Interpretable Glioma Classification via Cluster-Guided Multi-Omics Modelling

Gyeongsil Kim[1*], Daeho Lee[1*] and Jingwen Luo[1*]

[1*]Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, Berlin, 14195, Germany.

*Corresponding author(s). E-mail(s): kimg97@zedat.fu-berlin.de; dael93@zedat.fu-berlin.de; jingwen.luo@fu-berlin.de;

Abstract

Gliomas are a biologically and clinically diverse class of primary brain tumours. Traditional classifications based on histological grading—such as low-grade glioma (LGG) and glioblastoma (GBM)—provide a broad framework for diagnosis and prognosis. However, these categories often overlook the molecular heterogeneity of the disease, leading to overlapping boundaries between subtypes and challenges in clinical decision-making.

To address this issue, we developed a cluster-guided classification framework aimed at better reflecting the molecular diversity of gliomas while refining the LGG–GBM distinction. The goal was to improve both predictive performance and biological interpretability through an integrative, data-driven approach.

We applied MOFA+ to jointly model five omics layers—copy number variation (CNV), DNA methylation, RNA expression, protein expression, and single nucleotide variation (SNV)—to derive integrated latent representations. These were used in k-means-based consensus clustering to identify molecularly coherent subgroups. For classification, we trained 11 machine learning models and selected the most suitable one for each cluster. Biological interpretation was performed through functional annotation of clusters, SHAP-based feature importance analysis, and survival analysis.

The resulting cluster-specific classifiers outperformed global models in both predictive power and interpretability. Distinct pathway signatures and survival differences among clusters supported their biological and clinical relevance.

These findings shed light on the potential of combining unsupervised clustering with supervised modelling to refine glioma subtyping and support more personalised strategies in brain tumour research.

Keywords: Glioma, Multi-Omics Integration, Consensus Clustering, Interpretable Classification Framework

1 Background

Gliomas are a broad category of primary brain tumours that arise from glial cells, which are non-neuronal supportive cells in the central nervous system, including astrocytes, oligodendrocytes, and ependymal cells. They account for approximately 80% of malignant brain tumours in adults and are classified based on both their presumed cell of origin and their histological features [1].

Historically, gliomas have been divided into two comprehensive clinical categories: low-grade gliomas (LGGs), encompassing WHO grade II and III tumours, and glioblastoma multiforme (GBM), representing the most aggressive WHO grade IV tumours [1, 2]. However, this binary classification is rooted in histological appearance and prognosis, rather than molecular characteristics [2, 3].

While widely used, this legacy terminology has shown limitations in reflecting the biological complexity of gliomas. The 5th edition of the WHO Classification of Central Nervous System Tumours explicitly stated that the terms LGG and GBM are retained primarily for communication purposes and no longer correspond to distinct tumour entities [2].

Indeed, molecular overlaps between LGGs and GBMs have been identified. A seminal study by Ceccarelli et al. [3], using integrated TCGA data, revealed that LGGs comprise a mixed group of tumours with distinct molecular alterations, many of which blur the boundary with GBM. Similarly, Reuss et al. [4] found that certain histologically classified LGGs possess molecular features characteristic of GBM, such as IDH-wildtype astrocytomas, suggesting that some LGGs may exhibit aggressive behaviour typically associated with GBMs. These findings underscore the limitations of histology-based classification and highlight the need for more nuanced, data-driven approaches.

In response to this heterogeneity, molecular-based classification systems were introduced in the WHO 2016 and further refined in the WHO 2021 guidelines. Under these systems, gliomas are defined by key genetic alterations such as IDH mutation status, 1p/19q codeletion, and CDKN2A/B deletion. GBM is now molecularly defined as an IDH-wildtype astrocytic glioma exhibiting at least one of the following features: TERT promoter mutation, EGFR amplification, or combined chromosome 7 gain/chromosome 10 loss (+7/-10) [2]. Furthermore, IDH status has been proposed as a major determinant of glioma prognosis and classification [5].

Despite recent progress, key limitations remain in glioma classification. Many studies have relied on subgroups predefined by a limited set of molecular markers, rather than identifying subtypes through data-driven or unsupervised approaches. For example, Ceccarelli et al. [3] successfully delineated major glioma subtypes using TCGA data, but their method was guided by a fixed set of known alterations and did not involve classification strategies to validate or model the resulting groups.

In addition, unsupervised clustering and supervised classification methods are often applied in isolation, missing the opportunity to link data-driven subgroup discovery with predictive modelling. The iGlioSub framework by Ensenyat-Mendez et al. [6], for instance, used RNA expression and DNA methylation data to refine subtype classification, but the identified subgroups were not integrated into predictive models. As a result, their biological interpretation and clinical applicability remained limited.

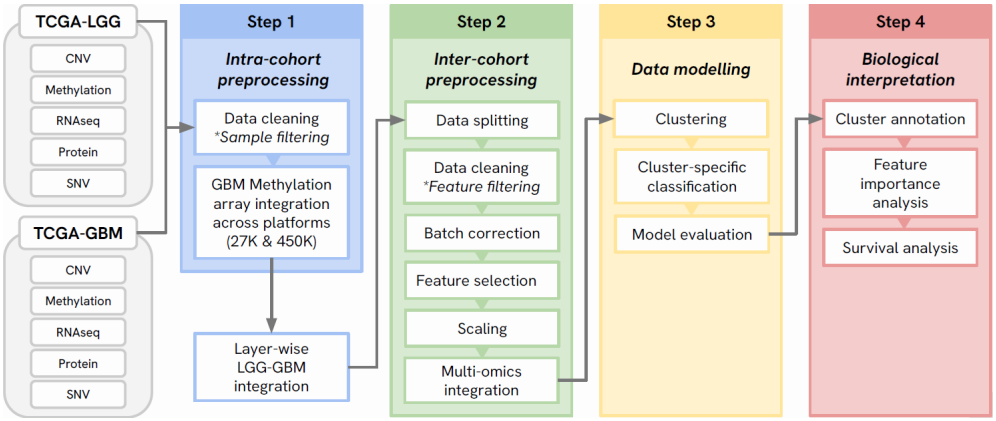


Fig. 1: Overview of the project

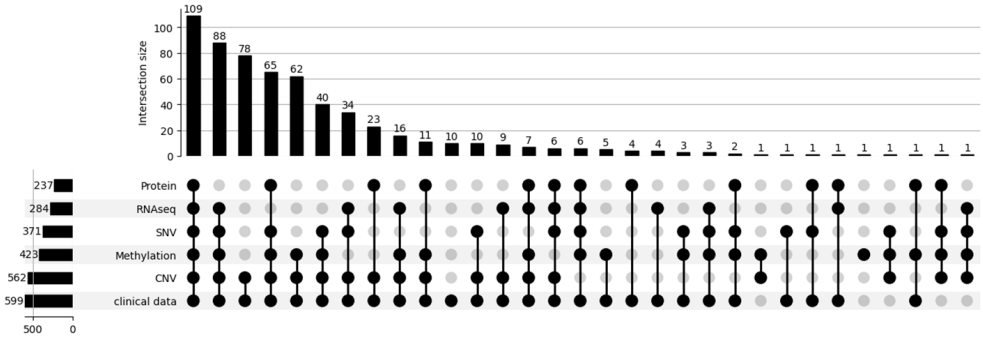
2 Goal

The goal of this project is to develop an interpretable classification framework that better reflects the molecular heterogeneity of gliomas, while refining the commonly used LGG-GBM categories. Specifically, we aimed to address the blurred molecular boundary between LGG and GBM, which often hinders both predictive accuracy and biological understanding.

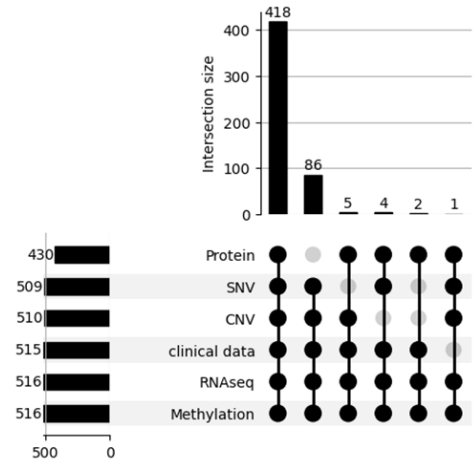
To this end, we propose an integrated strategy that combines unsupervised clustering with supervised learning to uncover and model latent glioma subgroups in a biologically meaningful way. By detecting molecularly similar patient groups based on shared profiles regardless of legacy clinical terms, we sought to enhance both classification performance and interpretability.

Our strategy consists of three main components: (1) Applying data-driven clustering to identify previously unrecognised glioma subgroups, (2) Integrating five omics layers—copy number variation (CNV), DNA methylation, RNA expression, protein expression and single nucleotide variation (SNV)—to reflect the systemic nature of tumour biology, (3) Developing cluster-specific classifiers to enable context-aware prediction and support biological interpretation via functional annotation and survival analysis.

An overview of our project is shown in Figure 1. The workflow consists of four main steps: intra-cohort preprocessing, inter-cohort integration, data modelling, and biological interpretation. Starting with multi-layered TCGA-LGG and TCGA-GBM datasets, we implemented a series of preprocessing steps to harmonise and integrate the data. This was followed by unsupervised clustering, development of cluster-specific predictive models, and biological interpretation through annotation, feature importance analysis, and survival modelling.



(a) GBM data



(b) LGG data

Fig. 2: An overview of the data intersection: (a) GBM and (b) LGG

3 Data and Preprocessing

3.1 Data

The primary datasets used in this project are TCGA-GBM [7] and TCGA-LGG [8] with 617 and 516 biospecimens, respectively. Five omics layers were used throughout the exploration and analysis: mutation, copy number variants (CNVs), and methylation data from the genomics layer; mRNA expression data from the transcriptomics layer; and protein expression data from the proteomics layer. These datasets were used in the subsequent stages of preprocessing, integration, modeling and anotation analysis.

To reduce preprocessing time, mutation, CNV, methylation, RNA expression and protein expression data were downloaded as preprocessed datasets from the UCSC Xena browser [9] and clinical data were directly downloaded from the GDC via Rscript

using a R package *easyTCGA* [10]. Further details, including data version numbers and download resources, are provided in Appendix A.

Both the CNV and protein expression datasets are formatted as genomic matrices, with rows representing gene-level identifiers and columns representing samples. CNV values indicate absolute copy numbers, while protein expression matrices contain normalized RPPA values. For mutation data, we obtained a file in the "Variant by Position" format, which records somatic mutations (SNPs and small INDELs) detected in each sample. This file was subsequently converted into a genomic matrix. The mRNA expression data are represented as gene expression counts and are also structured as genomic matrices. The methylation data consist of matrices containing beta values, and the clinical data include variables such as patient age, gender, histological type, tumor grade, and other relevant clinical characteristics.

An overview of the intersection of available samples across different omics layers is shown in Figure 2.

3.2 Intra-Cohort Preprocessing

This chapter outlines the intra-cohort preprocessing procedures applied to the multi-omics datasets. The primary goal of this step is to ensure internal consistency within each cohort by harmonizing sample identifiers and resolving discrepancies across data modalities.

3.2.1 Sample Filtering

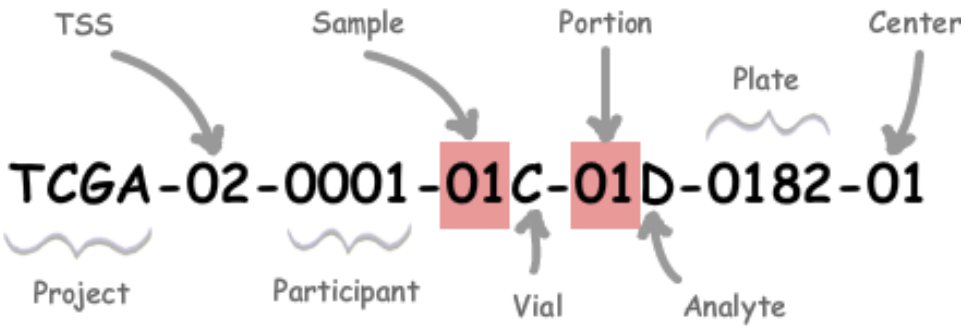


Fig. 3: TCGA barcode structure, illustrating how each part of the code corresponds to different levels of biospecimen information.

Effective integration of TCGA multi-omics data requires careful handling of several structural challenges, including sample redundancy, metadata inconsistencies, missing clinical annotations, and imbalanced missingness across omics layers. If left unaddressed, these issues can lead to data leakage, biased model evaluation, or unreliable

downstream analyses. To mitigate such risks, we applied a two-step sample filtering and harmonization procedure: (i) Sample Filtering through Identifier Harmonization and Redundancy Removal, and (ii) Filtering Samples for MOFA+ [11] Compatibility.

(i) Sample Filtering through Identifier Harmonization and Redundancy Removal

As illustrated in Figure 3 [12], each TCGA sample is labeled with a hierarchical barcode that encodes metadata such as project ID, participant ID, sample type, vial, and plate. However, the structure and depth of these barcodes vary across omics platforms. For example, RNA-seq data typically include plate-level identifiers, while mutation, CNV, methylation, and protein data usually extend only to the vial level. Additionally, individual participants may contribute multiple samples, and each sample can be associated with multiple plates, resulting in a 1:N:N relationship between participants, samples, and plates.

To resolve this heterogeneity, we truncated all barcodes to the vial level and used this identifier consistently across all omics layers. In GBM cohort RNA-seq data, where some samples were duplicated at the plate level, we retained only one representative plate per vial. Specifically, we selected the plate linked to the largest number of samples to minimize batch-specific biases. Summary statistics of this filtering process are presented in Table 1.

Following identifier harmonization, we selected a single representative sample per participant. In cases with multiple samples, we prioritized those associated with the largest total entries across omics types. If multiple candidates satisfied this condition, the sample with the lexicographically earliest vial ID was selected. This heuristic approach aimed to favor non-redundant, consistently profiled samples.

Furthermore, samples labeled with tissue type codes of 09 or higher—corresponding to normal tissues—had already been excluded in earlier filtering steps and therefore required no further action at this stage.

(ii) Filtering Samples for MOFA Compatibility

Although MOFA+ is designed to tolerate missing values and can handle up to 50% missingness per modality, its performance may deteriorate when individual samples exhibit low modality coverage or when missingness is unevenly distributed within a given modality.

To ensure that the data were suitable for latent factor inference and downstream classification, we implemented two additional filtering criteria. First, samples with fewer than a minimum number of available omics modalities were excluded. Second, for any given sample, if more than 50% of the features within a modality were missing (NA), that sample was also removed. These steps ensured that the retained cohort was composed of samples with sufficient and reliable information for multi-omics integration.

3.2.2 GBM Methylation Array Integration Across Platforms

The Cancer Genome Atlas (TCGA) has profiled DNA methylation in over 12,000 samples, spanning 34 cancer types, using two Illumina platforms: the Infinium HumanMethylation27 BeadChip (27k) and the HumanMethylation450 BeadChip (450k).

Table 1: Summary of RNA-seq samples and participant-level duplication

	Total number of RNA-seq samples	Number of samples from the same sample
GBM	372	173
LGG	516	0

While both arrays are based on bisulfite conversion and Illumina BeadArray technology, they differ substantially in probe chemistry (Infinium I vs. II) and genomic CpG coverage [13]. These differences necessitate careful handling of platform-specific variation in feature coverage and selection when integrating data across arrays.

To maximize sample utilization and ensure consistency of methylation feature inputs across platforms for downstream analyses, we performed a platform harmonization process on the TCGA-GBM methylation dataset. This process included three steps: (i) selection of platform-specific data in overlapping samples, (ii) probe harmonization, and (iii) transformation of beta values to M-values. Since all TCGA-LGG samples were measured using the 450k platform, we subsetted the LGG data based on the probe set retained after integrating the TCGA-GBM dataset, ensuring consistent input features across cohorts.

(i) Selection of Platform-Specific Data in Overlapping Samples

For GBM samples that were measured using both the 27k and 450k methylation arrays, we retained the 450k measurements for downstream analysis. This choice ensured consistency in data quality and platform compatibility, particularly given that all LGG samples were profiled using the 450k platform.

(ii) Probe Harmonization

To enable integrative analysis across platforms, we performed probe harmonization between the 27k and 450k arrays. Probes were matched based on their genomic coordinates rather than array-specific identifiers, ensuring accurate correspondence across platforms. Only probes that were unambiguously present on both arrays were retained, forming a unified feature set used in all subsequent analyses. This step ensured that the methylation data from GBM and LGG samples were aligned at the feature level, minimizing platform-related biases in downstream modeling.

(iii) Transformation of Beta-values to M-values

Given their distributional properties, M-values are statistically more suitable than beta values for differential methylation analysis [14]. The M-value is defined as the log2 ratio of the intensities of methylated versus unmethylated probes:

$$M_{value} = log_2 \frac{\beta}{1-\beta}$$

An M-value close to 0 indicates similar intensities of methylated and unmethylated signals, suggesting a partially methylated CpG site. Positive M-values indicate more methylated molecules, while negative values indicate the opposite.

3.3 Inter-Cohort Preprocessing

This section outlines the preprocessing steps performed to enable integrated analysis of multi-omics data from the GBM and LGG cohorts. The key procedures include feature alignment, data partitioning, feature filtering, batch correction, feature selection, and scaling. Each step was designed to reduce technical variation between cohorts and enhance the biological interpretability of downstream analyses. Through this process, we established a coherent framework for handling heterogeneous datasets within a unified analytical pipeline.

3.3.1 Feature Alignment

The feature alignment step consisted of two major components: (i) mapping features to a common identifier across omics layers, and (ii) transforming the data into a matrix format suitable for machine learning tasks.

(i) Mapping to a Common Identifier

To ensure consistency across omics modalities, we standardized feature names using Ensembl Gene IDs as the unified reference. Most data types, including RNA-seq, CNV, and SNV, were already annotated with Ensembl IDs. However, protein expression data generated via Reverse Phase Protein Array used a different naming convention, specifically Reverse Phase Protein Array (RPPA) targets [15]. To address this discrepancy, we utilized a publicly available mapping table linking RPPA targets to gene symbols [16]. Subsequently, we converted those gene symbols to Ensembl Gene IDs using standard gene annotation resources. The annotation data used for this mapping were included as part of the RNA-seq dataset obtained from the GDC data portal. Methylation probes were also mapped to the Ensembl gene names using an annotation file downloaded from the Xena browser. To map the methylation beta values to the gene level, we calculated the mean beta values for each gene promoter region. Details of the annotation file source are described in Appendix B

(ii) Matrix Formatting for Model Training

Most omics layers—including RNA-seq, CNV, methylation, and protein expression—were already provided in matrix format. For SNV data, a binary mutation matrix was constructed by computing the maximum Variant Allele Frequency (VAF) per sample–gene pair and binarizing the results. Missing values were imputed as zeros.

3.3.2 Data Splitting

To enable robust model training and evaluation, the complete cohort data were divided into four subsets: training, validation, internal test, and external test datasets. An overview of the data splitting strategy is presented in Figure 4.

To ensure the independence of the external test set, we utilized the Tissue Source Site (TSS) information provided by TCGA. Each TSS represents a distinct collection site or institution, and thus, samples from different TSS codes can be regarded as independently sourced. Based on this, we divided the TSSs for GBM and LGG separately using a 9:1 ratio and designated the 1-part subset as the external test dataset. The remaining 9 parts were used to construct the internal dataset, which was further

split into training, validation, and internal test sets at a ratio of 7:1.5:1.5, respectively, for both GBM and LGG.

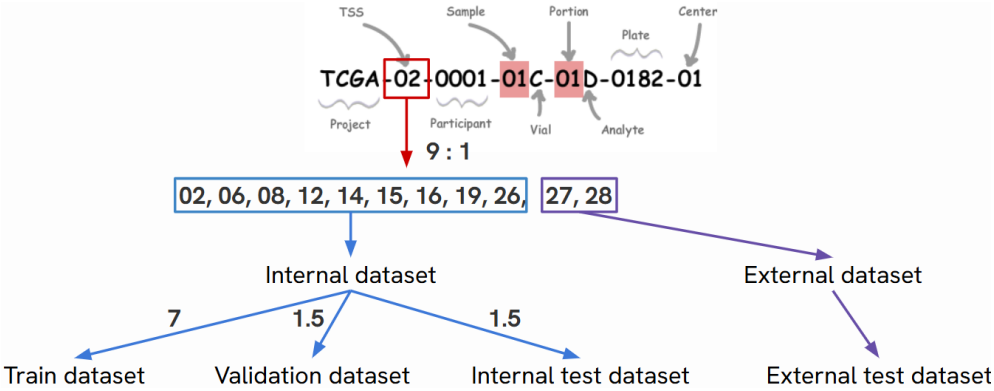


Fig. 4: An overview of the data splitting process. Samples were divided based on tissue source site codes from the TCGA barcode. A 9:1 ratio was used to separate the internal and external datasets. The internal dataset was further split into training, validation, and internal test sets in a 7:1.5:1.5 ratio. The external dataset was used as an external test dataset.

However, as illustrated in Figure 2, each sample contains a different combination of omics layers, resulting in heterogeneous data distributions depending on the omics type. Moreover, Figure 5 shows that the number of samples varies across TSS codes. If the omics composition of the divided subsets deviates substantially from the original distribution, it could introduce bias during model training. To mitigate this, we designed a two-step strategy for data partitioning.

(i) External Dataset Construction

For each of the GBM and LGG cohorts, we aimed to select a set of TSS codes comprising approximately 10% of the total samples. To avoid overly narrow selection and increase flexibility, we introduced a tolerance range of $\pm 5\%$, allowing candidate TSS combinations representing between 5% and 15% of the cohort. Among these, we computed the Kullback–Leibler (KL) divergence between each candidate's omics composition and the original distribution, and selected the TSS combination with the smallest divergence as the final external test dataset.

(ii) Internal Dataset Construction

From the remaining TSS codes not included in the external dataset, we constructed the internal dataset. For both GBM and LGG, samples were split within each omics combination at a ratio of 7:1.5:1.5 to form the training, validation, and internal test sets, respectively.

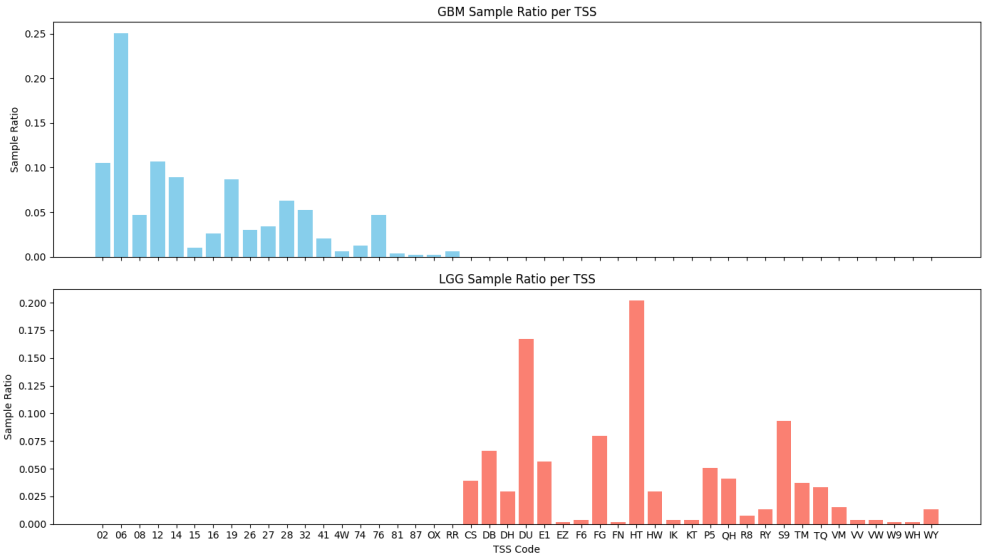


Fig. 5: Distribution of tissue source site (TSS) codes in the dataset. The bar plots show the sample ratio per TSS for GBM (top) and LGG (bottom)

3.3.3 Feature Filtering

To improve the signal-to-noise ratio and reduce potential artefacts, we applied two filtering strategies to exclude uninformative or confounding features from the analysis.

(i) Removal of Features with a High Proportion of Missing Values

The first strategy involved removing features that contained a substantial proportion of missing values. Feature filtering was performed separately for each omics layer to account for differences in data completeness. The goal was to retain only those features likely to carry meaningful biological signals.

The filtering threshold was defined based on the training datasets for GBM and LGG. Any feature with more than 50% missing values in either of the training sets was excluded. The same set of features was subsequently removed from the validation, internal test, and external test datasets to maintain consistency in the feature space across all subsets.

(ii) Removal of Features without a Statistically Significant Association between Disease and Sex

To assess potential demographic biases in the training datasets, we compared the gender distributions of the GBM and LGG cohorts using both Pearson's chi-square test and Fisher's exact test.

Patient identifiers corresponding to the training sets were extracted, and gender metadata was retrieved from TCGA clinical files. Samples with unknown gender annotations ($n = 4$) were excluded from the analysis.

A 2×2 contingency table was constructed to summarize the number of male and female samples in each cohort (GBM and LGG). Statistical significance of gender differences was first evaluated using Pearson's chi-square test with Yates' continuity correction. To ensure robustness under small sample conditions, Fisher's exact test was also performed.

Both tests yielded non-significant results ($p > 0.05$), indicating no statistically meaningful gender imbalance between the GBM and LGG training cohorts.

Accordingly, to prevent potential gender-related confounding effects, we excluded features corresponding to genes located on the X and Y chromosomes from the RNA-seq, methylation, CNV, SNV, and protein layers. This filtering was performed using the gene-to-chromosome annotation table provided by GDC, which maps Ensembl Gene IDs to their chromosomal locations. For methylation data, the annotation file downloaded from Xena also has each probe ID with their genomic coordinate.

(iii) Removal of Methylation Probes affected by SNPs

Methylation probes that contain known single-nucleotide polymorphisms (SNPs) within the probe body or at the single base extension (SBE) site may interfere with probe hybridization or extension efficiency, potentially affecting the accuracy and reliability of methylation measurements [17]. Probes overlapping with SNPs have been reported to produce spurious methylation signals or inconsistent measurements, particularly in genetically heterogeneous samples. To minimize potential technical artefacts and ensure the robustness of downstream analyses, we excluded SNP-associated probes based on publicly available annotation files [18]. Details of the annotation file are also provided in Appendix B. This filtering step constitutes an essential part of methylation data preprocessing and contributes to improving the reliability of differential analysis and feature selection.

3.3.4 Batch Correction

The goal of batch correction is to remove non-biological variations that may confound downstream analyses. In this study, we integrated DNA methylation data generated from two different platforms: Illumina 27K and 450K arrays. While this integration increased the overall data coverage, the two platforms are based on distinct experimental protocols and thus are prone to platform-specific technical artifacts.

To address this, we applied batch effect correction using the Python implementation of neuroCombat [19], a widely used tool for harmonizing high-dimensional data. Batch correction was performed separately for the GBM and LGG training datasets, with platform (27K vs. 450K) treated as the batch variable.

Notably, neuroCombat does not natively support the application of frozen parameters to new datasets. To overcome this limitation, we manually extracted the correction parameters estimated from the training data and applied them consistently to the validation, internal test, and external test datasets.

3.3.5 Feature Selection

To support integrative analysis with MOFA+, we performed feature selection for each omics layer based on feature-wise variance. This approach was chosen to address the

limitations of conventional methods such as differential expression (DEG) or differential methylation (DMR) analyses, which primarily capture average differences between GBM and LGG, and may fail to preserve within-label heterogeneity. In our preliminary experiments, DEG/DMR-based selection led to reduced interpretability of latent factors and unstable clustering performance in MOFA+.

To overcome these limitations, we first identified features that were commonly present in both the GBM and LGG training datasets. These shared features were then merged, and variance was calculated across the combined data. Based on this, we selected the top 1,000 high-variance features from the RNA-seq, CNV, and methylation datasets. Before this, RNA-seq data were transformed using $\log_2(\mathrm{CPM}+1)$ to reduce the influence of highly skewed count values and to stabilize variance, enabling more reliable selection. The same set of selected features was then consistently applied to the validation, internal test, and external test datasets to ensure feature alignment across all subsets.

This approach ensured a more balanced contribution of each omics layer during MOFA+ training and prevented overrepresentation from high-dimensional modalities, which could otherwise dominate the latent factor structure.

In contrast, all features from the protein and SNV datasets were retained without filtering. Protein data are inherently low-dimensional and were considered unlikely to introduce significant noise. For SNV data, we chose not to apply variance-based selection to avoid excluding rare but potentially important somatic mutations that may play critical roles in specific glioma subtypes. Retaining the full set of SNV features allowed us to better represent the molecular heterogeneity of glioma within the latent space of MOFA+.

3.3.6 Scaling

To ensure fair comparisons across samples and omics modalities, z-score normalization was applied to all datasets following batch correction. This step standardizes the range of each feature by removing the mean and scaling to unit variance, thereby reducing the influence of differing feature scales.

Z-score normalization was performed per feature, using the mean and standard deviation estimated from the training set. For the validation, internal test, and external test datasets, the same statistics computed from the training data were used to transform each corresponding feature. This reference-based scaling approach prevents data leakage and maintains comparability across all dataset splits.

3.4 Multi-Omics Integration

To integrate different types of modality into a single representation, a Multi-Omics Factor Analysis model (MOFA+) was applied to the training dataset. This state-of-the-art model was chosen for its ability to handle mosaic data patterns, where not all samples are measured across all omics layers. MOFA+ was used to learn a shared latent space across diverse omics layers. After training, the model parameters were manually extracted and used to embed the validation and test sets into the same reference latent space.

3.4.1 Integration Quality Assessment

The aim of this step was to identify the optimal number of latent factors for the MOFA+ model. To do this, models were trained with varying numbers of factors, ranging from 5 to 35, and the proportion of variance explained (VE) was calculated for each omics modality.

3.4.2 Model Selection

At least 60–70% of VE per layer was ideally expected, as this was assumed sufficient to capture meaningful biological structure. In cases where this threshold was not met, we evaluated whether the VE gains plateaued with increasing model complexity. A VE improvement of less than 0.05 per additional factor was used as a practical indicator of saturation. The simplest model that satisfied these conditions across the omics layers was selected for downstream analysis.

4 Methods

4.1 Data Modeling

4.1.1 Clustering

To identify biologically meaningful and robust patient subgroups from integrated multi-omics latent representations, we employed a bootstrap-based consensus clustering framework built upon K-means. This method is designed to account for the sensitivity of clustering to random initialization and sample-level variability, thus ensuring stable and reproducible patient stratification.

(i) Data Preprocessing

Prior to clustering, all latent features were subjected to Z-score normalization followed by standard scaling. To mitigate the influence of outliers on cluster center calculation, any sample with an absolute Z-score exceeding 3 within a bootstrap sample was excluded.

(ii) Bootstrap-Based Clustering and Consensus Matrix Construction

For each candidate number of clusters $k \in \{2, 3, \ldots, 7\}$, we performed 1000 bootstrap iterations. In each iteration, a bootstrap sample was drawn with replacement, and K-means clustering was applied using the *KMeans* method from *sklearn.cluster* module. To reduce instability due to random initialization, the number of K-means restarts (n_{init}) was firstly tested heuristically depending on k: for $k = 2$ to 4, $n_{\text{init}} = 10$; for $k = 5$ to 7, $n_{\text{init}} = 15$. For the final results, $n_{\text{init}} = 10$ were performed for all k.

The clustering labels from each iteration were used to update a co-assignment matrix that records the frequency with which each pair of samples was grouped into the same cluster. After all iterations, the co-assignment matrix was normalized to form the final consensus matrix, with values ranging from 0 to 1 representing pairwise clustering consistency.

(iii) Cluster Number Selection and Evaluation

To evaluate clustering quality and select the optimal number of clusters k^*, two complementary metrics were computed for each k: the consensus-based silhouette score and the within-cluster consensus score, using the method from *sklearn.metrics* module.

First, the silhouette score was computed using the consensus matrix. Specifically, we transformed the consensus matrix into a distance matrix by computing $1 -$ consensus, and used this precomputed distance in the silhouette calculation. Unlike traditional silhouette scores based on Euclidean distances, this variant captures structural stability and reproducibility rather than spatial cohesion.

Second, the within-cluster consensus score was computed by averaging the consensus values of all sample pairs within each cluster and then averaging across all clusters.

The two metrics were combined into a single score to balance structural quality and stability:

$$\text{CombinedScore}_k = \alpha \cdot \text{SilhouetteMean}_{\text{consensus}} + (1 - \alpha) \cdot \text{WithinConsensus}_k$$

We tested $\alpha \in \{0.2, 0.5, 0.8\}$ to explore different trade-offs. For each value, the k with the highest combined score was selected as the final number of clusters k^*.

(iv) Final Cluster Assignment

Given the selected number of clusters k^*, final cluster labels were assigned using either hierarchical clustering or spectral clustering on the consensus matrix. For hierarchical clustering, the consensus matrix was converted to a distance matrix, and clustering was performed using average, complete, or single linkage. The resulting dendrogram was cut at height corresponding to k^* clusters. These steps are implemented using *scipy.cluster.hierarchy* module. Spectral clustering was also applied using the consensus matrix as an affinity matrix, allowing for the identification of more complex, nonlinear cluster boundaries using the *SpectralClustering* method from *sklearn.cluster* module.

Final cluster centroids were computed from the labeled training data in the latent space. Validation and test samples were assigned to clusters by computing Euclidean distances to the centroids and assigning each sample to the nearest cluster.

To evaluate cluster size balance across datasets, we computed the Shannon entropy of the empirical cluster distribution. The entropy was normalized by the maximum possible entropy for k clusters to yield a normalized entropy score, which quantifies the degree of label imbalance and helps assess the evenness of cluster partitioning.

As a baseline method, we also applied K-means clustering directly to the full dataset without bootstrapping or consensus calculation. This conventional clustering approach was included for comparison in subsequent evaluations of cluster balance and predictive utility.

4.1.2 Cluster-Specific Classification

The objective of this step was to construct an optimal classifier for each cluster that is both accurate and interpretable. However, several challenges arose due to the heterogeneous nature of the data. Specifically, (1) the distribution of samples across clusters was imbalanced, and (2) some clusters contained a limited number of samples

or suffered from class imbalance, making it difficult to compare models reliably using standard performance metrics.

To address these issues, we employed a bootstrap-based model selection strategy. For each cluster, we performed 100 bootstrap resampling iterations on the corresponding training set. In each iteration, candidate models were trained using grid search-based hyperparameter tuning, and their performance was evaluated using the cluster-specific validation set. This allowed us to estimate the generalization ability of each model under distributional uncertainty. The best-performing model for each cluster was then selected based on the mean and variance of the F1-score across bootstrap iterations, prioritizing both predictive performance and stability.

We evaluated 11 candidate machine learning algorithms, including Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree (CART), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Quadratic Discriminant Analysis (QDA), CatBoost, XGBoost, and LightGBM (LGBM). The hyperparameter search space for each model is detailed in Table 2.

All models were implemented using the scikit-learn, xgboost, lightgbm, and catboost libraries. While default parameters were maintained for most configurations, hyperparameters were tuned within predefined grids specific to each algorithm.

Ultimately, for each cluster, we selected the model that achieved the highest bootstrap-averaged F1-score with the lowest variance. In cases where multiple models exhibited identical mean performance and variance, the more complex model was chosen to better capture the distribution of the test dataset.

4.1.3 Model Evaluation

To evaluate the generalization performance of the classifiers, we additionally constructed a baseline model trained on the entire dataset without incorporating any cluster information. This baseline model followed the same training and hyperparameter tuning procedure as the cluster-specific classifiers but was trained on all samples as a single pooled set.

Following model selection through bootstrap resampling, both the cluster-specific classifiers and the baseline model were trained on their respective training datasets. Hyperparameter tuning was performed using the validation set, based on the predefined search space described in Table 2, and the final model for each case was selected based on the highest F1 score achieved during validation.

Subsequently, the selected models were evaluated on both the internal and external test sets using four widely adopted performance metrics: accuracy, F1 score, precision, and recall.

To further assess the effectiveness of the cluster-based approach at a cohort-wide level, we introduced an evaluation strategy termed *Integration*. In this setting, predictions from all cluster-specific classifiers were aggregated and treated as a single output, which was then evaluated on the entire test dataset. This strategy enabled a direct comparison between the combined performance of the cluster-specific models and that of the baseline model trained on the full cohort.

Table 2: Hyperparameters for 11 candidate models

Model	**Hyperparameters**
Logistic Regression (LR)	`C = [0.001, 0.01, 0.1, 1, 10, 100]`, `penalty = {l1, l2}`, `solver = {liblinear, saga}`
Linear Discriminant Analysis (LDA)	`solver = {svd, lsqr, eigen}`
K-Nearest Neighbors (KNN)	`n_neighbors = [3, 5, 7, 11]`, `weights = {uniform, distance}`, `metric = {euclidean, manhattan}`
Decision Tree (CART)	`max_depth = [3, 5, 10, 20]`, `min_samples_split = [2, 5, 10]`, `criterion = {gini, entropy}`
Naive Bayes (NB)	`var_smoothing = [1e-9, 1e-8, 1e-7]`
Support Vector Machine (SVM)	`C = [0.1, 1, 10, 100]`, `kernel = {linear, rbf, poly}`, `gamma = {scale, auto}`, `degree = [2, 3, 4]`, `probability = True`
Random Forest (RF)	`n_estimators = [50, 100, 200]`, `max_depth = [None, 5, 10, 20]`, `min_samples_split = [2, 5]`, `criterion = {gini, entropy}`
Quadratic Discriminant Analysis (QDA)	`reg_param = [0.0, 0.1, 0.5, 1.0]`
CatBoost Classifier	`depth = [4, 6, 8]`, `learning_rate = [0.01, 0.05, 0.1]`, `iterations = [100, 200]`, `logging_level = Silent`
XGBoost Classifier	`max_depth = [3, 5, 7]`, `learning_rate = [0.01, 0.1, 0.3]`, `n_estimators = [100, 200]`, `subsample = [0.8, 1.0]`
LightGBM Classifier (LGBM)	`num_leaves = [31, 64, 128]`, `learning_rate = [0.01, 0.1, 0.3]`, `n_estimators = [100, 200]`, `boosting_type = {gbdt, dart}`

4.2 Biological Intepretation

4.2.1 Cluster Annotation

(i) Factor Annotation

Gene Set Enrichment Analysis (GSEA) was applied to the latent factors, forming the feature space of the chosen MOFA+ model, to characterise them at the functional level.

Public Knowledge Bases Curated MSigDB RData files provided by MOFAdata [20] were downloaded from the GitHub—specifically, MSigDB_v.6.0_C2_human, which includes curated gene sets from online pathway databases, PubMed publications, and expert knowledge, and MSigDB_v.6.0_C6_human, containing gene sets extracted from the Gene Ontology database.

Feature Mapping The discrepancies between the MOFA+ input feature space and public knowledge bases were resolved in different ways, depending on the layers. For the CNV, methylation, and RNA-seq layers, simple trimming was performed. For protein and SNV layers, gene symbols were mapped to Ensemble Gene Identifiers. In cases of 1:N mapping, a forced mapping to the first gene was applied as a practical solution, considering both the frequency of such cases and the potential for information loss.

Statistical Testing A parametric t-test with the Benjamini–Hochberg (BH) correction was performed. Gene set statistics were defined as the difference between mean factor values of two groups: one incorporating values associated with a functional term and the other comprising values not associated with the term.

Extraction of Top Enriched Terms Significantly enriched terms with adjusted p-values below 0.05 were selected. To avoid imbalance in contributions across omics layers, at most the top three enriched terms per layer per database were retained for each factor.

(ii) Identification of Discriminative Factors Across Clusters

t-test Assumption Check To determine which factors can significantly distinguish the clusters, pairwise comparisons were conducted across all three cluster combinations. Since the distribution of factor values was unknown, normality was first evaluated using the Shapiro–Wilk test across all 90 cluster–factor pairs (30 factors across 3 clusters). Based on the results, the Wilcoxon rank-sum test was used as a non-parametric alternative. Multiple testing correction was performed using the BH method. The assumption of equal variance was not assessed, because the normality was not satisfied.

Factor Assignment The one-vs-others strategy was selected to assign factors to a cluster. A factor was deemed discriminative for a cluster only if its values significantly differed in both pairwise comparisons. As an illustration, to assign a factor to cluster 0, significant differences must be observed in comparisons between cluster 0 vs. 1 and cluster 0 vs. 2. A p-value ≤ 0.05 was considered statistically significant.

4.2.2 Feature Importance Analysis

The aim of this was to identify and interpret the most informative features contributing to the predictive performance of each cluster-specific classification model. Except for random forest, most of the selected models did not provide built-in feature importance support. For this reason, SHapley Additive exPlanations (SHAP) analysis was chosen as a solution for its model-agnostic and interpretable nature, both locally and globally. The linear explainer was applied for logistic regression and models with a linear kernel; for others, the kernel explainer was employed.

4.2.3 Survival Analysis

To verify the hypothesis that clusters reflect a clinical prognosis and are meaningful beyond prediction, a Kaplan-Meier survival analysis was performed with pairwise log-rank tests, which is implemented using the Python library `lifelines`.

5 Results

5.1 Intra-Cohort Preprocessing

5.1.1 Sample Filtering

As an initial step in sample filtering, we resolved redundancy in RNA-seq data arising from plate-level replication. Figure 6 shows the number of samples per plate before and after filtering, reflecting our approach of retaining only one representative plate per vial—selected based on its frequency to reduce potential batch effects. Following this process, no duplicated entries from the same sample remained in either the GBM or LGG cohorts, as summarized in Table 3.

Table 3: Summary of removed RNA-seq plate-level duplicates during filtering

	Total number of RNA-seq samples	Number of samples from the same sample
GBM	285	0
LGG	516	0

After sample filtering, the final omics composition for each cohort is summarized in Figure 7, illustrating the number of samples available per modality and their intersections. In addition, Table 4 reports the total number of samples retained in each cohort following the full harmonization and filtering steps.

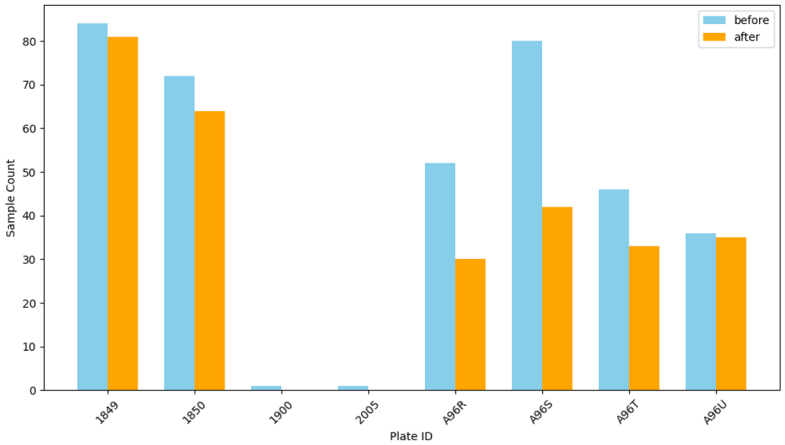
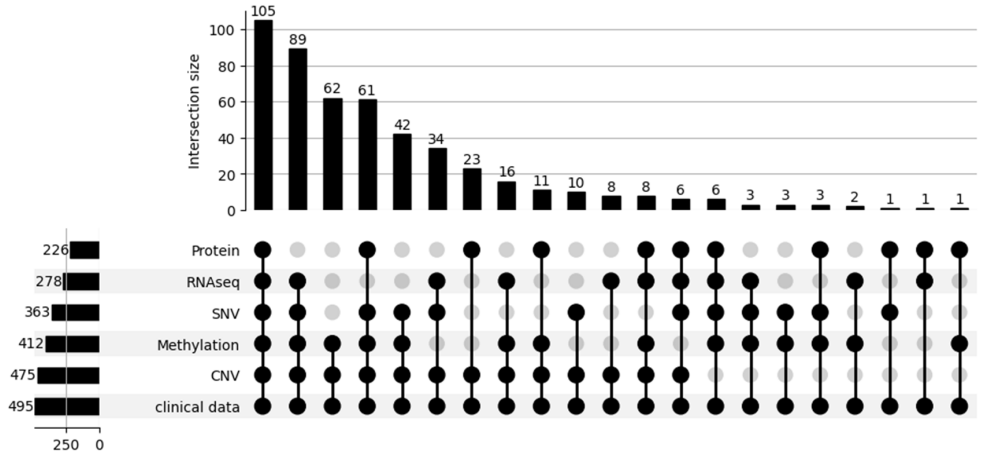
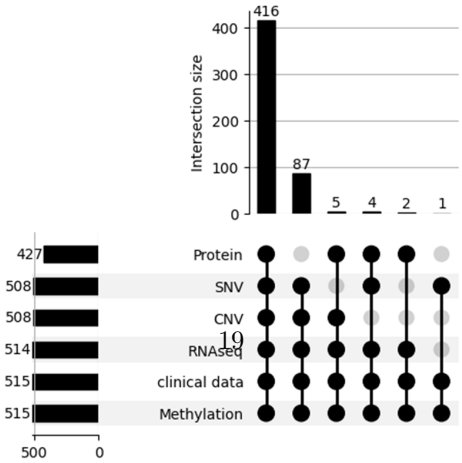


Fig. 6: GBM RNA-seq sample counts per plate before and after redundancy removal. To address plate-level duplication in the GBM cohort, only one representative plate per vial was retained—preferably the one associated with the largest number of samples—to minimize batch-specific effects. The plot illustrates how this filtering step effectively reduced redundancy while preserving overall sample coverage.



(a) GBM data



(b) LGG data

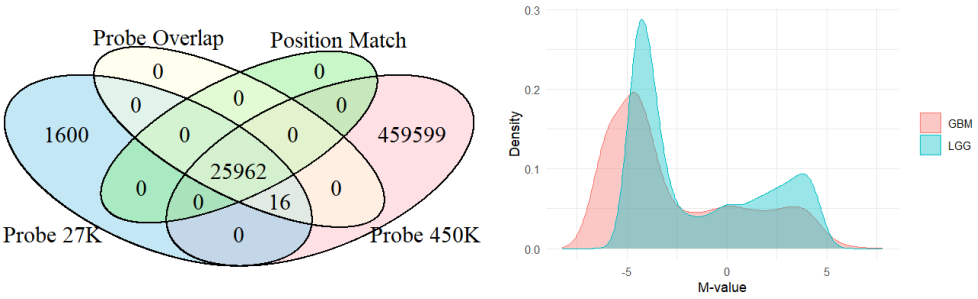
Fig. 7: Final intersection of available omics modalities for each GBM (a) and LGG (b) sample after data harmonization and filtering. The bar plots represent the number of samples sharing specific combinations of omics layers, including protein, SNV, CNV, RNA-seq, DNA methylation, and clinical data.

Table 4: Total number of samples per cohort after final sample filtering.

	GBM	LGG
Total sample count	495	515

5.1.2 GBM Methylation Array Integration Across Platforms

66. 9% (283 / 423) of the TCGA-GBM samples were profiled using the 27k array, which provides more limited genomic coverage. For the five samples that were measured on both platforms, we prioritized the 450k measurements for downstream analysis due to their broader coverage. As shown in the Venn plot Figure 8a, a total of 25,962 common probes were retained for the selection and analysis of the downstream characteristics. All of these probes are checked with the annotation file to ensure they are located in the promoter region (upstream 1500 bp and downstream 500 bp of TSS). Although we prioritized 450k samples where available, we restricted our analysis to the intersection of probes present on both arrays in order to maintain consistency across the full GBM cohort. For the TCGA-LGG cohort, the features were also subsets based on the integrated probe set derived from the GBM data set. Figure 8b shows the distribution of the M values of all remained probes in TCGA-GBM and TCGA-LGG after harmonization and transformation.



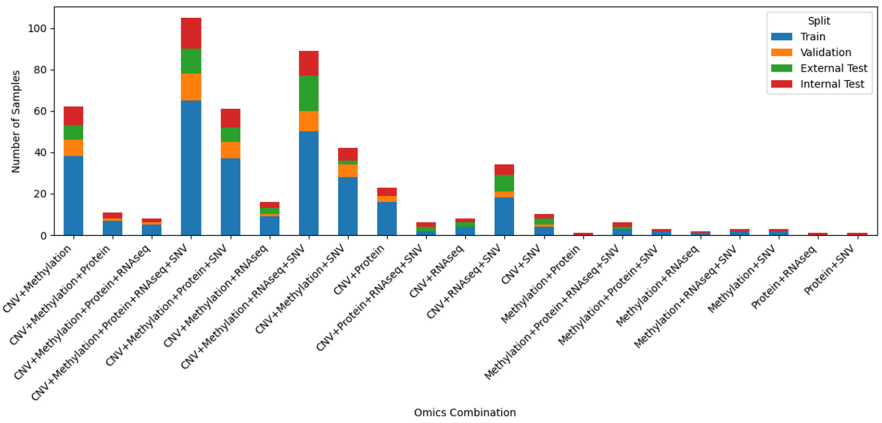
(a) Probe Harmonization (b) M-values Distribution

Fig. 8: Results after Methylation Array Integration: (a) venn plot of methylation probes from different array platform to show how many features are remaind after checking the identifier and genomic position of probes and (b) M values distribution density plot of GBM and LGG

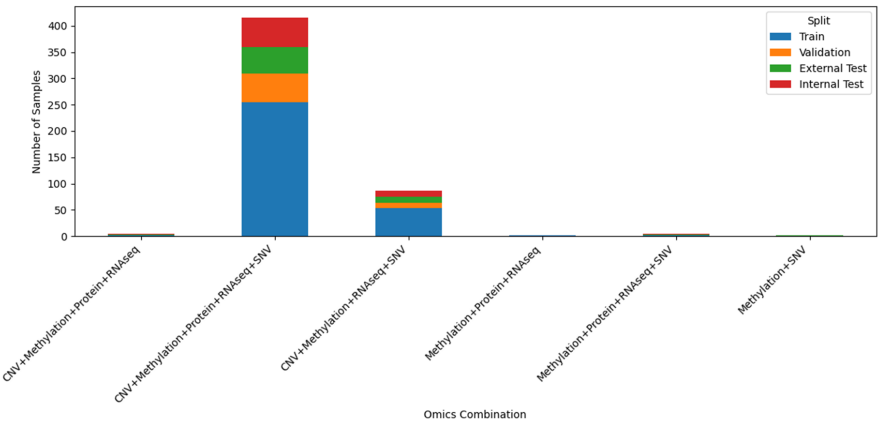
5.2 Inter-Cohort Preprocessing

5.2.1 Data Splitting

As shown in Figure 9, each omics combination is proportionally represented across the training, validation, internal test, and external test datasets. This confirms that the overall distribution of omics layers was preserved during the splitting process. Additionally, detailed statistics for each subset are summarized in Table 5, demonstrating a balanced allocation of samples across GBM and LGG cohorts.



(a) GBM data



(b) LGG data

Fig. 9: Distribution of omics combinations across data splits for (a) GBM and (b) LGG cohorts. Each stacked bar shows the number of samples per omics combination, separated by dataset split (training, validation, internal test, and external test). The figure confirms that the overall distribution of omics layers was preserved across all subsets.

Table 5: Result of data splitting across subsets for GBM and LGG

	Train	Validation	External Test	Internal Test
GBM	293	55	64	83
LGG	312	65	66	72

5.2.2 Feature Filtering

(i) Removal of Features with a High Proportion of Missing Values

Based on the 50% missing value threshold, a small proportion of features were removed from the CNV, methylation, and protein layers, while RNA-seq and SNV layers remained unaffected. The number and percentage of removed features for each omics layer are summarized in Table 6.

Table 6: Number of features removed based on missing value filtering

Omics Layer	Removed Features	Percentage of Total Features
CNV	4,983	8.22%
Methylation	18	0.09%
Protein	23	5.91%
RNAseq	0	0.00%
SNV	0	0.00%

(ii) Removal of Features without a Statistically Significant Association between Disease and Sex

Table 7 summarizes the number of sex-related features removed across omics layers for both GBM and LGG cohorts. For the CNV and SNV datasets, no features were mapped to sex chromosomes, and thus no filtering was required. For the RNA-seq data, genes located on the X and Y chromosomes were excluded. In the proteomics layer, RPPA targets were first mapped to their corresponding gene symbols, after which sex-linked features were identified and removed. For the methylation data, sex-related probes were excluded based on the manifest files provided by the UCSC Xena platform.

(iii) Removal of Methylation Probes affected by SNPs

According to the public annotation files, a total of 2,106 and 2,119 SNP-related probes were excluded in GBM and LGG dataset respectively.

5.2.3 Feature Selection

After completing all preprocessing steps, including feature filtering and selection, we retained a refined set of features across all omics layers for downstream integration.

Table 7: Number of sex-related features removed across omics layers in GBM and LGG datasets

Dataset	RNA-seq	CNV	SNV	Protein	Methylation
GBM	2,990	0	0	10	745
LGG	2,990	0	0	10	853

Specifically, we selected the top 1,000 high-variance features from the RNA-seq, CNV, and methylation datasets to reduce dimensionality while preserving informative signals. For the SNV data, no additional filtering was applied due to its inherent sparsity, resulting in a total of 10,008 retained features. In the proteomics layer, 366 features were retained based on the available RPPA measurements.

5.2.4 Batch Correction

After batch correction using neuroCombat, platform-driven clustering between 27K and 450K samples was effectively resolved, as shown in Figure 10. The silhouette score by batch dropped from more than 0.8 to near zero, confirming the successful removal of the platform effects. Although the validation, internal test, and external test datasets consisted of only 27K samples, the correction parameters estimated from the training set were explicitly applied to all three subsets to ensure consistency and comparability throughout the analysis.

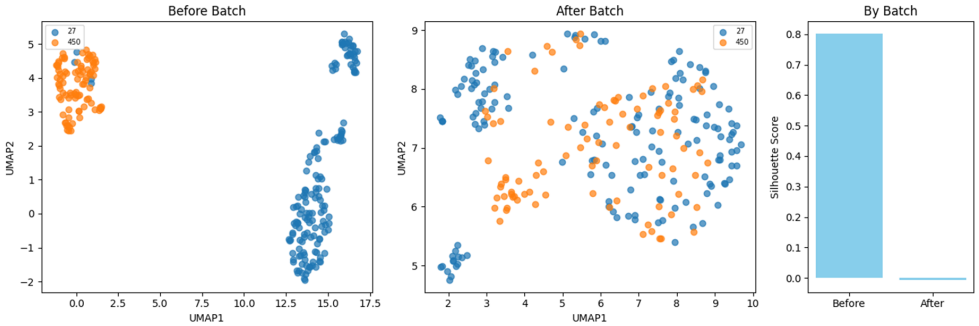


Fig. 10: Batch correction results for methylation (27K vs. 450K) training data. UMAP plots before (left) and after (middle) correction show improved mixing between batches. The silhouette score by batch (right) confirms the reduction in batch effect.

5.3 Multi-Omics Integration

Figure 11 shows the variance explained (VE) per omics layer as a function of the number of latent factors. Across most layers, the VE reached approximately 60–70% with 30 latent factors or plateaued with marginal gains below 0.05 per additional factor, indicating a saturation point. In particular, the CNV and methylation layers

exhibited diminishing returns beyond 20 factors, while the SNV layer showed relatively low VE overall, with minimal improvement after 30 factors.

Based on these observations, we selected the model with 30 latent factors as the final configuration. This model was deemed sufficient to capture biologically meaningful structures across all omics layers and was subsequently used to generate a unified latent representation for downstream analysis.

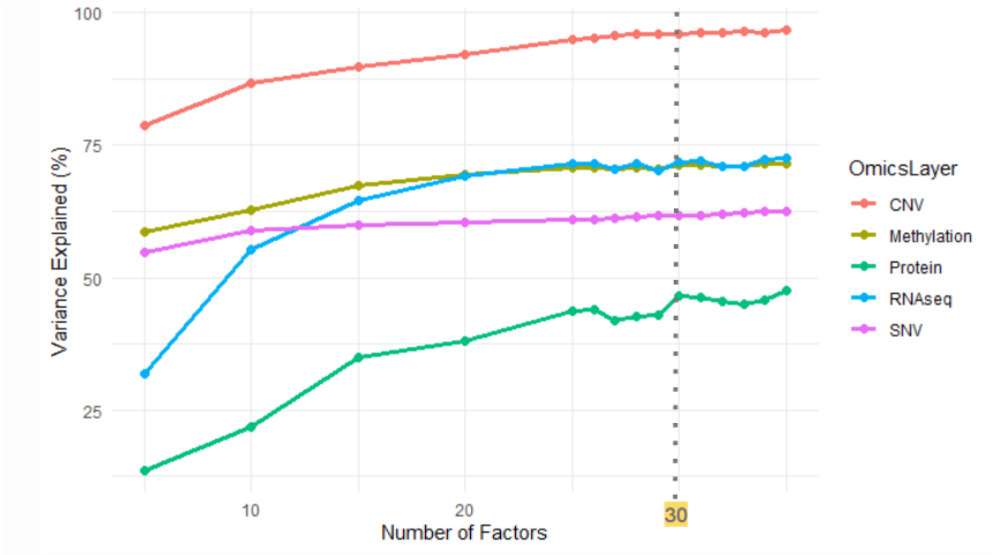
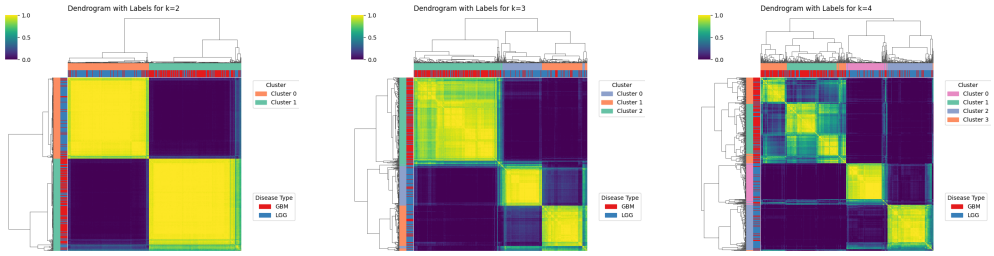


Fig. 11: Variance explained per omics layer as a function of the number of latent factors in MOFA+. The CNV and methylation layers showed high variance explained with increasing factor count, while SNV consistently explained less variance. Based on the observed saturation trend across layers, the model with 30 latent factors was selected for downstream analysis.

5.4 Data Modelling

5.4.1 Clustering

Consensus matrix heatmap with hierarchical clustering dendrogram for$k = 2, 3, 4$ is shown in Figure 12. It shows well-separated and stable clusters, with several dense blocks along the diagonal. The result of $k = 2$ (Figure 12a) clearly shows the two clusters enriched for GBM or LGG, respectively. For $k = 3$ (Figure 12b), it suggests substructure in the LGG enriched group. However, for $k = 4$ (Figure 12c), subgroups inside the GBM enriched cluster show less confidence in the results, while the substructure in the LGG enriched group remains.



(a) $k = 2$ (b) $k = 3$ (c) $k = 4$

Fig. 12: Consensus Matrix Heatmap. The heatmap visualizes the consensus scores between all pairs of samples across 1000 Bootstrap iterations. Values closer to 1 (yellow) indicate that the sample pair was frequently clustered together, whereas values near 0 (dark purple) indicate low co-clustering frequency. Hierarchical clustering dendrogram is displayed along the top and left axes. Moreover, samples are annotated with two types of bars. The top colour bar shows the cluster assignment labels, while the side colour bar indicates the disease types, where red denotes GBM and blue denotes LGG.

We selected $\alpha = 0.5$ to compute the combined score in order to apply a balanced weighting between structural separation and within-cluster consistency. As shown in Figure 13, $k = 3$ achieved the highest combined score under this setting.

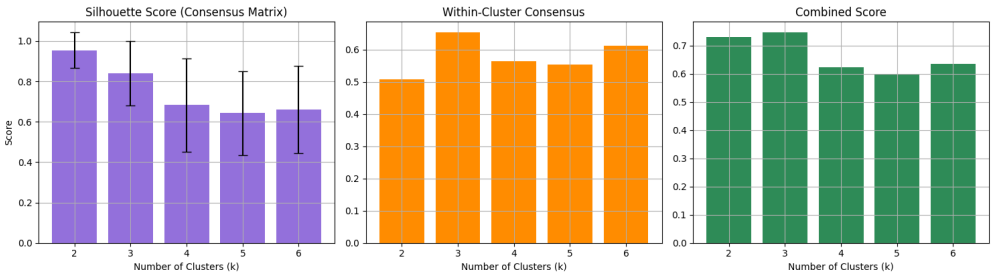


Fig. 13: Clustering evaluation metrics across different numbers of clusters (k). The silhouette score (left), within-cluster consensus (middle), and combined score (right) were used to assess clustering quality.

After determining $k = 3$, we evaluated clustering outcomes using various strategies applied to the consensus matrix: hierarchical clustering with average, complete, and single linkage, as well as spectral clustering. Figure 14 shows the normalized entropy scores across datasets for each method. All approaches except single linkage produced consistently high entropy values, indicating balanced cluster sizes.

To decide which clustering strategy to adopt, we conducted classification experiments using the assigned cluster labels as target variables. Hierarchical clustering with

complete linkage achieved the highest classification performance across all datasets. Based on this result, we selected the complete linkage method as our final clustering approach.

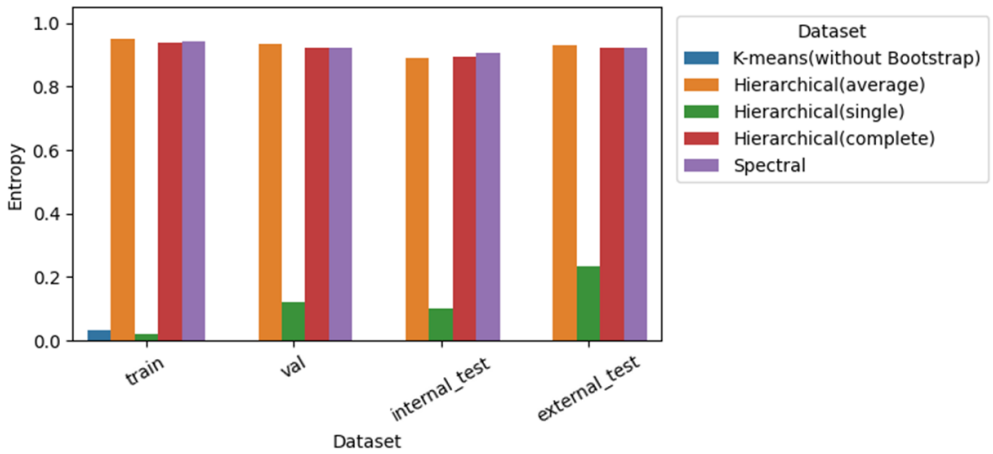


Fig. 14: Normalized entropy across different clustering methods ($k = 3$) for each dataset split. Higher entropy indicates that samples were assigned to multiple clusters across bootstrap iterations, suggesting more stable and unbiased clustering. Spectral and hierarchical (average and complete) clustering showed higher entropy across all splits, supporting their robustness in cluster assignment.

The Euclidean distance of validation, internal test, and external test is calculated in the latent space to the centroid of each cluster of the trained consensus matrix for clustering label assignment. The t-SNE plot of each dataset is shown in Figure 15. The group represents the GBM-dominant cluster 2, and it is obvious that the LGG dominant cluster 0 (blue) and 1 (yellow) share more similarity in the latent space.

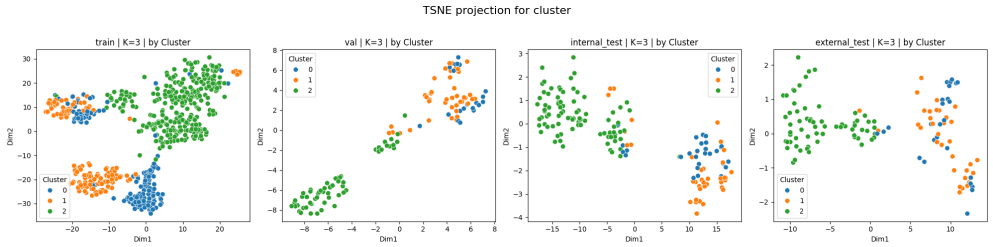


Fig. 15: t-SNE Projection of Clustering assignment in Validation/Internal Test/External Test ($k = 3$)

5.4.2 Cluster-Specific Classification

The results of model selection are presented in Figure 16, with detailed statistics summarized in Table 8. As shown, the SVM achieved the highest average F1-score for both Cluster 0 and Cluster 2, making it the model of choice in these cases. For Cluster 1, Logistic Regression (LR) outperformed other candidates and was selected accordingly.

The performance of the baseline model, trained on the entire dataset without clustering, is illustrated in Figure 17, with corresponding metrics also listed in Table 8. Among the baseline models, LR demonstrated the best F1-score and was thus selected for final evaluation.

The selected models were trained and tuned using their respective training datasets, and the optimal hyperparameters are summarized in Table 9. Notably, linear models were chosen for the baseline, Cluster 0, and Cluster 1, whereas Cluster 2 required a more complex, non-linear classifier—an SVM with a polynomial kernel—suggesting a more intricate data structure.

Finally, the selected models were evaluated on the internal and external test sets. As depicted in Figure 18, models for Cluster 0 and Cluster 1 consistently outperformed the baseline across all evaluation metrics (accuracy, F1-score, precision, and recall). The detailed performance metrics are reported in the rotated Table 10. Compared to the baseline, which yielded an internal F1-score of 0.892 and external F1-score of 0.919, the integration approach improved these scores to 0.964 and 0.977, respectively.

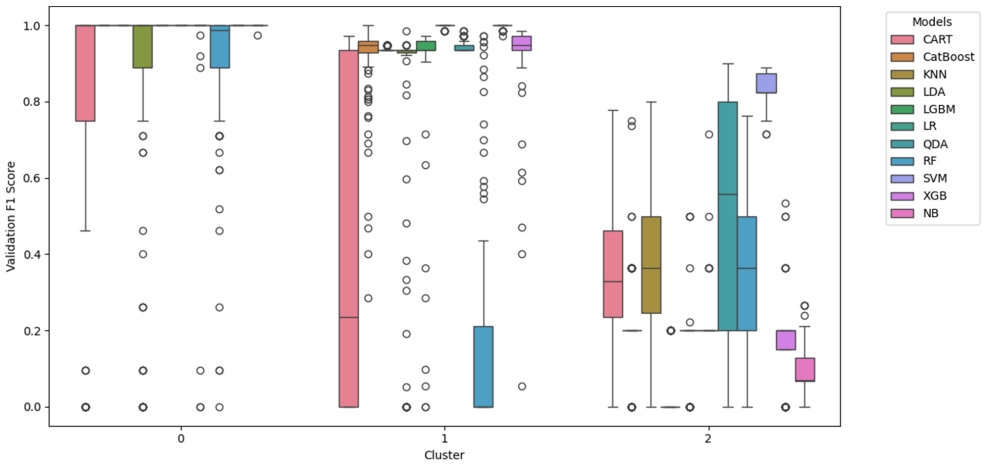


Fig. 16: Cluster-wise distribution of validation F1-scores from bootstrap evaluation. Each box represents the performance distribution of different models for a given cluster.

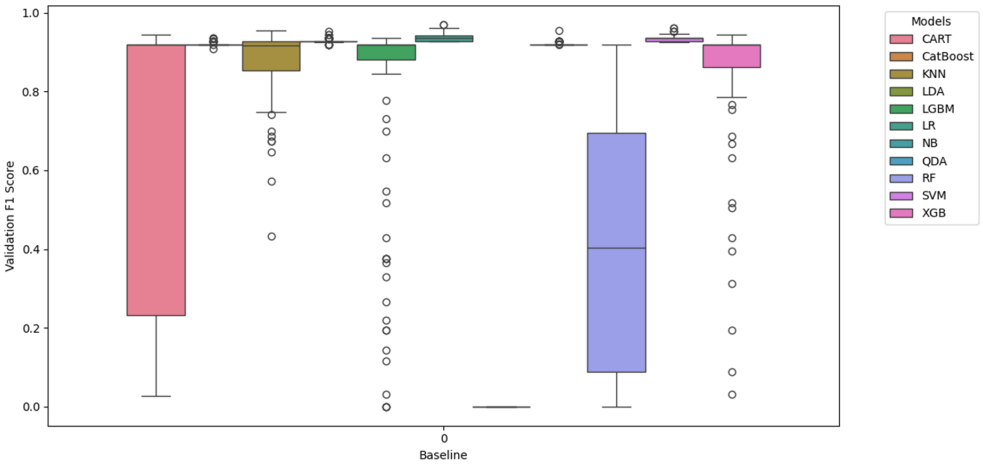


Fig. 17: Bootstrap-based validation F1-scores of the baseline model across different classifiers. Each boxplot shows the distribution of F1-scores for a given model.

Table 8: Bootstrap-averaged F1-score and variance for each model across clusters and baseline

Model	**Cluster 0**		**Cluster 1**		**Cluster 2**		**Baseline**	
	Mean	Var	Mean	Var	Mean	Var	Mean	Var
CART	0.825	0.333	0.456	0.462	0.353	0.166	0.692	0.344
CatBoost	1.000	0.000	0.908	0.121	0.209	0.135	0.921	0.004
KNN	1.000	0.000	0.938	0.005	0.382	0.176	0.878	0.087
LDA	0.815	0.343	0.768	0.339	0.026	0.068	0.928	0.005
LGBM	1.000	0.000	0.893	0.200	0.179	0.095	0.788	0.268
LR	1.000	0.000	**0.999**	**0.003**	0.215	0.067	**0.937**	**0.010**
NB	–	–	–	–	0.089	0.058	0.000	0.000
QDA	0.969	0.166	0.944	0.014	0.490	0.319	0.920	0.005
RF	0.896	0.199	0.173	0.296	0.358	0.168	0.411	0.314
SVM	**1.000**	**0.000**	0.999	0.004	**0.831**	**0.038**	0.935	0.007
XGB	0.999	0.003	0.923	0.128	0.166	0.114	0.847	0.176

Table 9: Selected Models and Tuned Hyperparameters for Baseline and Cluster-Specific Classifiers

Classifier	**Model**	**Hyperparameters**
Baseline	Logistic Regression	`C=100, penalty='l1', solver='liblinear'`
Cluster 0	Support Vector Classifier (SVC)	`C=0.1, degree=2, kernel='linear', probability=True`
Cluster 1	Logistic Regression	`C=1, solver='liblinear'`
Cluster 2	Support Vector Classifier (SVC)	`C=0.1, kernel='poly', probability=True`

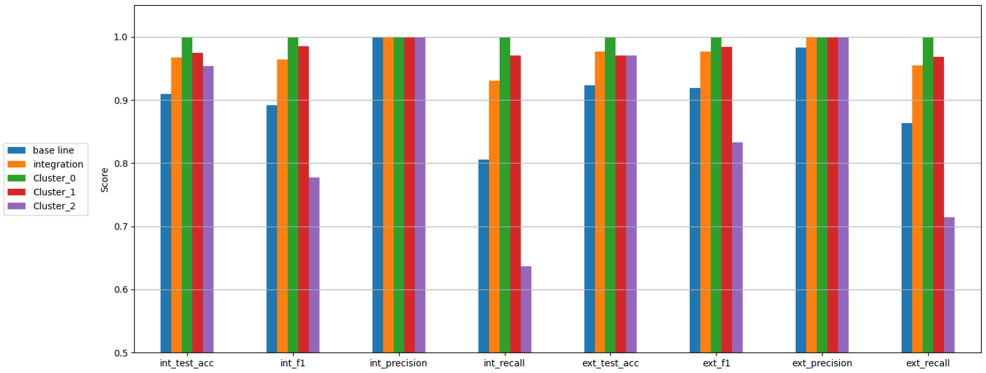


Fig. 18: Test performance comparison of baseline, cluster-specific, and integration models across multiple evaluation metrics. Bars represent scores for accuracy, F1, precision, and recall on internal and external test sets.

5.5 Biological Intepretation

5.5.1 Cluster Annotation

(i) Factor Annotation

To provide an interpretable summary at a glance, we present an oversimplified functional annotation in Table 11. Top-3-enriched terms were extracted per layer and database, and the following criteria steered the annotation: (1) term frequency was prioritised; (2) related functions were grouped into unified phrases; (3) cancer-related mechanisms were emphasised; (4) vague terms, e.g. pathways in cancer, were avoided in favour of specific biological processes; (5) concise two-part phrases were preferred; (6) insights were integrated across omics layers rather than limited to a single layer; and (7) links to other biological pathways, cell types, or regulatory contexts were included when evident, to spotlight potential functional associations beyond the original factor.

For transparency, the full GSEA result table underlying each oversimplified functional annotation is provided in the supplementary file. This spreadsheet contains the top enriched gene sets for each latent factor across different omics layers and served as the basis for the summarised functional terms shown in Table 11.

(ii) Identification of Discriminative Factors Across Clusters

t-test Assumption Check Multiple Shapiro-Wilk tests of factor values showed that they do not follow a normal distribution, as shown in Figure 19. Only 4 out of 90 cases—factors 8 for cluster 0, factors 11 for cluster 1 and factors 8 & 11 for cluster 2—showed p-values greater than 0.05, which prompted the use of the Wilcoxon rank-sum test, a non-parametric method of t-test, as a safer option for pairwise cluster comparisons.

Table 10: Internal and External Test Performance for Baseline, Integration, and Cluster-Specific Classifiers

Model	int_test_acc	int_f1	int_precision	int_recall	ext_test_acc	ext_f1	ext_precision	ext_recall
Base line	0.910	0.892	1.000	0.806	0.923	0.919	0.983	0.864
Integration	0.968	0.964	1.000	0.931	0.977	0.977	1.000	0.955
Cluster_0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Cluster_1	0.975	0.985	1.000	0.971	0.971	0.984	1.000	0.969
Cluster_2	0.954	0.778	1.000	0.636	0.971	0.833	1.000	0.714

Table 11: Oversimplified functional annotations for each latent factor

Factor	Functional Annotation
1	ECM remodelling & integrin signalling
2	Immune signalling & chromatin remodelling
3	Cell cycle control & sensory–metabolic signalling
4	Neural morphogenesis & tumour-linked signalling
5	Adaptive immunity & cell cycle–bone axis
6	Immune signalling & mitotic regulation
7	Epigenetic repression & mitochondrial signalling
8	Ubiquitin signalling & epithelial cell cycle control
9	Chromatin regulation & immune–metabolic interface
10	Epithelial differentiation & hormone–metabolic regulation
11	Immune adhesion & receptor–kinase signalling
12	Neuro-glial development & metabolic regulation
13	Neural ensheathment & detox–nutrient signalling
14	CNS development & immune–tumour regulation
15	Cell cycle arrest & hedgehog–Notch signalling
16	Epithelial renewal & cell cycle–YAP1 signalling
17	Immune activation & glial–cytokine signalling
18	Immune signalling & VEGF-mediated cell communication
19	Lipid metabolism & morphogen signalling
20	Glial–epithelial maturation & vitamin–lipid metabolism
21	Cell cycle–mTOR control & immune–autophagy crosstalk
22	Immune–mitotic coordination & tyrosine–STAT signalling
23	Neural tube patterning & lipid–viral regulation
24	Metabolic reprogramming & DNA damage–immune stress
25	Secreted signalling & replication–EGF regulation
26	Neural–electrical coupling & sex–TGFB signalling
27	Cell migration & YAP–GPCR angiogenic signalling
28	Lipid biosynthesis & integrin–mTOR–cytokine control
29	Mitosis–pyrimidine metabolism & mTOR–immune axis
30	Leukocyte apoptosis & bone–lipid–circadian integration

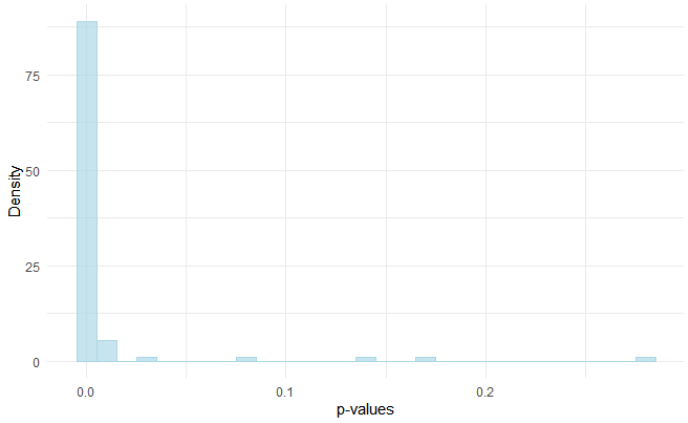
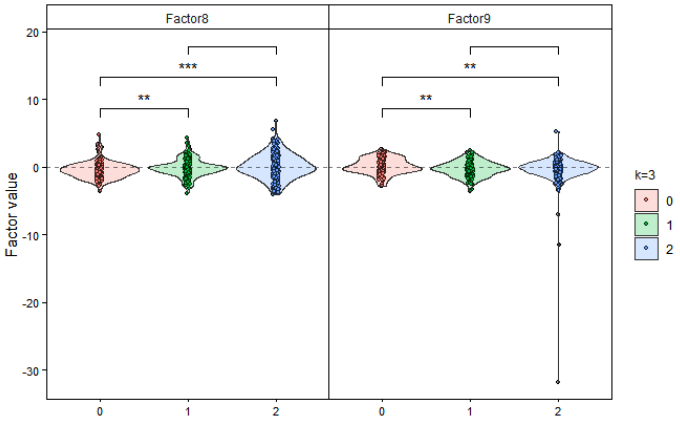
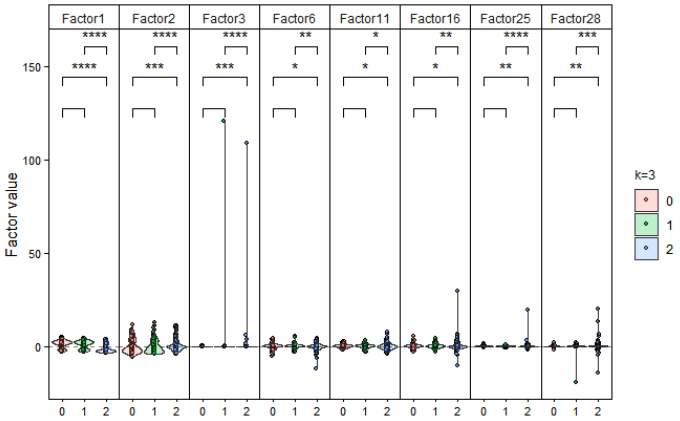


Fig. 19: P-value distribution of multiple Shapiro-Wilk tests. The null hypothesis assumes that the data follow a normal distribution. A p-value ≤ 0.05 indicates significant deviation from normality, allowing the null hypothesis to be rejected.

Factor Assignment The factors with significantly different values across clusters were shown in Figure 20. With the one-vs-others strategy to combine and interpret the pairwise comparison results, factors were assigned to Clusters 0 and 2, respectively, based on consistent significance in both pairwise comparisons. Accordingly, Cluster 0 was distinguished from the other clusters by Factors 8 and 9, both showing significant differences ($p \leq 0.01$ or lower) in pairwise comparisons across clusters (Figure 20a) and Cluster 2 was discriminated by a broader set of factors, including Factors 1, 2, 3, 6, 11, 16, 25, and 28, with several showing highly significant differences, indicating more distinct signals than Cluster 0 (****, $p \leq 0.0001$) (Figure 20b).

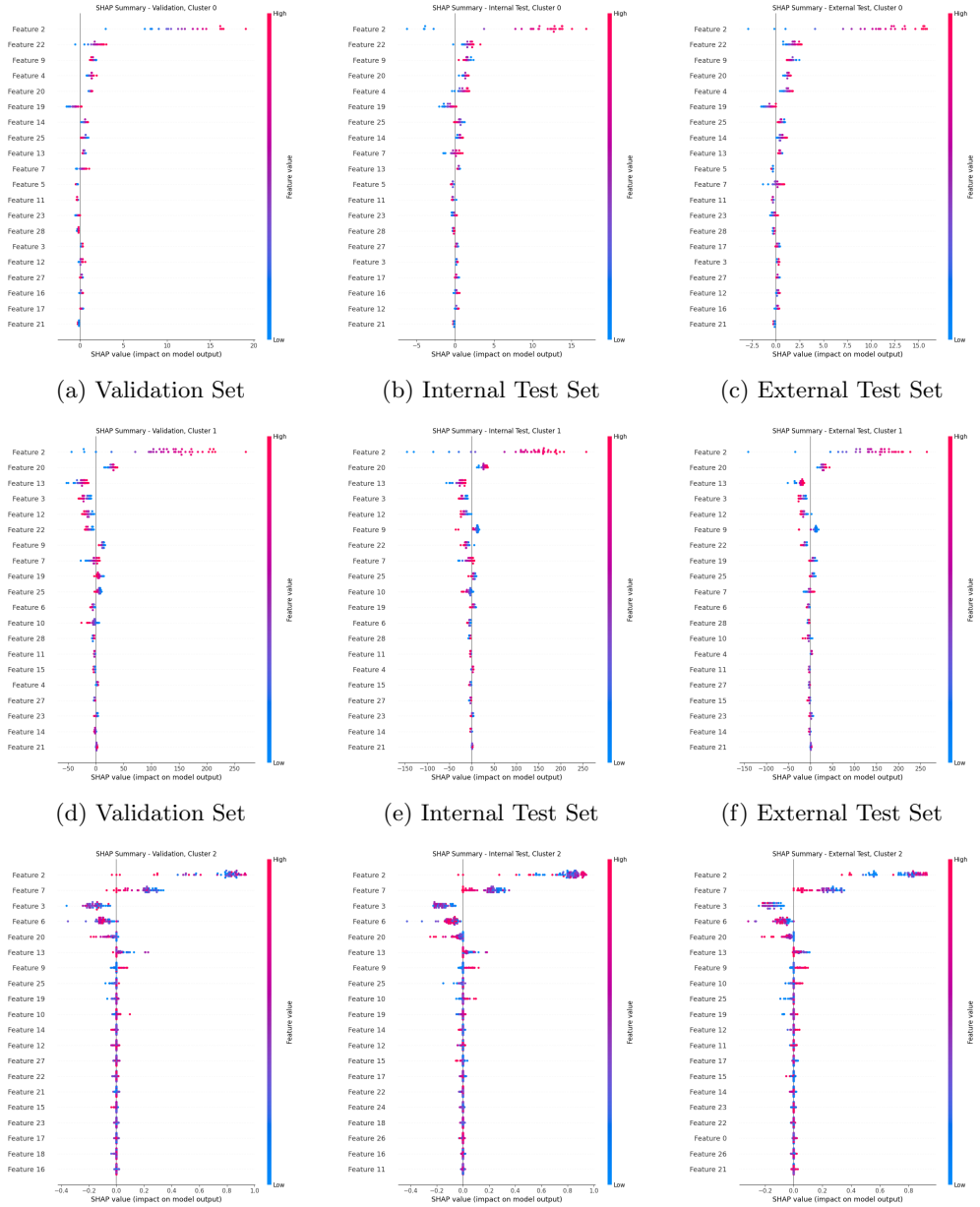


(a) Factors assigned to Cluster 0



(b) Factors assigned to Cluster 2

Fig. 20: Comparative analysis results of factor values across clusters. ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$. Cluster 1 is not shown, as no factors met significance in both pairwise comparisons: Cluster 1 vs. 0 and Cluster 1 vs. 2. Plotted factors were assigned to the cluster where they showed discriminative power.



(a) Validation Set (b) Internal Test Set (c) External Test Set

(d) Validation Set (e) Internal Test Set (f) External Test Set

(g) Validation Set (h) Internal Test Set (i) External Test Set

Fig. 21: SHAP analysis results across different data splits for Cluster 0 (top row), Cluster 1 (middle row) and Cluster 2 (bottom row). Results for the training set are not included, because SHAP values are based on conditional expectations over background (training) data used during model fitting.

5.5.2 Feature Importance

(i) SHAP Analysis

According to Figure 21, Factor 2 consistently showed the highest contribution to predictive power across all splits, implying a strong global signal that can distinguish LGG from GBM. Beyond this, feature importance patterns were consistent across data splits, especially for Cluster 1, Factors 20, 13, and 3 contributed most strongly (Figure 21d-21f), while for Cluster 2, Factors 7, 3, and 6 were most informative (Figure 21g-21i). In the case of Cluster 0, the top three features, Factors 22, 9, and 20, were identical in both internal and external test sets (Figure 21b and 21c). In the validation set, only a minor change was observed in the ranking between the third and fourth features, respectively, Factors 4 and 20 (Figure 21a). These locally important features suggest that certain factors drive prediction in a cluster-dependent manner.

(ii) Cluster-based Interpretation

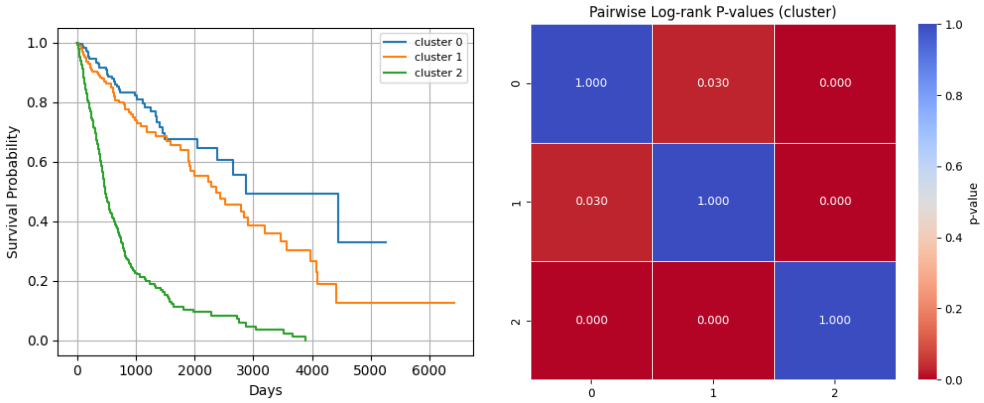
The LGG-dominant Cluster 0 was characterised by two prominent functional pathways: ubiquitin signalling (Factor 8) and chromatin–immune–metabolic regulation (Factor 9). Within this context, additional signals—namely, immune-related STAT signalling (Factor 22) and metabolic–differentiation pathways (Factor 20)—appeared to gain predictive value for distinguishing LGG from GBM.

The functional role of Factor 8 is supported by prior studies demonstrating that E3 ubiquitin ligases contribute to tumour growth by modulating EGFR and hypoxia-inducible factor-1α (HIF-1α) pathways [21]. Similarly, the influence of Factor 9 aligns with evidence showing that chromatin dysregulation in glioblastoma reprogrammes both immune responses and cellular metabolism [22]. These regulatory changes may enhance the discriminative power of immune signals such as STAT activation (Factor 22), which is known to be more predictive in the presence of altered upstream immune control [23].

In contrast, the GBM-dominant Cluster 2 exhibited greater biological complexity, being defined by eight distinct factors. Among them, Factors 1, 2, and 3 showed particularly strong statistical significance ($p \leq 0.001$), suggesting a coordinated interplay among extracellular matrix remodelling (Factor 1), immune–chromatin signalling (Factor 2), and cell cycle–sensory–metabolic control (Factor 3). This triad points to a functional axis involving ECM dynamics, immune modulation, and proliferative capacity—all of which are frequently dysregulated in GBM.

Factors 3 and 6 played a dual role in this cluster: they not only contributed to its biological identity but also served as key drivers in distinguishing LGG from GBM. Factor 6, associated with immune–mitotic signalling, appeared to exert a particularly strong predictive effect when co-activated with the proliferative signals of Factor 3. This interaction reflects previously reported immune–cell cycle co-regulation mechanisms in glioma progression [24], highlighting the importance of context-dependent synergy among pathways.

Furthermore, Factor 7 emerged as a top SHAP feature in Cluster 2 and may modulate this immune–proliferative dynamic through chromatin and mitochondrial regulation. This aligns with studies suggesting that mitochondrial–chromatin interactions influence both immune responses and cell proliferation in glioma [25].



(a) Kaplan-Meier survival Curve (b) Log-rank Tests Heatmap

Fig. 22: Survival Analysis of Clusters (K=3): (a) Kaplan-Meier survival curve of training data shows different and (b) pairwise log-rank tests p-values heatmap

Taken together, these findings emphasise that the predictive value of immune and metabolic features is highly dependent on their regulatory environment. Such dependencies are especially pronounced in clusters where proliferative, chromatin-based, and extracellular matrix signalling are already active.

Cluster 1 had no assigned factors, and therefore, context-specific interpretation was not performed.

5.5.3 Survival Analysis

The results of the Kaplan–Meier survival analysis with pairwise log-rank tests for the three identified clusters are presented in Figure 22, with survival curves shown in Figure 22a. Cluster 2 (green) is predominantly composed of GBM samples, while Cluster 0 (blue) and Cluster 1 (yellow) are largely enriched with LGG samples. The analysis reveals that Cluster 2 exhibits markedly lower survival probabilities compared to Clusters 0 and 1, highlighting the clinical relevance of the derived molecular subtypes. This trend is consistently observed in the validation, internal, and external test sets, as illustrated in Appendix C. Further examination of disease-specific survival curves within each cluster indicates that LGG samples assigned to Cluster 2 also display significantly poorer survival outcomes compared to LGG samples in Clusters 0 and 1 (Appendix C). Notably, while GBM patients generally show inferior survival compared to LGG patients, our findings demonstrate that even within the LGG subgroup, molecular stratification reveals prognostically distinct subpopulations.

As shown in Figure 22b, the pairwise log-rank tests ($\alpha = 0.05$) indicate statistically significant differences in survival across all three clusters. Notably, beyond the expected differences between the GBM-dominant Cluster 2 and the LGG-dominant Clusters 0 and 1, the survival outcome between Clusters 0 and 1 is also significantly different. This suggests that our integrative multi-omics approach successfully captured

molecular heterogeneity within LGG, identifying biologically and clinically distinct subpopulations with divergent prognoses.

6 Discussion

6.1 Leveraging High-Resolution Omics Technologies

In this multi-omics integration and modelling analysis, we primarily used publicly available data from the TCGA cohorts. Due to limitations in sequencing technology and cost at the time of data collection, some omics data, particularly those from earlier phases of the TCGA project, were produced using relatively lower-resolution platforms [26]. In our case, approximately two-thirds of GBM samples were probed using the 27k array, and a portion of transcriptomic data was profiled using microarray rather than high-throughput RNA sequencing.

With the recent advances in sequencing technologies and a significant cost reduction, high-resolution omics data are becoming increasingly accessible and scalable for large cohorts [27]. Among these, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to dissect cell-type-specific expression profiles in glioma, enabling characterization of intra-tumoral heterogeneity at unprecedented resolution [28]. A recent analysis of multi-omics single-cell profiling showed regional plasticity and epigenetic regulation of tumor cells, highlighting the complementary value of integrating scRNA, chromatin accessibility, and proteomic data in glioma studies [29].

Although no current large-scale cohort matches the sample size of TCGA while simultaneously providing single-cell RNA sequencing and other omics layers, integrating high-resolution technologies—such as scRNA-seq, spatial transcriptomics, or high-coverage methylation data (e.g., EPIC or WGBS)—into our multi-omics framework holds great promise. For instance, single-cell data can be incorporated during feature selection to inform bulk-level analysis, or leveraged during downstream annotation to uncover complex biological signals and refine subtype interpretations. Future studies that successfully combine such datasets are expected to enhance biological interpretability and improve the clinical relevance of molecular stratification.

6.2 Limitations of Feature Selection Method

We employed a simple feature selection method where we randomly selected the top 1,000 high-variance features from each omics layer. However, due to time constraints, this selection process was done arbitrarily. Despite this, the results were promising, suggesting that the approach was somewhat effective. However, some information was likely lost due to the arbitrary nature of the selection. Therefore, it will be important to explore more optimized feature selection methods in future studies to enhance the robustness of the results.

While high-variance features are clearly important, there are also features with low variance that may still have a significant impact on the analysis. These features may have been overlooked in our current approach. To address this, future studies could consider more advanced feature selection techniques, such as backward elimination or lasso regression. Both methods have the potential to help identify the most informative

features by iteratively removing less relevant ones or by applying regularization to penalize less significant features. These techniques could help ensure that features contributing meaningfully to the biological context are retained, even if their variance is not particularly high.

6.3 Omics Data Integration Method

6.3.1 Selection of Multi-Omics Integration Tools

In this study, we selected MOFA+ as our primary tool for multi-omics data integration. A key reason for this choice is its native support for mosaic-type data input, which enables effective modeling even when some samples have missing omics layers. Since its publication in 2020, MOFA+ has been widely adopted in the multi-omics community, with strong citation growth and usage.

By contrast, tools developed around the same period, such as DIABLO [30] require complete input matrices and are not suitable for mosaic settings. More recent tools like scMoMaT [31] and Auto-Attention Mosaic Integration tool [32] support mosaic integration at the single-cell level but were explicitly designed for single-cell resolution. Their direct application to bulk datasets, which lack cell-level structure, may reduce model performance and interpretability.

6.3.2 Omics Layer Utilization and Integration Strategy

We integrated five omics layers: RNA-seq, CNV, SNV, DNA methylation, and proteomics. To determine the optimal number of latent factors, we evaluated the Variance Explained per Omics Layer provided by MOFA+, identifying the elbow point where additional factors offered diminishing returns in explained variance.

In future work, comparative studies of omics-layer combinations (e.g., excluding low-contribution layers like proteomics) could enhance understanding of the added value provided by each modality in multi-omics integration.

6.4 Opportunities for Enhancing Clustering Strategies and Model Integration

There are two main limitations in the approach used in this study.

The first limitation concerns the clustering method. We applied consensus clustering to perform multiple clustering iterations to check if consistent results emerged. However, we only used a single clustering method in this study. This means we did not combine multiple clustering techniques to identify the optimal number of clusters or clustering structure. In fact, mixing multiple clustering methods could provide more reliable and precise clustering results. By applying different clustering algorithms together, we can obtain more robust cluster structures and more accurate outcomes. To address this, we should consider combining multiple clustering methods and integrating the results through a consensus clustering approach to obtain more stable and reliable clustering results.

The second limitation is that clustering and classification were performed independently. We selected clusters based on a score threshold and then applied separate

machine learning models to each cluster. However, since clustering and classification were conducted independently, the optimal k-value from clustering and the classification model were not aligned, potentially reducing performance. This lack of optimization between the clustering process and classification could limit the overall effectiveness of the model.

To resolve this issue, it is crucial to adopt an approach that optimizes both clustering and classification simultaneously, such as a mixture model. By using a mixture model, we can treat both clustering and classification as a unified model, allowing for more precise statistical modelling within each cluster. This will enable the creation of more tailored models for each subgroup, leading to more accurate predictions and better analysis.

6.5 Opportunities to Refine Biological Interpretation

6.5.1 Distinguishing Directionality of Feature Contributions

According to the MOFA+ documentation and the original publication [11], feature weights are necessary for better interpreting the sources of variance captured by each factor across omics layers. Taking Figure 23 as an example, Factor 1 includes both positively and negatively weighted genes. However, in the process of GSEA on factors, we jointly pooled features with both signs together, which is the loss of information. Before this analysis, separating features by the sign of their weight could provide greater interpretability by differentiating pathway activation and suppression. This approach represents a promising direction for enhancing biological insight from MOFA+ factors.

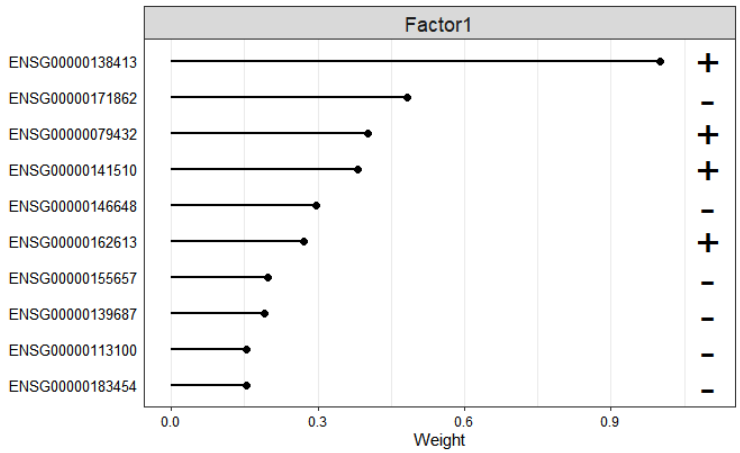


Fig. 23: Feature weights of the SNV layer for factor 1. Each line represents the weight of a feature, a gene in this case, which depicts the strength and direction of each feature's association with a latent factor. A positive weight indicates a positive contribution to the factor, whereas a negative weight indicates an inverse relationship.

6.5.2 Reflecting Layer-Specific Contributions

Each factor in MOFA+ integrates signals from multiple omics layers, but the contribution of each modality varies considerably, as shown in Figure 24. For example, Factor 1 is mainly shaped by methylation and RNA expression, whereas others are more influenced by CNV or SNV. However, in the current enrichment analysis, we treated all features equally, regardless of their omics origin. There is room for improvement by taking into account the layer-specific contributions, such as weighting features based on their source. This could help the interpretation better reflect the biological mechanisms driving each factor.

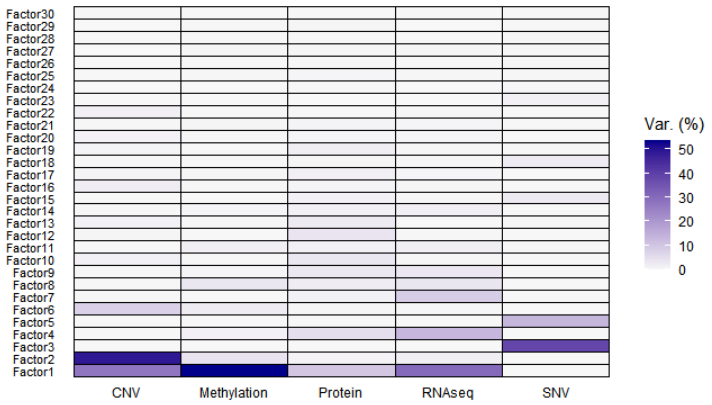


Fig. 24: Omics-layer contributions to each latent factor of the selected MOFA+ model. Colour intensity indicates the proportion of variance explained by each data modality, reflecting the extent to which each omics layer drives the variation captured by individual factors. In the MOFA+ framework, features with positive weights tend to be higher in samples with positive factor values, while features with negative weights tend to be higher in samples with negative factor values.

6.5.3 Expanding Enrichment Term Inclusion

We have only utilised the top three enriched terms per layer and database for factor annotation. While this kept things manageable, we may unintentionally omit other highly significant terms with comparable p-values. Especially when many pathways are statistically robust and biologically associated, limiting the results to a fixed top-n has the potential to slow down the identification of a seed, which serves as a starting point for future work. As a possible improvement, it would make sense to apply an adaptive threshold, such as including all terms below a specific false discovery rate (FDR) cutoff. It could help broaden the interpretation without compromising statistical rigour, particularly for factors which have multiple enriched pathways.

6.6 Use of Independent External Datasets

In the initial phase of data collection, we identified two other publicly available glioma cohorts independent from TCGA, including the Chinese Glioma Genome Atlas (CGGA) [33], and a recent study by Chouleur et al. [34], which profiled 160 IDH-mutant glioma patients who underwent surgery at the University of Milan between May 2012 and June 2018.

Upon further inspection, we found that the CGGA dataset provides three types of omics data: WES, RNA-seq (available as two batches: RNA693 and RNA325), and 27k methylation arrays. However, the methylation data contain 151 samples, among which 150 lack clinical subtype annotations (e.g., LGG vs GBM). For RNA-seq data, significant batch effects are present, requiring correction across cohorts. SNV data is only available in pre-binarized gene–sample matrices indicating mutation presence, without access to raw variant calls (e.g., MAF files), making it infeasible to reproduce the mutation matrix used for TCGA unless raw FASTQ files are reprocessed. For the Chouleur et al. dataset, only transcriptomic and proteomic data are publicly available. The number of samples with both modalities and complete clinical annotation is limited. Moreover, two key omics layers (CNV and methylation), which explained most of the total variance in our MOFA+ model, are not available in either dataset.

Due to these constraints, neither the CGGA nor the Chouleur et al. datasets were included in the current integrative analysis. To ensure the robustness and independence of our test data, we defined a fully independent external test set based on the 'source' field available in TCGA metadata (see Methods), thereby minimizing data leakage risks during integrative model training and validation.

7 References

[1] Louis, D.N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W.: The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta neuropathologica **131**(6), 803–820 (2016)

[2] Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H., Pfister, S.M., Reifenberger, G., *et al.*: The 2021 who classification of tumors of the central nervous system: a summary. Neuro-oncology **23**(8), 1231–1251 (2021)

[3] Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M., *et al.*: Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell **164**(3), 550–563 (2016)

[4] Reuss, D.E., Kratz, A., Sahm, F., Capper, D., Schrimpf, D., Koelsche, C., Hovestadt, V., Bewerunge-Hudler, M., Jones, D.T., Schittenhelm, J., *et al.*: Adult idh wild type astrocytomas biologically and clinically resolve into other tumor entities. Acta neuropathologica **130**(3), 407–417 (2015)

[5] Eckel-Passow, J.E., Lachance, D.H., Molinaro, A.M., Walsh, K.M., Decker, P.A., Sicotte, H., Pekmezci, M., Rice, T., Kosel, M.L., Smirnov, I.V., *et al.*: Glioma groups based on 1p/19q, idh, and tert promoter mutations in tumors. New England Journal of Medicine **372**(26), 2499–2508 (2015)

[6] Ensenyat-Mendez, M., Íñiguez-Muñoz, S., Sesé, B., Marzese, D.M.: igliosub: an integrative transcriptomic and epigenomic classifier for glioblastoma molecular subtypes. BioData mining **14**(1), 42 (2021)

[7] Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, Z., Berman, S.H., *et al.*: The somatic genomic landscape of glioblastoma. Cell **155**(2), 462–477 (2013) https://doi.org/10.1016/j.cell.2013.09.034

[8] Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M., *et al.*: Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell **164**(3), 550–563 (2016) https://doi.org/10.1016/j.cell.2015.12.028

[9] Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., *et al.*: Visualizing and interpreting cancer genomics data via the xena platform. Nature Biotechnology **38**(6), 675–678 (2020) https://doi.org/10.1038/s41587-020-0546-8

[10] Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M.: Toward a shared vision for cancer genomic data. New England Journal of Medicine **375**(12), 1109–1112 (2016) https://doi.org/10.1056/NEJMp1607591

[11] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., Stegle, O.: Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome biology **21**(1), 111 (2020)

[12] National Cancer Institute: The cancer genome atlas (tcga) barcode. Genomic Data Commons (2016). https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/

[13] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.-B., *et al.*: High density dna methylation array with single cpg site resolution. Genomics **98**(4), 288–295 (2011) https://doi.org/10.1016/j.ygeno.2011.07.007

[14] Du, P., Zhang, X., Huang, C., Jafari, N., Kibbe, W.A., Hou, L., Lin, S.M.: Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics **11**(1), 587 (2010) https://doi.org/10.1186/1471-2105-11-587

[15] Genomic Data Commons: Reverse Phase Protein Array (RPPA) Data. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/RPPA_intro/

[16] MD Anderson Functional Proteomics RPPA Core Facility: RPPA Expanded Antibody List (Updated). https://www.mdanderson.org/content/dam/mdanderson/documents/core-facilities/Functional%20Proteomics%20RPPA%20Core%20Facility/RPPA_Expanded_Ab_List_Updated.xlsx

[17] Meeks, G.L., Henn, B.M., Gopalan, S.: Genetic differentiation at probe snps leads to spurious results in meqtl discovery. Communications Biology **6**(1), 1295 (2023) https://doi.org/10.1038/s42003-023-05446-5

[18] Zhou, W.: Infinium Methylation Annotation Resources. https://zwdzwd.github.io/InfiniumAnnotation/. Accessed: 2025-07-30 (n.d.)

[19] Fortin, J.-P.J., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T.: Harmonization of cortical thickness measurements across scanners and sites. NeuroImage **167**, 104–120 (2018)

[20] Argelaguet, R., Velten, B., Arnol, D., Buettner, F., Huber, W., Stegle, O.: MOFAdata: Data package for Multi-Omics Factor Analysis (MOFA). https://bioconductor.org/packages/MOFAdata. R package version 1.24.0 (2025). https:

//doi.org/10.18129/B9.bioc.MOFAdata

[21] Scholz, N., Kurian, K.M., Siebzehnrubl, F.A., Licchesi, J.D.: Targeting the ubiquitin system in glioblastoma. Frontiers in oncology **10**, 574011 (2020)

[22] McClellan, B.L., Haase, S., Nunez, F.J., Alghamri, M.S., Dabaja, A.A., Lowenstein, P.R., Castro, M.G., et al.: Impact of epigenetic reprogramming on antitumor immune responses in glioma. The Journal of clinical investigation **133**(2) (2023)

[23] Ou, A., Ott, M., Fang, D., Heimberger, A.B.: The role and therapeutic targeting of jak/stat signaling in glioblastoma. Cancers **13**(3), 437 (2021)

[24] Elguindy, M.M., Young, J.S., Ho, W.S., Lu, R.O.: Co-evolution of glioma and immune microenvironment. Journal for ImmunoTherapy of Cancer **12**(12), 009175 (2024)

[25] Chen, L., Zhang, H., Shang, C., Hong, Y.: The role and applied value of mitochondria in glioma-related research. CNS neuroscience & therapeutics **30**(12), 70121 (2024)

[26] National Cancer Institute: Technologies used in TCGA. https://www.cancer.gov/ccg/research/genome-sequencing/tcga/using-tcga-data/technology. Accessed: 2025-07-29 (n.d.)

[27] Lim, B., Choi, I., Kim, K., Park, J.: Advances in single-cell omics and multiomics for high-resolution molecular profiling. Experimental & Molecular Medicine **56**(1), 1–12 (2024) https://doi.org/10.1038/s12276-024-01186-2

[28] Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., *et al.*: Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science **344**(6190), 1396–1401 (2014) https://doi.org/10.1126/science.1254257

[29] Wang, X., Sun, Q., Liu, T., Lu, H., Lin, X., Wang, W., Liu, Y., Huang, Y., Huang, G., Sun, H., et al.: Single-cell multi-omics sequencing uncovers region-specific plasticity of glioblastoma for complementary therapeutic targeting. Science Advances **10**(47) (2024) https://doi.org/10.1126/sciadv.adn4306

[30] Rohart, F., Gautier, B., Singh, A., Lê Cao, K.-A.: mixomics: An r package for 'omics feature selection and multiple data integration. PLOS Computational Biology **13**(11), 1005752 (2017) https://doi.org/10.1371/journal.pcbi.1005752

[31] Li, J., Tong, R., Cao, J., Xiong, L., Tan, L., Lin, X., Chen, Y., Yuan, Y., Lu, X., Chen, P., *et al.*: scmomat: single-cell multi-omics mosaic integration using matrix tri-factorization. Nature Communications **14**(1), 1273 (2023) https://doi.org/10.1038/s41467-023-36676-w

[32] Zhao, Y., Xu, M., Wang, L., Hu, K., Zhang, T., Li, Z., Wang, H., Yu, W., Cheng, J., Wang, Z.: Single-cell mosaic integration and cell state transfer with auto-scaling self-attention mechanism. Nature Machine Intelligence (2024) https://doi.org/10.1038/s42256-024-00844-x . in press

[33] Zhao, Z., Zhang, K., Wang, Q., Li, G., Zeng, F., Zhang, Y., Wu, F., Chai, R., others, Jiang, T.: Chinese glioma genome atlas (cgga): A comprehensive resource with functional genomic data from chinese glioma patients. Genomics, Proteomics & Bioinformatics **19**(1), 1–12 (2021) https://doi.org/10.1016/j.gpb.2020.10.005

[34] Chouleur, T., Etchegaray, C., Villain, L., Lesur, A., Bello, L., al.: A strategy for multimodal integration of transcriptomics, proteomics, and radiomics data for the prediction of recurrence in patients with idh-mutant gliomas. International Journal of Cancer **157**(3), 573–587 (2025) https://doi.org/10.1002/ijc.35441

8 Author contributions

8.1 Gyeongsil Kim

Gyeongsil Kim was responsible for: **1)** selecting glioma datasets (TCGA LGG/GBM) to prioritise sufficient sample size, **2)** reviewing literature on glioma to support methodological decisions, **3)** proposing a cluster-based classification framework as the core research idea to tackle the heterogeneity of glioma, **4)** designing the complete analysis workflow: preprocessing, modelling, and biological interpretation, **5)** defining preprocessing steps (e.g. quality control, scaling, feature selection, multi-omics integration), **6)** proposing a TSS-based data split strategy for internal vs. external validation, **7)** identifying limitations of batch correction under TSS splitting, **8)** performing statistical tests (association between sex and disease) for feature selection, **9)** assessing multi-omics integration quality to select the optimal integration model, **10)** designing robust clustering steps with the mitigation of initialisation sensitivity and consensus clustering, **11)** proposing a unified metric combining silhouette and consensus scores to select the number of clusters, **12)** exploring a range of clustering methods and distance metrics, **13)** designing cluster-specific classification steps, **14)** outlining steps for biological interpretation: cluster annotation, feature importance, and survival analysis, **15)** implementing GSEA-based cluster annotation using latent factors, **16)** analysing feature importance taking the cluster into account, **17)** developing a unified logging system for transparent coordination.

ChatGPT was utilised in the course of the project for the following purposes: **1)** identifying and addressing logical gaps between methodological steps while designing the overall analytical workflow, **2)** refining the oversimplification approach for GSEA on latent factors, **3)** generating visualisations using `ggplot2`, **4)** converting plain text, plots and tables into LaTeX-compatible format, and **5)** improving the clarity, grammar, and phrasing of written content, in conjunction with Grammarly, to ensure precise and natural academic expression.

8.2 Daeho Lee

Daeho Lee was responsible for: **1)** Downloading and organizing multi-omics data (RNA-seq, CNV, SNV, protein, and methylation arrays) for the TCGA GBM and LGG cohorts. **2)** Preprocessing and harmonizing RNA-seq, CNV, SNV, and protein data, including the removal of RNA-seq duplicates and selection of representative samples. **3)** Aligning TCGA barcodes at the vial level across omics layers and standardizing feature identifiers. **4)** Filtering samples with excessive missingness and constructing a mosaic-structured input matrix for MOFA+. **5)** Selecting top high-variance features within each omics layer for dimensionality reduction. **6)** Designing and applying batch correction, and extending transformations to validation and test sets. **7)** Proposed and implemented a data splitting strategy using KL divergence to minimize distributional differences across omics layers, dividing samples into training, validation, and test sets (7:1.5:1.5) while maintaining balanced omics composition. **8)** Implementing MOFA+ for joint latent factor modeling across five omics layers. **9)** Manually applying MOFA+ model parameters (e.g., weights, scaling) to validation and test sets to address

generalization limitations. **10)** Debugging the consensus clustering pipeline, including correction of the `n_init` parameter in K-means. **11)** Applying Z-score normalization and excluding outliers ($|Z| > 3$) to improve clustering stability. **12)** Evaluating clustering quality using silhouette scores and normalized entropy to determine optimal k. **13)** Proposing and developing a bootstrap-based evaluation framework for robust classifier selection within clusters. **14)** Training, validating, and testing 11 classifiers per cluster, with model selection based on the mean and variance of F1-scores. **15)** Assessing model performance on internal and external test sets using accuracy, precision, recall, and F1-score. **16)** Developing and integrating SHAP-based explainability modules, along with visualizations of feature contributions.

ChatGPT was used to: **1)** generate Python-based plots and LaTeX tables, **2)** check grammar and identify ambiguous or overly complex expressions in the manuscript, **3)** provide assistance with formatting, structuring, and refinement of scientific text.

8.3 Jingwen Luo

Jingwen Luo was responsible for: **1)** Downloading multi/omics data for GBM and LGG from Xena and GDC portal, exploratory analysis of multi-omics coverage and clinical variable completeness, identifying sample-level overlap and platform-specific missingness. **2)** Intra-data preprocessing including SNV MAF value based gene-sample matrix integration, methylation array integration, clinical data integration. **3)** Inter-data preprocessing including feature mapping for methylation data, feature selection for mutation and methylation data, batch correction for methylation array platform. **4)** Clustering code implementation with several trails to find the optimal hyper-parameters and method. **5)** Biological interpretation including comparison of clusters to known clinical and molecular variables, survival analysis. **6)** Exploration of independent external data set.

ChatGPT was used to: **1)** issue-identification during code implementation **2)** assistance to converting data to LaTeX tables, **3)**assistance to check grammar and expressions in the report writing.

Appendix A Data Source and Availability

Table A1: Details of Data Source and Availability

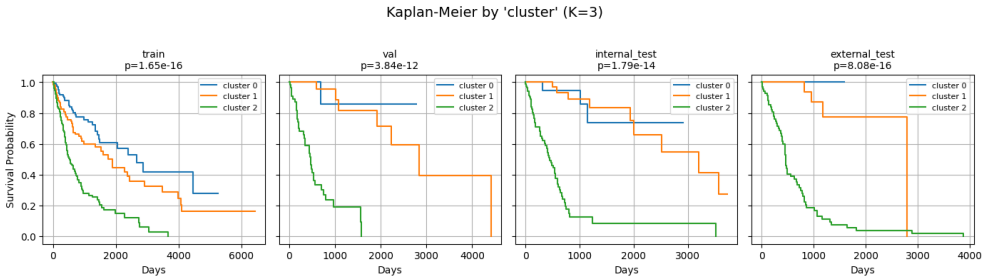
Omics	Data Source	Version
SNV	UCSC Xena (GBM somatic mutation (SNPs and small INDELs))[1]	07-30-2024
	UCSC Xena (LGG somatic mutation (SNPs and small INDELs))[2]	08-05-2024
CNV	UCSC Xena (GBM Gene Level Copy Number (ABSOLUTE))[3]	05-13-2024
	UCSC Xena (LGG Gene Level Copy Number (ABSOLUTE))[4]	05-10-2024
Methylation	UCSC Xena (GBM Illumina Human Methylation 27)[5]	05-13-2024
	UCSC Xena (GBM Illumina Human Methylation 450)[6]	05-13-2024
	UCSC Xena (LGG Illumina Human Methylation 450)[7]	05-10-2024
RNAseq	UCSC Xena (GBM STAR - Counts)[8]	09-26-2024
	UCSC Xena (LGG STAR - Counts)[9]	05-10-2024
Protein Expression	UCSC Xena (GBM normalized RPPA value)[10]	09-07-2024
	UCSC Xena (LGG normalized RPPA value)[11]	09-07-2024
Clinical	Download via R script	06-10-2025

Appendix B Annotation Files

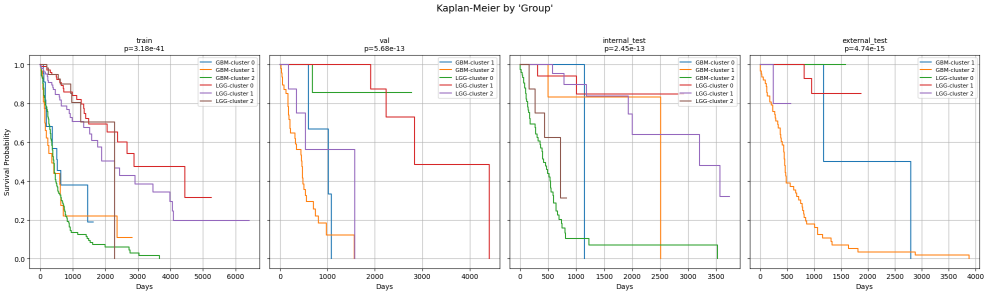
Table B2: Details of Annotation Files

Filename	Description	Source
gencode.v36.annotation.gtf.gz	Gene annotation file based on GENCODE v36	Xena [9]
HM27.hg38.manifest.gencode.v36.tsv.gz	Probe-to-gene annotation for Illumina Human-Methylation27 array	Xena [9]
HM450.hg38.manifest.gencode.v36.tsv.gz	Probe-to-gene annotation for Illumina Human-Methylation450 array	Xena [9]
HM450.hg38.snp.tsv.gz	SNP overlap information for probes on the HumanMethylation450 array	ref [18]

Appendix C Kaplan-Meier Survival Plot



(a) Kaplan-Meier survival Curves within each Cluster (All Dataset)



(b) Disease-specific Survival Curves within each Cluster

Fig. C1: Kaplan-Meier survival Curves (a) within each cluster in Training (first), Validation (second), Internal Test (third), External Test (fourth); (b) Disease-specific within each cluster in Training (first), Validation (second), Internal Test (third), External Test (fourth)