Interpretable Glioma Classification via Cluster-Guided Multi-Omics Modelling

# 1. Development Environment

For Python notebook( .ipynb) file:

* We performed the majority of this project using Google Colab for accessibility and ease of use. So, you do not need to download the .ipynb files.
* All notebooks can be executed directly within the shared Google Colab environment.
* Since Colab resets the environment every time it restarts, each notebook includes its own package installation cells.
* Therefore, you do not need to manually set up any Python environment.

For R Markdown (.Rmd) files:

* Certain steps in our workflow required the R programming due to library limitations or data format compatibility.
* R Markdown (.Rmd) files cannot be executed on Colab.
* These files must be downloaded and run locally in an appropriate R environment.

# 2. Folder Structure

## 2.1 Code folder

* 0.1\_Data\_collection\_GDC\_download.Rmd
  + Omics data collection from the GDC portal.
  + Note: To reduce preprocessing time, omics data were also manually downloaded from the UCSC Xena browser. This code is kept for potential use of data from the latest version without being preprocessed from the Xena source.
* 0.2\_Data\_collection\_GDC\_clinical\_info.Rmd
  + Clinical data collection from the GDC portal
* 0.3\_R\_code-clinical\_data\_convert\_to\_csv.ipynb
  + Converts clinical data downloaded from 0.2.Data\_collection\_TCGA\_clinical\_info.Rmd from \*.Rmd format into .csv.
* 1\_Data\_preprocessing-gene\_mapping.ipynb
  + Attempts to perform gene name mapping across each omics layer by converting probes to gene symbols.
  + Initially intended to unify feature names across modalities.
  + Note: MOFA+ requires probe-level data, so this step is no longer used in the pipeline. However, the notebook is retained for potential future gene-level analysis.
* 1.2\_Data\_preprocessing\_snv.Rmd
  + Preprocess mutation files to a matrix of SNVs with rows representing gene-level identifiers and columns representing samples for both the GBM and LGG datasets
* 2.1\_RNA\_seq\_sample\_filtering\_by\_plate.ipynb:RNA\_seq.ipynb
  + Addresses plate duplication issues in RNA-seq data when multiple entries exist for a single sample.
* 2.2\_filtering\_sex\_related\_gene\_and\_protein\_gene\_name\_mapping.ipynb
  + Filters out sex-related genes from all omics layers.
  + Additionally, for the proteomics data, it maps features from target names to corresponding gene names for consistency.
* 3.1\_Intra-cohort\_preprocessing\_sample\_filtering.ipynb
  + Sample alignment: TCGA sample identifiers differ slightly across omics layers. This script standardizes them using the shortest consistent format.
  + Remove incomplete cases: Samples lacking data in one or more omics layers are excluded.
  + Remove low-information samples: Samples with more than 50% missing values (NA) in any modality are removed, even if all modalities are present.
* 3.2\_TCGA\_preprocess\_methylation.Rmd
  + Probe filtering for methylation data in both GBM and LGG datasets, including missing data handling, sex-related probe filtering, and SNP-related probe filtering.
  + Sample filtering of the GBM dataset to exclude data from the same samples from different probing platforms.
  + Integrate GBM methylation data and convert Beta values to M values.

* 4.1\_Data\_spliting.ipynb
  + Splits the dataset into training, validation, internal test, and external test sets.
* 4.2.1\_Data\_cleaning\_high\_NA\_feature\_filtering.ipynb
  + Removes features (e.g., genes, probes) that are missing in more than 50% of samples within each omics layer.
* 4.2.2\_Data\_cleaning\_sex\_disease\_association\_test.ipynb
  + Statistical tests for any association between the disease (LGG and GBM) and sex, using chi-square and Fisher’s exact tests.
* 4.2.3\_Data\_cleaning\_Feature\_selection\_methylation.Rmd
  + Feature selection (DMP/DMR) between the GBM and LGG datasets.
  + Feature mapping of probe-sample matrix to gene-sample matrix.
  + Note: DMP and DMR are not used as the input of the MOFA+ model due to the potential filtering of high-variance features, so this step is no longer used in the pipeline. After filtering the high-variance features of methylation data, the probe-sample matrices of GBM and LGG are used for the omics data integration. The gene-sample matrix format of methylation data is used for the annotation analysis.
* 4.3\_batch\_correction.ipynb
  + Performs batch effect correction for DNA methylation datasets (450k vs 27k platforms).
* 4.4\_feature\_selection.ipynb
  + Performs feature selection based on variance thresholding across samples.
* 5.1\_Multi-Omics\_data\_integration.ipynb
  + Integrates all omics data using MOFA+ (Multi-Omics Factor Analysis), which captures latent factors shared across modalities.
* 5.2\_Multi-Omics\_integration\_quality\_assessment.Rmd
  + Assesses how much variance was explained across all omics layers in order to select a MOFA+ model with the optimal number of factors, and shows which omics layers contributed to which factors.
* 6\_Clustering\_kmeans.ipynb
  + Perform Consensus Clustering on the Training set and assign cluster label of validation/internal/external test dataset by trained consensus matrices centroid.
  + Visualization of cluster assignment confidence
* 7-8\_Classfication\_and\_model\_evaluation\_SHAP.ipynb
  + Contains code for clustering-based classification of glioma subtypes using the integrated multi-omics data.
  + SHAP explanation using classification models.
* 9.1.1\_Cluster\_Annotation.ipynb
  + Compare clusters against known variables. Statistical associations between clusters and variables are tested.
* 9.1.2\_Cluster\_annotation\_factor\_analysis.Rmd
  + Identifies discriminative factors across clusters and performs Gene Set Enrichment Analysis (GSEA) to biologically annotate each factor.
* 9.2\_Survival\_Analysis.ipynb
  + Kaplan-Meier survival analysis with log-rank tests

## 2.2 Data folder

* 0.reference
  + gencode.v36.annotation.gtf.gz
  + HM27.hg38.manifest.gencode.v36.tsv.gz
  + HM450.hg38.manifest.gencode.v36.tsv.gz
* 0.clinical\_rdata\_to\_csv
  + Clinical data converted from .Rmd to .csv via `0.3.R\_code-clinical\_data\_convert\_to\_csv.ipynb`
* 1\_raw\_data
  + Contains GBM and LGG data downloaded from TCGA (GDC, Xena) along with clinical data from step 0.3
* 2\_gene\_name\_mapping\_data
  + Output from `1. Data\_preprocessing-gene\_mapping.ipynb` Mapping of probe IDs to gene names for each omics layer
* 3\_RNA\_seq\_sample\_filtering\_by\_plate
  + Results from: `2. RNA\_seq\_sample\_filtering\_by\_plate.ipynb` and `2.2\_filtering\_sex\_related\_gene\_and\_protein\_gene\_name\_mapping.ipynb`
* 4\_Intra-cohort-preprocessing
  + Output of `3.1.Intra-cohort\_preprocessing\_Data\_cleaning(sample\_filtering).ipynb` Sample alignment, missing data filtering, and cohort cleaning
* 5\_Inter-cohort\_preprocessing
  + 5.1.data\_spliting: Results of `4.1.Data\_spliting.ipynb`
  + 5.2.data\_cleaning(Feature\_filtering): Output from `4.2.Data\_cleaning(feature\_filtering).ipynb`
  + 5.3.1.batch\_correction: Output from `4.3\_batch\_correction.ipynb`
  + 5.3.2.probe\_to\_Ensembl: Output from `4.2.3\_Data\_cleaning\_Feature\_selection\_methylation.Rmd`
  + 5.4.feature\_selection : Results of `4.4\_feature\_selection.ipynb`
  + 5.5Multi-omics\_integration: Output from `5.1\_Multi-Omics\_data\_integration.ipynb`
* 6\_Unsupervised\_clustering
  + Output of `6\_Clustering\_kmeans.ipynb`
  + Kmeans\_30mixed: clustering results using 30-factor latent space from MOFA+
  + …
* 7\_classification
  + base\_line: 100x bootstrap classification results using baseline features
  + mixed30: 100x bootstrap results using 30-factor latent space from MOFA+
* 9\_Biological\_Interpretation
  + 9.1\_Cluster\_Annotation: Inputs for 9.1.1\_Cluster\_Annotation.ipynb and 9.1.2\_Cluster\_annotation\_factor\_analysis.Rmd
  + 9.2\_SHAP: result image of sharp implementation

# 3. Usage

When you click on any Python notebook (.ipynb) file in this repository, it will automatically open in **Google Colab**.

You can simply run all cells in the notebook from top to bottom.

* Accessing Shared Data via Google Drive

Since the necessary data files are stored in a shared Google Drive folder, you must first mount your Drive within Colab.

At the beginning of each notebook, run the following code block to connect to your Google Drive:

from google.colab import drive

drive.mount('/content/drive')

* Running Locally

If you prefer to run the notebooks on your **local machine**, simply **replace** all instances of the Colab path: /content/drive/MyDrive/

with the local path to your project directory, such as:

/home/your\_username/Ds\_project/

Make sure that all file paths are updated accordingly.

* 0.\**\_Data\_collection\_\**.Rmd
  + 0.1\_Data\_collection\_GDC\_download.Rmd
  + 0.2\_Data\_collection\_GDC\_clinical\_info.Rmd
  + Download .Rmd files and run the code locally.
* 1.2\_Data\_preprocessing\_snv.Rmd
  + Download this .rmd file along with the following input files from DS\_project/Data/1\_raw\_data/:
    - TCGA-GBM.somaticmutation\_wxs.tsv.gz
    - TCGA-LGG.somaticmutation\_wxs.tsv.gz
  + Change the input path and run the code.
* 3.2\_TCGA\_preprocess\_methylation.Rmd
  + Download this .rmd file along with the following reference files from DS\_project/Data/0.reference/:
    - HM450.hg38.snp.tsv.gz
  + Download this .rmd file along with the following input files from DS\_project/Data/1\_raw\_data/:
    - TCGA-GBM.methylation270.tsv.gz
    - TCGA-GBM.methylation450.filtered.tsv.gz
    - TCGA-LGG.methylation450.filtered.tsv.gz
  + Change the input path and run the code.
* 4.2.3\_Data\_cleaning\_Feature\_selection\_methylation.Rmd
  + Download this .rmd file along with the following reference files from DS\_project/Data/0.reference/:
    - gencode.v36.annotation.gtf.gz
    - HM27.hg38.manifest.gencode.v36.tsv.gz
  + Download this .rmd file along with the following input files from DS\_project/Data/0.reference/5\_Inter-cohort\_preprocessing/5.3.1.batch\_correction/:
    - \*\_methylation\_450\_27\_batch\_corrected\_\*.csv
    - \*\_methylation\_450\_27\_batch\_corrected\_\*.csv
  + Identify the outpath in .Rmd, the output of the gene-sample methylation matrix will be generated to the defined outpath.
  + Change the input path and run the code.
* 5.2\_Multi-Omics\_integration\_quality\_assessment.Rmd
  + Download this .rmd file along with the following input files:
    - DS\_project/Data/5\_Inter-cohort\_preprocessing/5.5Multi-omics\_integration/trained\_model\_factor\_[5,10,15,20,25,26,27,28,29,30,31,32,33,34,35].hdf5
  + Create an input folder in the same directory as the .rmd file and move the .hdf5 files into it.
  + Run the code.
* 9.1.2\_Cluster\_annotation\_factor\_analysis.Rmd
  + Download this .rmd file along with the following input files:
    - model
      * DS\_project/Data/5\_Inter-cohort\_preprocessing/5.5Multi-omics\_integration/trained\_model\_factor\_30mixed.hdf5
    - clinical datasets
      * DS\_project/Data/0.clinical\_rdata\_to\_csv/TCGA-GBM-clinicalXML\_patient.csv
      * DS\_project/Data/0.clinical\_rdata\_to\_csv/TCGA-LGG-clinicalXML\_patient.csv
    - clustering results
      * DS\_project/Data/6\_Unsupervised\_clustering/kmeans\_30mixed/k\_[2,3,4,5]/train\_cluster\_labels\_consensus\_complete.csv
    - gene ID mapping
      * DS\_project/Data/2\_gene\_name\_mapping\_data/chromosome\_probe\_gene.csv
    - functional databases
      * DS\_project/Data/9\_Biological\_Interpretation/9.1\_Cluster\_Annotation/MSigDB\_v6.0\_C2\_human.RData
      * DS\_project/Data/9\_Biological\_Interpretation/9.1\_Cluster\_Annotation/MSigDB\_v6.0\_C5\_human.RData
  + Create an input folder in the same directory as the .rmd file.
  + Move the model file, clinical datasets and functional databases into the input folder
  + Move the clustering results into the input/kmeans\_30mixed/k\_[2,3,4,5] folders.
  + Move the gene ID mapping file into the input/gene\_name\_mapping folder
  + Run the code.