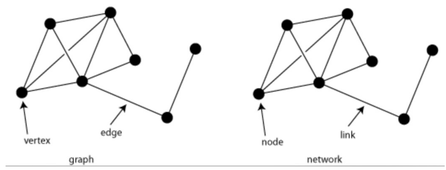
Enhancing Disease-Gene Association Discovery Using Large Language Models for Link Prediction

By, Esha Verma
Mentor: Zubair Qazi. Su/Wu Labs

Scripps Research

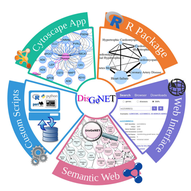
Background

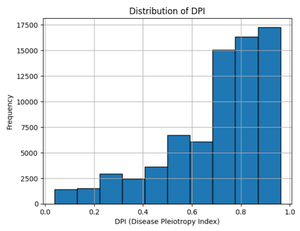
- Networks are a **type of data structure** can be applied to many domains to demonstrate relationships between entities
- Our study: **Gene – Disease association network**
  - Genes and diseases each as nodes in the graph
  - Edge is represented with GDAs (Gene-disease associations)
    - Indicates a causal relationship where mutations or alterations in a gene are known to cause a disease



Data

- **DisGeNet**
  - Databased used to explore gene-disease relationships, which can help researchers understand
    - Disease mechanisms
    - Potential drug targets
    - Biomarkers for diagnostics
  - Aggregates data from Scientific Literature, Clinical Databases, Genetic Studies and Gene-Disease Databases

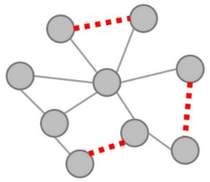




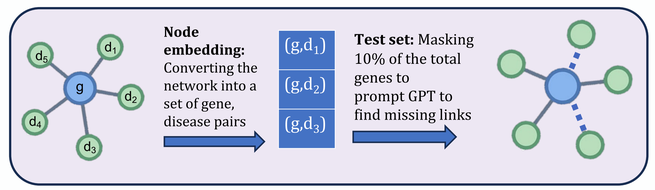
Basic Metrics
- Node Count: 15035
- Edge Count: 73469
- Average Degree: 9.7
- Density : 0.00065
- Connected Components: 209

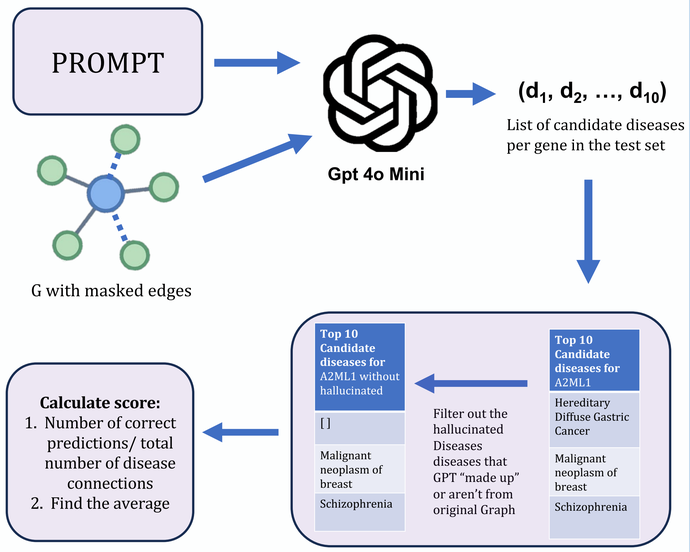
Task

- **Link Prediction**
  - Predict new edges between entities in a graph
- Graph Machine Learning
  - SOTA Methods → GCNs(Graph Convolutional Networks)
- Our focus : **Large Language Model Utilization**
  - Prompt an LLM to predict disease-gene associations given information about the already existing links in Graph G



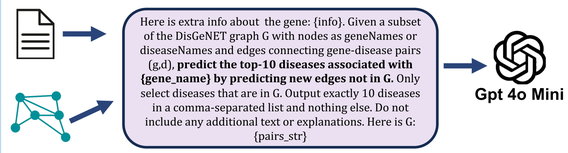
Methodology





Prompt Engineering

Second Iteration: Extra gene info perprompt extracted from NCBI



Third Iteration: Chain of Thought Prompting Technique

Reasoning Process:
Step 1: Identify the known diseases associated with {gene_name} in the graph G.
Step 2: Identify other genes in the graph G that are associated with these diseases
Step 3: Analyze the connectivity pattern of these related genes and diseases to infer potential new associations for {gene_name}.
Step 4: Rank the potential new diseases based on their connection strength and relevance to {gene_name
Step 5: Select the top-10 diseases from this ranking.

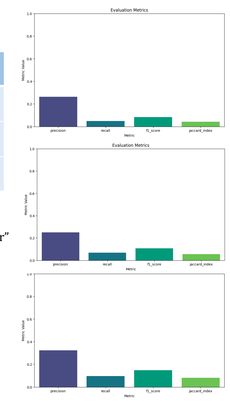
Results/Discussion

Common accuracy metrics:

	Precision	Recall	F1 Score	Jaccard Index
First Iteration	0.263	0.049	0.083	0.043
Second Iteration	0.249	0.0679	0.1067	0.056
Third Iteration	0.324	0.0965	0.148	0.08

Token Similarity Metrics(LLM specific)
Example of tokens : "Breast Cancer" → "Breast" + "Cancer"

	BLEU	ROGUE1	ROGUE2	ROGUEL
First Iteration	0.026	0.151	0.0563	0.122
Second Iteration	0.0466	0.183	0.0883	0.1417
Third Iteration	0.058	0.204	0.108	0.1582



Discussion / Conclusion

- Overall, the pre based trained GPT may not be suited well for Disease-gene association tasks.
- However, the performance improved as we integrated more information specific to the task AND a reasoning process for the LLM.

Future Steps

1. Concept Matching for the hallucinations
2. Improved scoring method based on Token Similarity
3. Prompt Engineering – find the right amount of context to provide to GPT
4. Fine tune a smaller model for better performance/ test other LLMs (Llama, Gemini, etc)
5. Test and compare with Graph Neural Network(GNN) based approach

References

1. Hu, Taojun, and Xiao-Hua Zhou. "Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions." *arXiv preprint arXiv:2404.09135* (2024).
2. Lu, Haohui, and Shahadat Uddin. "Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends." *Healthcare*. Vol. 11. No. 7. MDPI, 2023.
3. Li, Juanhui, et al. "Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking." *Advances in Neural Information Processing Systems* 36 (2024).
4. Shu, Dong, et al. "Knowledge Graph Large Language Model (KG-LLM) for Link Prediction." *arXiv preprint arXiv:2403.07311* (2024).
5. Zhu, Jing, et al. "Pitfalls in link prediction with graph neural networks: Understanding the impact of target-link inclusion & better practices." *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 2024.
6. He, Zhongmou, et al. "LinkGPT: Teaching Large Language Models To Predict Missing Links." *arXiv preprint arXiv:2406.04640* (2024).
7. Jin, Mingyu, et al. "ProLLM: Protein Chain-of-Thoughts

**Thank you to Teresa Hill, Dr. Nicholson, my mentor Zubair Qazi, and the Su/Wu labs for giving me this opportunity and guiding me throughout my project.