KNOWLEDGE GRAPH DATABASE FOR COVID-19 DRUG DISCOVERY

Joey (Aryaman) Dubey

Guided by: Reynold Cheng and Xiaodong LI

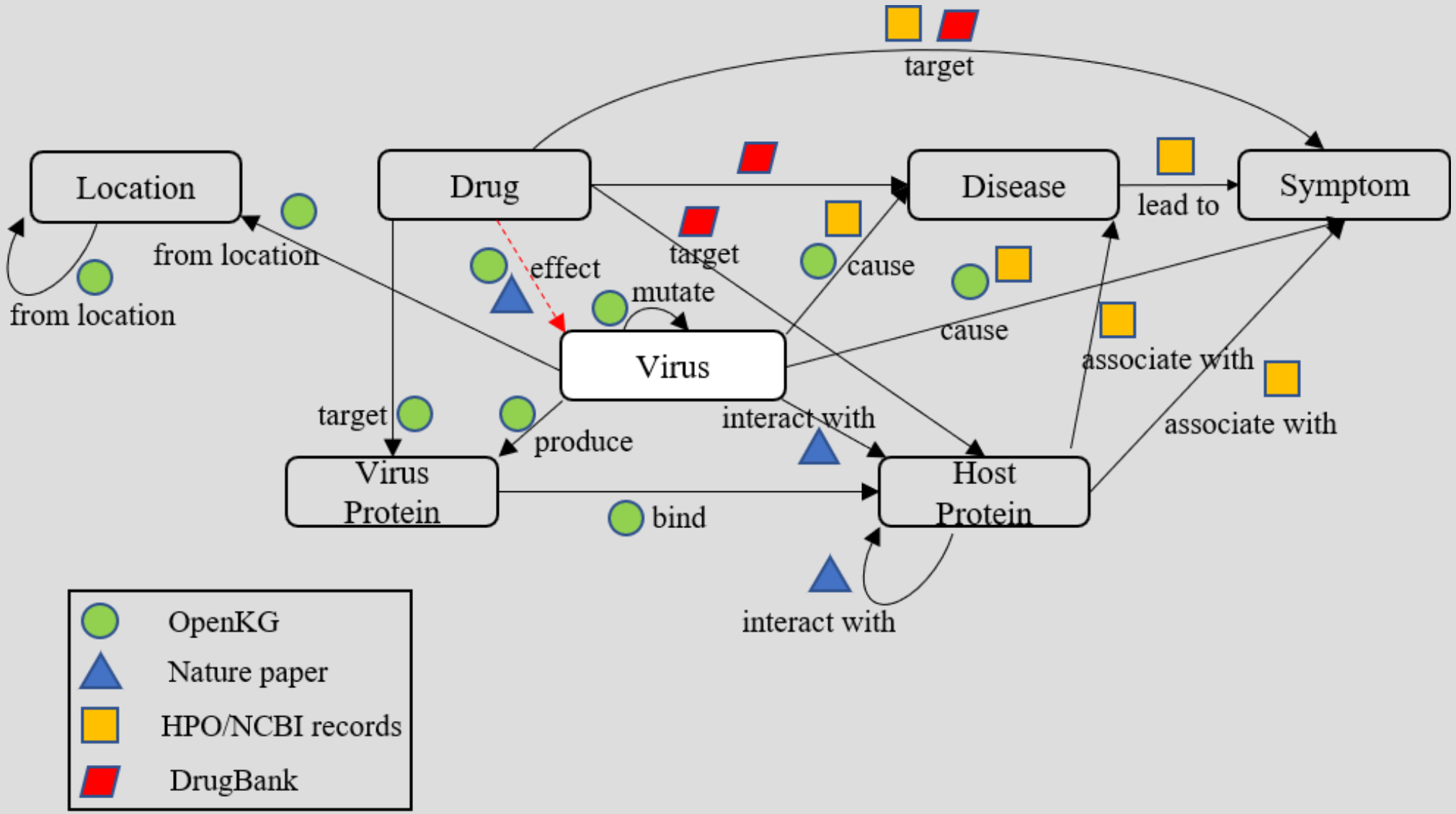
MOTIVATION

- Very relevant topic. SARS-CoV-2 outbreak (COVID-19) has become a pandemic
- No effective COVID-19 drug treatment known yet
- Discovery of new drugs very time consuming + expensive
- Repurposing of existing drugs using network based strategies effective

KNOWLEDGE GRAPH METHOD

- Combining different types of related data: higher prediction making accuracy
- Network-based data models: Heterogenous Information Networks (HINs)
- HINs used in various link prediction applications
- Eg of HIN: Knowledge Graph – used for our COVID-19 drug prediction

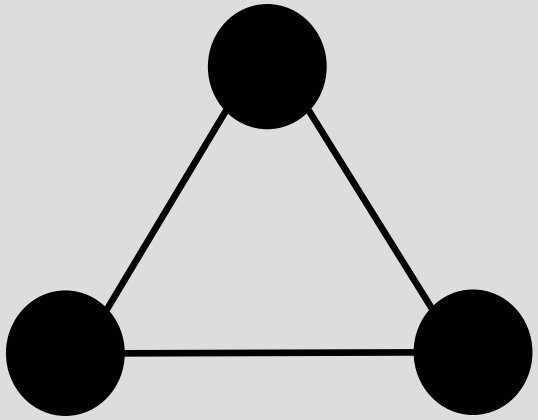
KNOWLEDGE GRAPH SCHEMA



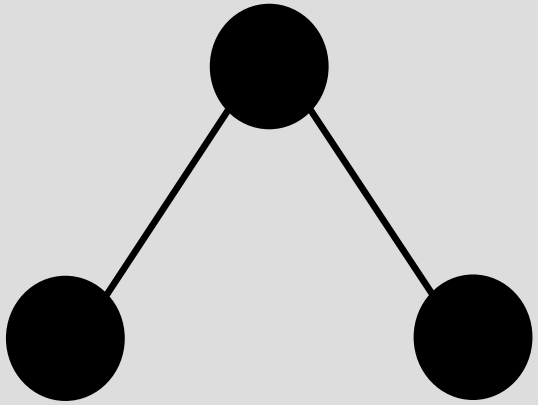
1. MOTIF MATCHING

- Knowledge Graph too complex & large (several thousand nodes + edges)
- Decompose recursively into small subgraph patterns of size *k* nodes: called motifs
- For small values of *k*, some fast and efficient motif enumeration algorithms exist
- We generate motifs with $k = \{1, 2, 3, 4\}$.
  *(k = 5 to be implemented in future)

MOTIF TYPES (3-NODE)



i. Triangle



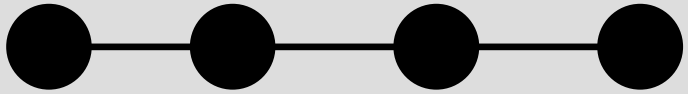
ii. 3-Star

ALGORITHM (MOTIF MATCHING)

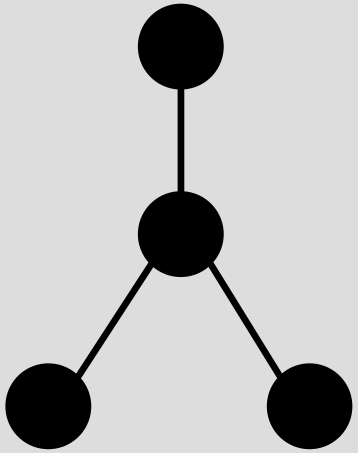
3 node motifs:

```
get seed node i
for node i:
        get neighbours j, k of i
        if edge(j, k) exists:
                return triangle
        else:
                return 3-star
```

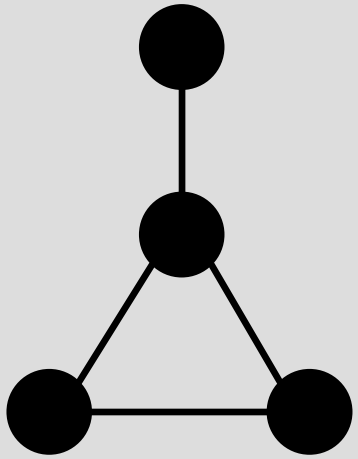
MOTIF TYPES (4-NODE)



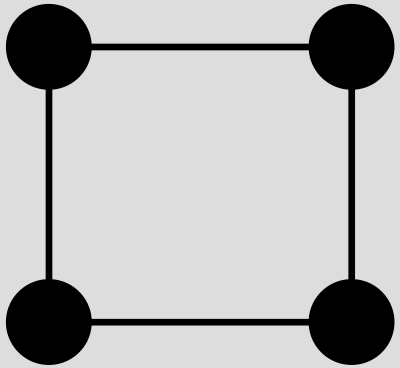
i. 4-Path



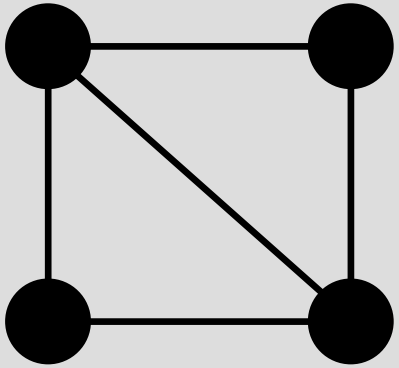
ii. 4-Star



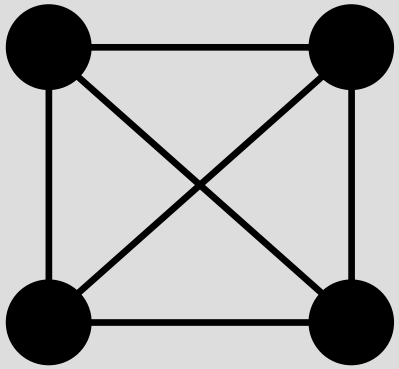
iii. Tailed Triangle



iv. Rectangle



v. Diamond



vi. 4-Clique

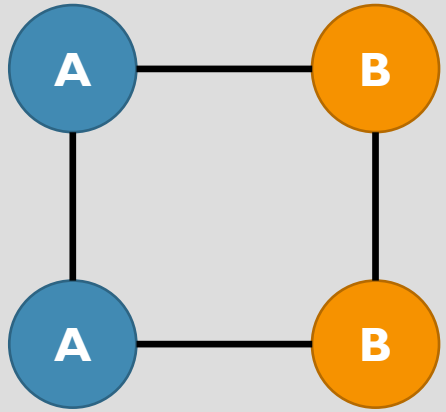
ALGORITHM (MOTIF MATCHING)

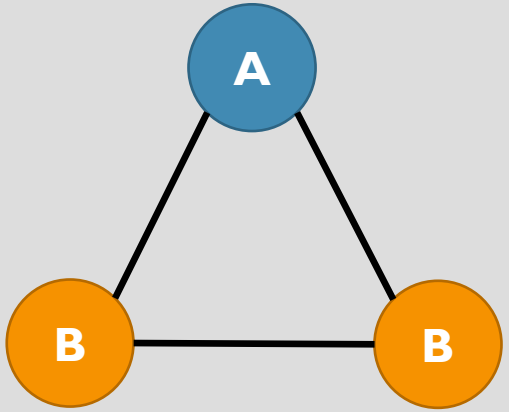
4 node motifs

```
if maxDegree == 3:
        get seed node i
        for node i:
                get neighbours j, k, p of i
                if edge(j, k) exists:
                        if edge(j, p) exists:
                                if edge(k, p) exists:
                                        return 4-clique
                                else:
                                        return diamond
                        else:
                                return tailedTriangle
                else:
                        return 4-star
else:
        get first seed node i
        get neighbours j, k of i
                get neighbour p of second seed j:
                        if edge(k, p) exists:
                                return rectangle
                        else:
                        return 4-path
```

ALGORITHM (MOTIF MATCHING)

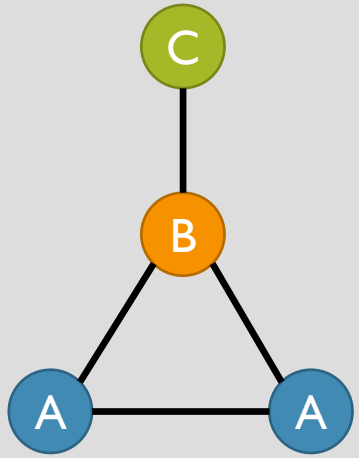
- Motifs can have multiple labels of same kind
- Overcounting risk
- Example: Rectangle motif with labels A & B, triangle motif with labels A & B





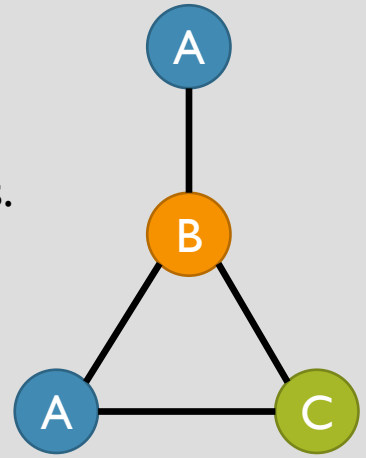
ALGORITHM (MOTIF MATCHING)

- Too avoid overcounting risk, we use concept of **orbits**
- Nodes whose configuration (degree, neighbours) remain same if swapped with each other, are considered in same orbit
- Example: consider tailed triangles with labels A, B, C as shown



The configuration of label A nodes remains same, if swapped. Same orbit. **Overcounting risk**.

The configuration of label A nodes changes, if swapped. Different orbits. **No overcounting risk**.



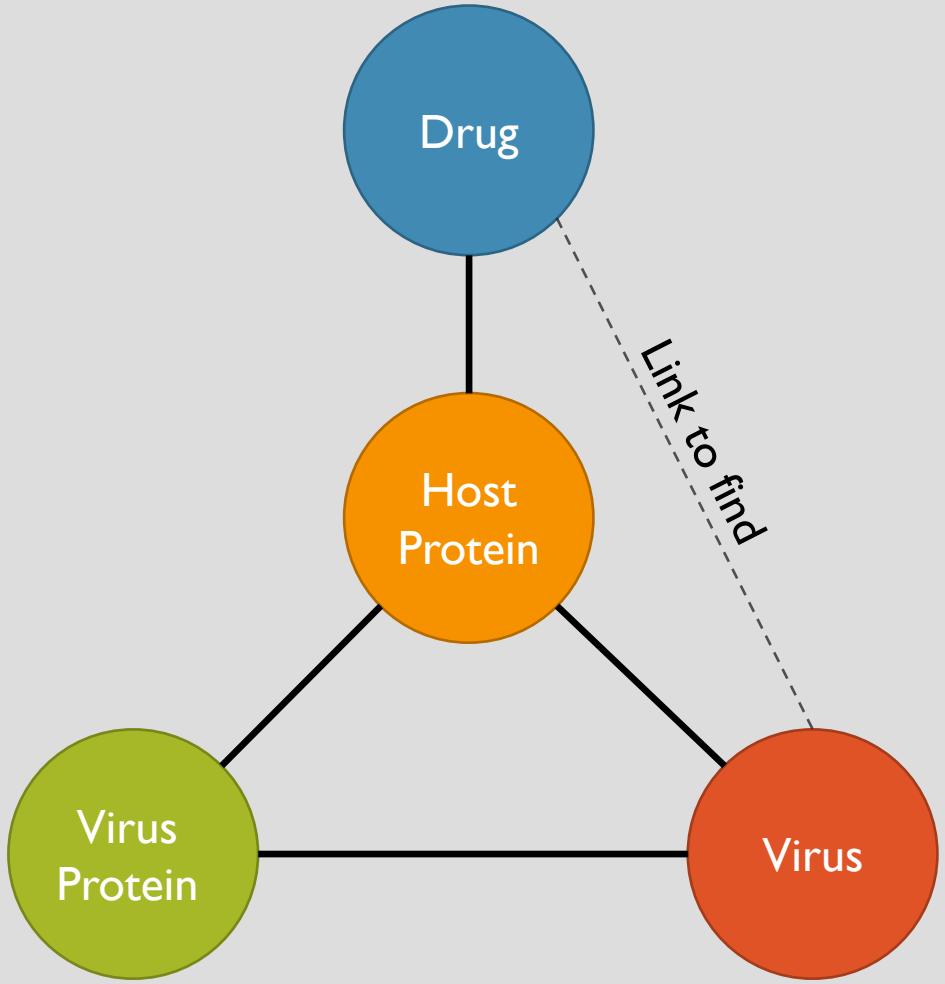
2. MOTIF BASED LINKED PREDICTION

- Predict which link is likely to appear in graph
- Prediction made by studying topological features of graph edges & nodes
- Most features rely on neighbourhood of nodes
- Higher order motifs ($k > 3$) ➔ higher prediction accuracy
- **Motif feature vector** generation

2. MOTIF BASED LINKED PREDICTION

- Nodes in knowledge graph: Drug, Disease, Host Protein, Symptom, Virus, VirusProtein, Strain, Location

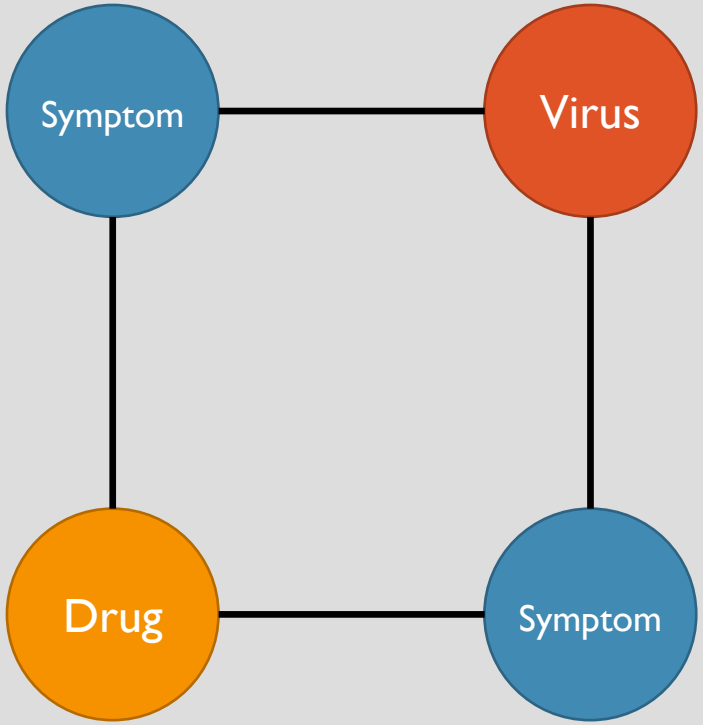
- Link to be predicted:
  Drug (?) <====> Virus (SARS-CoV-2)



Hypothetical motif for visualisation

LINK PREDICTION

- Current approach: manually select "interesting motifs"
  (must contain "Drug" node & "Virus" node)
- Higher motif score → greater chance of Drug-Virus link existence
- Present algorithm: measure frequency of particular drugs in "interesting" motifs; higher frequencies → better candidate drugs

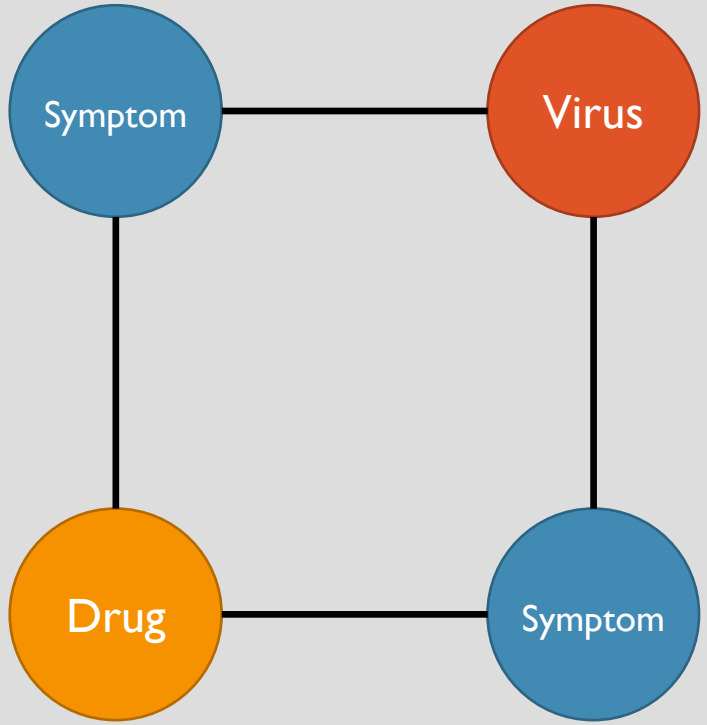


LINK PREDICTION

Consider the rectangle motif depicted in previous slide.

- Occurs 174 times in KG
- Check frequencies of drugs occurring in motif:
  Drug A frequency: 164
  Drug B frequency: 10

- Algorithm predicts higher likelihood of Drug-Virus link for Drug A.

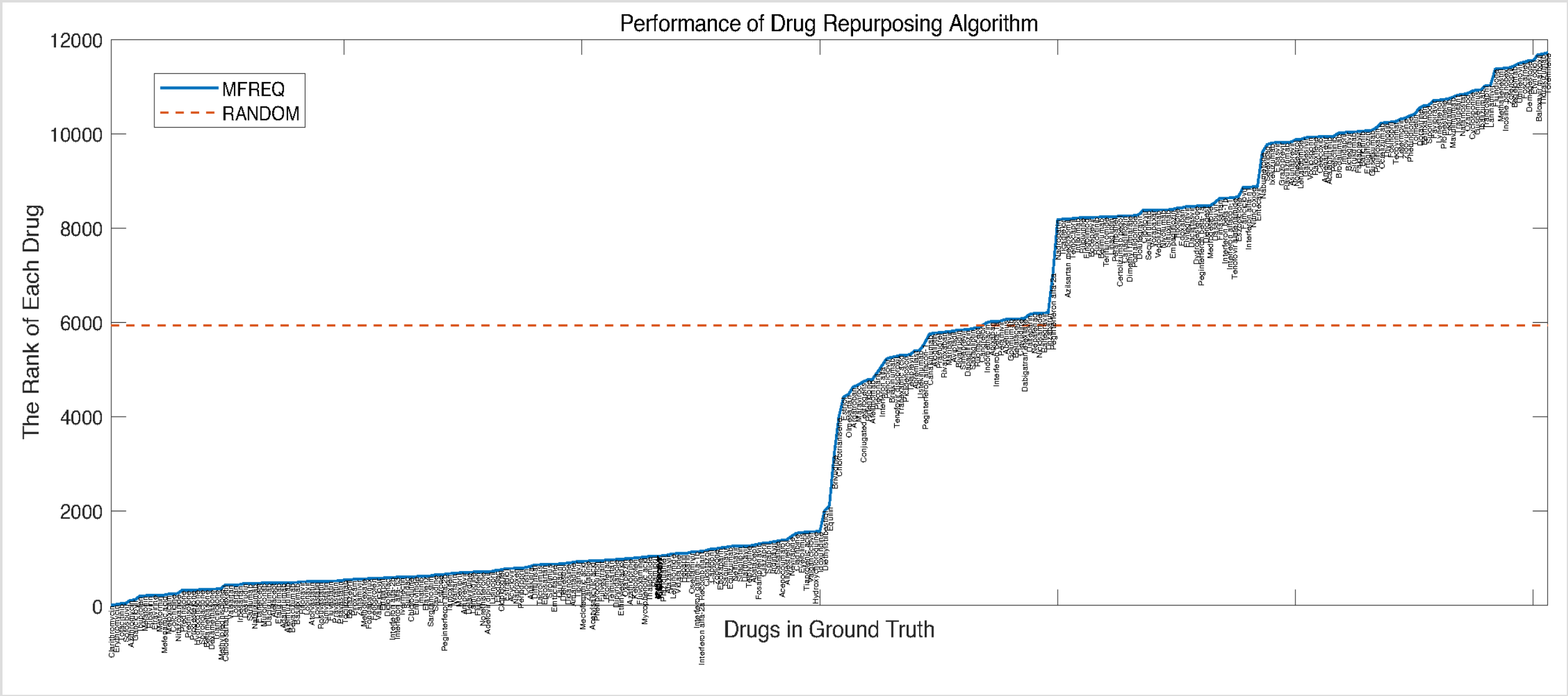


LINK PREDICTION

- Previous step gives some interesting results for further use

Present algorithm, next step:

- use “interesting” motifs to generate motif feature vector
- For each drug, sum all the corresponding motif feature vector elements
- Resulting sum is drug’s score
- Higher sum > greater chance of Drug-Virus link


Performance of Drug Repurposing Algorithm
MFREQ
RANDOM
The Rank of Each Drug
Drugs in Ground Truth
0
2000
4000
6000
8000
10000
12000

LINK PREDICTION

Future plan:

- Motif Frequency algorithm outperforms random algorithm
- But can be more accurate
- Develop another algorithm
- Apply deep learning models using KG motifs to train new algorithm

THANK YOU