Instruction to run the DL4H\_Team21.ipynb

To avoid download issues, use **gdown 4.6.0**. Newer versions like 5.1.0 may cause errors related to file access permissions and link retrievals. See “Handling Errors”.

**Option A**: Downgrade to gdown 4.6.0 **BEFORE** running any code block.

1. If your gdown version is **NOT** 4.6.0, downgrade to gdown 4.6.0 before proceeding.

| #Comment out this line: #!pip install --upgrade --no-cache-dir gdown  #Add this line: !pip install gdown==4.6.0 |
| --- |

1. Select Runtime -> Disconnect and delete runtime
2. Run all the code blocks. If you encounter any access error, refer to the 'Handling Errors' section for troubleshooting steps.

**Option B**: Use the latest gdown version.

1. If you prefer the latest version 5.1.0, modify the commands to handle large downloads more effectively **BEFORE** running any code block.

| !pip install --upgrade --no-cache-dir gdown  #Add the following commands to zip all the files, download it then unzip.  import gdown id = "1-th6FSfnpyajMPMuAdxVN8yNGk9lTy3M" gdown.download(id=id, output="myfile.zip") !unzip "myfile.zip" |
| --- |

1. Ensure to comment out the other download commands in the notebook to prevent conflicts.
   1. Download data

| # !gdown "https://drive.google.com/drive/folders/10H8Oh-ri7FZRzq4MlJB3z4LIaWbr1Xs6?usp=drive\_link" -O "/content/data" --folder |
| --- |

* 1. Download Project Code

| # !gdown "https://drive.google.com/drive/u/2/folders/1bJRIjPv\_L3jDCU-lB0hTL9ZSeelsF0GG" -O "/content/code" --folder |
| --- |

* 1. Download Model Checkpoint

| # !gdown 1QEVQ11\_qa2O9v5n18i3v6EuaCGiHJzDu -O "/content/mlp\_checkpoint.pt" |
| --- |

* 1. Download Training and Validation Embeddings

| # !gdown "https://drive.google.com/drive/folders/1nmiHkps12g9OfmKKyoqXmgNVFG2lOGVl?usp=drive\_link" -O "/content/mlp\_embeddings" --folder |
| --- |

* 1. Download Model Evaluation Results

| # !gdown "https://drive.google.com/drive/folders/1KntQomsoZxzi0unLnLQsMsAqlcuCuIjE?usp=drive\_link" -O "/content/model\_eval" --folder |
| --- |

* 1. Download Training and Validation Loss Data

| # !gdown "https://drive.google.com/drive/folders/1EvRBWYqIUF8r3hPBEhcbgFOrd6NlcoM2?usp=drive\_link" -O "/content/losses" --folder |
| --- |

1. Replace /content/data, /content/losses, /content/code, /content/mlp\_checkpoint.pt, /content/model\_eval to point to the unzipped folder
2. In section 3.2.2.3 Statistics

| #Update the file paths in the code, changing them from /content/data/... to /content/tar\_files/data/....  #training\_data = '/content/data/training.txt' #val\_data = '/content/data/val.txt' #test\_data = '/content/data/test.txt' training\_data = '/content/tar\_files/data/training.txt' val\_data = '/content/tar\_files/data/val.txt' test\_data = '/content/tar\_files/data/test.txt' |
| --- |

1. In section 3.4.2 Computational Requirements

| #model\_log\_filenames = { # 'MLP1': '/content/losses/mlp1.csv', # 'MLP2': '/content/losses/mlp2.csv', # 'MLP3': '/content/losses/mlp3.csv', # 'GCN1': '/content/losses/gcn1.csv', # 'GCN2': '/content/losses/gcn2.csv', # 'GCN3': '/content/losses/gcn3.csv', #} model\_log\_filenames = {  'MLP1': '/content/tar\_files/training loss\_validation loss\_computation\_time/mlp1.csv',  'MLP2': '/content/tar\_files/training loss\_validation loss\_computation\_time/mlp2.csv',  'MLP3': '/content/tar\_files/training loss\_validation loss\_computation\_time/mlp3.csv',  'GCN1': '/content/tar\_files/training loss\_validation loss\_computation\_time/gcn1.csv',  'GCN2': '/content/tar\_files/training loss\_validation loss\_computation\_time/gcn2.csv',  'GCN3': '/content/tar\_files/training loss\_validation loss\_computation\_time/gcn3.csv', } |
| --- |

1. In section 3.4.4 Demonstration of the Training Process Using a Sample Subset

| #%cd /content %cd /content/tar\_files/ |
| --- |

1. In section 4.1.1.1 Model Checkpoint

| #%cd "/content/code" %cd "/content/tar\_files/code" |
| --- |

| #path\_molecules = "/content/data/ChEBI\_definitions\_substructure\_corpus.cp" #path\_token\_embs = "/content/data/token\_embedding\_dict.npy" path\_molecules = "/content/tar\_files/data/ChEBI\_definitions\_substructure\_corpus.cp" path\_token\_embs = "/content/tar\_files/data/token\_embedding\_dict.npy" |
| --- |

| #CHECKPOINT = '/content/mlp\_checkpoint.pt' CHECKPOINT = '/content/tar\_files/2024\_4\_8\_epoch40\_sample100\_mlp1/test\_outputfinal\_weights.40.pt' |
| --- |

| #!python3 code/ranker\_test.py mlp\_embeddings test\_embeddings --test |
| --- |

| !python3 tar\_files/code/ranker\_test.py tar\_files/2024\_4\_8\_epoch40\_sample100\_mlp1/embeddings test\_embeddings --test |
| --- |

1. In section 4.1.1.2 Losses

| #filename = '/content/losses/mlp1.csv' filename = '/content/tar\_files/training loss\_validation loss\_computation\_time/mlp1.csv' |
| --- |

1. In section 4.1.2.2 Losses

| #filename = '/content/losses/gcn1.csv' filename = '/content/tar\_files/training loss\_validation loss\_computation\_time/gcn1.csv' |
| --- |

1. 4.2.1 MLP-ensembles, GCN-ensembles and All-ensembles

| #model\_result\_file\_map\_1 = { # 'MLP1': '/content/model\_eval/mlp1\_eval.csv', # 'MLP2': '/content/model\_eval/mlp2\_eval.csv', # 'MLP3': '/content/model\_eval/mlp3\_eval.csv', # 'GCN1': '/content/model\_eval/gcn1\_eval.csv', # 'GCN2': '/content/model\_eval/gcn2\_eval.csv', # 'GCN3': '/content/model\_eval/gcn3\_eval.csv', # 'MLP-Ensemble': '/content/model\_eval/ensemble\_mlp1\_mlp2\_mlp3\_eval.csv', # 'GCN-Ensemble': '/content/model\_eval/ensemble\_gcn1\_gcn2\_gcn3\_eval.csv', # 'All-Emsemble': #'/content/model\_eval/ensemble\_mlp1\_mlp2\_mlp3\_gcn1\_gcn2\_gcn3\_eval.csv', #} model\_result\_file\_map\_1 = {  'MLP1': '/content/tar\_files/model evaluation results/mlp1\_eval.csv',  'MLP2': '/content/tar\_files/model evaluation results/mlp2\_eval.csv',  'MLP3': '/content/tar\_files/model evaluation results/mlp3\_eval.csv',  'GCN1': '/content/tar\_files/model evaluation results/gcn1\_eval.csv',  'GCN2': '/content/tar\_files/model evaluation results/gcn2\_eval.csv',  'GCN3': '/content/tar\_files/model evaluation results/gcn3\_eval.csv',  'MLP-Ensemble': '/content/tar\_files/model evaluation results/ensemble\_mlp1\_mlp2\_mlp3\_eval.csv',  'GCN-Ensemble': '/content/tar\_files/model evaluation results/ensemble\_gcn1\_gcn2\_gcn3\_eval.csv',  'All-Emsemble': '/content/tar\_files/model evaluation results/ensemble\_mlp1\_mlp2\_mlp3\_gcn1\_gcn2\_gcn3\_eval.csv', } |
| --- |

1. In section 4.2.2 Hybrid Ensembles

| #file\_name = f'/content/model\_eval/{file\_name}' file\_name = f'/content/tar\_files/model evaluation results/{file\_name}' |
| --- |

1. In section 4.3 Ablation Studies

| #model\_result\_file\_map\_2 = { # 'MLP1': '/content/model\_eval/mlp1\_eval.csv', # 'GCN1': '/content/model\_eval/gcn1\_eval.csv', # 'Ablation add dropout rate 0.5 MLP': '/content/model\_eval/ablation\_add\_dropout\_rate\_0.5\_mlp\_eval.csv', # 'Ablation max pool GCN': '/content/model\_eval/ablation\_max\_pool\_gcn\_eval.csv', # 'Ablation reduce batch size 16 MLP': '/content/model\_eval/ablation\_reduce\_batch\_size\_16\_mlp\_eval.csv', # 'Ablation reduce learning rate MLP': '/content/model\_eval/ablation\_reduce\_learning\_rate\_mlp\_eval.csv', # 'Ablation remove convolutional layer 2 GCN': '/content/model\_eval/ablation\_remove\_convolutional\_layer\_2\_gcn\_eval.csv', # 'Ablation remove molecule encoder hidden layer 2 MLP': '/content/model\_eval/ablation\_remove\_molecule\_encoder\_hidden\_layer\_2\_mlp\_eval.csv', # 'Ablation remove nomralization layers MLP': '/content/model\_eval/ablation\_remove\_normalization\_layers\_mlp\_eval.csv', } model\_result\_file\_map\_2 = {  'MLP1': '/content/tar\_files/model evaluation results/mlp1\_eval.csv',  'GCN1': '/content/tar\_files/model evaluation results/gcn1\_eval.csv',  'Ablation add dropout rate 0.5 MLP': '/content/tar\_files/model evaluation results/ablation\_add\_dropout\_rate\_0.5\_mlp\_eval.csv',  'Ablation max pool GCN': '/content/tar\_files/model evaluation results/ablation\_max\_pool\_gcn\_eval.csv',  'Ablation reduce batch size 16 MLP': '/content/tar\_files/model evaluation results/ablation\_reduce\_batch\_size\_16\_mlp\_eval.csv',  'Ablation reduce learning rate MLP': '/content/tar\_files/model evaluation results/ablation\_reduce\_learning\_rate\_mlp\_eval.csv',  'Ablation remove convolutional layer 2 GCN': '/content/tar\_files/model evaluation results/ablation\_remove\_convolutional\_layer\_2\_gcn\_eval.csv',  'Ablation remove molecule encoder hidden layer 2 MLP': '/content/tar\_files/model evaluation results/ablation\_remove\_molecule\_encoder\_hidden\_layer\_2\_mlp\_eval.csv',  'Ablation remove nomralization layers MLP': '/content/tar\_files/model evaluation results/ablation\_remove\_normalization\_layers\_mlp\_eval.csv', } |
| --- |

1. In section 4.4.1 Ensemble Strategy

| #model\_result\_file\_map\_3 = { # 'MLP1+GCN1': '/content/model\_eval/ensemble\_mlp1\_gcn1\_eval.csv', # 'Ensemble max rank': '/content/model\_eval/additional\_experiment\_ensemble\_max\_rank\_eval.csv', # 'Ensemble weighted rank average': '/content/model\_eval/additional\_experiment\_ensemble\_weighted\_rank\_average\_eval.csv', } model\_result\_file\_map\_3 = {  'MLP1+GCN1': '/content/tar\_files/model evaluation results/ensemble\_mlp1\_gcn1\_eval.csv',  'Ensemble max rank': '/content/tar\_files/model evaluation results/additional\_experiment\_ensemble\_max\_rank\_eval.csv',  'Ensemble weighted rank average': '/content/tar\_files/model evaluation results/additional\_experiment\_ensemble\_weighted\_rank\_average\_eval.csv', } |
| --- |

1. Select Runtime -> Disconnect and delete runtime
2. Run all the code blocks. If you encounter any access error, refer to the 'Handling Errors' section for troubleshooting steps.

**Handling Errors**

When using a newer version of gdown, such as version 5.1.0, we might occasionally encounter issues with file retrieval. These problems often manifest as errors stating 'Failed to retrieve file URL' and suggest that you cannot retrieve the public link of the file.

| Failed to retrieve file url:  Cannot retrieve the public link of the file. You may need to change  the permission to 'Anyone with the link', or have had many accesses.  Check FAQ in https://github.com/wkentaro/gdown?tab=readme-ov-file#faq. You may still be able to access the file from the browser:  https://drive.google.com/uc?id=15yKVBcBdL1-Q4AnupidOrxMaw1DEOYKI but Gdown can't. Please check connections and permissions. |
| --- |

We have confirmed that all files in the shared Google Drive folder are publicly accessible. However, issues may arise due to the accumulation of cookies, especially if Google issues a warning about a cookie that does not specify a particular download. This can lead Google to block the download, misleading our code into perceiving a permissions problem. Visit the [GitHub Discussion page](https://github.com/wkentaro/gdown/issues/43) for more information.

Therefore, for increased stability, we provided two strategies above, “Option A” and “Option B”, to manage these downloading issues effectively. Both methods have been thoroughly tested through over 10 trial runs on our end without any issues. However, if the error persists, follow these steps:

1. Limit Downloads Per Runtime: Google seems to impose limits on the number of files and the total size of files you can download in a single runtime. Attempting more than one download per file can trigger these restrictions.
2. Reset the Runtime: Navigate to ‘Runtime’ -> ‘Disconnect and delete runtime’ to clear your current runtime.
3. Retry the Preferred Option: If you initially chose Option A and encountered an error, switch to Option B for its different approach, as Option A depends significantly on using gdown version 4.6.0. Conversely, if Option B fails, attempt Option A.