Graph Methods for COVID-19 Response

William L. Hamilton

Why graph learning methods?

- Our response to the COVID-19 pandemic involves three key types of data:
  1. Biomedical treatment data
     - Biomedical knowledge graphs
     - Multiomic interaction and expression data
     - Results from in-vitro assays and clinical trials
  2. Epidemiological network data
     - Social networks of infected individuals
     - Global transportation/flight networks
     - Geospatial networks and population-level infection rates
  3. Supply chain networks
     - Transportation networks
     - Geospatial supply and demand data

Why graph learning methods?

- Our response to the COVID-19 pandemic involves three key types of data:
  1. Epidemiological network data
  2. Biomedical treatment data
  3. Supply chain networks

- These datasets all involve **heterogeneous and relational structures:**
  - Social networks
  - Transportation networks
  - Molecular graph structures
  - Multi-omic interaction networks
  - Knowledge graphs
  - Supply chain networks

Possible applications

- Possible applications using this data include:
  1. Computational drug design
     - Can we design better antivirals to target COVID-19?
  2. Computational treatment design
     - Can we design better treatment strategies using existing drugs?
  3. Epidemiological forecasting
     - Can we better predict how and where infection rates will change over time?
  4. Demand forecasting and supply chain optimization
     - Can we forecast COVID-19 related demands to optimize supply chains?
  5. Outbreak tracking and tracing
     - Can we model and predict infection risk at the individual level?

Possible applications

- Possible applications using this data include:

1. Computational drug design
   - Can we design better antivirals to target COVID-19?
2. Computational treatment design
   - Can we design better treatment strategies using existing drugs?
3. Epidemiological forecasting
   - Can we better predict how and where infection rates will change over time?
4. Demand forecasting and supply chain optimization
   - Can we forecast COVID-19 related demands to optimize supply chains?
5. Outbreak tracking and tracing
   - Can we model and predict infection risk at the individual level?

Computational drug design

- **Task:** Generate a novel molecule that can act as an effective anti-viral agent.

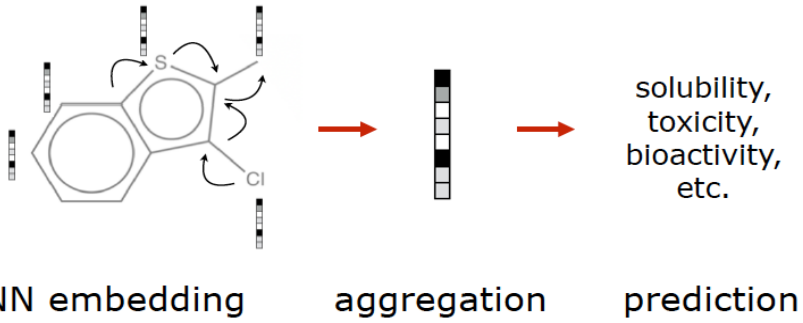
- Sub-problem 1: Molecule representation and property prediction
  - How can we effectively represent molecules as vector embeddings?
  - How can we predict molecular properties based on their learned representations?

- Sub-problem 2: Molecule generation and search
  - How can we generate molecules that have particular properties?
  - How can we effectively search over the space of possible molecules?

Image credit: Wengong Jin

Molecule representations

- Use GNNs on molecule graph structure.
- Baselines and other approaches:
  - LSTMs on sequence representations (i.e., SMILES strings).
  - Handcrafted features (e.g., molecular fingerprints)
- GNNs of choice include
  - graph isomorphism networks (GINs; Xu et al., 2019),
  - message-passing networks (MPNNs; Gilmer et al., 2017),
  - and hierarchical motif-based GNNs (Jin et al, 2020)



Jin et al., 2020's hierarchical approach

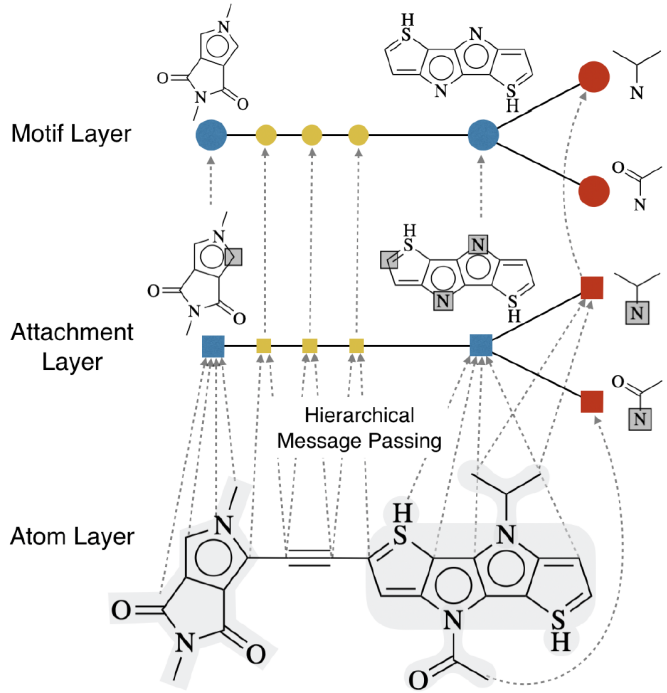


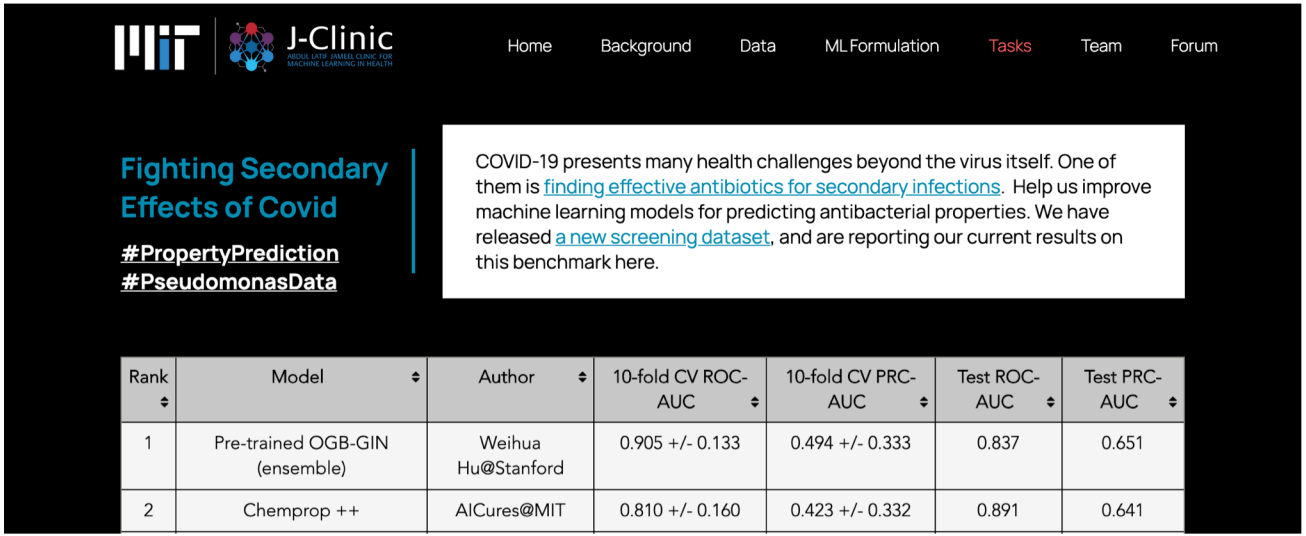
Image credit: Wengong Jin

Molecule representations

- Developing better molecule representations and models to predict properties is still an open challenge.
- The MIT AICures groups is currently running a challenge related to COVID-19:

You can participate!

https://www.aicures.mit.edu/tasks



Rank	Model	Author	10-fold CV ROC-AUC	10-fold CV PRC-AUC	Test ROC-AUC	Test PRC-AUC
1	Pre-trained OGB-GIN (ensemble)	Weihua Hu@Stanford	0.905 +/- 0.133	0.494 +/- 0.333	0.837	0.651
2	Chemprop ++	AICures@MIT	0.810 +/- 0.160	0.423 +/- 0.332	0.891	0.641

Image credit: Gómez-Bombarelli et al., 2018

Generating useful molecules

How do we generate molecules that might be good for drug discovery?

Approach 1: Latent space optimization

1. Train VAE to encode and decode molecule representations.
2. Train separate model (e.g., neural network or Gaussian process) to predict properties from latent embeddings.
3. Search in the latent space for points that have the desirable properties and decode.

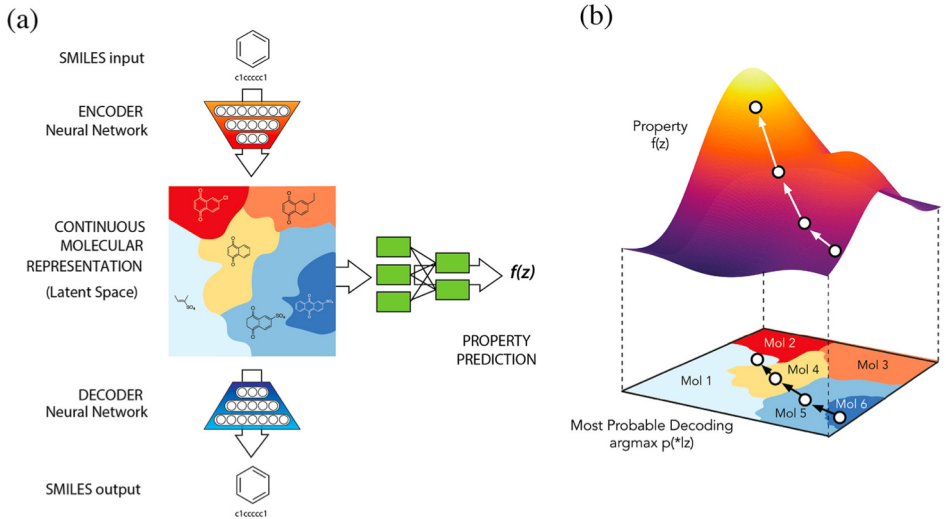


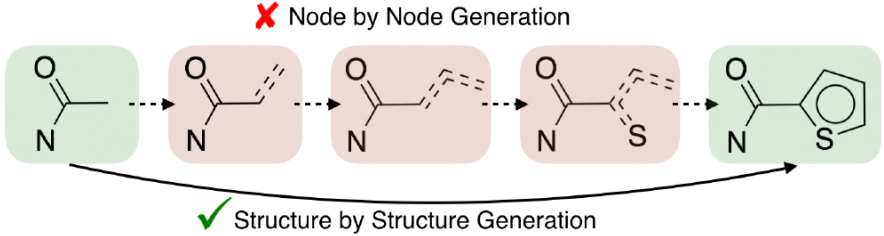
Image credit: Jin et al., 2018

Generating useful molecules

How do we generate molecules that might be good for drug discovery?

Approach 1: Latent space optimization

- Hierarchical encoders/decoders generally outperform simple GNNs/LSTMs in this setting.
- Current state-of-the-art: https://arxiv.org/pdf/2002.03230.pdf



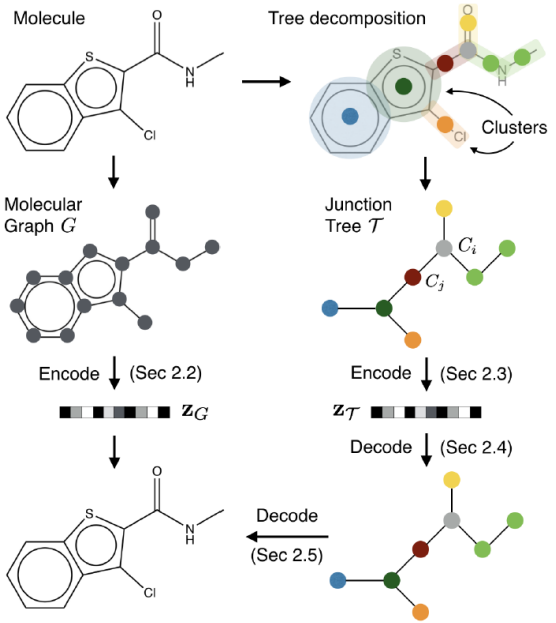


Image credit: Maksym Korablyov

Generating useful molecules

How do we generate molecules that might be good for drug discovery?

Approach 2: Search and reinforcement learning

- Treat the generation problem as a search problem.
- Actions correspond to adding motifs or atoms to the molecule.
- Rewards correspond to the properties and/or binding energy of the (partially) generated molecule.
- Current work: https://mila.quebec/en/ai-society/exascale-search-of-molecules/

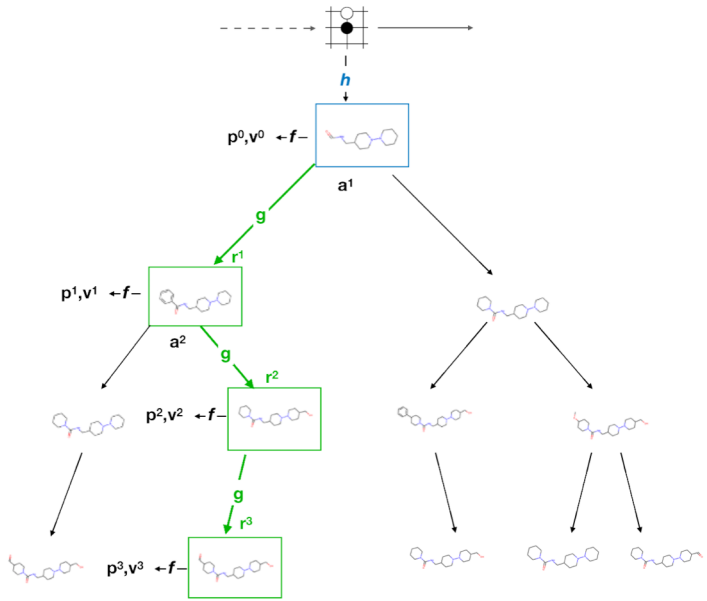


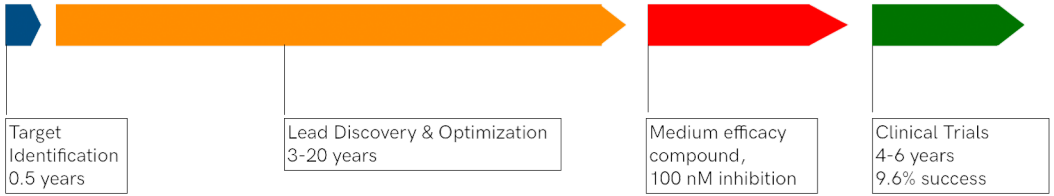
Image credit: Maksym Korablyov

Computational drug design: Challenges

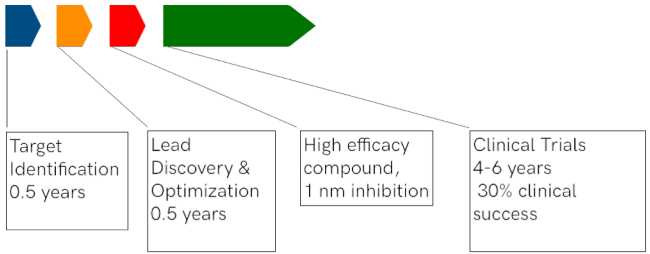
- Can use physics simulations or existing prediction models to evaluate candidate molecules. Still expensive and noisy.

- Computational drug design shortens the time to find promising drugs, but development is still **very time consuming.**

Typical Drug Development pipeline



LambdaZero aim Drug Development pipeline



Possible applications

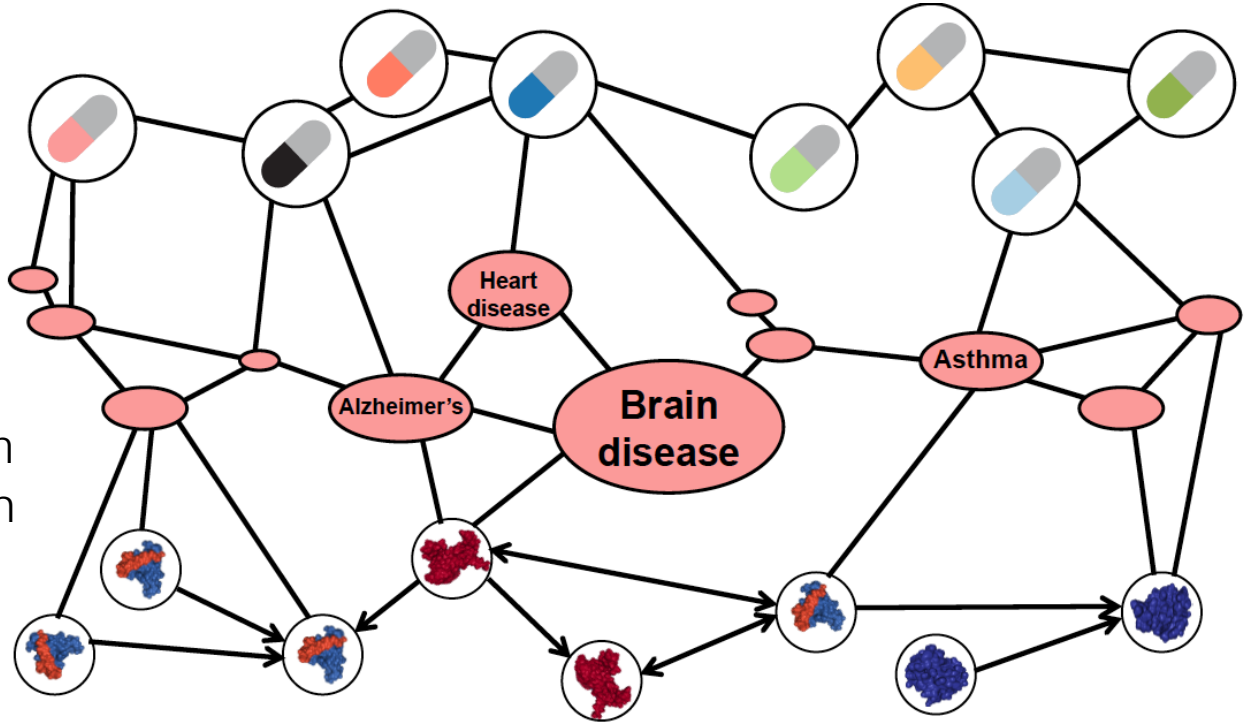
- Possible applications using this data include:
  1. Computational drug design
     - Can we design better antivirals to target COVID-19?
  2. Computational treatment design
     - Can we design better treatment strategies using existing drugs?
  3. Epidemiological forecasting
     - Can we better predict how and where infection rates will change over time?
  4. Demand forecasting and supply chain optimization
     - Can we forecast COVID-19 related demands to optimize supply chains?
  5. Outbreak tracking and tracing
     - Can we model and predict infection risk at the individual level?

Computational drug repurposing

- **Task:** Predict whether existing drugs will be useful to treat COVID-19
- Approach 1 (Structure-based):
- Leverage known information about the viral protein structure
  - Benefit: Leverages biological knowledge and can be highly specific
  - Challenges:
    - Difficult when we do not know the 3D structure.
    - High-risk of viral mutations rendering the drug ineffective.
  - Methodology for Approach 1 is analogous to the drug design problem.
- Approach 2 (Network-based):
- Leverage knowledge of biological interactions between drugs, diseases, and proteins
  - Benefit: Does not rely on only on known protein structures. Quite general.
  - Challenges:
    - Model predictions can be difficult to explain and interpret.
    - Difficult to integrate structural knowledge of biological processes

Network-based drug repurposing

- Drug-efficacy depends on various interactions:
  - Protein-protein
  - Disease-drug
  - Drug-drug
  - Disease-disease
  - …
- Drug repurposing is a form of **relation prediction** within this interaction network.



Network-based drug repurposing

- Drugs and diseases can be related based on their respective protein interactions.
- More overlap between subgraphs indicates stronger possible relationship.
- Identifying repurposing candidates based on protein network interaction is already happening for COVID-19 [Zhu et al,., Cell Discovery ‘20]

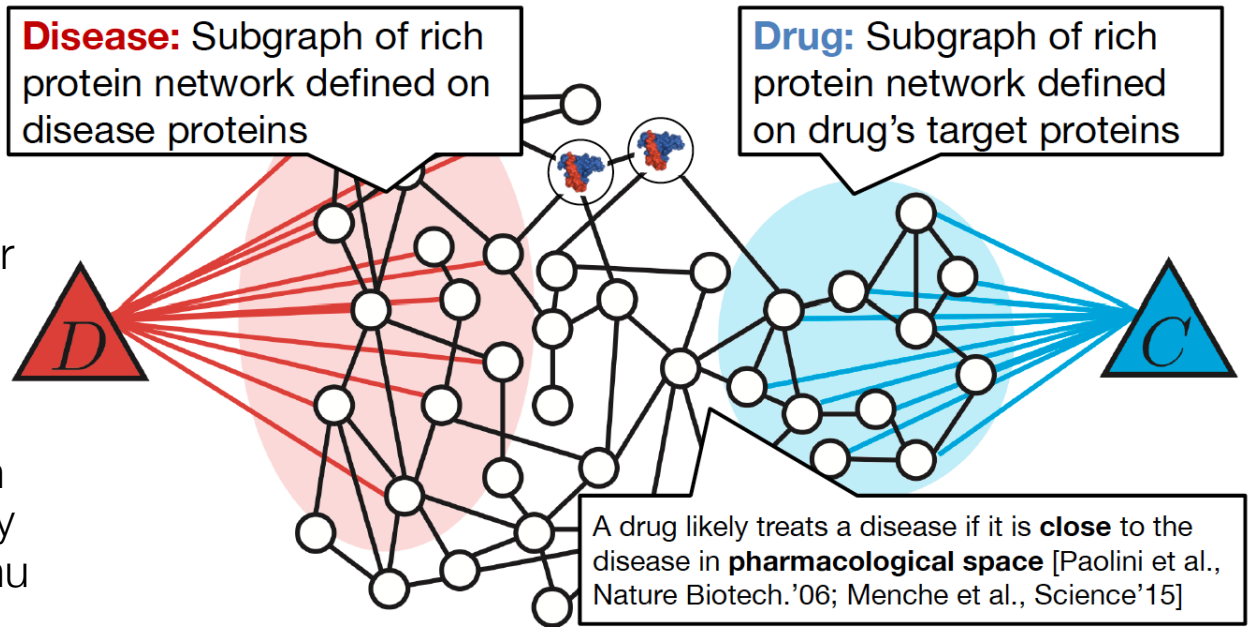
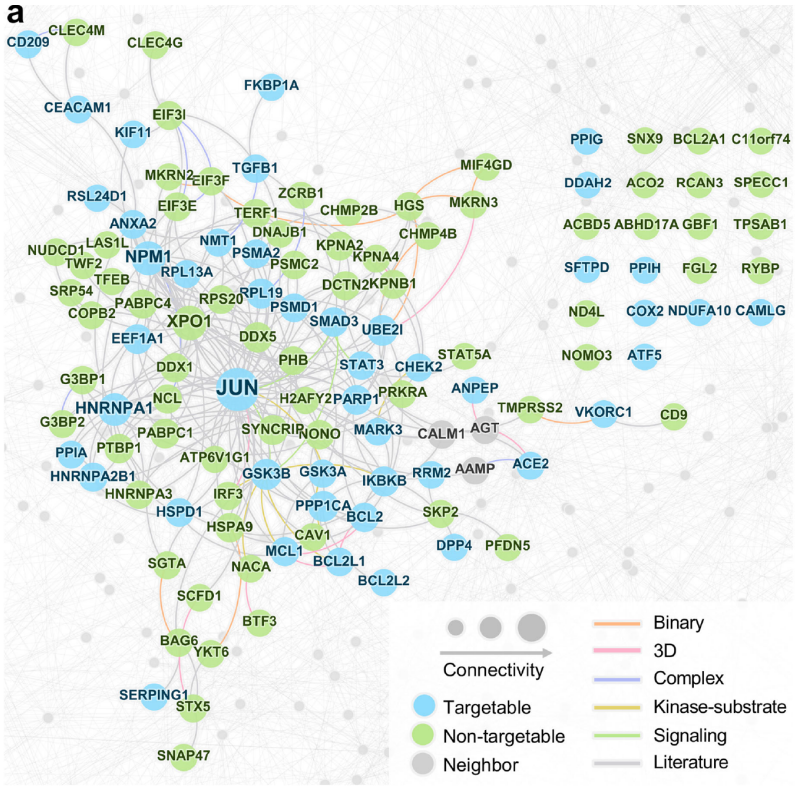


Image and data/analysis credit: Zhou et al., Cell Discovery '20

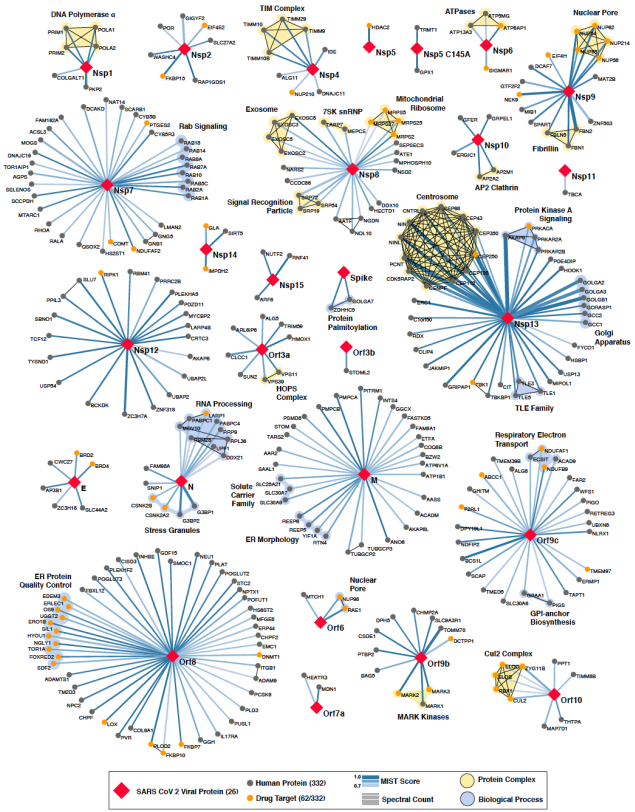
COVID-19 preliminary work [Zhou et al., 2020]

- Assembled set of proteins likely to interact with SARS-CoV-2 based on other known coronaviruses:
  - 4 different human coronaviruses (HCoVs): SARS-CoV-1, MERS-CoV, HCoV-229E, and HCoV-NL63
  - One mouse and one avian coronavirus
- All these proteins are embedded within the known human interactome.
- Identified possible drug candidates based on statistical overlap between drug and SARS-CoV-2 interaction subgraphs.



COVID-19 preliminary work [Gordon et al., 2020]

- Used affinity-purification mass spectrometry to identify proteins that interact with SARS-CoV-2 in human cells.
- Found 332 high confidence protein interactions between SARS-CoV-2 proteins and human cell proteins.
- Identified 67 druggable proteins (using existing drugs). These are either virus proteins or human proteins that directly interact with virus proteins.
- Released data in supplementary material: https://www.biorxiv.org/content/10.1101/2020.03.22.002386v3



Repurposing with GNNs

- Existing work on network-based COVID-19 drug repurposing uses on statistical heuristics (Zhou et al., 2020) or biological domain knowledge (Gordon et al., 2020).

- Can we do even more with graph representation learning?

- Basic idea:
  - Use GNNs to embed diseases and drugs.
  - Predict disease-drug interactions based on learned embeddings.

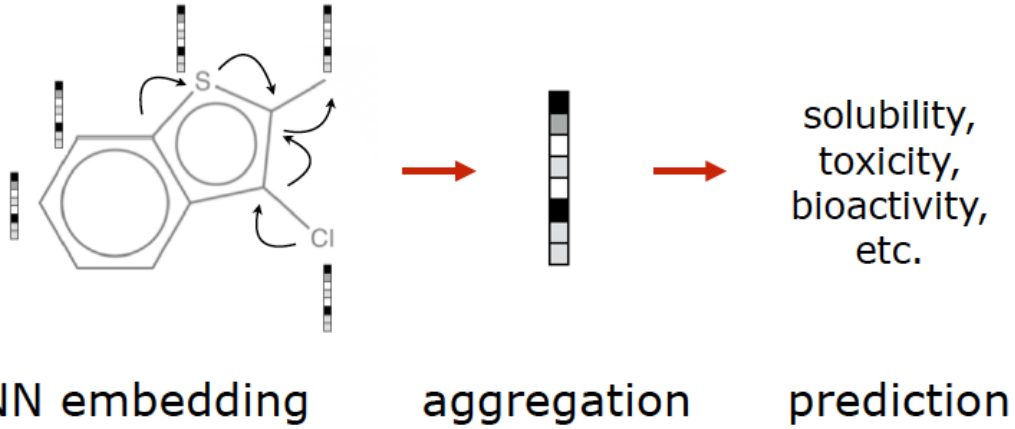
Repurposing with GNNs: Design decisions

- Representing drugs and diseases:
  - Option 1: Represent diseases and drugs as subgraphs (i.e., sets) of protein nodes
    - Can leverage pooling methods and set-based learning methods.
    - Message-passing and other GNN operations defined only over proteins.
    - Drug-disease interactions are used in the loss function (but not in message-passing)
  - Option 2: Heterogenous and multi-relational graph approach
    - Drugs and diseases are represented as nodes.
    - Add edges between drugs and the proteins they interact with (same for diseases).
    - Can add other types of biological interactions (e.g., drug-drug interactions)
- Output model and loss function
  - Option 1: Treat it as a classification task
    - E.g., input to a final MLP layer is concatenation of drug embedding and disease embedding. Output is probability of interaction, which is evaluated using a binary cross entropy loss.
  - Option 2: Treat it as a relation prediction task
    - Use any edge scoring function from graph representation learning (e.g., dot product) to score likelihood of interaction between drug node and disease node. Train using contrastive (e.g., max-margin) loss.

Image credit: Wengong Jin

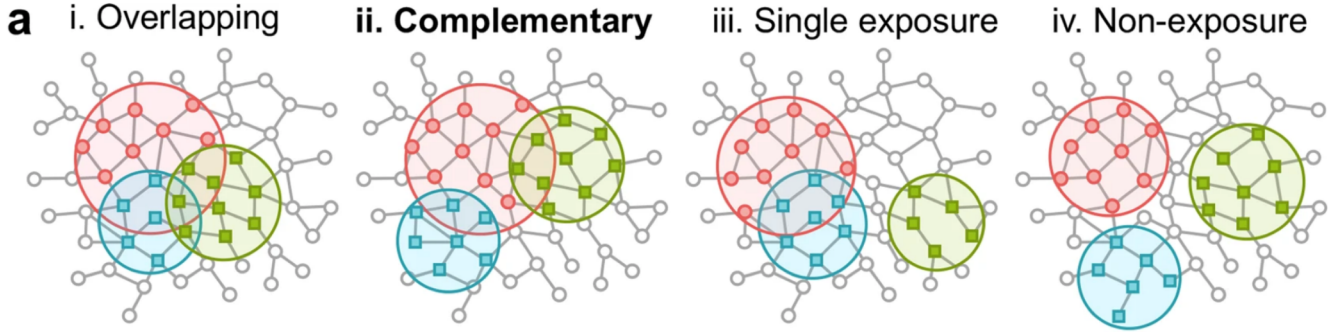
Repurposing with GNNs: Molecule structure

- Recall: Drugs can be represented as molecule graphs.
- We can run GNNs on molecule graphs to get drug embeddings.
- Could be combined with the PPI network information.
  - E.g., use molecule embeddings to get initial features for drug nodes in the interaction network).



Repurposing with GNNs: Combinations

- Often drug in combinations (i.e., cocktails) have synergistic effects.
- Machine learning methods can be especially useful here.
  - There are around 10^4 approved drugs, which could feasibly all be screened in-vitro within months for COVID-19.
  - But there are 10^8 possible pairs of drugs, which is far too many to test in the lab!
- Finding good combination candidates often relies on the notion of complimentary exposure (in the drug-protein interaction subgraphs):



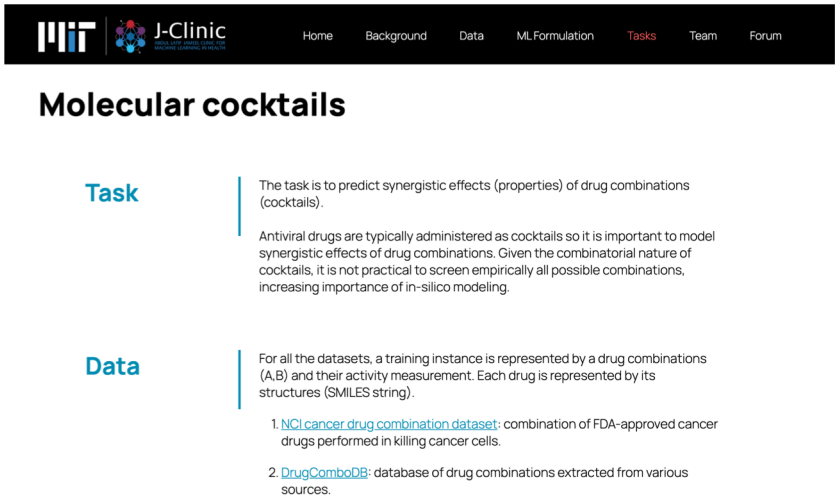
Repurposing with GNNs: Combinations

- Often drug in combinations (i.e., cocktails) have synergistic effects.
- From a ML perspective, we simply need to work with pairs of drug embeddings and modify our model/loss accordingly!

MIT AICures has also released data for predicting the efficacy of drug combinations.

https://www.aicures.mit.edu/tasks

The data is not specific to COVID-19, but the models we develop will be useful once more data on COVID-19 is available!



Repurposing with GNNs: Data sources

- The COVID-19 interactome data released by Gordon et al., 2020: https://www.biorxiv.org/content/10.1101/2020.03.22.002386v3
- The property prediction and combination efficacy prediction tasks released by the MIT AICures team: https://www.aicures.mit.edu/tasks
- COVID-19 genome data: https://www.gisaid.org/
- The Stanford BioSNAP dataset, which contains large biomedical knowledge graph extracted and cleaned from public sources: http://snap.stanford.edu/biodata/

Possible applications

- Possible applications using this data include:
  1. Computational drug design
     - Can we design better antivirals to target COVID-19?
  2. Computational treatment design
     - Can we design better treatment strategies using existing drugs?
  3. Epidemiological forecasting
     - Can we better predict how and where infection rates will change over time?
  4. Outbreak tracking and tracing
     - Can we model and predict infection risk at the individual level?
  5. Demand forecasting and supply chain optimization
     - Can we forecast COVID-19 related demands to optimize supply chains?

Epidemiological forecasting

- **Task:** Forecast infection rates, hospitalization rates, and other COVID-19 impacts.
- **Benefits:** Allow regions to better prepare, plan, and allocate resources.
- **Risks and ethical concerns:**
  - Expert epidemiologists have released models, and additional predictions might confuse the public.
  - Public data is unreliable.
- **How could GNNs help?**
  - Leverage heterogenous data (e.g., Twitter, social network, news)
  - Do not require strong underlying assumptions.
- **Possible data sources:**
  - Johns Hopkins CSSE Data: https://github.com/CSSEGISandData/COVID-19
  - Links and data discussed at: https://www.kaggle.com/c/covid19-global-forecasting-week-2
  - More links and data: https://www.wolframcloud.com/obj/examples/COVID19Resources

Outbreak tracking and tracing

- **Task:** Predict infection risk at the individual level based.
- **Benefits:** Allow more precise self-isolation directives.
- **Risks and ethical concerns:**
  - Would require a large amount of potentially sensitive data.
  - Risks of not being dismantled after COVID-19.
- **How could GNNs help?**
  - Leverage heterogenous data (e.g., Twitter, social network, news)
  - Do not require strong underlying assumptions.
- **Possible data sources:**
  - Several governments and academic groups have discussed the possibility of releasing privacy preserving applications (e.g., https://www.howwefeel.org/, https://science.sciencemag.org/content/early/2020/03/30/science.abb6936)

Demand forecasting for supply chain

- **Task:** Forecast demand for different materials, supplies, and products.

- Risks and ethical concerns:
  - No obvious ethical risks, but some data might be private.

- How could GNNs help?
  - Leverage heterogenous data (e.g., Twitter, social network, news)
  - Do not require strong underlying assumptions.

- Possible data sources:
  - Impactful data would require substantial governmental or industry collaboration.

Spatio-temporal GNNs

- Epidemiological forecasting
- Outbreak tracking and tracing
- Demand forecasting and supply chain optimization

What do they have in common?

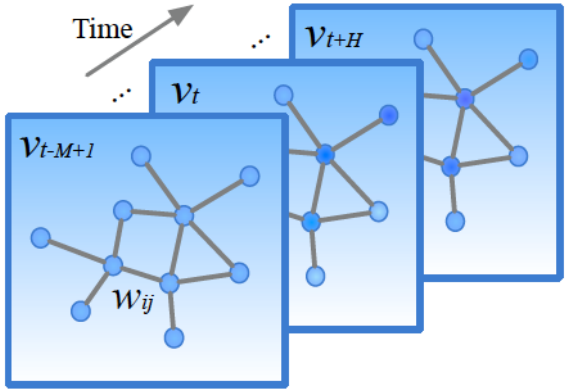
- Heterogenous relational data
- Temporal information and changes
- Node-level predictions

Spatio-temporal GNNs are a useful model for all three of these problems!

Image credit: Yu et al., 2018

Spatio-temporal GNNs

- Assumptions:
  - Use nodes to represent locations, individuals, etc.
  - Edges represent relationships (e.g., contacts between individuals or travel flow between locations).
  - Node attributes and edges can change across time-steps.
- Model intuition:
  - Apply a GNN at each time-step.
  - Add recurrent connections between time-steps.
  - For example,



Message from neighbors at time-step t in layer k

Neighbor embeddings at time-step t

$$\mathbf{m}_u^{(k,t)} = \text{AGGREGATE}(\{\mathbf{h}_v^{(k-1,t)}, \forall v \in \mathcal{N}_t(u)\})$$

$$\mathbf{h}_u^{(k,t)} = \text{UPDATE}(\mathbf{h}_u^{(k-1,t)}, \mathbf{h}_u^{(k,t-1)}, \mathbf{m}_u^{(k,t)})$$

New embedding at time-step t in layer k

Previous embedding from layer (k-1) at the same time-step

Embedding from same layer at the previous time-step

Message from neighbors at time-step t in layer k