Databases

DisGeNET

Contains 2+ million gene-disease associasions, provides evidence scores an

You need to create a free account to download data.

OMIM

Focuses on Mendelian disorders with detailed phenotypic descriptions.
Contains 15000+ entries with well established gene-disease relationships.

GWAS Catalog

Contains 4000+ published genome-wide association studies.
Focuses on complex diseases and traits with SNP-level associations.

None-associative Databases

NCBI Gene

Comprehensive gene information including sequences, functions, and literat

Ensembl

Genome browser providing detailed gene annotations, variants, and comparat

初步模型结构

主要以Graph Convolution Neural Network为主，时间最充裕的情况可以对比一下两种encoder和两种decoder的不同组合，最不充裕的情况

Encoder：
- 2-layer GCN (See `Literature_Study/PGCN.pdf`)，那里@sbx对data preprocessing和模型结构有比较详细的描述，并且比原论文好读。
- 9-block × 3-layer GCN + 9 × residual connection (See `Proj_Related_Papers/DGHNN.pdf` 和 `Proj_Related_Papers/DGHNN_Supp.docx`)，这篇论文很短并且好多之前已经总结过了，所以@sbx没有做总结，不用管其中hypergraph的部分

Decoder:
- TF-Transformer (See `Proj_Related_Papers/DGHNN.pdf` 和 `Proj_Related_Papers/DGHNN_Supp.docx`)
- 1-layer GCN (See `Literature_Study/PGCN.pdf`)

建议组合是 `Proj_Related_Papers/DGHNN.pdf` 和 `Proj_Related_Papers/DGHNN_Supp.docx` 提出的结构，因为这篇文章是2025年的，@sbx没有整理是因为没时间了，但是它比较新，并且看上去效果也不错。代码方面可以参考

Baseline Choices:
- DBN (See `Literature_Study/RBM_DBN.pdf`)，这个可以@sbx来写，不是传统的机器学习/network方法，是深度学习方法，但是也比较老了，应该可以作baseline
- Traditional ML/Network methods: **欢迎补充**

目前最急需做的事

1. **确定上述模型选择**
2. 数据集的选择/大小/类型，因为@sbx认为他总结到的两篇文章的data都是异质的，不是很好获取和融合，所以需要尽快
3. 根据获得的数据确定**Data Preprocessing**的方法，因为@sbx认为他总结到的两篇文章的预处理方法都不简单。没有预处理得到的输入数据格式，代码都写不了。