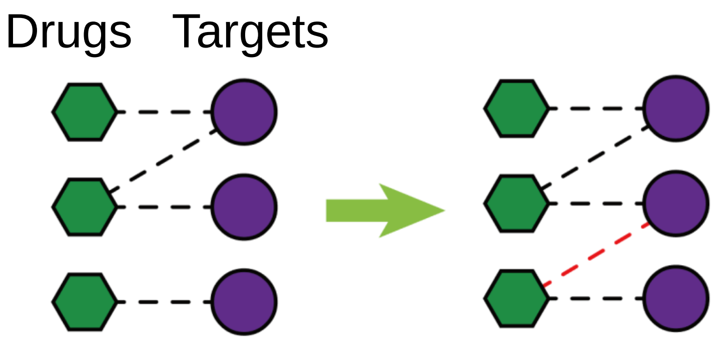
Homework: Drug repositioning for SARS-CoV-2: DTI/CPI prediction

2020.4.28

Basic concepts in DTI / CPI prediction



- Drugs: a subset of compounds that are approved or in clinical trails
- Targets: a subset of proteins that are druggable or disease-related
- DTI (binary): chemical interaction
- Affinity (scalar): strength of the interaction

- Drug-target interaction (DTI) prediction: drug repositioning
  - Goal: predict new links in the drug-target interaction network
  - Information: knowledge graphs
  - Scale: thousands of drugs/targets

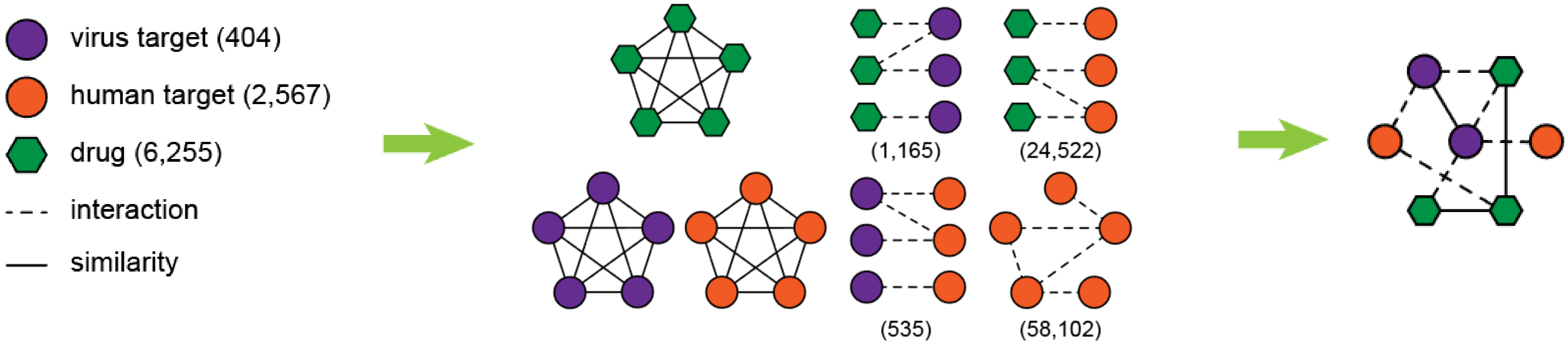
- Compound-protein interaction (CPI) prediction: drug screening/ repositioning
  - Goal: classification/regression
  - Information: molecular compositions of proteins and compounds
  - Scale: often involves much more compounds

Input of DTI prediction: the heterogeneous network

Types of nodes and edges

Individual networks

Heterogeneous network

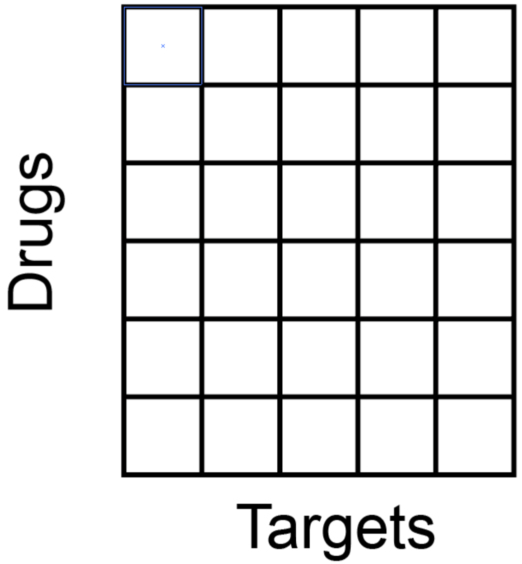


Interactions: binary values

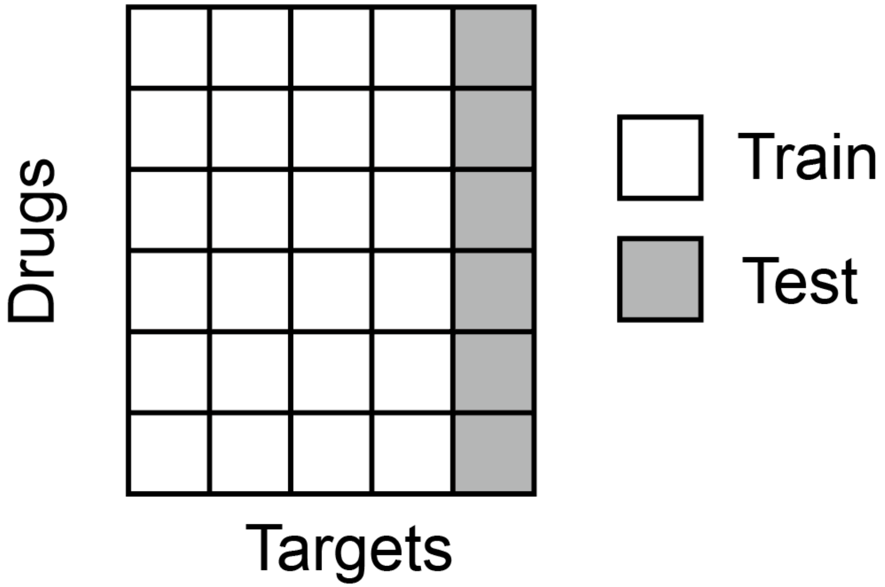
Similarities: scalar values

Evaluate DTI model using the virus target-drug network

The individual networks are stored in matrix format, and edges in dictionaries, e.g.,

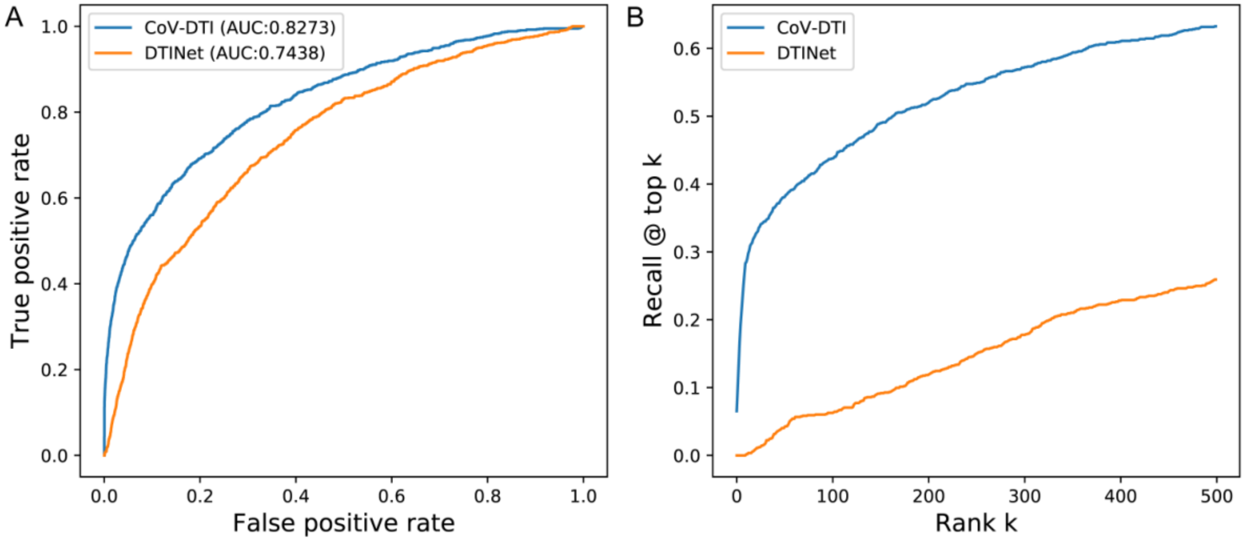


Perform cross-validations according to the target dimension:



Evaluate DTI model using the virus target-drug network

If all the positive and negative virus DTIs are used, please use AUC and recall@top k (true positives in top k predictions) for evaluation:



If negative DTIs are sampled in a fixed (and relatively balanced) ratio, e.g., 1:1 or 1:10, please use AUC and AUPR for evaluation.

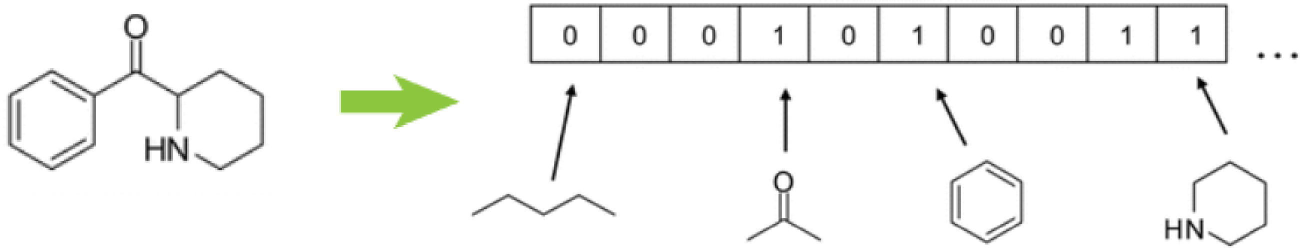
Input of CPI prediction

- Classification: list of (proteins, compounds, 0/1 indicating interaction)
- Regression: list of (proteins, compounds, scalar value indicating affinity)

- Proteins are represented by primary amino-acid sequences
- Compounds are represented by InChIs (International Chemical Identifiers)

- Protein encoding and feature extraction:
  - One-hot, word2vec, learnable embeddings, …
  - CNN, RNN, …

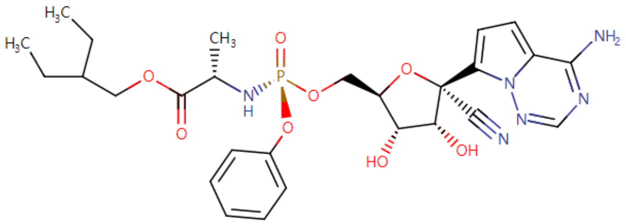
- Compound encoding and feature extraction:
  - Fingerprint: substructures are hashed into bit-vectors (http://www.rdkit.org/docs/source/rdkit.Chem.rdMolDescriptors.html#rdkit.Chem.rdMolDescriptors.GetMorganFingerprintAsBitVect)



  - SMILES (simplified molecular input line entry specification) string

    CCC(CC)COC(=O)[C@H](C)N[P@](=O)(OC[C@H]1O[C@](C#N)([C@H](O)[C@@H]1O)C1=CC=C2N1N=CN=C2N)OC1=CC=CC=C1

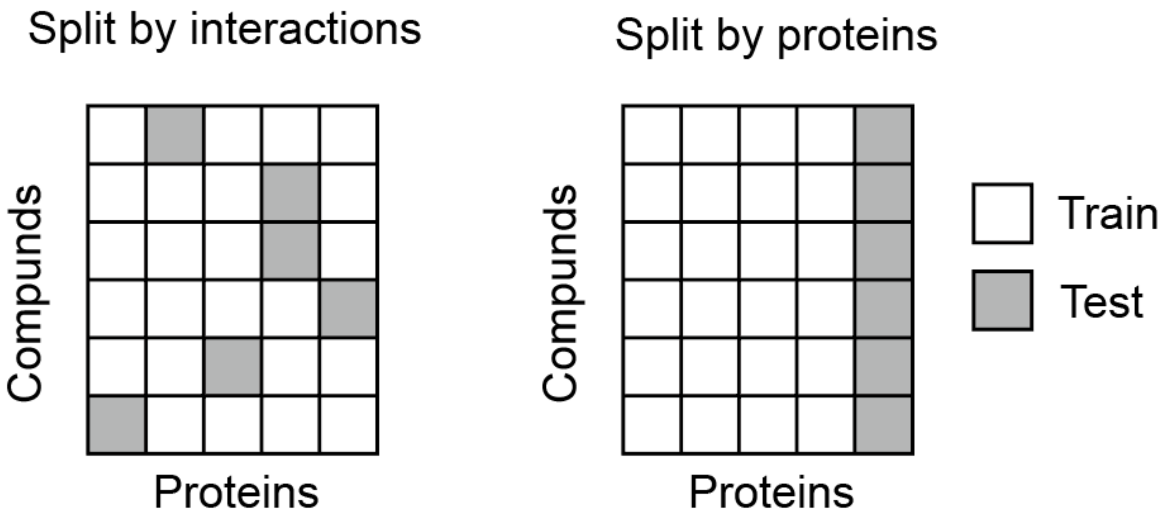
  - Graph:



  - Models: CNN, RNN, graph neural networks, …

Evaluate CPI model

- Perform cross-validations according to pairs or proteins:



- Classification metrics: AUC, AUPR
- Regression metrics: root mean squared error (RMSE), Pearson’s correlation

Predict drugs for SARS-CoV-2 proteins

- Choose one task (DTI or CPI) to build your model

- Predict potential active drugs among 6255 candidates, for 2-3 viral proteins

- Possible evaluation through recently reported active drugs: remdesivir, chloroquine, nitazoxanide, nafamostat, favipiravir… (https://doi.org/10.1038/s41422-020-0282-0)

- Present top 10 drugs, and their drug names, original indications and original targets (All the information can be found at https://www.drugbank.ca/)