About Paper:

Modeling polypharmacy side effects with graph convolutional networks

Marinka Zitnik, Monica Agrawal, Jure Leskovec

Reviewed by: Eric Wang
University of Virginia

https://qdata.github.io/deep2Read/

Bioinformatics, 2018
ISMB 2018

OXFORD

Modeling Polypharmacy Side Effects with Graph Convolutional Networks

Marinka Zitnik [1], Monica Agrawal [1] and Jure Leskovec [1,2,*]

[1] Department of Computer Science, Stanford University, Stanford, CA, USA
[2] Chan Zuckerberg Biohub, San Francisco, CA, USA

* To whom correspondence should be addressed.

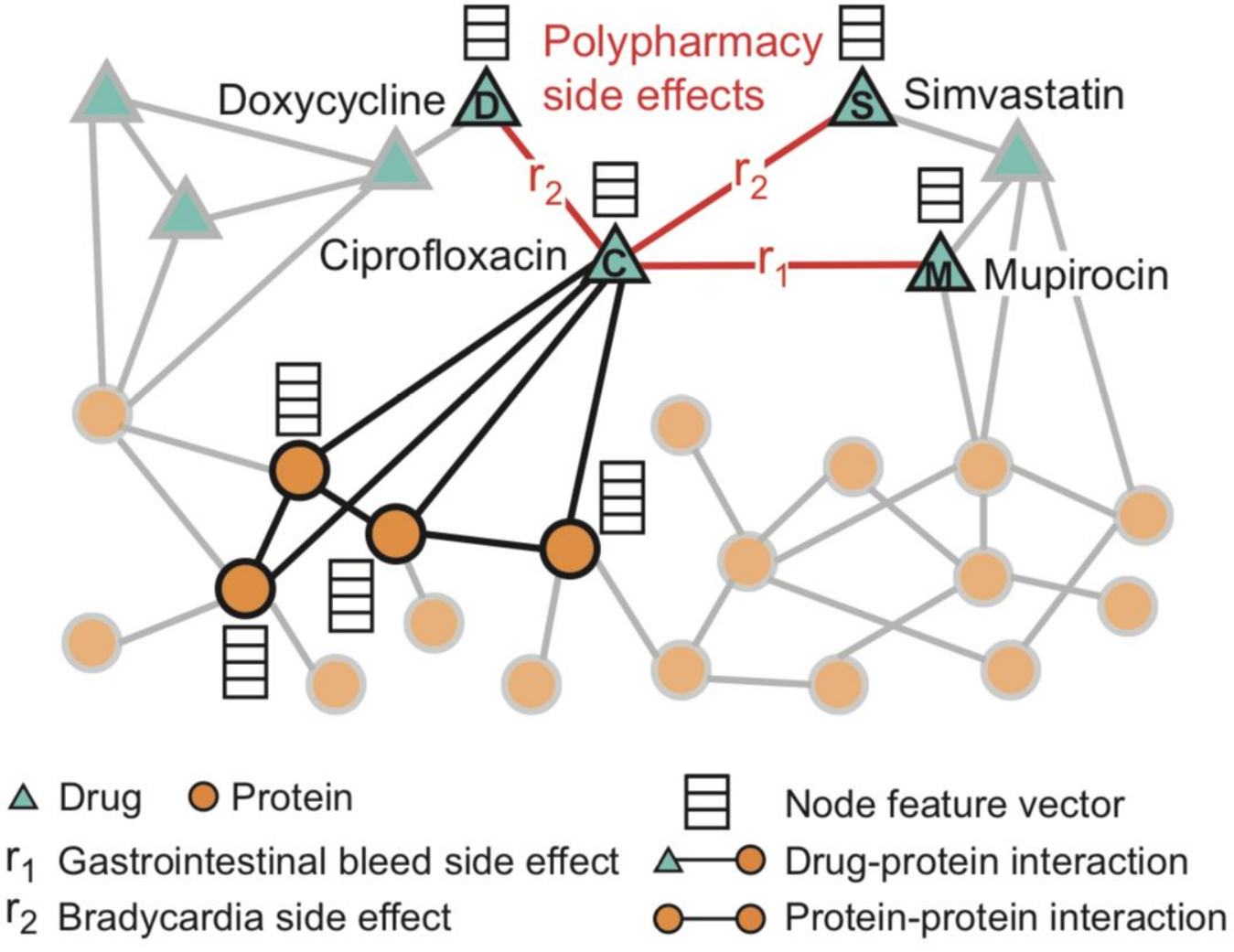
Introduction

- The use of drug combinations, termed **polypharmacy**, is common to treat patients with complex diseases or co-existing conditions. However, a major consequence of polypharmacy is a much higher risk of adverse side effects for the patient.
- Polypharmacy side effects emerge because of **drug-drug interactions**, in which activity of one drug may change, favorably or unfavorably, if taken with another drug. The knowledge of drug interactions is often limited because these complex relationships are rare, and are usually **not observed in relatively small clinical testing**.

- Discovering polypharmacy side effects thus remains an important challenge with significant implications for patient mortality and morbidity.

Present work

- *Decagon*, a method for predicting side effects of drug pairs. We model the problem by constructing a large **two-layer multimodal graph** of protein-protein interactions, drug-protein interactions, and drug-drug interactions (*i.e.*, side effects).
- Each drug-drug interaction is labeled by a different edge type, which signifies the type of the side effect. We then develop a new **multirelational edge prediction model** that uses the multimodal graph to predict drug-drug interactions as well as their types. Our model is a convolutional graph neural network that operates in a multirelational setting.


Polypharmacy
side effects
Doxycycline
D
S
Simvastatin
r2
r2
Ciprofloxacin
C
r1
M
Mupirocin
Drug
Protein
Node feature vector
r1 Gastrointestinal bleed side effect
Drug-protein interaction
r2 Bradycardia side effect
Protein-protein interaction

Observations

1. There is a wide range in **how frequently** certain side effects occur in drug combinations. We find that more than **53%** of polypharmacy side effects are known to occur in **less than 3%** of the documented drug combinations. As a result, **predicting rarer side effects** becomes a challenging task, and thus it is important to develop an **end-to-end approach** such that the model is able to share information and learn from all side effects at once.

2. Polypharmacy side effects do not appear independently of one another in drug combinations, suggesting that **joint modeling over multiple side effects** can aid in the prediction task. A prediction model should leverage **dependence between side effects** and be able to re-use the information learned about the molecular basis of one side effect to better understand the molecular basis of another side effect.

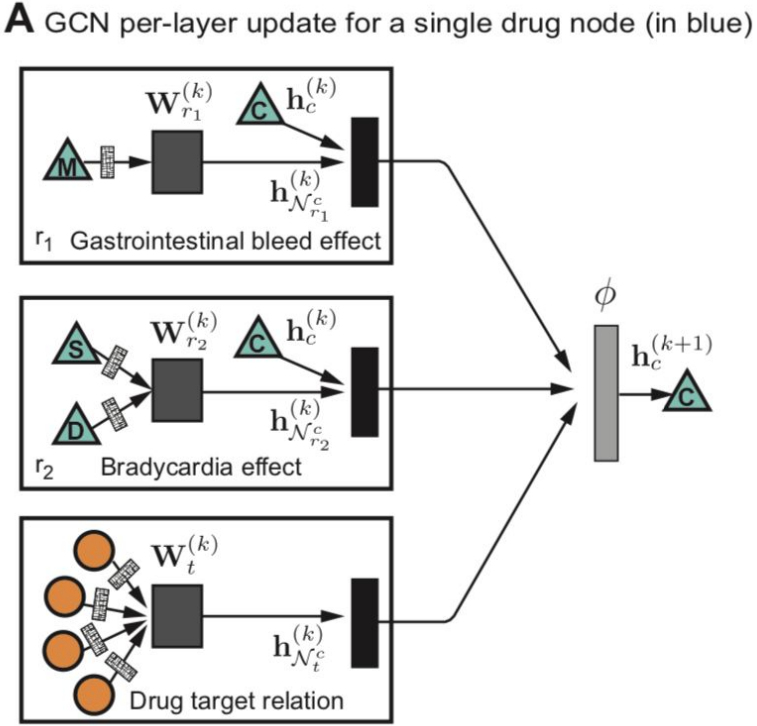
3. the relationship between proteins targeted by a drug pair and occurrence of side effects
    a. More than 68% of drug combinations have zero target proteins in common, suggesting it is important to use protein-protein interaction information to “connect” different proteins targeted by different drugs.
    b. **Random drug pairs** have smaller overlap in targeted proteins than **co-prescribed drugs**
    c. This trend is **unequally** observed across different side effects

Graph Convolutional Encoder

In each layer, *Decagon* propagates latent node feature information across edges of the graph, while taking into account the type (relation) of an edge (Schlichtkrull *et al.*, 2017). A single layer of this neural network model takes the following form:

$$\mathbf{h}_i^{(k+1)} = \phi\left(\sum_r \sum_{j \in \mathcal{N}_r^i} c_r^{ij} \mathbf{W}_r^{(k)} \mathbf{h}_j^{(k)} + c_r^i \mathbf{h}_i^{(k)}\right), \qquad (1)$$

where $\mathbf{h}_i^{(k)} \in \mathbb{R}^{d(k)}$ is the hidden state of node v_i in the k-th layer of the neural network with $d^{(k)}$ being the dimensionality of this layer's representation, r is a relation type, and matrix $\mathbf{W}_r^{(k)}$ is a relation-type specific parameter matrix. Here, ϕ denotes an non-linear element-wise activation function (*i.e.*, a rectified linear unit), which transforms the representations to be used in the layer of the neural model, c_r^{ij} and c_r^i are normalization constants, which we choose to be symmetric $c_r^{ij} = 1/\sqrt{|\mathcal{N}_r^i||\mathcal{N}_r^j|}$ and $c_r^i = 1/|\mathcal{N}_r^i|$ with $\mathcal{N}_r^i$ denoting the set of neighbors of node v_i under relation r. Importantly note that the sum in Eq. 1 ranges only over the neighbors $\mathcal{N}_r^i$ of a given node i and thus the computational architecture (*i.e.*, the neural network) is different for every node. Figure 3A



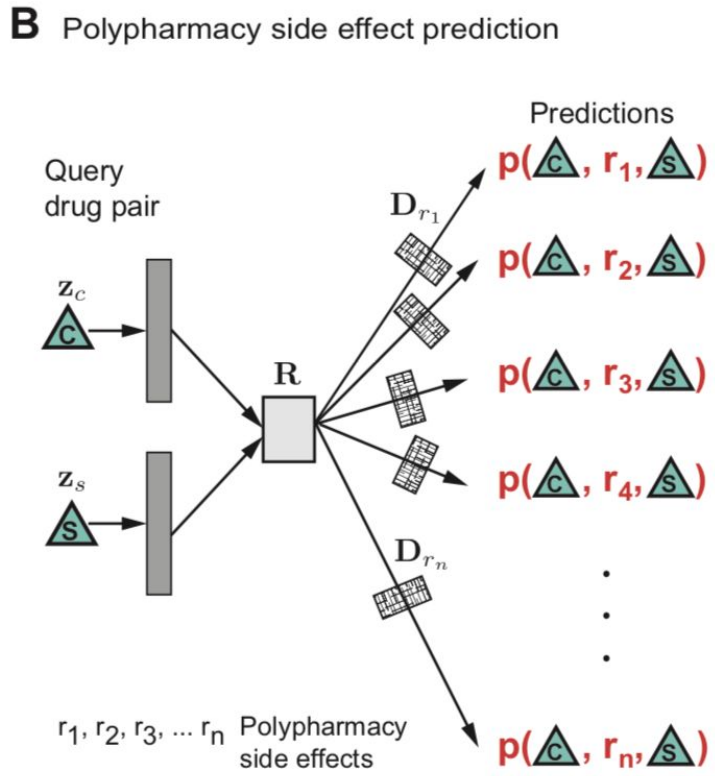
Tensor Factorization Decoder

The goal of decoder is to reconstruct labeled edges in G by relying on learned node embeddings and by treating each label (edge type) differently. In particular, decoder scores a (v_i, r, v_j)-triple through a function g whose goal is to assign a score $g(v_i, r, v_j)$ representing how likely it is that drugs v_i and v_j are interacting through a relation/side effect type r (Figure 3B). Using embeddings for nodes i and j returned by *Decagon*'s encoder (Section 4.1) $\mathbf{z}_i$ and $\mathbf{z}_j$, the decoder predicts a candidate edge (v_i, r, v_j) through a factorized operation:

$$g(v_i, r, v_j) = \begin{cases} \mathbf{z}_i^T \mathbf{D}_r \mathbf{R} \mathbf{D}_r \mathbf{z}_j & \text{if } v_i \text{ and } v_j \text{ are drugs} \\ \mathbf{z}_i^T \mathbf{M}_r \mathbf{z}_j & \text{if } v_i \text{ and } v_j \text{ are both proteins, or,} \\ & v_i \text{ and } v_j \text{ are a protein and a drug} \end{cases} \quad (2)$$

followed by the application of a sigmoid function σ to compute probability of edge (v_i, r, v_j):

$$p_r^{ij} = p((v_i, r, v_j) \in \mathcal{R}) = \sigma(g(v_i, r, v_i)). \quad (3)$$



RSC Publishing MOLECULAR OMICS
Article | Journal | Search | Submit

Mol Omics. 2018 Jun 1; 14(3): 197–209.
Published online 2018 Jun 7. doi: 10.1039/c8mo00027a

PMCID: PMC6115748
PMID: 29876573

PTMscape: an open source tool to predict generic post-translational modifications and map modification crosstalk in protein domains and biological processes†

Ginny X. H. Li,[a] Christine Vogel,[b] and Hyungwon Choi*[a,c]

Introduction

- Protein post-translational modifications (PTMs) regulate cellular functions in various ways: catalyzing enzymatic activities, conferring substrate specificity to control allosteric interactions, mediating interactions with other molecules such as DNA, co-factors, and lipids, and localizing proteins to organelles.
- With advances in enrichment techniques for PTMs, high-resolution mass spectrometry (MS) has now become the method of choice to experimentally detect and quantify major PTMs at a proteome scale. A wealth of PTM data arising from tandem MS/MS experiments has been curated and shared in public databases such as **PhosphoSitePlus (PSP),3 PHOSIDA,4 and Uniprot,5** and some major PTMs such as **phosphorylation** and **ubiquitination** have been mapped for multiple species.
- For instance, as of December 2017, the PSP database described ~240 000 phosphorylation and ~22 000 ubiquitination sites for >20 000 different human proteins.

Performance evaluation of the linear SVMs across five PTM types. The number of true positive sites used in the 10-fold cross-validation is about half the amount of data present in the PSP database after removal of redundant protein sequences and those that do not have secondary structure information from SPIDER3. AUC – area-under-the-curve; MCC – the highest Matthew's correlation coefficient at all score thresholds; sensitivity/specificity at score threshold corresponding to the highest MCC value

PTM type	No. of proteins	No. of PSP sites	Window size 25			
			AUC	MCC	Sensitivity	Specificity
Acetylation (K)	3729	10 479	0.66	0.25	0.61	0.64
Methylation (K)	1521	2566	0.74	0.39	0.61	0.76
Ubiquitination (K)	4874	22 592	0.64	0.22	0.67	0.54
SUMOylation (K)	1020	2996	0.77	0.42	0.63	0.79
Methylation (R)	2301	5450	0.79	0.47	0.62	0.84
Phosphorylation (S)	8510	76 008	0.74	0.36	0.70	0.66
Phosphorylation (T)	6982	28 359	0.72	0.33	0.66	0.66
Phosphorylation (Y)	6097	18 645	0.70	0.30	0.72	0.58