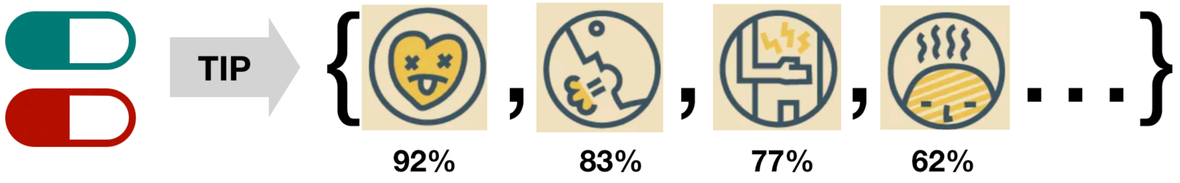
# Tri-graph Information Propagation for Polypharmacy Side Effect Prediction


Hao Xu[1,*], Shengqi Sang[2,3], and Haiping lu[1]

[1]Department of Computer Science, University of Sheffield, UK
[2]Department of Physics and Astronomy, University of Waterloo, Canada
[3]Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada
[*]current address: Department of Chemical Engineering, Queen's University, Canada
hx9@queensu.ca, s4sang@uwaterloo.ca, h.lu@sheffield.ac,uk


**Motivation**: In the treatment of **complex or simultaneous disease**, patients often have to take multiple drugs concurrently, called **polypharmacy**. This often causes additional side effects, i.e. polypharmacy side effect (PoSE) due to interactions between drugs, causing patients to face *higher unwanted health risks and undesired medical expenses.* Graph convolutional network (GCN) [3] is an emerging approach for graph representation learning. GCN-based drug representation learning has shown improved performance in PoSE prediction [6].

**Objective**: Given a pair of drugs, predict the types and probabilities of side effects that they cause.

**Challenge**: Representation learning on biomedical graphs is challenging:

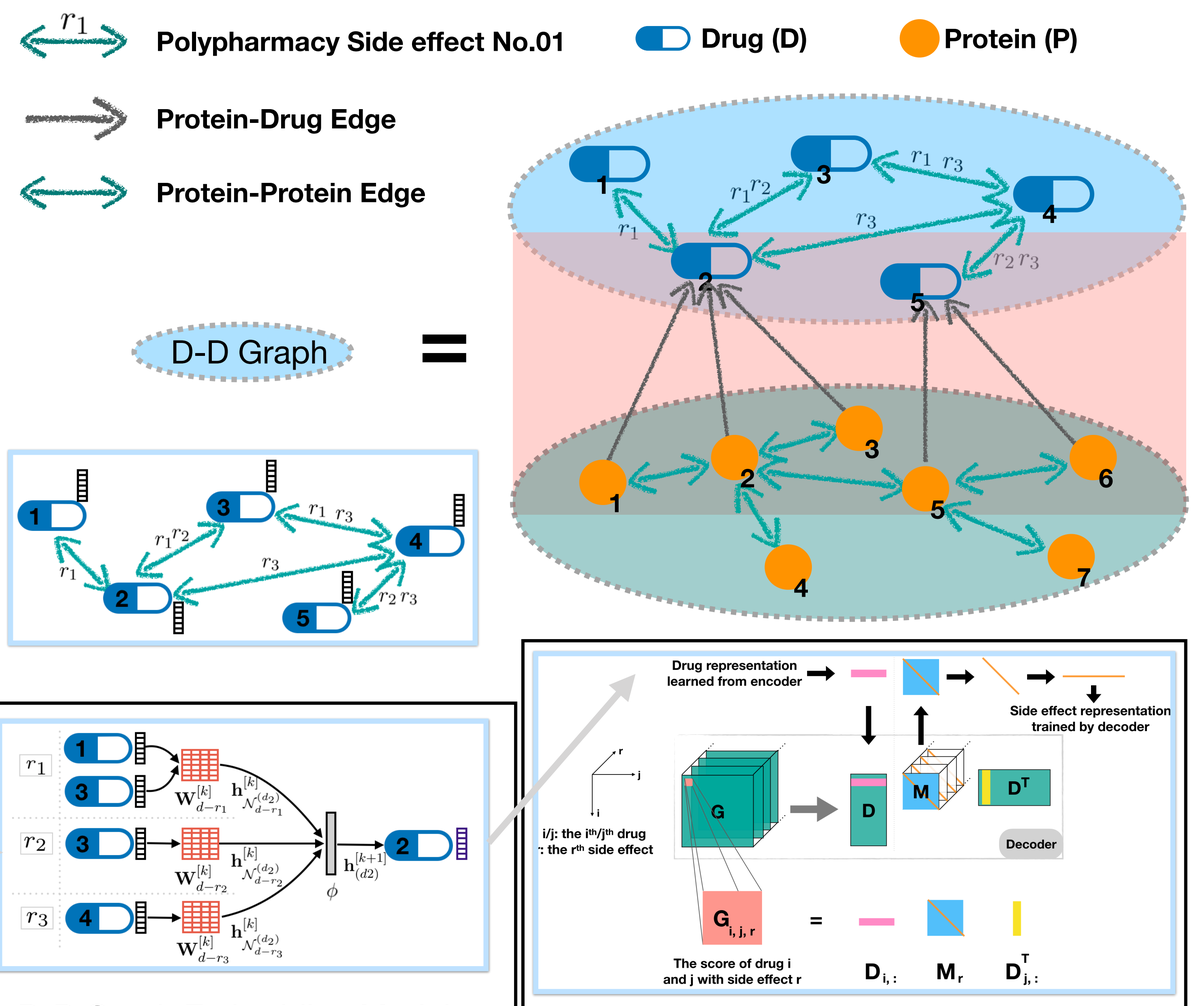
- The biomedical graph is always huge, sparse, heterogeneous, multi-relational, biased and uncompleted
- The existing GCN algorithm with neighbourhood aggregation operation has both high computational cost and high memory demand.

## 1. Dataset [2]

| Graph Name | Nodes | Edges | Unique Labels |
|---|---|---|---|
| D-D Graph | 645(D) | 63473 | 1317 |
| P-D Graph | 3648(P), 284(D) | 18690 | 1 |
| P-P Graph | 19081(P) | 715612 | 1 |

## 2. Problem Formulation

We consider POSE prediction task as a **graph completion problem** which aims to find the **undiscovered edges** and **labels** on the graph. It can also be considered as a *multi-relational link prediction* or a *multi-label classification* problem as well.

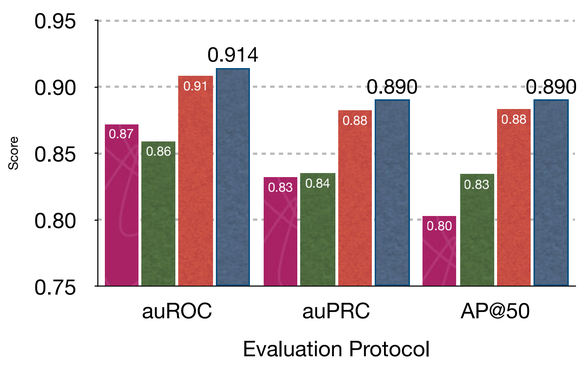



## 4. Tri-graph Information Propagation (TIP) Method

Inspired by the existing models [1,4,6] and motivated by their limitations, we propose the TIP model for **improving prediction accuracy,** and **time** and **space efficiency.**

TIP has four steps via three subgraphs:
(1) learn protein embedding on the P-P graph
(2) propagate such embedding to the D-D graph via the P-D graph
(3) learn the final drug embedding
(4) predict the side effects on the D-D graph.

## 5. Performance Comparison




| | Peak GPU Memory Usage (G) | Training Time per Epoch |
|---|---|---|
| Decagon[6] | ~> 28 | ~> 9600 |
| DistMult[5] | 9.25 | 41 |
| R-GCN[4] | 10.49 | 82 |
| TIP | 9.47 | 115 |

| Model | ARCT | AUPRC |
|---|---|---|
| $dTIP_D$ | DDM-NN | 0.791 |
| $dTIP_P$ | PPM-GGM-NN | 0.746 |
| TIP-cat | PPM-GGM-DDM-DF | 0.889 |
| TIP-sum | PPM-GGM-DDM-DF | **0.890** |

DistMult, R-GCN, dTIP_D learn from drug-drug interaction only

## 6. Three key benefits:

- **Flexibility**. TIP embeds proteins and drugs into **different spaces** of possibly different dimensions. Two ways to combine protein and drug information in the P-D graph: concatenation and convolution (or sum).
- **Efficiency**. Separate embedding of proteins and drugs can greatly improve the time (83×) and space (3×) efficiency of GCN-based representation learning and information propagation for them,
- **Accuracy**. More focused learning of drug representation makes better use of available data sources and can lead to improved POSE prediction, e.g., by 7.2% in our experiments.

## 7. Conclusion Highlight

- **Pharmacological information** does contain drug-drug interaction information. However, additional pharmacological information in TIP-sum only improves the performance slightly.
- Information propagation from PPM to GGM can be considered as learning a higher-level representation of a subset of proteins, which captures the relationship between proteins, and between proteins and drugs.
- Even if the model does not have access to pharmacological information, it can predict the side effects with molecular origins very well.

## 8. Future Work - Understand

- Compare predictions between the different TIP models (PPM, PPM-DDM, etc.) and explain why different subgraphs improve predictions.
- Interpreting mechanisms of side effects based on signalling, regulation, and metabolism.


In Yang's Computational Systems Biology Lab
Queen's University
Laurence Yang
laurence.yang@queensu.ca


## Reference:


[1] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine LearningVolume 70, pages 1263–1272. JMLR. org, 2017.
[2] Sagar Maheshwari Marinka Zitnik, Rok Sosic and Jure Leskovec. BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata, August 2018.
[3] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling.
[4] Modeling relational data with graph convolutional networks. In European Semantic Web Conference, pages 593–607. Springer, 2018.
[5] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575, 2014.
[6] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13):i457–i466, 2018.