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
```

What we should do

1. **What disease are we going to study? Is it Mendelian or not?** (We can only pick one)
2. **What DNN should we use**: MLP(x, too simple), CNN, GNN(hard, requires biological prerequisites, see this article)
3. Use database above to search for the disease to find list of associated genes with high evidence scores. You should also check whether it has **multiple supporting publications.** We may:
   - Check OMIM for Mendelian forms(单基因型) of the disease (If the disease is Mendelian)
   - Use GWAS Catalog to find SNPs linked to the disease and map them to candidate genes
   -
4. When preprossessing the data, we must:
   - **Process SNPs**: Map SNPs to genes using reference genome
   - **Data Balancing**: Use SMOTE or ADASYN to address class imbalance
   - **Feature Scaling**: Apply Min-Max scaling or Standardization to features
   - **Normalization**: Normalize gene expression data using log transformation or z-score normalization
5. When training the DNN, we **must** adopt:
   - **Stratified k-fold cross-validation**: Ensures equal representation of classes in each fold
   - **Evaluation metrics**: Use AUC-ROC, precision-recall curve, accuracy, F1-score
   - **Early stopping**: Prevent overfitting by monitoring validation loss
   - **Regularization**: Apply L2 regularization or batch normalization