Knowledge Graphs & Drug Repurposing

Preparatory Work for the Master Thesis 2024-25

Siddharth Sahay

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Outline

1. Drug Repurposing
2. Knowledge Graphs & Knowledge Graph Embeddings
3. State-of-the-art
4. Biomedical KGs
5. Evaluation metrics
6. Main challenges
7. Thesis roadmap

Drug Repurposing and Discovery

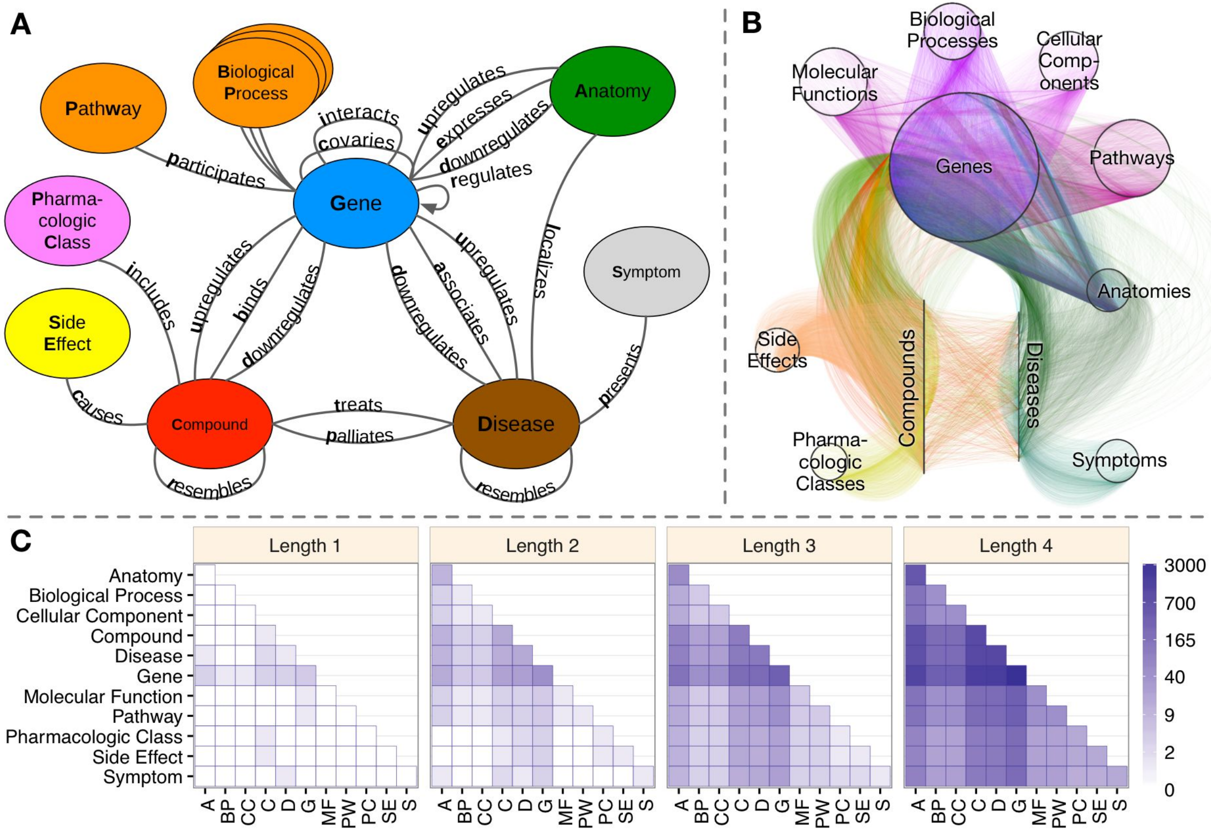
- ~7000 rare diseases; <6% have approved therapy
- $2.5B and 10+ years per rug
- Repurposing can cut costs and save time, drastically
- Drug-disease search space is huge
- KGs + KGEs organise and explore this space

Knowledge Graphs

- Consists of triples (head, relation, tail)
- Example: (Luke Skywalker, SonOf, Darth Vader)

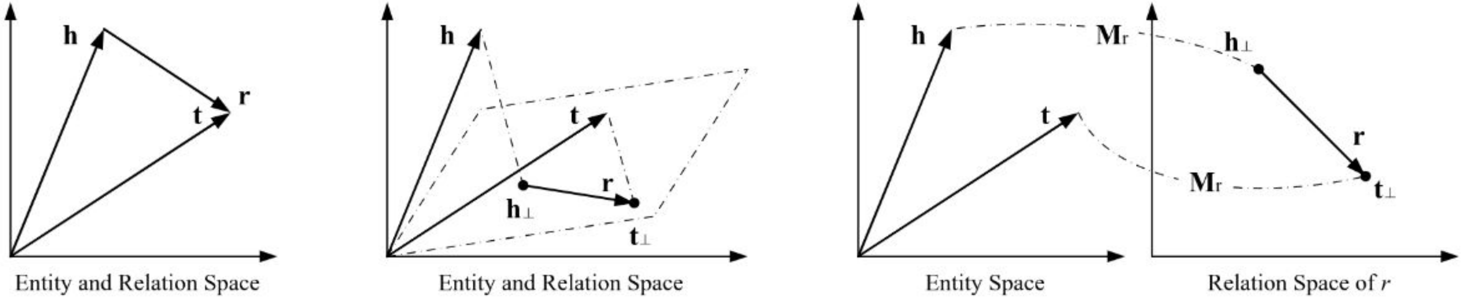
- Nodes: drugs, diseases, genes, pathways, edges
- Edges: relationships between these nodes
- Great for human intuition and biomedical knowledge representation

Hetionet KG


A
Pathway
Biological Process
Anatomy
Pharma-cologic Class
Gene
Symptom
Side Effect
Compound
Disease
participates
interacts
covaries
upregulates
expresses
downregulates
regulates
includes
upregulates
binds
downregulates
downregulates
associates
upregulates
localizes
presents
causes
treats
palliates
resembles
resembles
B
Molecular Functions
Biological Processes
Cellular Comp-onents
Genes
Pathways
Anatomies
Side Effects
Compounds
Diseases
Pharma-cologic Classes
Symptoms
C
Length 1
Length 2
Length 3
Length 4
Anatomy
Biological Process
Cellular Component
Compound
Disease
Gene
Molecular Function
Pathway
Pharmacologic Class
Side Effect
Symptom
A
BP
CC
C
D
G
MF
PW
PC
SE
S
3000
700
165
40
9
2
0

Knowledge Graph Embeddings

- Project relations into high-dimensional vector space
- Easier for ML models to use for link prediction
- Various methods: scoring function-based, path-based and semantic matching models

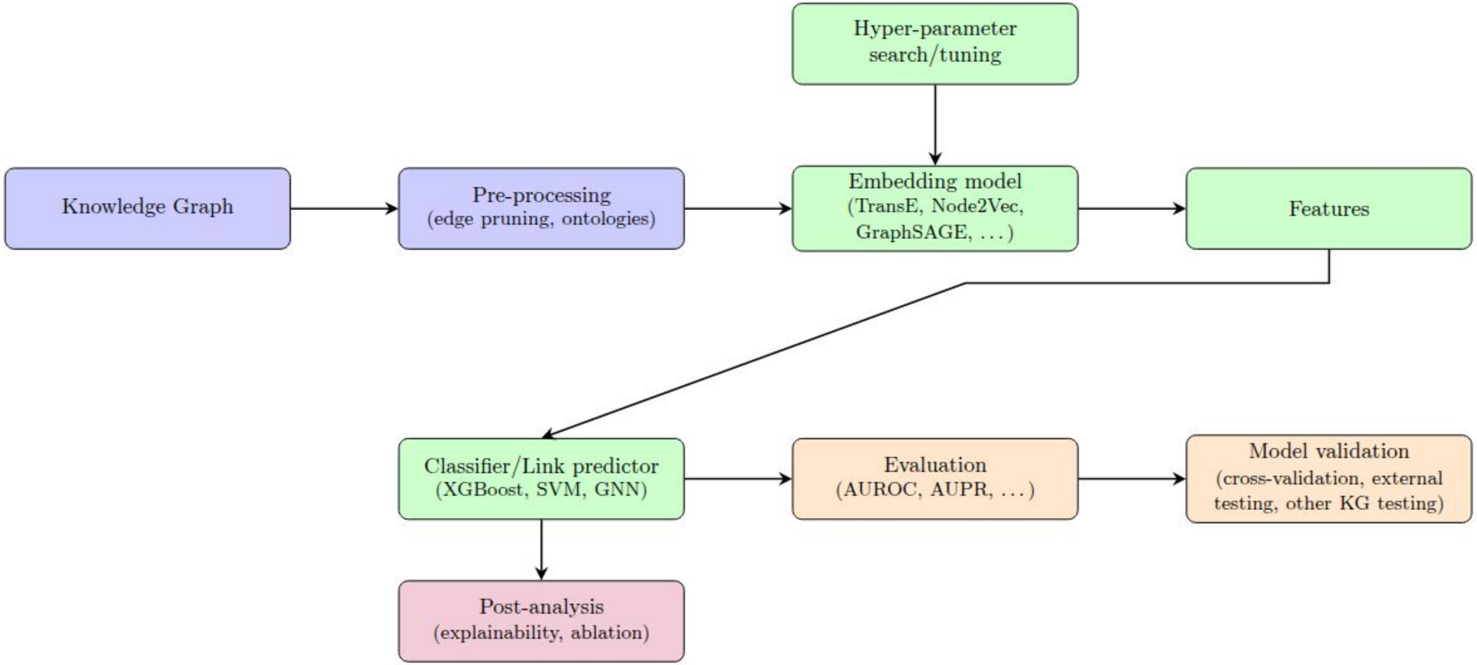


(a) TransE.

(b) TransH.

(c) TransR.

Generalised pipeline for Drug Repurposing


Hyper-parameter search/tuning
Knowledge Graph
Pre-processing (edge pruning, ontologies)
Embedding model (TransE, Node2Vec, GraphSAGE, . . .)
Features
Classifier/Link predictor (XGBoost, SVM, GNN)
Evaluation (AUROC, AUPR, . . .)
Model validation (cross-validation, external testing, other KG testing)
Post-analysis (explainability, ablation)

State of the art

- Traditional ML methods: DT2Vec+
- Random-walk based: DREAMwalk, AnyBURL
- Deep Learning (GNN) based: GDRNet, DRAGNN, EKGDR, DTD-GNN
- LLM based: DrugChat, MoCoSA, LMKE
- Other: RPath, PoLo, GNBR

eXplainable AI & Interpretability

- XAI makes ML models more transparent and understandable
- Many methods:
  - Path-based reasoning
  - Subgraph extraction
  - Logical pattern recognition
  - Attention interpretation with GATs
  - Counterfactual reasoning

Key biomedical KGs

- Hetionet
- PharMeBINet
- Bioteque
- Clinical Knowledge Graph (CKG)
- BOCK
- Many more that are publicly available

Evaluation metrics

- AUROC, AUPR
- Hits@K
- Mean Rank and Mean Reciprocal Rank

Limitations

- Bias towards PPI
- Data incompleteness
- Scalability
- Beyond second-order neighbourhoods
- Interpretability

Thesis roadmap

1. Baseline benchmarking
    a. Systematically compare pipelines
    b. KGEs: TransE, DistMult, random-walk based
    c. Classifiers: XGBoost, SVMs, GNNs
2. Optimisations and Oligogenic extension
    a. Hyperparameter search
    b. Integration with BOCK
3. Designing a novel method
    a. Fill all gaps in baseline
    b. Experiment further with GNNs
4. Testing and writing

References