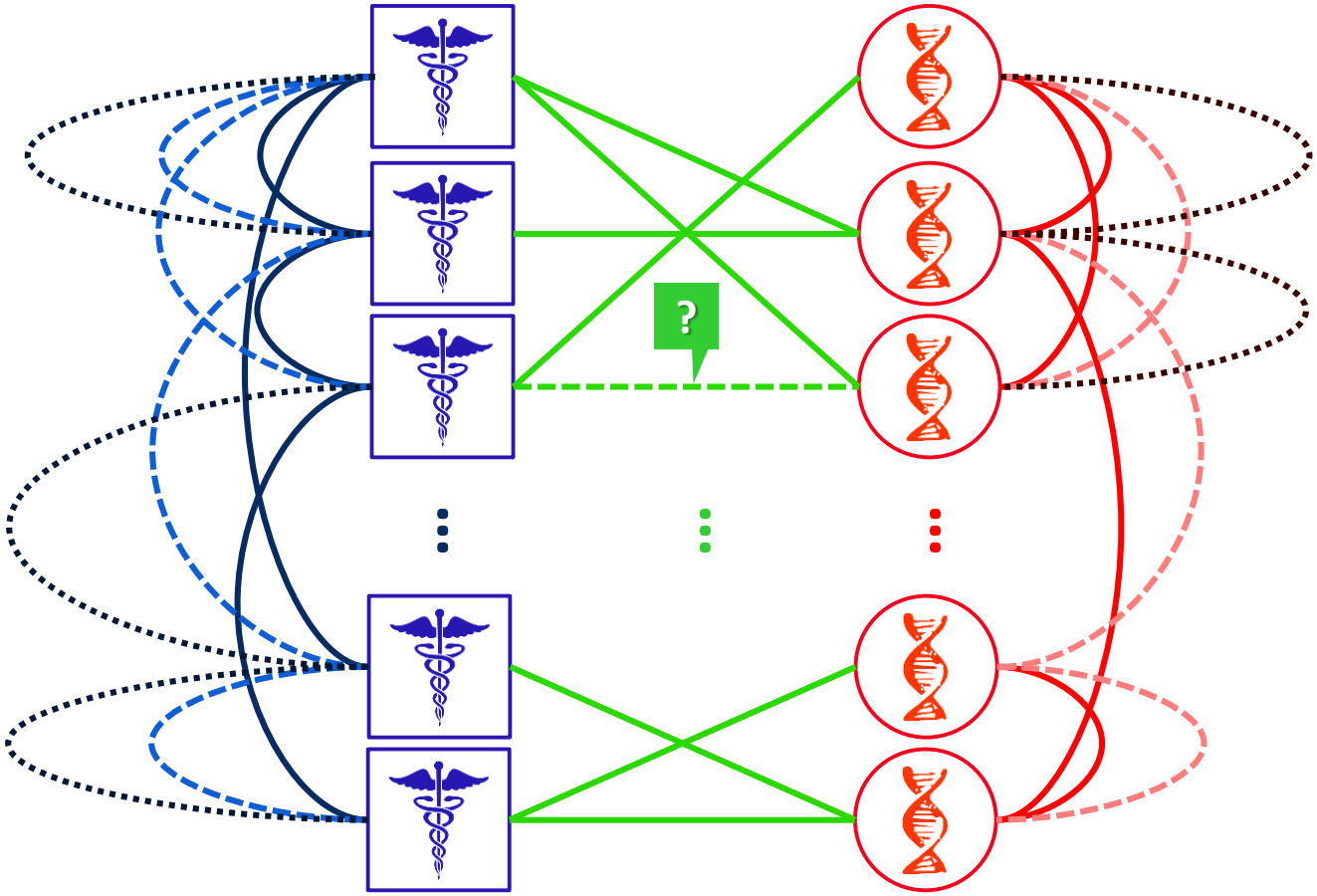
# Drug-Target Interaction Network + Similarities



# Multiple Similarities



# D-T Interaction Network + Multiple Similarities

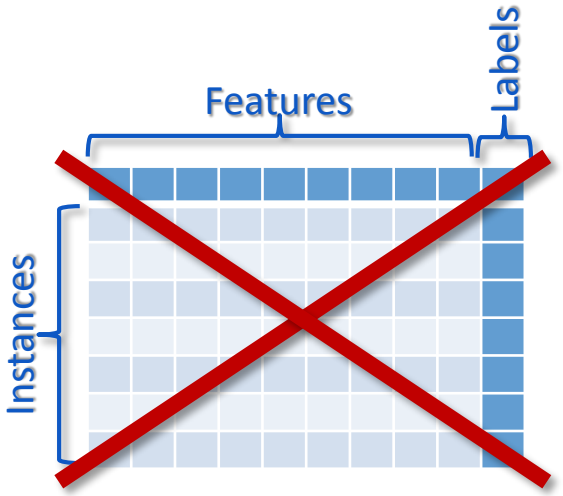


# Drug-Target Interaction Prediction

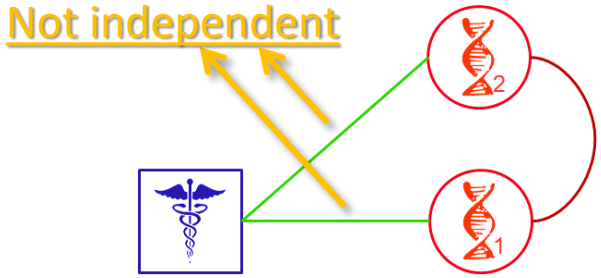

- Data:
  - Drug-target interaction network
  - Set of drug-drug similarities
  - Set of target-target similarities
- Task:
  - Link Prediction (New drug-target interactions)

# Challenges

- Data is not originally flat:
  - Classifiers need a set of features and instances.
  - Instances: all interactions in the network (pairwise)
    or only interaction of one drug or target.
  - Features: Feature engineering




- Not Independent and Identically Distributed (IID): Interactions depend on each other (a drug tends to interact with similar targets)



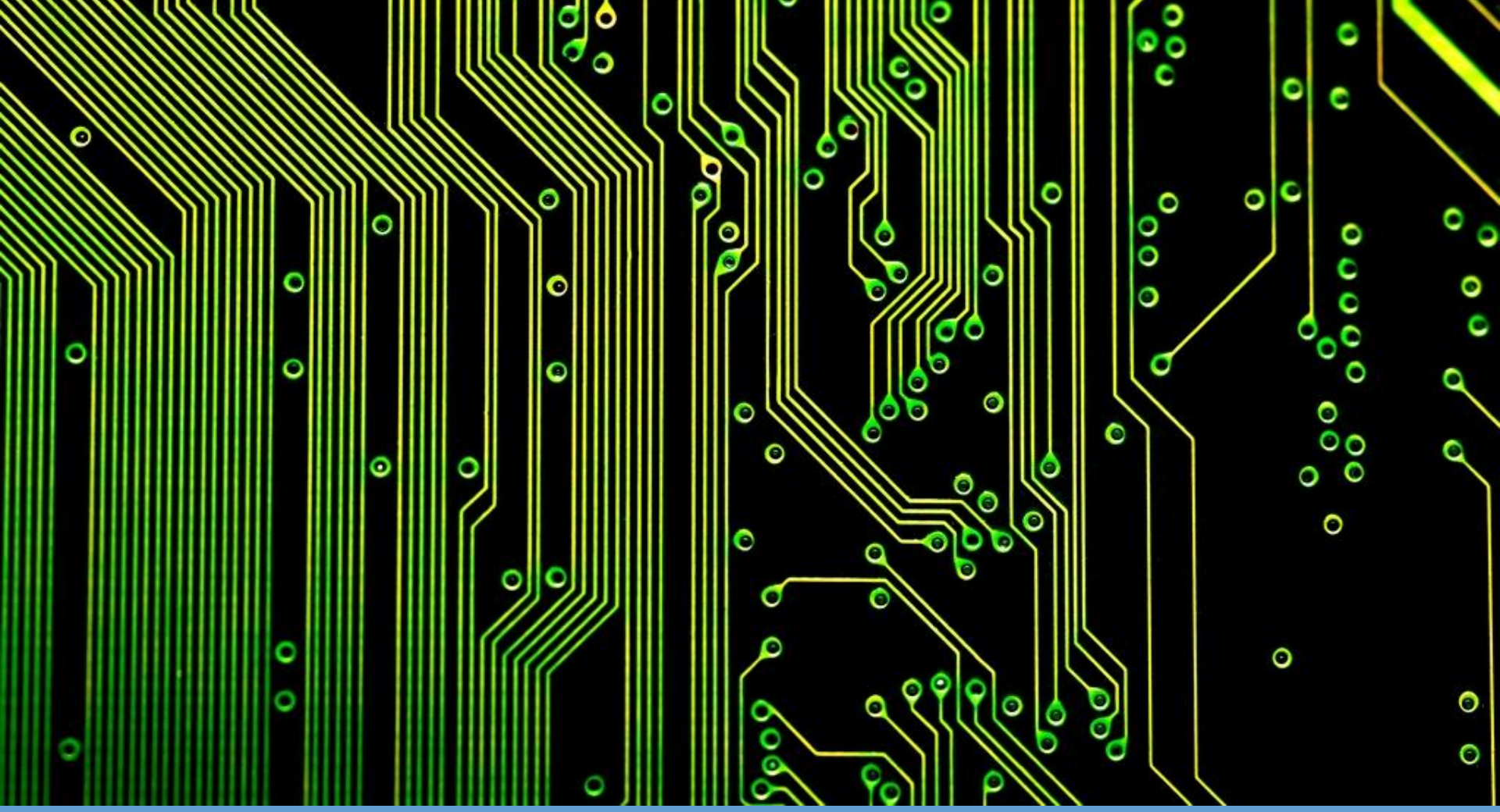

- Multi-relational:
  - Drug-Target Interactions
  - Different Drug-Drug Similarities
  - Different Target-Target Similarities


Probabilistic
Similarity
Logic

# Probabilistic Similarity Logic

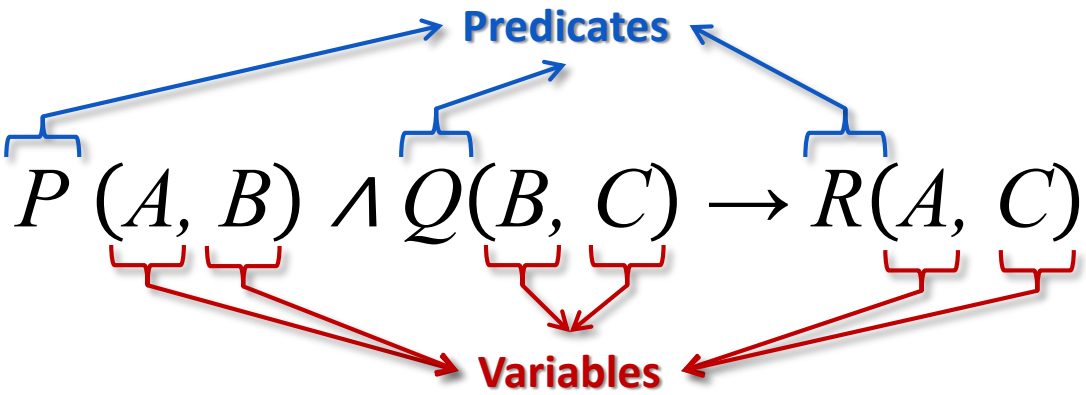
# Probabilistic Similarity Logic (PSL)

- Declarative language based on logic to express collective probabilistic inference problems.
  - Logical foundation
  - Probabilistic foundation
  - Weight Learning



# Logic Foundation

# General Rules

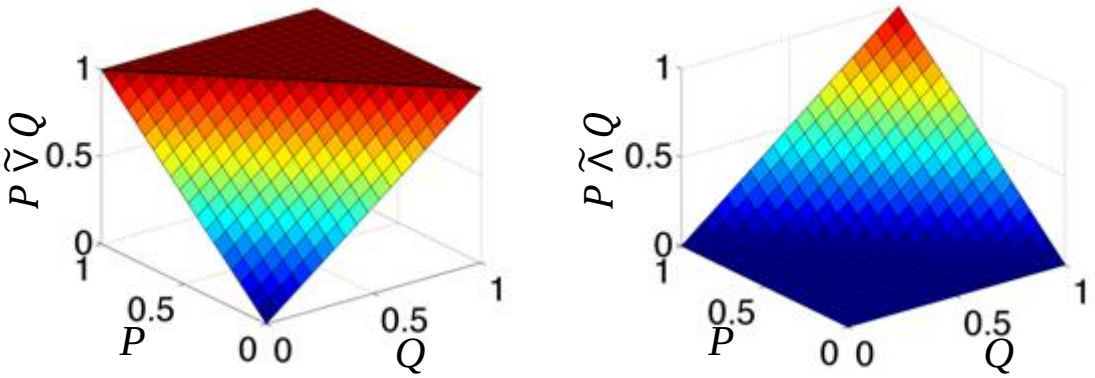



e.g., $Interacts(D, T_2) \wedge SimilarTarget(T_1, T_2) \rightarrow Interacts(D, T_1)$

- Can use predicate to define relations between variables.
  e.g. Interacts(D, T)
- Grounding: Instantiation of predicates with data.
  e.g. Interacts(acetaminophen, cox2)
- Groundings have a soft-truth values between [0, 1]

# Lukasiewicz t-norm and co-norm

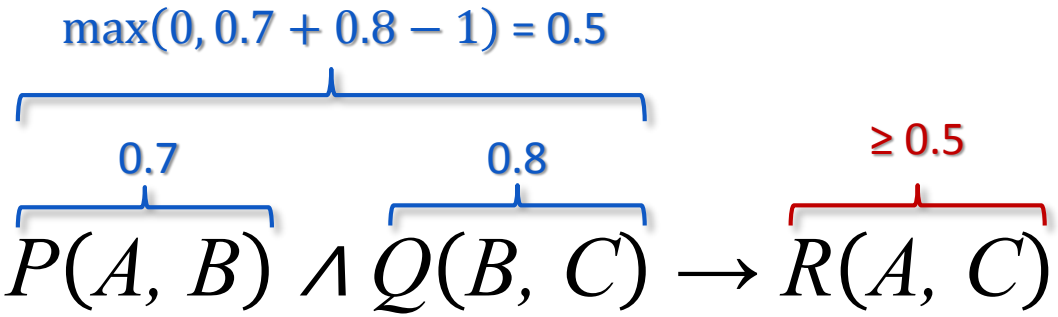
$$P(A, B) \wedge Q(B, C) \rightarrow R(A, C)$$




- $P \tilde{\wedge} Q = max(0, P + Q - 1)$
- $P \tilde{\vee} Q = min(1, P + Q)$
- $\neg P = 1 - P$

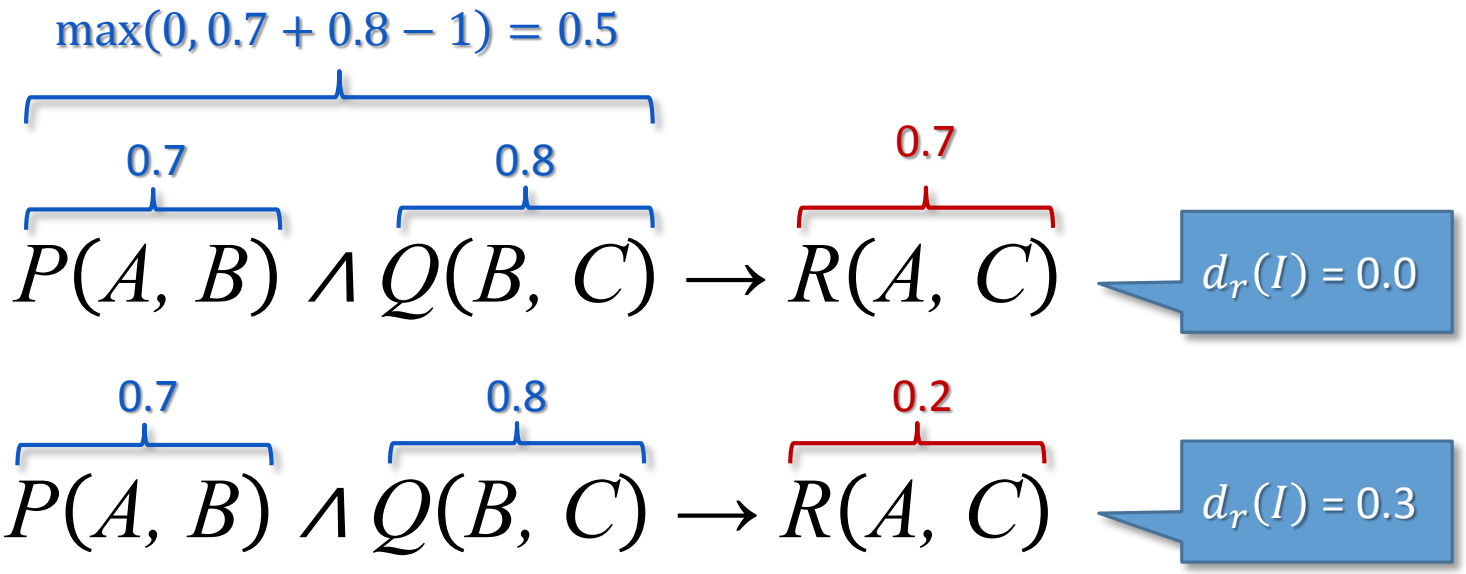
# Satisfaction

- Interpretation (I) : an assignment of soft-truth values to a set of groundings.
- Rule satisfaction: $r_{body} \rightarrow r_{head}$ is satisfied when $I(r_{body}) \leq I(r_{head})$



# Distance to Satisfaction

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



# Rule Weights

$$w : P(A, B) \wedge Q(B, C) \rightarrow R(A, C)$$

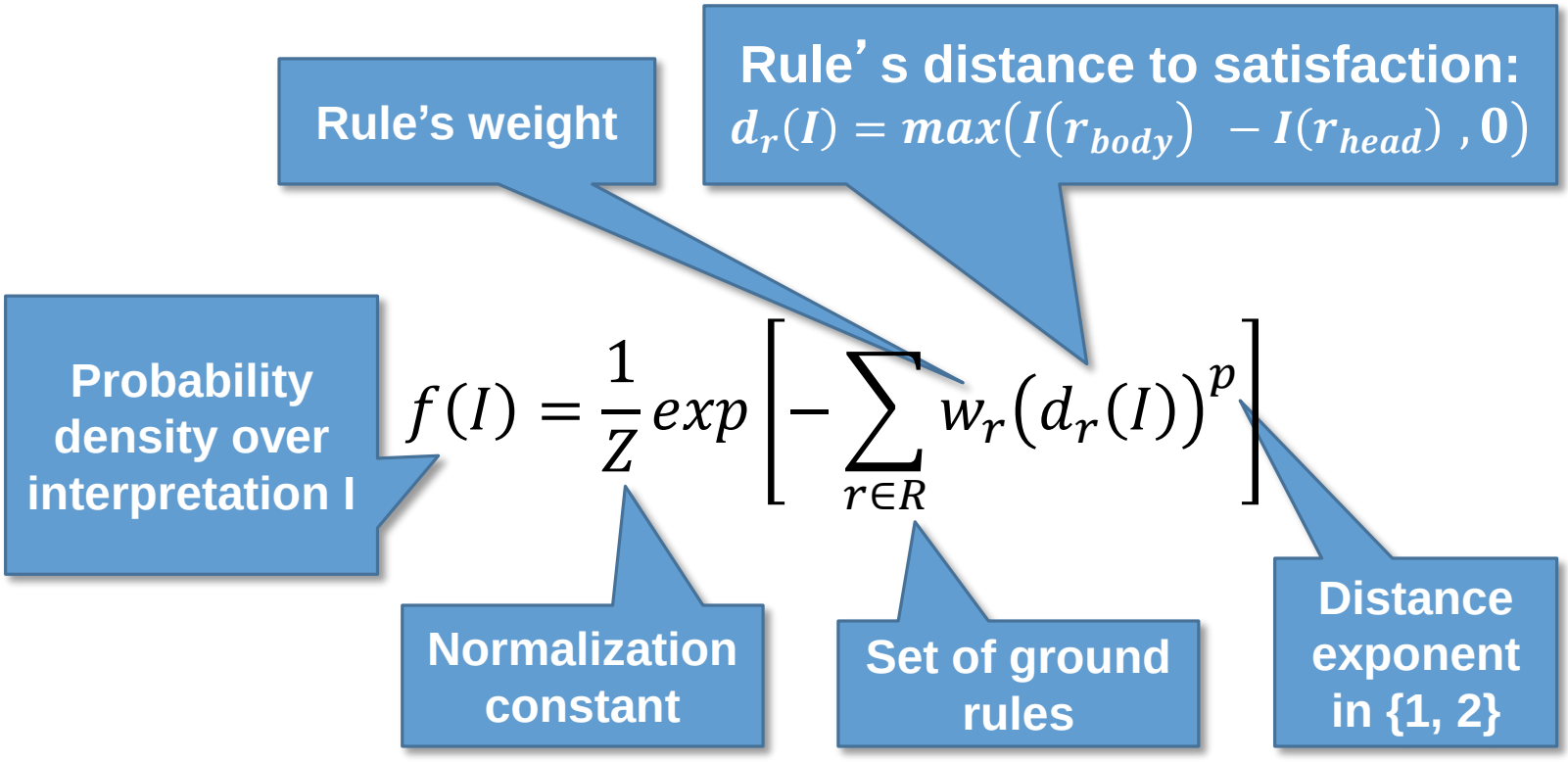
- Rule can have weights which corresponds to importance of the rule.
  - Can come from domain knowledge
  - Can be learned from data

# Review

- PSL program + Dataset → Set of ground rules
- Some groundings (predicates) have known truth values and some have unknown truth values.
- Every Interpretation of unknown groundings (predicates) → different weighted distances to satisfaction
- How to decide which Interpretation is best?

# Probabilistic Foundation

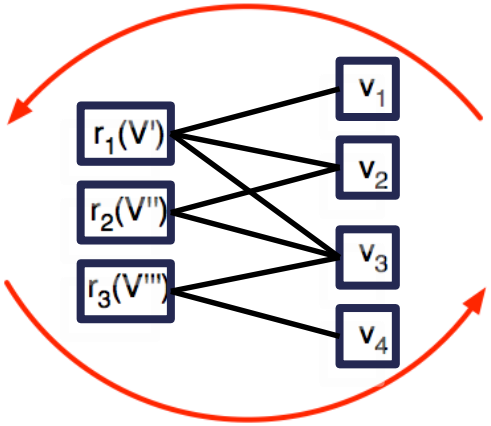
# Probabilistic Model



# Inferring Most Probable Explanations

- Given a set of observed groundings infer the values of unknown groundings
- e.g., Given a set of drug-target interactions + a set of D-D and T-T similarities infer the value of other interactions.

- Convex optimization: perform inference using the alternating direction method of multipliers (ADMM) [Bach et al., NIPS 2012]
- Fast, scalable, and straightforward
- Optimize sub-problems (ground rules) independently.

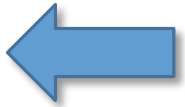



# Weight Learning

# Weight Learning

$$w : P(A, B) \wedge Q(B, C) \rightarrow R(A, C)$$

- Learn the weights from training data
- Various methods:
  - Approximate maximum likelihood [Broecheler et al., UAI 10]
  - Maximum pseudo-likelihood
  - Large-margin estimation



# PSL Summary

- Design probabilistic models using declarative language
  - Syntax based on **first-order logic**
- Inference of most-probable explanation is fast **convex optimization** (ADMM)
- Learning algorithms for training rule weights from labeled data.

# Drug-Target Interaction Prediction with PSL

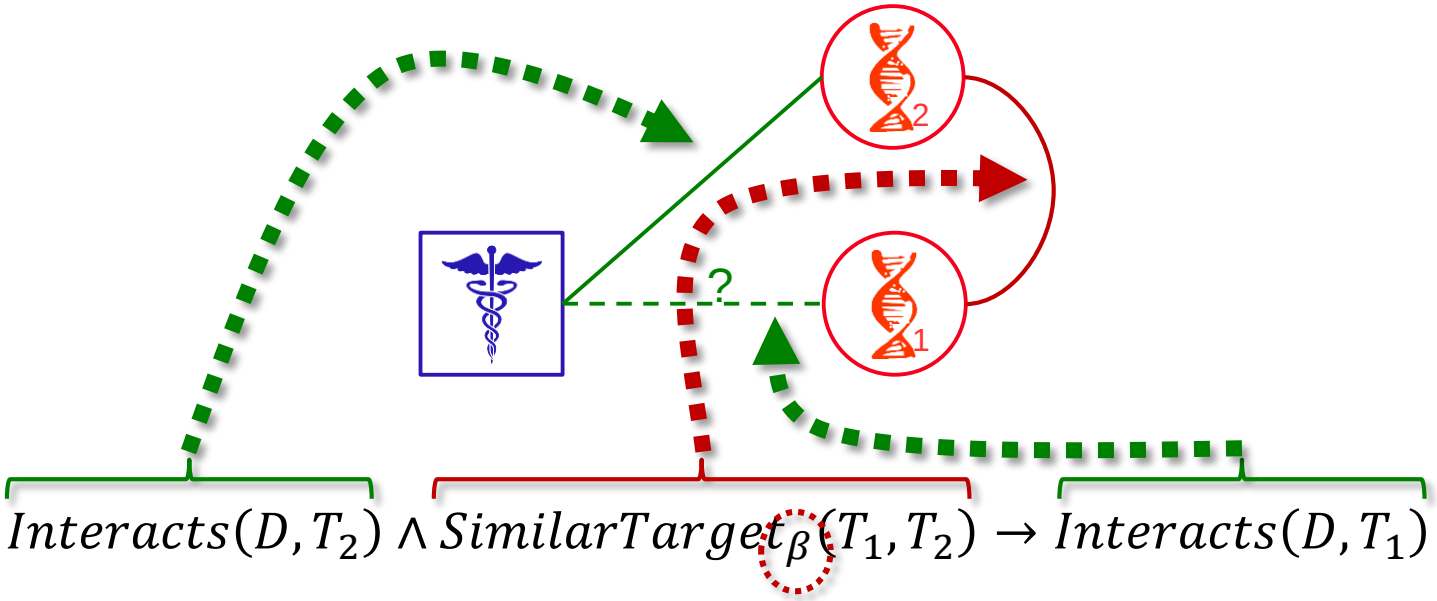
# Predicates

- $\mathrm{Interacts}(\mathrm{D}, \mathrm{T})$
- $\mathrm{SimilarTarget}_{\beta}(\mathrm{T}_1, \mathrm{T}_2)$
  - e.g. $\beta$ can be Sequence-based, PPI-network-based, Gene Ontology-based.
- $\mathrm{SimilarDrug}_{\alpha}(\mathrm{D}_1, \mathrm{D}_2)$
  - e.g. $\alpha$ can be Chemical-based, Ligand-based, Expression-based, Side-effect-based, Annotation-based.


# Drug-Target Interaction Prediction Rules

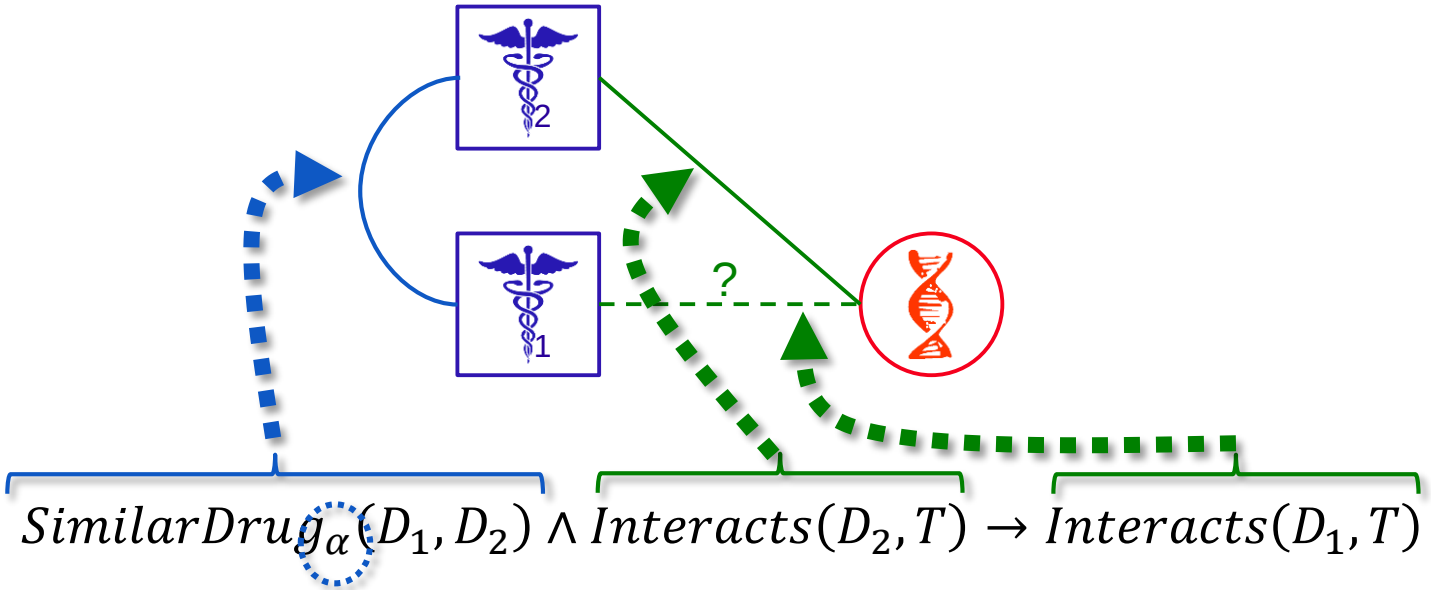
# Triad-based rules (Targets)

- Drugs tend to interact with similar targets
  (friend of friend is a friend)



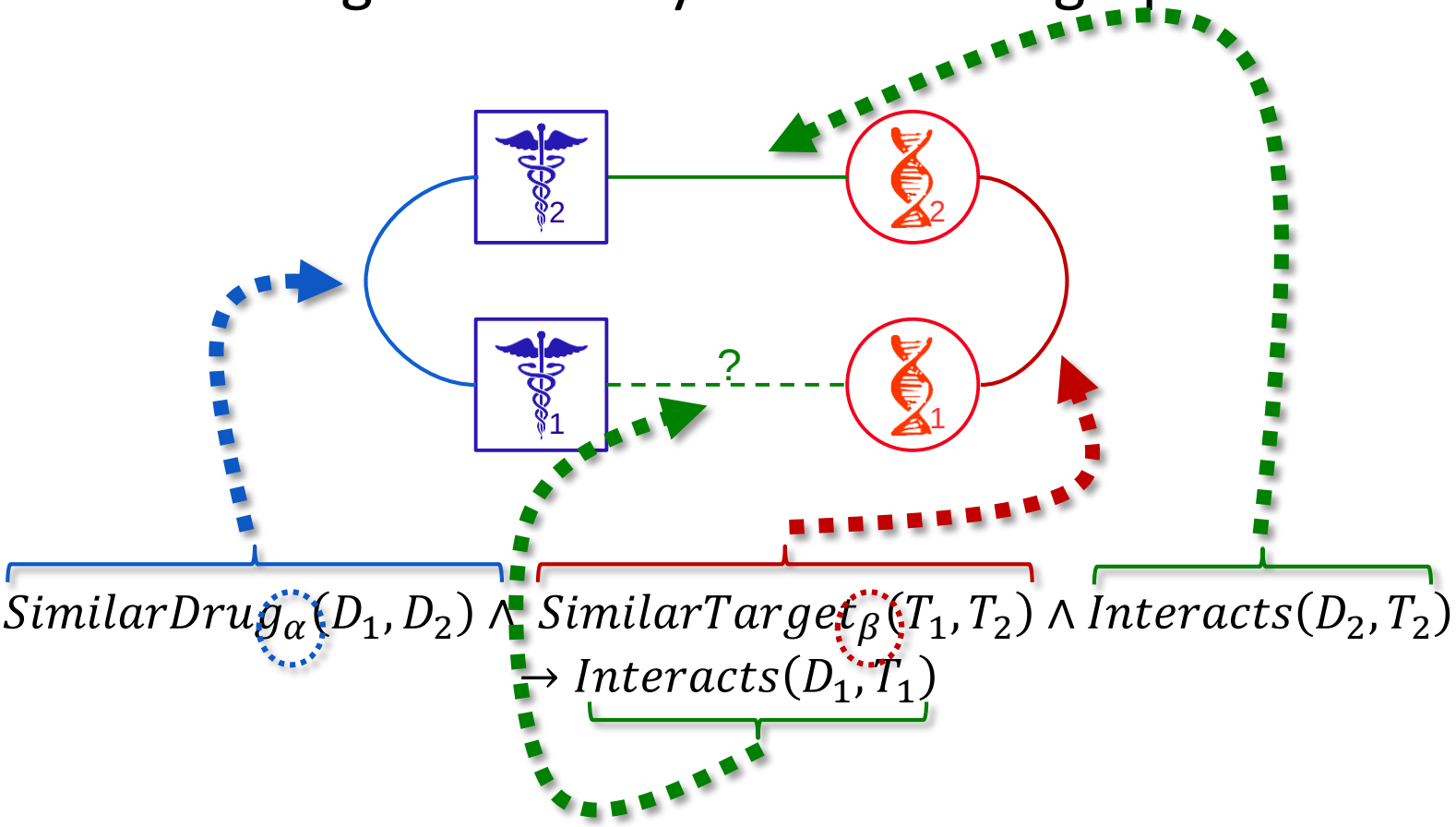
# Triad-based rules (Drugs)

- Targets tend to interact with similar drugs
  (friend of friend is a friend)



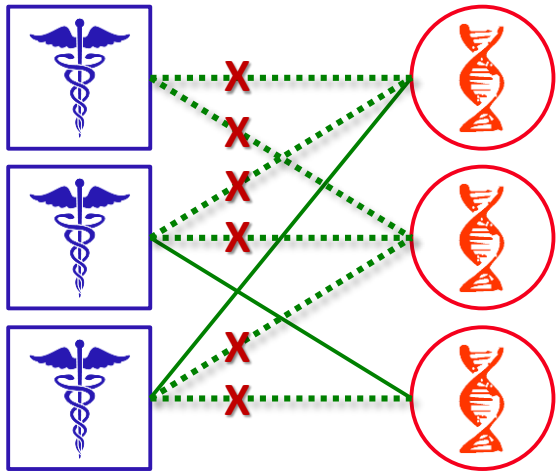
# Tetrad-based Rules (Similar Edges)

- Similar edges are likely to form in a graph



# Negative Prior

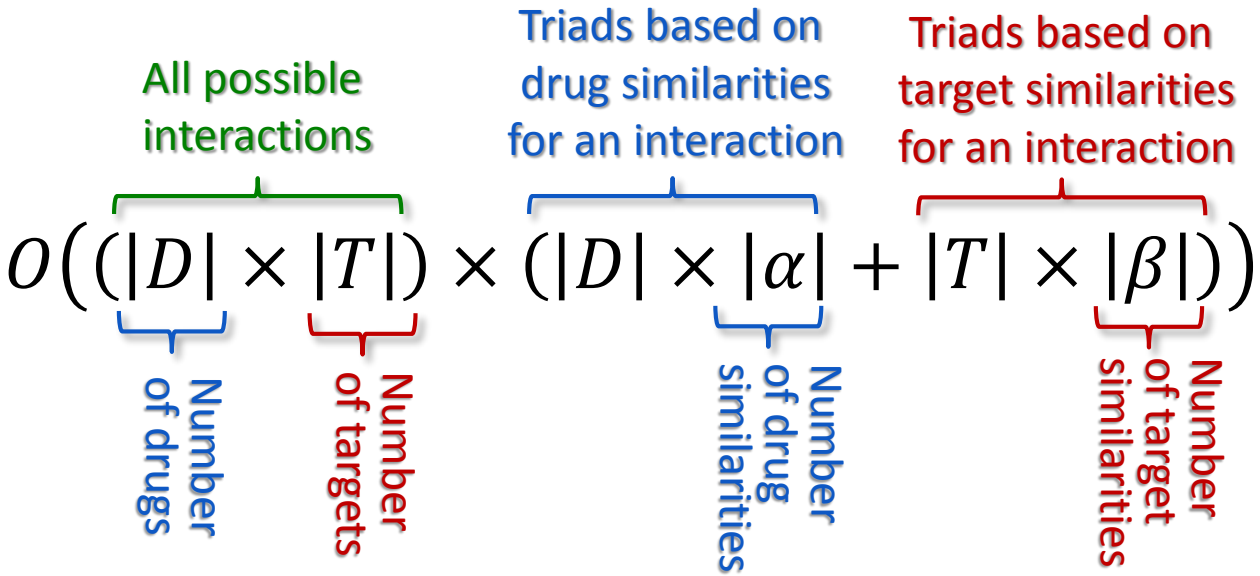
- Negative prior indicates “Interacts” predicate is most likely false
- i.e., most drugs and targets do not interact



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

# Size of the problem

- Total ground triad-based rules can be:




- e.g., in our experiments it was **180M**
- For tetrad-based rules the situation is even worst!

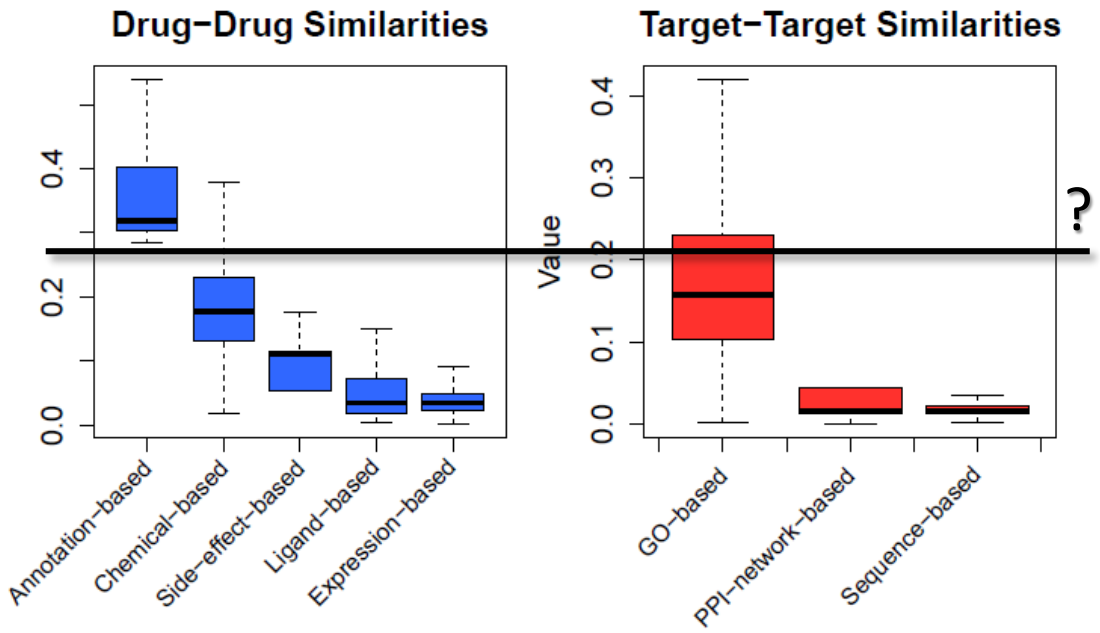

# Blocking

# Blocking

- Limit some of the rules from being grounded
- Ignore some of the less significant similarities between drugs and between targets.

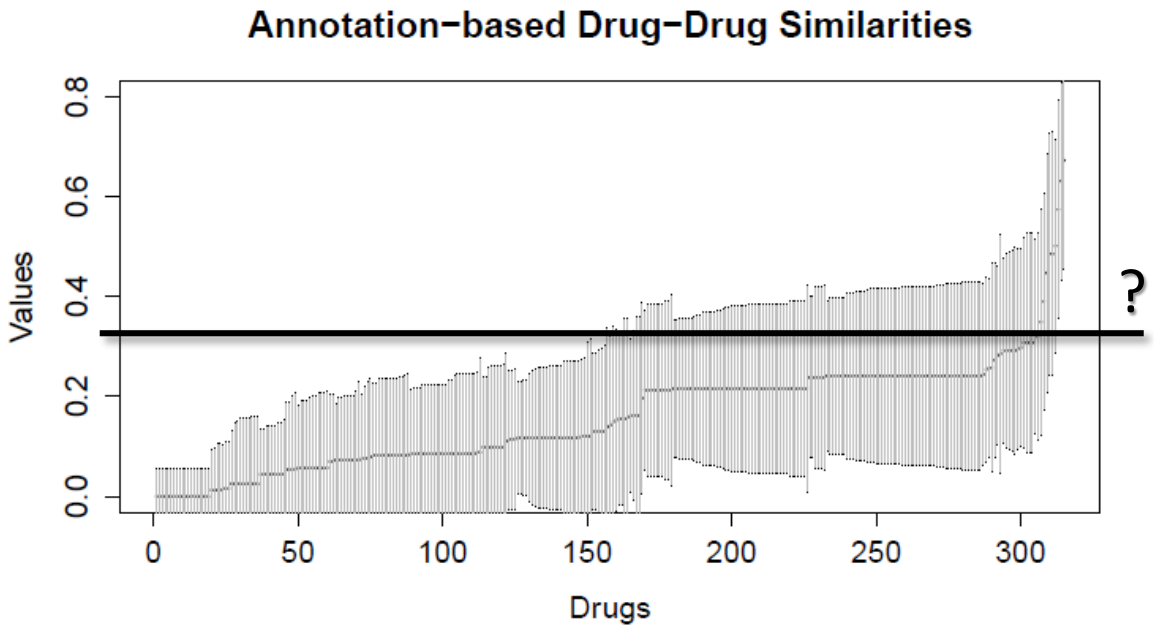
# Same Threshold for All Similarities

- Fixed threshold either ignores most of the values in one similarity or includes most of the values from the other



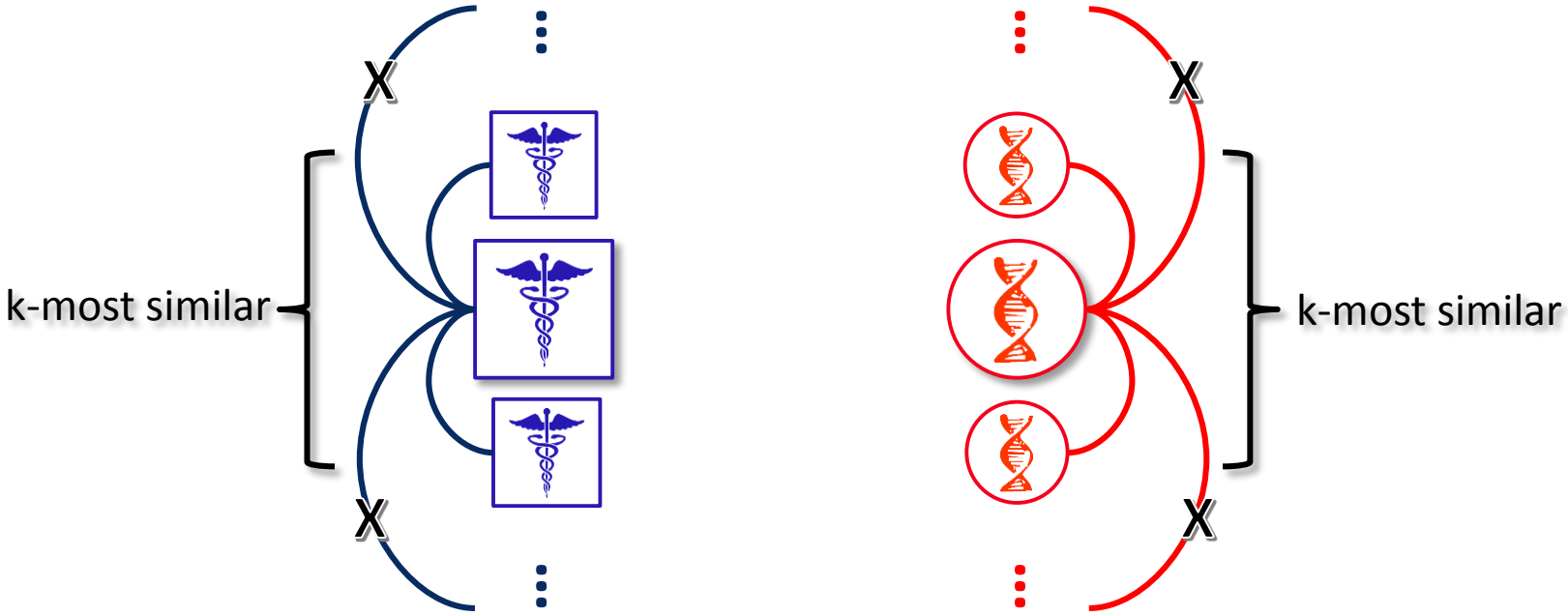
# A Threshold for Each Similarity

- Same problem for individual drug or target!



# K-Nearest Neighbors-based

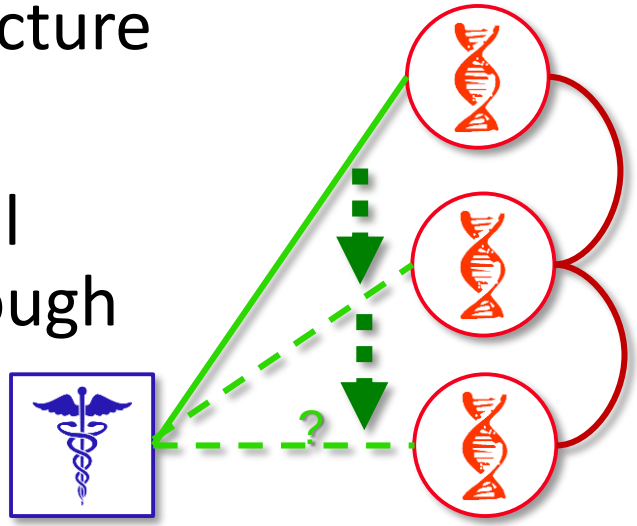
- Preserve the k-highest values in each similarity for each drug and each target and set the others to zero.



# PSL Advantages

# PSL Advantages

- PSL captures the original structure
- Collective Inference (No IID assumption): Results in global information propagation through the network.
- Inference based on Interpretable rules




- Class Imbalance: PSL can handle huge class-imbalance problems in link prediction problems.

# Experimental Evaluation

# Dataset

- 315 Drugs
- 250 Targets
- Interaction: [Knox et al. 2011]
  - 1,306 observed interactions
  - 78,750 possible interactions
- Similarities: [Perlman et al. 2011]
  - 3 target-target similarities
  - 5 drug-drug similarities

# Drug-Drug Similarities [Perlman et al. 2011]

- **Chemical-based:**
  - Jaccard similarity of the SMILES fingerprints
- **Ligand-based:**
  - Jaccard similarity between protein receptor families extracted via matched ligands with drugs SMILES
- **Expression-based:**
  - Spearman correlation of gene expression responses to drugs using Connectivity Map.
- **Side-effect-based:**
  - Jaccard similarity between drugs side-effects from SIDER
- **Annotation-based:**
  - Semantic Similarity of Drugs based on the World Health Organization ATC classification system

# Target-Target Similarities [Perlman et al. 2011]

- **Sequence-based:**
  - Smith-Waterman sequence alignment scores
- **Protein-protein interaction network-based:**
  - The distance in the protein-protein interactions network using all-pairs shortest path.
- **Gene ontology-based:**
  - Semantic similarity between Gene ontology annotations

# Triad Rules

| | Rule | AUROC |
|---|---|---|
| Drug-Drug Similarity | Annotation-based | 0.685 ± 0.026 |
| | Chemical-based | 0.714 ± 0.030 |
| | Ligand-based | **0.751 ± 0.030** |
| | Expression-based | 0.584 ± 0.025 |
| | Side-effect-based | 0.614 ± 0.030 |
| Target-Target Similarity | PPI-network-based | 0.816 ± 0.026 |
| | GO-based | 0.608 ± 0.029 |
| | Sequence-based | **0.842 ± 0.019** |
| **All rules (similarities)** | | **0.931 ± 0.018** |

# Triad Rules: Comparison with reported results

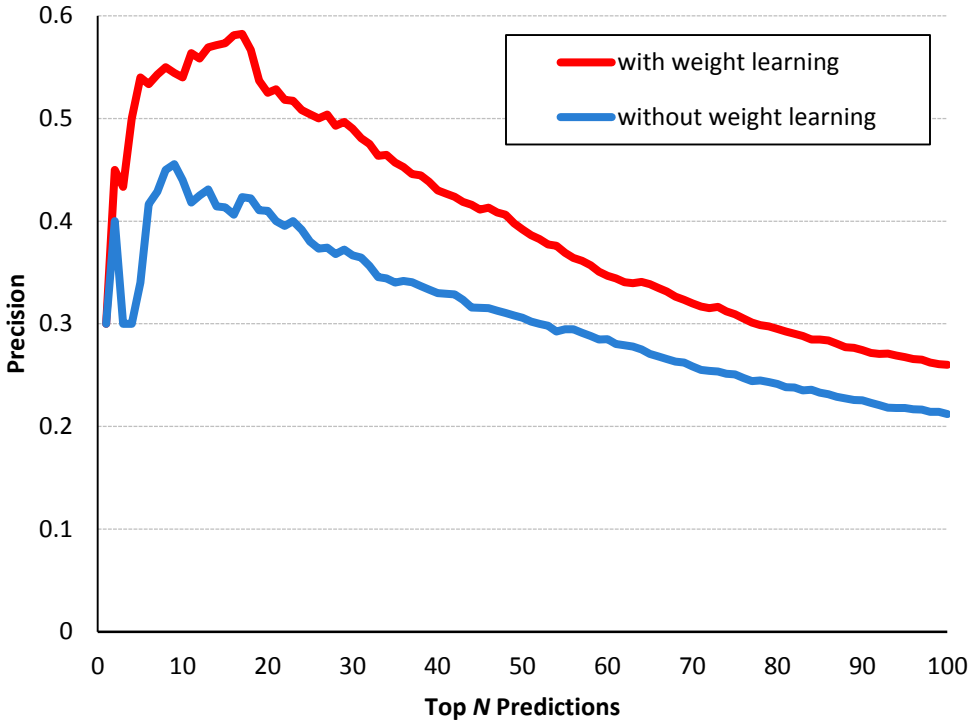
| Method | AUROC | Condition |
|---|---|---|
| PSL | 0.931 ± 0.018 | Without Sampling (10 Fold C.V.) |
| Perlman et al. 2011 | 0.935 | With Sampling (Reported Results) |
| Yamanishi et al. 2008 | 0.884 | |
| Bleakley et al. 2009 | 0.814 | |

# Triad Rules: Blocking and Weight Learning

| Condition | AUROC | | |
|---|---|---|---|
| | K=5 | K=15 | K=30 |
| All weights fixed | 0.926 ± 0.016 | 0.929 ± 0.020 | 0.923 ± 0.021 |
| + Weight learning | 0.930 ± 0.016 | **0.931 ± 0.018** | 0.924 ± 0.21 |

| Condition | Time to Complete (10-folds) | | |
|---|---|---|---|
| | K=5 | K=15 | K=30 |
| All weights fixed | 12 mins | 3 h | 9 h |
| + Weight learning | 1 h | 10 h | 28 h |

# Triad Rules:
# Precision of Top 100 Predictions



# Triad and Tetrad based rules

| Method | AUROC with k=5 |
|---|---|
| Triad-based Rules | 0.930 ± 0.016 |
| Tetrad-based Rules | 0.796 ± 0.025 |
| Triad-based & Tetrad-based | 0.913 ± 0.017 |

# Conclusion

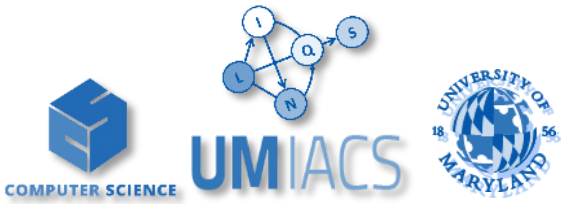
- Identified challenges of network-based drug-target interaction prediction.
- Described PSL framework to address them:
  - Captures original network structure
  - Is a declarative language to implement different rules
  - Performs collective inference (No IID assumption)
  - Weight learning based on training data
- Matched performance of the state-of-the-art with simple triad-based rules.
- The proposed method can easily be applied to other tasks with similar structures.

# Thank you

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

Shobeir Fakhraei*, Louiqa Raschid, Lise Getoor
University of Maryland, College Park, MD, USA

**http://psl.cs.umd.edu**



# Refrences


- 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, Feb. 2011
- Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. “DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. ” Nucleic Acids Res. Jan 2011
- 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, Jul 2008.
- K. Bleakley and Y. Yamanishi. “Supervised prediction of drug target interactions using bipartite local models.” Bioinformatics, Sep. 2009
- Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor, Hinge-loss, “Markov Random Fields: Convex Inference for Structured Prediction”, Uncertainty in Artificial Intelligence (UAI) 2013
- Stephen H. Bach, Matthias Broecheler, Lise Getoor, and Dianne P. O’Leary, “Scaling MPE Inference for Constrained Continuous Markov Random Fields with Consensus Optimization”, Advances in Neural Information Processing Systems (NIPS) 2012