Graph Neural Networks for Drug Development

Marinka Zitnik
marinka@hms.harvard.edu

Stanford Engineering
Computer Science
Drug Development

1. Step 1: Design and Discovery
2. Step 2: Preclinical Research
3. Step 3: Clinical Research
4. Step 4: FDA Review
5. Step 5: Post-Market and Safety Monitoring

10,000 compounds
250 compounds
5 compounds
1 compound

12-16 years, ~$1 billion to $2 billion

Marinka Zitnik - Harvard - marinka@hms.harvard.edu
Opportunities for AI in Drug Development

Step 1: Design and Discovery
- Support decision-making for a new drug in the laboratory

Step 2: Preclinical Research
- Answer basic questions about safety and animal testing

Step 3: Clinical Research
- Predict if drug is safe & effective to test on people, find new uses for drugs

Step 4: FDA Review
- Automatic document review to make a decision to approve the drug or not

Step 5: Post-Market and Safety Monitoring
- Detect adverse and safety issues in real time using electronic health data
Why is it so challenging to realize this vision?

Finding drugs for disease treatments relies on several types of interactions, e.g., drug-target, protein-protein, drug-drug, drug-disease, disease-protein pairs.

Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, Information Fusion 2019
Today’s Talk

Step 1: Design and Discovery
- Support decision-making for a new drug in the laboratory

Step 2: Preclinical Research
- Answer basic questions about safety and animal testing

Step 3: Clinical Research
- Predict if drug is safe & effective to test on people, find new uses for drugs

Step 4: FDA Review
- Automatic document review to make a decision to approve the drug or not

Step 5: Post-Market and Safety Monitoring
- Detect adverse and safety issues in real time using electronic health data

Marinka Zitnik - Harvard - marinka@hms.harvard.edu
New tricks for old drugs

Faced with skyrocketing costs for developing new drugs, researchers are looking at ways to repurpose older ones — and even some that failed in initial trials.

Goal: Find which diseases a drug (new molecule) could treat
What drug treats what disease?

Drugs

Diseases

Goal: Predict what diseases a new molecule might treat

“Treats” relationship

Unknown drug-disease relationship

Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, Information Fusion 2019
Key Insight: Subgraphs

**Disease:** Subgraph of rich protein network defined on disease proteins

**Drug:** Subgraph of rich protein network defined on drug’s target proteins

A drug likely treats a disease if it is close to the disease in **pharmacological space** [Paolini et al., Nature Biotech.'06; Menche et al., Science’15]

**Idea:** Use the paradigm of embeddings to operationalize the concept of closeness in pharmacological space
Predicting Links Between Drug and Disease Subgraphs

**Task:** Given drug $C$ and disease $D$, predict if $C$ treats $D$

**Input data**

Our method (SUGAR)

**Predictions**

**Task:** 1) Learn embeddings for $C$’s and $D$’s subgraphs 2) Use embeddings to predict probability that $C$ treats $D$
Neural Message Passing

$p(\mathcal{C}, \mathcal{D})$

Edge decoder

Subgraph encoder

Aggregate information from subgraphs

Aggregate information from neighbors

Marinka Zitnik - Harvard - marinka@hms.harvard.edu
We need drug repurposing dataset

- Protein-protein interaction network culled from 15 knowledge databases with 19K nodes and 350K edges.


- Side information on drugs, diseases, proteins, etc.: Molecular pathways, disease symptoms, side effects.
## Predictive Performance

**Task:** Given a disease and a drug, predict if the drug could treat the disease

<table>
<thead>
<tr>
<th>Approach</th>
<th>AUPRC</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method (SUGAR)</td>
<td>0.851</td>
<td>0.888</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graphlets [Bioinformatics’13]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-directional random walks [Bioinformatics’16]</td>
</tr>
<tr>
<td>Heterogeneous graph inference [Bioinformatics’14]</td>
</tr>
<tr>
<td>Drug-disease closeness [Nat. Commun.’17]</td>
</tr>
<tr>
<td>Drug-disease dispersion [Nat. Commun.’17]</td>
</tr>
<tr>
<td>Gene-based network overlap [Nat. Commun.’17]</td>
</tr>
</tbody>
</table>

**Up to 49% improvement**

**Up to 172% improvement**
## Drug Repurposing at Stanford

**Task:** Predict if an existing drug can be repurposed for a new disease

<table>
<thead>
<tr>
<th>Drug</th>
<th>Disease</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-acetyl-cysteine</td>
<td>cystic fibrosis</td>
<td></td>
</tr>
<tr>
<td>Xamoterol</td>
<td>neurodegeneration</td>
<td></td>
</tr>
<tr>
<td>Plerixafor</td>
<td>cancer</td>
<td></td>
</tr>
<tr>
<td>Sodium selenite</td>
<td>cancer</td>
<td>Rank: 36/5000</td>
</tr>
<tr>
<td>Ebselen</td>
<td>C difficile</td>
<td>Rank: 10/5000</td>
</tr>
<tr>
<td>Itraconazole</td>
<td>cancer</td>
<td>Rank: 26/5000</td>
</tr>
<tr>
<td>Bestatin</td>
<td>lymphedema</td>
<td>Rank: 11/5000</td>
</tr>
<tr>
<td>Bestatin</td>
<td>pulmonary arterial hypertension</td>
<td>Rank: 16/5000</td>
</tr>
<tr>
<td>Ketaprofen</td>
<td>lymphedema</td>
<td>Rank: 28/5000</td>
</tr>
<tr>
<td>Sildenafil</td>
<td>lymphatic malformation</td>
<td>Rank: 26/5000</td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>pulmonary arterial hypertension</td>
<td>Rank: 46/5000</td>
</tr>
<tr>
<td>Benzamil</td>
<td>psoriasis</td>
<td>Rank: 114/5000</td>
</tr>
<tr>
<td>Carvedilol</td>
<td>Chagas’ disease</td>
<td>Rank: 9/5000</td>
</tr>
<tr>
<td>Benserazide</td>
<td>BRCA1 cancer</td>
<td>Rank: 41/5000</td>
</tr>
<tr>
<td>Pioglitazone</td>
<td>interstitial cystitis</td>
<td>Rank: 13/5000</td>
</tr>
<tr>
<td>Sirolimus</td>
<td>dystrophic epidermolysis bullosa</td>
<td>Rank: 46/5000</td>
</tr>
</tbody>
</table>
Feedbacks for the AI Loop

Will Benzamil treat psoriasis?

a. Heterogeneous biomedical network
b. Deep graph convolutional model
c. Predictions

Psoriasis
- Ebselen, $p = 0.96$
- Bestatin, $p = 0.84$
- Benzamil, $p = 0.76$
- Sirolimus, $p = 0.54$

What data can explain these predictions?

Drug, Disease, Protein, Molecular pathway, Drug side effect

“Olfactory signaling” pathways
“Innate immune response” pathways

Expert panel
Feedbacks for the AI Loop

Will Benzamil treat psoriasis?

a. Heterogeneous biomedical network
b. Deep graph convolutional model
c. Predictions

Psoriasis
- Ebselen: $p = 0.96$
- Bestatin: $p = 0.84$
- Benzamil: $p = 0.76$
- Sirolimus: $p = 0.54$
...

What data can explain these predictions?

Expert panel
Explaining GNN Predictions

Key idea:
- Summarize where in the data the model “looks” for evidence for its prediction
- Find a small subgraph most influential for the prediction
GNN Explainer: Key Idea

- **Input:** Given prediction \( f(x) \) for node/link \( x \)
- **Output:** Explanation, a small subgraph \( M_x \) together with a small subset of node features:
  - \( M_x \) is most influential for prediction \( f(x) \)
- **Approach:** Learn \( M_x \) via counterfactual reasoning
  - **Intuition:** If removing \( v \) from the graph strongly decreases the probability of prediction \( \Rightarrow v \) is a good counterfactual explanation for the prediction
"Why did you predict that this molecule will have a mutagenic effect on Gram-negative bacterium S. typhimurium?"
Today’s Talk

Step 1: Design and Discovery

- Support decision-making for a new drug in the laboratory

Step 2: Preclinical Research

- Answer basic questions about safety and animal testing

Step 3: Clinical Research

- Predict if drug is safe & effective to test on people, find new uses for drugs

Step 4: FDA Review

- Automatic document review to make a decision to approve the drug or not

Step 5: Post-Market and Safety Monitoring

- Detect adverse and safety issues in real time using electronic health data

Marinka Zitnik - Harvard - marinka@hms.harvard.edu
Polypharmacy

Patients take multiple drugs to treat complex or co-existing diseases

46% of people over 65 years take more than 5 drugs

Many take more than 20 drugs to treat heart diseases, depression or cancer

15% of the U.S. population affected by unwanted side effects

Annual costs in treating side effects exceed $177 billion in the U.S. alone

[Ernst and Grizzle, JAPA’01; Kantor et al., JAMA’15]
Unexpected Drug Interactions

Co-prescribed drugs

Task: How likely will a particular combination of drugs lead to a particular side effect?

Side Effects

3% prob., 2% prob.
Why is modeling polypharmacy hard?

Combinatorial explosion
- >13 million possible combinations of 2 drugs
- >20 billion possible combinations of 3 drugs

Non-linear & non-additive interactions
- Different effect than the additive effect of individual drugs

Small subsets of patients
- Side effects are interdependent
- No info on drug combinations not yet used in patients
Setup: Multimodal Networks

Mode 1
e.g., drugs

- E.g., Specific type of drug-drug interaction ($r_1$)

Mode 2
e.g., proteins

- E.g., drug-target interaction ($r_4$)
- E.g., protein-protein interaction ($r_5$)

Edge type $i$
Node types

Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 2018
Our Approach: Decagon

1. **Encoder**: Take a multimodal network and learn an embedding for every node.

2. **Decoder**: Use the learned embeddings to predict typed edges between nodes.
Encoder: Propagate Neighbors

Generate embeddings based on local network neighborhoods separated by edge type

1) Determine a node’s computation graph for each edge type

2) Learn how to transform and propagate information across computation graph

Example for edge type $r_3$:

1st order neighbor of $v$

2nd order neighbor of $v$

Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 2018
Decoder: Weighted, Typed Edges

**Input:** Embeddings of two nodes, C and S

**Output:** Predicted edges, new discovered relationships

\[ p(\triangle_C, r_1, \triangle_S) = \sigma(z_C^T D_{r_1} R D_{r_1} z_S) \]
\[ p(\triangle_C, r_2, \triangle_S) = \sigma(z_C^T D_{r_2} R D_{r_2} z_S) \]
\[ p(\triangle_C, r_3, \triangle_S) = \sigma(z_C^T D_{r_3} R D_{r_3} z_S) \]
\[ p(\triangle_C, r_4, \triangle_S) = \sigma(z_C^T D_{r_4} R D_{r_4} z_S) \]

Tensor factorized model captures dependences between different edge types

**Parameter weight matrices**

\( z_C, z_S \)

\( R, D_{r_1}, D_{r_2}, D_{r_3}, D_{r_4} \)

Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics* 2018
We need polypharmacy dataset

Objective:
Capture molecular, drug, and patient data for all drugs prescribed in the U.S.

We build a unique dataset:
- 4,651,131 drug-drug edges:
  Patient data from adverse event system, tested for confounders [FDA]
- 18,596 drug-protein edges
- 719,402 protein-protein edges:
  Physical, metabolic enzyme-coupled, and signaling interactions

Drug and protein features:
drugs' chemical structure, proteins' membership in pathways

A polypharmacy network with over 5 million edges and over 1,000 different edge types

Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 2018
We apply our deep approach to the polypharmacy network

E.g.: How likely will Simvastatin and Ciprofloxacin, when taken together, break down muscle tissue?
Results: Side Effect Prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC</th>
<th>AP@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method (Decagon)</td>
<td>0.834</td>
<td></td>
</tr>
<tr>
<td>RESCAL Tensor Factorization [Nickel et al., ICML'11]</td>
<td>0.693</td>
<td></td>
</tr>
<tr>
<td>Multi-relational Factorization [Perros, Papalexakis et al., KDD'17]</td>
<td>0.705</td>
<td></td>
</tr>
<tr>
<td>Shallow Network Embedding [Zong et al., Bioinformatics'17]</td>
<td>0.725</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.731</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>0.476</td>
<td>0.567</td>
</tr>
</tbody>
</table>

Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics* 2018
New Predictions

First AI method to predict side effects of drug combinations, even for combinations not yet used in patients

Next: Can the method generate hypotheses and give:

- **Doctors** guidance on whether it is a good idea to prescribe a particular combination of drugs to a particular patient
- **Researchers** guidance on effective wet lab experiments and new drug therapies with fewer side effects
New Predictions

Approach:
1) Train deep model on data generated prior to 2012
2) How many predictions have been confirmed after 2012?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Drug 1</th>
<th>Drug 2</th>
<th>Side effect</th>
<th>Evidence found</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pyrimethamine</td>
<td>Aliskiren</td>
<td>Sarcoma</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Tigecycline</td>
<td>Bimatoprost</td>
<td>Autonomic</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Telangiectases</td>
<td>Omeprazole</td>
<td>Dacarbazine</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Tolcapone</td>
<td>Pyrimethamine</td>
<td>Blood brain</td>
<td></td>
</tr>
</tbody>
</table>

Case Report
Severe Rhabdomyolysis due to Presumed Drug Interactions between Atorvastatin with Amlodipine and Ticagrelor

<table>
<thead>
<tr>
<th>Rank</th>
<th>Drug 1</th>
<th>Drug 2</th>
<th>Side effect</th>
<th>Evidence found</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Anagrelide</td>
<td>Azelaic acid</td>
<td>Cerebral thrombosis</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Atorvastatin</td>
<td>Amlodipine</td>
<td>Muscle inflammation</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Aliskiren</td>
<td>Tioconazole</td>
<td>Breast inflammation</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Estradiol</td>
<td>Nadolol</td>
<td>Endometriosis</td>
<td></td>
</tr>
</tbody>
</table>
Clinical Validation of New Predictions

Drug interaction markers, lab values, and many other surrogates

Robert Martin
22 Feb 1953  Male

<table>
<thead>
<tr>
<th>Medication</th>
<th>Brand</th>
<th>Dose</th>
<th>Frequency</th>
<th>Quantity</th>
<th>Refills</th>
<th>Condition</th>
<th>Provider</th>
<th>Prescribed</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Renew by</th>
</tr>
</thead>
<tbody>
<tr>
<td>beclomethasone</td>
<td>QVAR HFA</td>
<td>2 puffs</td>
<td>bid</td>
<td>12</td>
<td></td>
<td>Asthma</td>
<td>Barnes</td>
<td>19 Feb 2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>chlorthalidone</td>
<td></td>
<td>25 mg</td>
<td>1 daily</td>
<td>90</td>
<td>3</td>
<td>Hypertension</td>
<td>Barnes</td>
<td>19 Sep 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>insulin glargine</td>
<td>Lantus</td>
<td>28 u</td>
<td>daily</td>
<td>90</td>
<td>11</td>
<td>Diabetes</td>
<td>Ballard</td>
<td>19 Nov 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>metformin</td>
<td></td>
<td>1000 mg</td>
<td>1 bid</td>
<td>180</td>
<td>3</td>
<td>Diabetes</td>
<td>Barnes</td>
<td>4 Mar 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>naproxen</td>
<td>Aleve</td>
<td>500 mg</td>
<td>1 bid</td>
<td>90</td>
<td>0</td>
<td>Rheumatoid arthritis</td>
<td>Barnes</td>
<td>4 Mar 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>prednisone</td>
<td></td>
<td>20 mg</td>
<td>2 d x5d pm</td>
<td>84</td>
<td>0</td>
<td>Asthma</td>
<td>Barnes</td>
<td>12 Sep 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Sep 2013</td>
</tr>
<tr>
<td>zolpidem</td>
<td></td>
<td>5 mg</td>
<td>1 hs</td>
<td>90</td>
<td>0</td>
<td>Insomnia</td>
<td>Barnes</td>
<td>15 Mar 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22 Sep 2013</td>
</tr>
<tr>
<td>simvastatin</td>
<td></td>
<td>40 mg</td>
<td>1 daily</td>
<td>84</td>
<td>0</td>
<td>High cholesterol</td>
<td>Belden</td>
<td>19 Mar 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30 Sep 2013</td>
</tr>
<tr>
<td>terbinafine</td>
<td></td>
<td>250 mg</td>
<td>1 daily</td>
<td>84</td>
<td>0</td>
<td>Onychomycosis</td>
<td>Foote</td>
<td>30 Jul 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19 Oct 2013</td>
</tr>
</tbody>
</table>
Today’s Talk

Step 1: Design and Discovery
- Support decision-making for a new drug in the laboratory

Step 2: Preclinical Research
- Answer basic questions about safety and animal testing

Step 3: Clinical Research
- Predict if drug is safe & effective to test on people, find new uses for drugs

Step 4: FDA Review
- Automatic document review to make a decision to approve the drug or not

Step 5: Post-Market and Safety Monitoring
- Detect adverse and safety issues in real time using electronic health data

Marinka Zitnik - Harvard - marinka@hms.harvard.edu
Complex, interconnected datasets are transforming science and medicine.

Graph ML can unlock these datasets.

Physical instruments facilitate discoveries.

Instruments for modern, data-intensive sciences.

Thank you!
And thanks to my collaborators:
Jure Leskovec, Russ B. Altman, Will Hamilton, Rex Ying, Monica Agrawal, Dylan Bourgeois, Jiaxuan You, Evan Sabri Eyuboglu

Papers, data & code

\texttt{cs.stanford.edu/~marinka}
\texttt{snap.stanford.edu/biodata}

\begin{center}
\begin{tabular}{cc}
\textbf{HARVARD UNIVERSITY} & \textbf{WE’RE HIRING!} \\
\end{tabular}
\end{center}

Students and postdocs for projects in machine learning on biomedical data

marinka@hms.harvard.edu