



DEPARTMENT OF
MATHEMATICS

CAROLINA FERREIRA PÉ-CURTO

BSc in Mathematics

DECIPHERING GLIOMA HETEROGENEITY THROUGH MULTI-VIEW CLUSTERING:

TOWARDS IMPROVED DISEASE MOLECULAR
SUBTYPING AND PERSONALISED MEDICINE

MASTER IN MATHEMATICS AND APPLICATIONS

NOVA University Lisbon
September, 2024

DECIPHERING GLIOMA HETEROGENEITY THROUGH MULTI-VIEW CLUSTERING:

TOWARDS IMPROVED DISEASE MOLECULAR
SUBTYPING AND PERSONALISED MEDICINE

CAROLINA FERREIRA PÉ-CURTO

BSc in Mathematics

Adviser: Marta Belchior Lopes
Assistant Researcher, FCT-NOVA

Co-adviser: Roberta Coletti
Researcher, FCT-NOVA

Examination Committee

Chair: Isabel Cristina Maciel Natário
Associate Professor, FCT-NOVA

Rapporteur: Rui Miguel Carrasqueiro Henriques
Assistant Professor, Instituto Superior Técnico

Adviser: Marta Belchior Lopes
Assistant Researcher, FCT-NOVA

MASTER IN MATHEMATICS AND APPLICATIONS

NOVA University Lisbon

September, 2024

**DECIPHERING GLIOMA HETEROGENEITY THROUGH MULTI-VIEW CLUSTERING:
TOWARDS IMPROVED DISEASE MOLECULAR SUBTYPING AND PERSONALISED
MEDICINE**

Copyright © Carolina Ferreira Pé-Curto, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

ACKNOWLEDGEMENTS

The path to completing this dissertation would not have been possible without the help of various people and I would like to appreciate their support.

To my adviser Marta Lopes and co-adviser Roberta Coletti, I am extremely thankful for your guidance and encouragement this past year. Without your wisdom, meticulous attention to detail, and patience in answering even my most trivial biological questions, my thesis would not have reached the level of depth and quality it has today.

I want to express my deepest gratitude to my family, especially my parents, brother, and grandparents. Thank you for your unwavering support and for always being there to encourage me. The values you have instilled in me have shaped who I am today, and your belief in my potential has been a constant source of strength. I truly could not have reached this milestone without you.

I am also thankful to my friends for listening and distracting me when I most needed it. A special note of appreciation goes to Ana and Diana, who have significantly influenced my life in and outside of university. Your kind words, supportive nature, playful humor, and our love for music have brightened the past year.

Each of you played a fundamental role no matter how small in making this thesis and I am profoundly grateful for every word.

” *“If you are the smartest person in the room, then you
are in the wrong room.”*

— **Confucius**, Philosopher

ABSTRACT

Gliomas are the most common type of brain tumor, associated with poor prognosis and high mortality. Despite advances, the classification of gliomas remains unresolved, emphasizing the need for further disease subtyping to better stratify patients. In this study, three multi-view clustering methods—CIMLR, intNMF, and moCluster—were employed to cluster glioma patients, using transcriptomics and methylomics data. The best obtained partition is closely aligned with the 2021 WHO CNS classification. However, the 3 clusters produced by intNMF exhibited distinct survival distributions, whose significance should prompt further biological validation studies that could potentially support a revision of the current classification of glioma types.

Key molecular features identified were cross-referenced with existing glioma literature, validating both established and recently discovered biomarkers. To improve robustness, we applied consensus clustering to integrate the partitions generated by the multi-view methods. The resulting consensus solution produced three clusters that mirrored the 2021 WHO glioma classification, reinforcing the validity of this approach.

While this consensus cluster is compatible with current glioma knowledge, there is still a need for further research to refine glioma subtypes and improve personalized cancer treatment. The ultimate goal remains to enhance prognosis and patient outcomes.

Keywords: Glioma, Multi-omics, Multi-view clustering

RESUMO

Os gliomas são o tipo mais comum de tumor cerebral, associado a um mau prognóstico e a uma elevada mortalidade. Apesar dos avanços, a classificação dos gliomas continua por resolver, salientando a necessidade de uma maior subcategorização da doença para melhor separar os pacientes. Neste estudo, foram utilizados três métodos de agrupamento *multi-views* - CIMLR, intNMF e moCluster - para agrupar doentes com gliomas, utilizando dados transcriptómicos e metilómicos. O melhor agrupamento obtido assemelha-se à classificação da WHO CNS de 2021. No entanto, os 3 clusters produzidos do intNMF apresentaram distribuições de sobrevivência distintas, cuja significância deve desencadear mais estudos biológicos de confirmação que potencialmente validem uma revisão da classificação atual dos tipos de glioma.

As principais características moleculares identificadas foram comparadas com a literatura existente sobre gliomas, validando tanto os biomarcadores estabelecidos como os recentemente descobertos. Para melhorar a robustez, aplicou-se *consensus clustering* para integrar as partições geradas pelos métodos de *multi-view*. A solução de *consensus* produziu 3 grupos que refletem a classificação de glioma da WHO CNS de 2021, reforçando a validade desta abordagem.

Embora o *consensus cluster* seja compatível com o conhecimento atual do glioma, ainda há uma necessidade de mais estudos para refinar os tipos de glioma e melhorar o tratamento personalizado do cancro. O objetivo final continua a ser melhorar o prognóstico e os resultados dos pacientes.

Palavras-chave: Glioma, Multi-ómica, Multi-view clustering

CONTENTS

List of Figures	viii
List of Tables	x
Acronyms	xii
1 Introduction	1
1.1 Problem Statement	1
1.2 Motivation and Goals	2
1.3 Structure	3
2 Literature Review	4
2.1 Clustering Methods	4
2.1.1 Hierarchical Clustering	4
2.1.2 Nonhierarchical Clustering	5
2.2 Multi-view Clustering Methods	11
2.2.1 Matrix Factorization Methods	11
2.2.2 Nonnegative Matrix Factorization	15
2.2.3 Graph-based Methods	19
2.3 Consensus Clustering Methods	24
2.3.1 Median Partition	24
2.3.2 Object Co-occurrence	25
2.4 Clustering Validation Measures	29
2.4.1 Internal Validation	29
2.4.2 External Validation	31
3 Methodology	33
3.1 Pipeline	33
3.2 Data Collection and Preprocessing	33
3.3 Multi-view Clustering Methods	35

3.4	Consensus Clustering Methods	35
3.5	Evaluate Clustering Performance	37
3.6	Implementation Details	38
4	Results and Discussion	39
4.1	Data Visualization	39
4.2	Multi-view Clustering	39
4.2.1	CIMLR	39
4.2.2	intNMF	51
4.2.3	moCluster	62
4.2.4	Comparison	80
4.3	Consensus Clustering	82
4.3.1	3 Methods	82
4.3.2	2 Methods	85
5	Conclusion	90
	Bibliography	92
	Appendices	
A	CIMLR	110
A.1	Parameter Tuning	110
A.2	Different Seed - 3 Clusters	111
B	intNMF	113
B.1	5 Clusters	113
B.1.1	Relevant Features	115
B.1.2	Survival Analysis	116
B.2	6 Clusters	119
B.3	Convergence Study	120
B.3.1	3 Clusters	120
B.3.2	4 Clusters	120
B.3.3	5 Clusters	122
B.3.4	6 Clusters	122
B.4	Different Seed - 3 Clusters	124
C	moCluster	125
C.1	Gap statistic	125
C.1.1	2 Joint Latent Variables	125
C.1.2	6 Joint Latent Variables	130

LIST OF FIGURES

3.1	Thesis pipeline	34
4.1	t-SNE 2D Projection of: a) Transcriptomics. b) Methyloomics.	40
4.2	CIMLR result with 3 clusters.	41
4.3	Separation cost (y-axis) for different numbers of clusters (x-axis).	43
4.4	CIMLR result with 4 clusters.	44
4.5	Wordcloud with the genes mapped from the significant CpGsites with: a) 3 clusters. b) 4 clusters.	46
4.6	Kaplan-Meier curves for the CIMLR method with: a) 3 clusters. b) 4 clusters.	49
4.7	Kaplan-Meier curves with the official WHO glioma classification.	50
4.8	intNMF result with 3 clusters.	52
4.9	Normalized coefficients for the intNMF method with 3 clusters.	54
4.10	intNMF result with 4 clusters.	57
4.11	Normalized coefficients for the intNMF method with 4 clusters.	58
4.12	Kaplan-Meier curves for the intNMF method with: a) 3 clusters. b) 4 clusters.	61
4.13	Consensus principal components: a) Relative contribution of variance from each data set. b) Explained and cumulative variance per principal component.	63
4.14	Elbow plot of the consensus principal components.	63
4.15	moCluster method with 2 JLV and 3 clusters: Silhouette plot.	65
4.16	moCluster method with 2 JLV and 3 clusters: Gap statistic.	65
4.17	Correlation heatmap between non-sparse and sparse 2 consensus principal components.	66
4.18	Metrics for 3-means with non-sparse and sparse consensus principal components.	67
4.19	moCluster method with 2 sparse JLV and 3 clusters: Silhouette plot.	68
4.20	Stacked bar chart of RNA and DNA counts and type by principal component with sparsity set as 0.50.	69
4.21	moCluster method with 6 JLV and 3 clusters: Silhouette plot.	73
4.22	Gap statistic for the 6 latent variables obtained through moCluster.	74
4.23	moCluster method with 6 JLV and 5 clusters: Silhouette plot.	75

4.24	Correlation heatmap between non-sparse and sparse 6 consensus principal components.	75
4.25	Metrics for 5-means with non-sparse and sparse consensus principal components.	76
4.26	moCluster method with 6 sparse JLV and 5 clusters: Silhouette plot. . . .	77
4.27	Stacked bar chart of RNA and DNA counts and type by principal component with sparsity set as 0.85.	78
4.28	Kaplan-Meier curves for the moCluster method.	79
4.29	Heatmap of the co-association matrix from the 3 MVC methods with dendrogram.	82
4.30	Dendrogram from the average-link hierarchical clustering of the co-association matrix with 3 methods.	83
4.31	Consensus clustering result with 3 methods and 5 clusters: Silhouette plot.	84
4.32	Kaplan-Meier curves for the 5 predicted consensus clusters.	85
4.33	Heatmap of the co-association matrix from the 2 MVC methods with dendrogram.	86
4.34	Dendrogram from the average-link hierarchical clustering of the co-association matrix with 2 methods.	86
4.35	Consensus clustering result with 2 methods and 3 clusters: Silhouette plot.	88
4.36	Kaplan-Meier curves for the 3 predicted consensus clusters.	88
B.1	intNMF result with 5 clusters.	114
B.2	Normalized coefficients for the intNMF method with 5 clusters.	116
B.3	intNMF method with 5 clusters.	118
B.4	intNMF result with 6 clusters.	120
B.5	Value of the objective function in each iteration (black) and the absolute difference between iterations (coral).	121
B.6	Convergence study for intNMF with 4 clusters.	121
B.7	Convergence study for intNMF with 5 clusters.	122
B.8	Convergence study for intNMF with 6 clusters.	123
C.1	Gap statistic for the moCluster method with 2 JLV and varying sparsity.	125
C.4	Gap statistic for the moCluster method with 6 JLV and varying sparsity.	130

LIST OF TABLES

2.1	Advantages and disadvantages of clustering methods.	10
2.2	A summary of multi-view clustering methods.	23
2.3	Example for building the hypergraph.	28
2.4	A summary of consensus clustering methods.	29
2.5	Contingency matrix of two partitions: a represents the number of pairs that are clustered together in both partitions; b the number of pairs that are in the same group in P_b but in different groups in P_a ; c the number of pairs that are in the same group in P_a but in different groups in P_b ; and d the number of pairs that are in different groups in both partitions.	32
4.1	CIMLR result with 3 and 4 clusters.	40
4.2	Mapping to genes the significant CpG sites obtained through CIMLR for: a) 3 clusters. b) 4 clusters.	47
4.3	Days until death after diagnosis of the astrocytoma patients in cluster 2 from the CIMLR method with 3 clusters.	50
4.4	intNMF result with 3 clusters.	51
4.5	Mapping to genes the CpG sites specific to cluster 2 (from the 3 cluster output), with a normalized coefficient on the specific coefficient matrix bigger than 0.5.	53
4.6	intNMF method with 4 clusters.	56
4.7	Mapping to genes the CpG sites specific to cluster 2 (from the 4 cluster output), with a normalized coefficient on the specific coefficient matrix bigger than 0.5.	60
4.8	moCluster method with 2 JLV and 3 clusters.	64
4.9	moCluster method with 2 sparse JLV and 3 clusters.	67
4.10	Top 10 of the 2 principal component-specific variables ordered by absolute coefficient score.	69
4.11	moCluster method with 6 JLV and 3 clusters.	72
4.12	Comparison of cluster 1 for 2 and 6 latent variables partitions.	73
4.13	moCluster method with 6 JLV and 5 clusters.	74

4.14	Cumulative explained variance for 6 consensus principal components out of 15 with sparsity parameters ranging from 0.40 to 0.75.	75
4.15	moCluster method with 6 sparse JLV and 5 clusters.	77
4.16	Comparison of the multi-view clustering methods across various cluster validation measures.	81
4.17	Consensus clustering result with 3 methods and 5 clusters.	84
4.18	Consensus clustering result with 2 methods and 9 clusters.	87
4.19	Consensus clustering result with 2 methods and 3 clusters.	87
A.1	Tuning number of neighbors for the CIMLR method with 3 clusters. . . .	110
A.2	Tuning number of neighbors for the CIMLR method with 4 clusters. . . .	112
A.3	CIMLR result with 3 clusters for a different seed.	112
B.1	intNMF result with 5 clusters.	114
B.2	intNMF result with 6 clusters.	119
B.3	intNMF result with 3 clusters.	124

ACRONYMS

ARI	Adjusted Rand Index (<i>pp.</i> 32, 35)
BH	Benjamini-Hochberg (<i>pp.</i> 48, 50, 62, 80, 85, 116)
CC	Consensus Clustering (<i>pp.</i> 24, 26, 87)
CIMLR	Cancer Integration via Multikernel Learning (<i>pp.</i> 11, 20–22, 35, 38, 39, 42, 43, 45–48, 50–53, 58, 60, 61, 64, 68, 70, 78, 80, 81, 84, 85, 87, 90)
CNS	Central Nervous System (<i>pp.</i> 1–3, 18, 19, 21, 34, 39, 61, 71, 90)
COCA	Cluster of Clusters Analysis (<i>p.</i> 26)
CPC	Consensus Principal Component (<i>pp.</i> 62, 74, 76, 78)
CPCA	Consensus Principal Component Analysis (<i>pp.</i> 13, 14)
CPI	Cluster Prediction Index (<i>p.</i> 18)
CSPA	Cluster-based Similarity Partitioning Algorithm (<i>pp.</i> 27, 36)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise (<i>p.</i> 6)
EAC	Evidence Accumulation Theory (<i>p.</i> 36)
GBM	Glioblastoma Multiforme (<i>pp.</i> 1, 2, 5, 18, 19, 22, 34, 40, 42, 43, 45–51, 53, 55, 56, 60–62, 64, 67, 70–73, 77, 79, 80, 83–87, 89, 90, 114, 118, 119)
HGPA	HyperGraph Partitioning Algorithm (<i>pp.</i> 27, 36)
IDH	Isocitrate DeHydrogenase (<i>pp.</i> 1, 18, 55)
intNMF	integrative Nonnegative Matrix Factorization (<i>pp.</i> 17, 34, 35, 38, 39, 51, 56, 68, 70, 78, 80, 81, 84, 85, 87, 90, 113, 120)
JLV	Joint Latent Variables (<i>pp.</i> 13, 14, 62, 68, 71, 72, 78–80)

LGG	Low-Grade Glioma (<i>pp. 1, 19, 21, 22, 41–43, 48–50, 53, 61, 62, 64, 70–73, 80, 85, 87, 89, 90, 114, 118, 119</i>)
MCLA	Meta-CLustering Algorithm (<i>pp. 28, 36</i>)
MF	Matrix Factorization (<i>p. 13</i>)
MOFA	Multi-Omics Factor Analysis (<i>pp. 15, 35</i>)
MONET	Multi-Omic NETworks in gliomas (<i>pp. 3, 91</i>)
mRNA	messenger RNA (<i>pp. 55, 70, 71</i>)
MVC	Multi-View Clustering (<i>pp. 11, 16, 83</i>)
NMF	Nonnegative Matrix Factorization (<i>pp. 11, 15–17, 24, 113</i>)
NMI	Normalized Mutual Information (<i>p. 32</i>)
NNDSVD	Nonnegative Double Singular Value Secomposition (<i>p. 17</i>)
PC	Principal Component (<i>pp. 68, 80</i>)
PCA	Principal Component Analysis (<i>p. 14</i>)
RNA-Seq	RNA-Sequencing (<i>p. 15</i>)
SIMLR	Single-cell Interpretation via Multikernel Learning (<i>p. 20</i>)
SNF	Similarity Network Fusion (<i>pp. 11, 19, 20</i>)
t-SNE	t-distributed Stochastic Neighbor Embedding (<i>pp. 21, 39</i>)
TCGA	The Cancer Genome Atlas (<i>pp. 2, 5, 18, 19, 33, 71</i>)
WCSS	Within Cluster Sum of Squares (<i>p. 8</i>)
WHO	World Health Organization (<i>pp. 1–3, 18, 19, 21, 22, 34, 39, 42, 46, 48, 51, 61, 62, 64, 71, 88, 90, 118</i>)

INTRODUCTION

1.1 Problem Statement

Glioma is the most common form of central nervous system (CNS) tumor that originates from glial cells, with about 3 out of 10 brain tumors diagnosed as gliomas [2, 3]. The classification of glioma types has seen major changes in recent years. In 2016, the World Health Organization (WHO) Classification of Tumors of CNS introduced the isocitrate dehydrogenase (IDH) mutation and the 1p/19q-codeletion as defining features alongside the preexisting tumor grade and histological features (states of tissues as observed under a microscope), resulting in the classification of adult-type diffuse gliomas into 3 main types (glioblastoma, oligodendroglioma, astrocytoma), and oligoastrocytoma, which diagnosis was discouraged [4, 5]. In 2021, the WHO published an updated version of the CNS tumors classification guidelines. Currently, 3 adult-type diffuse gliomas are diagnosed: glioblastoma (GBM), oligodendroglioma, and astrocytoma. The previous diagnosis of oligoastrocytoma, with mixed characteristics of both astrocytoma and oligodendroglioma, was relocated to these two glioma types [5, 6].

Tumors can also be classified according to WHO grade varying from 1 to 4, with the higher grade corresponding to a worse prognosis, indicating increased aggressiveness and a greater likelihood of recurrence or metastasis [6]. Low-grade gliomas (LGG) are characterized by WHO grades 2, or 3, and these tumors are considered slow-growing compared to high-grade gliomas, yet, within a decade, more than 70% of LGG can transition to a higher grade or become aggressive in behavior [6, 7]. Specifically, some astrocytomas and oligodendrogliomas can be classified as LGG since their grades vary between 2 and 4. GBM is consistently graded as WHO grade 4, since it has an exceptionally poor prognosis, with a median survival of only 10 to 14 months and a starkly low percentage of patients, merely 3 to 5%, surviving beyond three years [8]. This highlights the need for intensive treatment strategies and a deeper knowledge of glioma types, particularly GBM.

Tumors are considered complex rearrangements of different molecular levels, and the individual study of single-omics, such as genomics (study of genes), transcriptomics (gene expression), and proteomics (protein expression), is considered insufficient to fully

understand and personalize cancer treatment [9]. With the evolution of technology, in particular next-generation sequencing, several multi-view molecular data have been collected, and the joint integration of multi-omics data is now considered essential to fully understand the complex biological phenomena that could not previously be explained through single-omics [9–11]. As multi-omics is characterized by data sourced from different omics, a natural consequence is the heterogeneity (continuous, binary,...) and high dimensionality of data, with diverse biological variables (columns) but with a limited number of samples (rows), which induces statistical and computational challenges [12]. While these traits may pose challenges, multi-omics data is now considered a standard in biology, and an invaluable tool that with machine learning pioneered a new era in neuro-oncology [12, 13]. In particular, the availability of large-scale multi-omics datasets is attributed to an important part of the research on GBM, allowing the identification of omics drivers in GBM progression [13].

1.2 Motivation and Goals

As mentioned previously, gliomas are the most frequent CNS tumors yet are characterized by a poor prognosis, with a high mortality rate. This makes the glioma classification (i.e., patient stratification) problem still relevant and unresolved, prompting the need to explore clustering methods to decipher glioma heterogeneity and support molecular subtyping. In light of this, multiple state-of-the-art integrative multi-view clustering algorithms were implemented, and a consensus clustering method was employed, to aggregate and find a general agreement between the algorithms, allowing for a more robust solution. Furthermore, the mentioned analysis is possible due to the existence of public multi-omics datasets, such as The Cancer Genome Atlas (TCGA). TCGA is a vast repository with various omics, like transcriptomics and genomic data, for various cancers, including gliomas. Notably, the classification of glioma patients is not in line with the 2021 WHO classification, however, the labels were updated by Marta Lopes et al. [5], ensuring that the patient stratification used in this thesis reflects the latest criteria.

The mentioned patient stratification can then single out new biomarkers or further validate previously known biomarkers, deepening our knowledge of gliomas. This comprehensive approach is expected to contribute to a more nuanced understanding of glioma heterogeneity and offer insights to develop personalized therapeutic strategies for patients sharing the same molecular characteristics. Furthermore, a survival analysis is performed to extend the comprehension of the clusters besides molecular profiling. By integrating survival data with molecular subtyping, we seek to elucidate the clinical relevance of the identified subtypes and their implications for patient prognosis.

This thesis aims to help decipher glioma heterogeneity with multi-view clustering by uncovering and identifying patient groups, a significant contribution to personalized cancer treatment. To do so a pipeline was developed to integrate and cluster multi-omic data from different sources, including data preparation and preprocessing, studying and

applying state-of-the-art multi-view integrative clustering methods, survival analysis based on selected molecular features, feature ranking for relevance, and biological input for further research. It's worth noting that the mentioned pipeline is not exclusive to gliomas and can be expanded to other multi-omics datasets.

The present dissertation is inserted in the scientific project "MONET - Multi-omic networks in gliomas", funded by the Foundation for Science and Technology (PTDC/CCI-BIO/4180/2020). Aligned with the project goal, the thesis objective is to unveil subgroups among patients based on their molecular profiles within categories established by the 2021 WHO CNS tumors classification by employing multi-view clustering methods, the first multi-omics unsupervised within the MONET project.

1.3 Structure

This Masters dissertation is divided into 5 chapters: Introduction, Literature Review, Methodology, Results and Discussion, and Conclusion.

The present chapter, Introduction, focuses on contextualizing the relevant disease, glioma tumors, the classification problem, and why discovering and studying potential new grouping of patients is pertinent. The goals and expected contributions of the thesis are also highlighted.

The second chapter details the relevant literature for the problem, including single-view clustering methods, multi-view clustering methods, consensus clustering, and clustering validation measures.

The third section elaborates on the chosen methodology, which encompasses justifying the omics datasets studied, the selection of the multi-view clustering methods, and the consensus clustering method, with careful consideration of their known strong points and limitations. Implementation details are also presented to allow reproducibility.

In the fourth chapter, the results of the chosen methodologies are displayed, complemented with an interpretation of the findings. The results include confusion matrices, several plots, tables with the cluster validation measures, survival analysis to evaluate the clusters' prognosis value, and relevant features were studied and cross-referenced with existing literature. Each of these has descriptive comments explaining trends and observations, to provide a clear and objective overview of the findings. Furthermore, the highlighted features are compared to conclusions drawn in the literature, known biomarkers, and potential new findings.

Finally, the conclusions, and limitations of the present thesis, as well as pathways for future work are contemplated.

LITERATURE REVIEW

The present section is divided into single-view clustering, multi-view clustering, consensus clustering methods, and cluster validation measures.

2.1 Clustering Methods

Clustering is an unsupervised learning technique designed to identify inherent groups within data by minimizing dissimilarity within a group and simultaneously maximizing it between groups. As an unsupervised method, it eliminates the need for labeling data, making it particularly advantageous in terms of time and cost efficiency.

Clustering techniques can be defined in various ways with the most common ones being hierarchical (or connectivity-based) and nonhierarchical clustering. The latter category is broad, therefore centroid-based and density estimation methods are highlighted in this section since they have a widespread adoption, versatility, and effectiveness in handling different types of datasets and clustering scenarios. Yet the category includes multiple others such as a mixture of distribution [14] and grid-based [15] clustering. Additionally, graph-based clustering methods are considered relevant in various biological problems and networks and are commonly used when dealing with gliomas [16].

2.1.1 Hierarchical Clustering

Hierarchical clustering can be categorized into bottom-up (or agglomerative) and top-down (or divisive) approaches. In the bottom-up method, each observation initially forms its cluster, with subsequent recursive pairing of observations. Conversely, top-down clustering starts with a single cluster with all observations and iteratively splits them into smaller clusters [14]. The results of clustering can be viewed through dendrograms, which are graphical representations of the hierarchical relationships between data points. Vertical lines indicate the distance between clusters with longer lines indicating greater dissimilarity, and horizontal lines connect data points that belong to the same cluster. The most common way to find the ideal number of clusters is determined by cutting the

dendrogram at the point where a significant increase in distance between clusters occurs, indicating the formation of distinct groups [14].

The initial step involves creating a distance or dissimilarity matrix, often utilizing measures such as the Euclidean distance. The choice between bottom-up and top-down approaches depends on the desired granularity of clusters, with the bottom-up method typically used for identifying smaller clusters and the top-down method for larger clusters [17].

Hierarchical clustering is a well-known clustering technique that has several applications and is employed in earlier glioma studies [18]. In particular, in 2009, the method was used in a GBM dataset with 189 patients, and gene dosage and clinical data from TCGA, and was able to identify 2 major subgroups with differences in the mean number of landscape gene alterations, such as mutations, amplifications, deletions, and other genomic changes [19].

2.1.2 Nonhierarchical Clustering

2.1.2.1 Centroid-Based Clustering

Centroid-based clustering aims to partition observations into k clusters by minimizing the sum of distances from the points within a cluster to their centroid [17]. It is essential to highlight that employing the Euclidean distance in centroid-based clustering may not be optimal for datasets with high dimensionality, potentially leading to rapid convergence issues [17]. This challenge becomes particularly pronounced in the context of multi-omics datasets. Despite this, centroid-based methods have demonstrated robust performance in high-dimensional datasets by incorporating dimensionality reduction techniques and carefully considering inputs derived from low-dimensional representations [20].

K-means clustering stands out as one of the most prominent techniques within this category, with a well-established methodology. The method has been applied to earlier gliomas studies [21] and is present in state-of-the-art multi-view clustering methods.

K-means Clustering

The objective function for K-means clustering, often referred to as the “within-cluster sum of squares” or “inertia”, is to find the partition $S = S_1, \dots, S_k$ of the observations $(x_1, \dots, x_n)$, that solves this minimization problem:

$$\min_S \sum_{i=1}^k \|C_i - x_i\| \quad (2.1)$$

where C_i is the centroid of cluster i and x_i is the i -th data point [22].

The K-means clustering algorithm provides various initialization strategies. One approach is to randomly select k data points from the dataset as the initial cluster centroids. Alternatively, the K-means++ initialization method [23] begins by randomly selecting the

first centroid from the data points. Subsequent centroids are chosen from the remaining data points, with a probability distribution favoring points farther away from existing centroids, often leading to improved convergence and cluster quality.

After initialization, each data point is assigned to the nearest cluster based on the smallest distance, typically measured using the Euclidean distance, from the data point to the cluster centroid [18]. The centroids are then updated by computing the mean of the data points assigned to each cluster. The algorithm iterates until a new centroid coincides with a previous one, indicating consistent assignment of data points to the same clusters.

2.1.2.2 Density-Based Clustering

The density-based clustering method is distinctive in its approach to identifying clusters based on areas of higher statistical density within the dataset. Unlike traditional algorithms, not all observations must be assigned to clusters, and points outside the optimized clusters are treated as noise [14]. This noise-handling capability makes density-based clustering particularly robust in the presence of outliers and irregular cluster shapes [17].

A noteworthy exemplar of this method is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, which has been applied in multi-omic analyses for unsupervised clustering, particularly to identify subtypes within single cancers [24].

DBSCAN

DBSCAN introduces the concept of core points, which are data points within high-density regions, and border points, which lie on the periphery of clusters. Specifically, a data point is considered a core point if there are at least MinPts points within a maximum distance set by a neighborhood of ϵ , including the point itself. MinPts represent the minimum number of points required to form a dense region. Thus, a data point is a core if a dense region is formed within its ϵ -neighborhood. A border point, on the other hand, is a point within the neighborhood of ϵ but has fewer than MinPts points within its neighborhood [25].

The algorithm initiates by selecting an arbitrary point and identifying all points that are density-reachable from it. Point P is considered density-reachable from point Q if there exists a sequence of points starting from Q and ending at P , where each point in the sequence is within the ϵ -neighborhood of the previous point. If the chosen point is a core point, it becomes a cluster. In the case of a border point, it is incorporated into the existing cluster due to its proximity within ϵ , but no additional points are explored for this cluster. Points that do not meet the core or border criteria are labeled as noise. The algorithm then proceeds to the next point for further exploration [17, 25].

2.1.2.3 Graph-Based Clustering

Graph-based clustering offers a powerful approach to uncovering patterns and relationships within complex datasets. By representing data as a graph, where nodes represent data points and edges represent relationships or similarities, insights into the intrinsic structure of the data can be obtained. This is particularly useful in scenarios where traditional methods may struggle to reveal hidden patterns. Graph-based clustering methods have been applied in various fields, including web page classification [26], worldwide air transportation [27], and in bioinformatics [28–30].

A considerable amount of data can be transformed into a graph using similarity graphs, $G(V, E)$ where V is the set of vertices (each represents a data point) and E is the set of edges. However, when utilizing similarity or dissimilarity functions for this purpose, it is important to note that while these functions are effective in encoding dispersed data, they may not accurately capture the true nature or structure of the data, potentially leading to errors.

To evaluate the quality of clustering, it is essential that within the same cluster, data points exhibit similarity, while data points in different clusters should be dissimilar. When applied to graphs, this implies the goal of maximizing the weight of edges connecting nodes within clusters and minimizing the weight of edges between clusters [31].

A well-known approach to graph-based clustering is spectral clustering, easily implemented, solved efficiently with standard linear algebra methods, and has outperformed traditional approaches [31].

Spectral Clustering

There are many variations of spectral clustering algorithms. Consider the unnormalized spectral clustering [31] which takes as input a similarity matrix, S , and the number k of clusters to construct. Assume there are n data points, so S is a matrix of $n \times n$. The main steps are as follows:

1. Construct the similarity graph with weighted edges, $G(V, E)$. One possible example is using the Gaussian kernel as a similarity function. Let W be the weighted adjacency matrix, a $n \times n$ matrix indicating which nodes are connected: if no edge exists the entry is 0, otherwise, it is the weight of the edge. Define D the degree matrix, a diagonal matrix indicating the number of edges each node is connected to.
2. Compute the unnormalized Laplacian, L :

$$L = D - W$$

Note that the unnormalized Laplacian is a symmetric, positive semi-defined matrix, and the smallest eigenvalue of L is zero which corresponds to the first eigenvector. Therefore, there are n nonnegative, real values, and crescent eigenvalues.

3. Compute the first k eigenvectors $v_1, \dots, v_k$ of L .
4. Let $U_{n \times k}$ be the matrix with the first k eigenvectors as columns and let u_i be the i -th row of U . Note that this step reduces the dimensionality from $n \times n$ to $n \times k$.

$$U_{n \times k} = \begin{bmatrix} | & & | \\ v_1 & \dots & v_k \\ | & & | \end{bmatrix} = \begin{bmatrix} - & u_1 & - \\ & \vdots & \\ - & u_n & - \end{bmatrix}$$

Apply K-means clustering algorithm to the $u_i, i = 1, \dots, n$, and obtain the $C_1, \dots, C_k$ clusters.

The output of the method are clusters $A_1, \dots, A_k$ with $A_i = \{j | u_j \in C_i\}$.

Additionally, various extensions of graph-based clustering for multi-view data have also been implemented throughout the years [32–35].

2.1.2.4 Determining k

Determining the optimal number of clusters is a crucial aspect of clustering. Three prominent methods for this purpose are the elbow method, the average silhouette method, and the gap statistic method [17]:

- **Elbow Method:** This technique computes the intra-cluster variation (Within-Cluster Sum of Square, WCSS) for various potential clustering results. The optimal number of clusters is identified at the inflection point where a decrease in WCSS becomes less pronounced. At this juncture, increasing the number of clusters does not significantly reduce WCSS.
- **Silhouette Method:** This method calculates a silhouette value for each data point, indicating how well it is placed within its assigned cluster concerning others, with higher values indicative of more effective clustering performance [36]. The metric definition is detailed in [subsubsection 2.4.1.3](#). The ideal number of clusters is the one that maximizes the average silhouette value.
- **Gap Statistic Method:** The gap statistic, similarly to the silhouette method, evaluates clustering results by comparing the intra-cluster variation of each clustering outcome with that of a randomly generated Monte Carlo sample. After clustering the data points and computing the within-dispersion measure W_{obs} , B reference datasets are created, and clustered, and their within-dispersion measures W_{ref} are computed. The estimated gap statistic is then calculated, where k denotes the potential number of clusters:

$$\text{Gap}(k) = \frac{1}{B} \sum_b \log(W_{ref}) - \log(W_{obs})$$

The optimal number of clusters, k , is chosen as the smallest value for which the following condition holds:

$$\text{Gap}(k) \geq \text{Gap}(k + 1) - s_{k+1}$$

where s_k represents the standard deviation of the $\log(W_{ref})$ values across the B reference datasets, adjusted by a correction factor [37].

These methods collectively provide a comprehensive framework for determining the number of clusters in clustering, enhancing the reliability of the clustering outcomes [17]. Despite these approaches being widely acknowledged, it is crucial to note that selecting the right number of clusters remains a challenging task. The choice often involves a trade-off between model simplicity and the ability to capture meaningful patterns in the data.

Table 2.1: Advantages and disadvantages of clustering methods.

Category	Example	Advantages	Disadvantages
Hierarchical	Average linkage (Agglomerative technique) [14]	1. Inherent structural hierarchy, making interpretation more straightforward 2. No requirement of cluster count [14] 3. Process preserves essential information about data point integration in clusters	1. Static recursive approach: once a data point is assigned to a cluster, the model does not explore other potential assignments [14, 17] 2. Influence of dissimilarity matrix in cluster results [14]
Centroid-Based	K-means [18]	1. Simplicity of implementation 2. Computational efficiency 3. Recognition of distinct functional forms and relative sizes [18]	1. Difficulty clustering non-convex shapes and varying sizes 2. Number of clusters as hyperparameter [18]
Density-Based	DBSCAN [25]	1. No requirement of cluster count [25] 2. Recognition of clusters with non-convex or arbitrarily shaped boundaries [25] 3. Selection of hyperparameters comes with guidelines [25] 4. Noise Handling [25] 5. Resilience to outliers [25]	1. Tuning of two hyperparameters 2. Requires careful supervision since different combinations of the hyperparameters seriously affect the clustering results [17, 25] 3. Difficulty in handling varying density clusters
Graph-Based	Spectral Clustering [31]	1. Representation of relationships between the data points clustered [38] 2. Flexible tool implemented in various fields 3. Reduction in recalculations if data is increasing over time due to graph convergence [38]	For Spectral Clustering: 1. Solving eigenvalue problem, which is computationally expensive and numerically unstable for large and sparse graphs [31] 2. Number of clusters as hyperparameter [31] 3. Tuning of parameters for the similarity graph [31]

The advantages and disadvantages of the clustering methods addressed in this subsection can be consulted in [Table 2.1](#).

2.2 Multi-view Clustering Methods

Multi-view clustering (MVC) methods are a type of analysis that integrates multiple data sources or “views” (in this particular case, multiple omics) to group similar objects based on the information of the various views, which further helps explain a phenomenon since different views can have complementary information [39].

Various strategies have been employed for multi-view clustering. An earlier procedure was to join features from different views into a single view and apply clustering methods. However, since the grouping might be biased toward a view with a dominantly large number of features in the concatenation, this does not guarantee that all the views are considered when clustering [39]. Another possible approach was to separately cluster each data matrix and then manually integrate the clusters, which is a subjective integrative method and challenging since they may not be clearly correlated [40].

Thus, more methods based on multi-view integration have emerged, aiming to grasp the shared patterns in different views and improve clustering performance. A characteristic across multi-view datasets is the high dimensionality that could also pose a challenge to successful integration, specifically for multi-omics datasets there is a high number of features with a low number of samples. Hence, many algorithms focus on dimensionality reduction or similarity matrices for clustering [41]. Some known examples of dimensionality reduction algorithms are based on matrix factorization, like iCluster [42], iCluster+ [43], and moCluster [44], and nonnegative matrix factorization (NMF) such as intNMF [11]. On the other hand, Similarity Network Fusion (SNF) [45], and Cancer Integration via Multikernel Learning (CIMLR) [40] are cases of similarity matrices-based clustering algorithms.

2.2.1 Matrix Factorization Methods

Let $X_1, \dots, X_N$ be the input matrices, each representing a different omic. Each X_i matrix has a size of $n \times J_i$ where n is the number of subjects (ie, patients) in the sample, J_i is the number of variables in omic i , and N is the number of omics considered. The main goal of matrix factorization (MF) is to decompose a matrix into the product of two matrix factors such that their multiplication is a good approximation:

$$X_i \approx U V_i, \quad i = 1, \dots, N \quad (2.2)$$

with $U \in \mathbb{R}^{n \times k}$ a common basis matrix, $V_i \in \mathbb{R}^{k \times J_i}$ a specific coefficient matrix, k the latent space dimension (in this case the number of clusters). It’s important to note that the original data is projected to a latent space, where the columns of U represent directions or features in the latent space. The coefficients in the corresponding column of V_i , $i = 1, \dots, N$ indicate how each basis vector contributes to each sample in the original data.

2.2.1.1 iCluster

iCluster is one of the early multi-omics integration methods that identifies clusters through the joint estimation of latent variables. The primary objectives of iCluster involve achieving cohesive sample clusters while concurrently pinpointing the multi-omics characteristics that play a role in shaping these joint sample clusters [42, 46].

iCluster decomposes multiple continuous data matrices into the product of two matrices [42]:

$$X_i^T = V_i U + \mathcal{E}_i, \quad i = 1, \dots, N \quad (2.3)$$

where V_i is the coefficient matrix of dimension $J_i \times (k - 1)$, U is the latent component matrix of dimension $(k - 1) \times n$ that connects the N omics, and $\mathcal{E}_i$ is a random error matrix with dimension $n \times J_i$, mean zero and a diagonal covariance matrix $\Psi_i = \text{diag}(\sigma_{i,1}^2, \dots, \sigma_{i,J_i}^2)$ (representing the variances unique to each data type after accounting for the correlation between the omics). The model assumes each data matrix is row-centered, which is why the intercept terms are not included in the expression [42].

To solve a likelihood-based solution, it is necessary to assume that $U \sim N(0, I)$, hence, the iCluster model can be considered a Gaussian latent variable model. The method is based on the Expectation-Maximization algorithm to solve the maximum likelihood estimates of U and Ψ . Then, a K-means clustering is applied on $\hat{E}[U|X_i]$, with $Z \sim N(0, I)$ to recover the class indicator matrix. Additionally, a penalty to induce sparsity in $U_i, i = 1, \dots, N$ is implemented, which improves the clustering performance by reducing the model variance and allows for feature selection.

While the iCluster method has reasonably good clustering results when the omic data follows the underlying Gaussian assumptions [46], it still only models continuous variables. To fill this gap, it was created iCluster+, which applies different models to the data matrix according to its type of data, namely, logistic regression for binary variables, multinomial logistic regression for multcategory variables, or Poisson regression for count variables [44]. This means that, for example, if the data matrix is binary, the logistic regression models the likelihood of each object belonging to a cluster, and in turn allows for a more accurate capture of the underlying data distributions (instead of using a Gaussian model for all data), which results in a more accurate weight matrix and clustering assignment [42, 43].

Both the mentioned methods are parametric, assuming an underlying data distribution. However, when working with multi-omics, the data can exhibit diverse distributions depending on the technology used to measure the biological process. This variability poses a challenge in modeling all datasets comprehensively in a single integrative analysis [11]. Data transformations are also feasible to address non-Gaussian variables. For instance, DNA methylation beta values, ranging from 0 to 1 and not normally distributed, can be successfully modeled using linear regression in iCluster models after applying the logit transformation [46].

iCluster and iCluster+ utilize an iterative Expectation-Maximization algorithm, a method known for its flexibility but lacking guaranteed convergence to a deterministic or optimal solution.

Moreover, the iCluster+ parameters can be difficult to tune since the sparsity parameters are not independent across datasets. Modifying the sparsity parameter for one dataset influences the number of selected variables in the other datasets. Therefore, tuning the parameters of iCluster+ to achieve a similar number of selected variables as other methods posed a challenge. The authors of iCluster+ recommend performing cross-validation to select both the number of subgroups and penalties. However, iCluster+ is already considered one of the slowest methods, and cross-validation would be highly time-consuming [41].

2.2.1.2 moCluster

moCluster's approach to the MF problem utilizes consensus principal component analysis (CPCA) to decompose the data matrix X_i into two components: a shared latent matrix denoted as U and a loading sparse matrix represented by V_i . Yet note that both iCluster and moCluster leverage a linear combination of data features to establish a collection of joint latent variables (JLV) that capture essential patterns within the multi-omics dataset, as illustrated in Equation 2.3. This divergence in methodology emphasizes moCluster's reliance on CPCA for the decomposition process, distinguishing it from the iCluster methods.

CPCA is a statistical technique used to identify and analyze the shared patterns among multiple datasets, without merging them into a single integrated dataset. To apply the method it is necessary to center and scale the data matrices, as well as weighting each data matrix by the reverse of the first eigenvalue which allows the different matrices to contribute comparable variance to the JLVs.

A JLV is represented by the $\vec{u}$ and it is initially randomly chosen. The algorithm for CPCA is looped until its convergence as follows:

1. Compute $\vec{v}_i$, which represents the variable loadings of X_i and serves as the feature coefficient vector, as

$$\vec{v}_i = X_i^T \cdot \frac{\vec{u}}{\vec{u}^T \vec{u}}.$$

2. Normalize $\vec{v}_i$ to $\|\vec{v}_i\| = 1$.
3. moCluster modification: $\vec{v}_i = \text{soft}(\vec{v}_i)$.
4. Calculate $\vec{u}_i$ the latent variables for each X_i :

$$\vec{u}_i = X_i \cdot \vec{v}_i^T.$$

5. Combine all scores in matrix $U = [\vec{u}_1, \dots, \vec{u}_N]$.

6. Compute the weight $\vec{w}$:

$$\vec{w} = \frac{U^T \cdot \vec{u}}{\vec{u}^T \cdot \vec{u}}.$$

This gives the contribution of the latent variables to the score matrix U .

7. Normalize $\vec{w}$ to $\|\vec{w}\| = 1$.

8. Determine the JLV $\vec{u}$:

$$\vec{u} = U \cdot \vec{w}.$$

For each dimension or JLV, the final step is to compute the feature coefficient vector $\vec{v}_i$ as shown in step 1 and deflate the data matrix with the obtained JLV, $X_i = X_i - \vec{u} \cdot \vec{v}_i$ [47]. Then, the higher-order JLVs are computed with the deflated data, representing the residual data or variance not captured previously.

An advantage of calculating the deflation step with the JLVs is their orthogonality [44, 47], helping to capture distinct patterns or relationships between sets of variables.

In the moCluster method, the initial JLV $\vec{u}$ is randomly chosen. Furthermore, the CPCA is modified to implement feature selection to the method and induce sparsity in the feature coefficient vector $\vec{v}_i$, increasing the interpretation of the biological meaning of the JLVs. A soft-thresholding operator between step 4 and 5 is added, $\text{soft}(x, a) = \sin(x) (|x| - a)_+$, meaning that larger values of a results in more coefficients of $\vec{v}_i$ being set to zero and conversely [44].

After computing the JLV, a permutation step test is implemented to evaluate the concordant or discordant structures between the omics, and the samples are shuffled within each dataset independently of the others. This maintains the structure within each dataset but disrupts any potential relationships or patterns that may exist across datasets. Then, CPCA is performed on the permuted datasets. Similar to PCA the importance of each JLV can be measured by the pseudoeigenvalues that represent the explained variance. When the eigenvalues from the original data sets are significantly higher than the permuted ones, these represent the concordant structures between the omics, implying there is a pattern across the different omics datasets that couldn't be explained by randomness. Since including extra JLVs allows the identification of divergent patterns across the omics, unique to specific datasets, it is preferable to maintain the JLVs according to the permuted dataset, and the number of JLVs to keep can be determined through the "elbow" point in a scree plot [44].

Finally, moCluster applies traditional algorithms, e.g. hierarchical (see 2.1.1) or K-means clustering (Section 2.1.2.1) to the original JLVs, using the optimal number of clusters obtained through the study of permuted omics.

The method moCluster has been implemented in different multi-omics studies, including colorectal cancer [44], breast cancer [48], murine liver, liver cancer, and obesity data sets [41], and its high clustering performance is highlighted in all mentioned articles.

2.2.1.3 MOFA

Multi-Omics Factor Analysis (MOFA) is a state-of-the-art method based on matrix decomposition, aiming to reveal a shared axis of variation between the different omics layers. MOFA is based on Equation 2.3, with the error term expressing the view-specific residual noise term, whose definition depends on the data type.

The method is based on a probabilistic Bayesian framework [49], where prior distributions are assigned to the unobserved variables to incorporate prior knowledge before analyzing the actual data [50]. Specifically:

- The common basis matrix U has a standard normal prior, meaning that the elements of U are assumed to follow a normal distribution with mean zero and variance one, reflecting a baseline assumption about the latent factors.
- The weight matrices are given a sparsity prior (defined by the MOFA authors [49]), which allows for feature selection, by reducing the number of variables with contributing weights.

The noise model for continuous data, similar to the iCluster method, is the Gaussian noise model assuming independent and identically distributed heteroscedastic residuals $\mathcal{E}_i$, meaning that the variance of the residuals can vary across observations. Additionally, MOFA allows the input of binary and count data, for which the Gaussian model is not appropriate. The noise model for binary data is the Bernoulli model, and for the count data such as RNA-Seq (RNA-Sequencing) data, the Poisson model is employed [49].

By employing these different noise models, MOFA effectively integrates and analyzes diverse types of omics data, providing a comprehensive understanding of the shared and distinct patterns across multiple omics layers [49].

Some drawbacks of the MOFA method include the capability of only modeling linear relationships, and making assumptions about the dependencies across cells [51]. MOFA+ was designed with these issues in mind, still aiming to reconstruct a low-dimensional representation of the data. This expansion can still only capture moderate non-linear relationships, requires multi-modal measurements from the same set of cells, and assumes no relationships between variables in the prior distributions. The latter is limiting since genomic features interact in complex regulatory networks [51].

MOFA has been used frequently in the study of cancer, namely, it was able to identify relevant clinical markers in chronic lymphocytic leukaemia [49] and it has been used to identify axes of variation in GBM [52].

2.2.2 Nonnegative Matrix Factorization

NMF is considered a valuable instrument across a wide range of fields, including motion segmentation, text mining, pattern recognition research, and cancer subtypes [39, 46, 53]. The particularity of NMF is that X_i , U , and V_i have all nonnegative entries [39, 41].

Additionally, Equation 2.2 with the previous nonnegative restriction is a version of NMF designed specifically for multi-view data, whereas the original NMF was developed for single-view data, represented as $X \approx U V$.

Since the aim is to obtain a good approximation of X_i , Equation 2.2 can be rewritten as [54]:

$$\min_{U \geq 0, V_1 \geq 0, \dots, V_N \geq 0} \|X_i - UV_i\|_F^2 \quad (2.4)$$

where $\|\cdot\|_F$ represents the Frobenious norm.

Subsequently, different traditional clustering methods can be formulated for MVC using NMF [39]. For example, the K-means objective function for MVC is

$$\begin{aligned} \min_{U, V_1, \dots, V_N} \sum_{i=1}^N (\alpha_i)^\gamma \|X_i - UV_i\|_{2,1}^2 \\ \text{s.t. } U_{o,n} \in \{0, 1\}, \sum_{n=1}^N U_{o,n} = 1, \forall i = 1, \dots, N \end{aligned} \quad (2.5)$$

where α_i is the weight of the i -th view, γ is a parameter to control the weight distribution, and the $\ell_{2,1}$ norm is the mixed norm, where the ℓ_2 norm is applied row-wise, and then the ℓ_1 norm is applied column-wise. Regarding the clustering extension, the U matrix can be interpreted as the cluster memberships, and the $V_i, i = 1, \dots, N$ as the cluster centroids.

NMF can learn meaningful representations that effectively capture the latent components of the data. Moreover, clustering methods grounded in NMF offer scalability for large-scale datasets, due to efficient optimization algorithms, and the nonnegative constraint could introduce a certain level of sparsity in both U and $V_i, i = 1, \dots, N$. However, it is essential to acknowledge that clustering algorithms maintain their challenges. For instance, the K-means variant using NMF is limited to clusters with a spherical shape relative to the cluster center, and the sensitivity to different initializations can result in varying clusters.

Many NMF methods rely on nonconvex optimization formulations, since the objective function is convex when concerning U but nonconvex when U and V_i are considered together [11]. This nonconvex nature makes these methods susceptible to converging to local optimum, particularly in scenarios involving missing values and outliers [39]. The local optimum can be optimized through numerical methods, using several initializations of U and $V_i, i = 1, \dots, N$, and then choosing the one that minimizes the objective function [11].

NMF methods prove to be efficient tools for handling large datasets, but they often face challenges in scenarios involving missing values [55].

Another drawback is that NMF is limited to handling nonnegative matrices, which can pose challenges in certain scenarios, especially when dealing with omics data that may include negative values. In such situations, a suggested approach is to shift the data in a positive direction, although the biological impact of this shift may not be well understood. Nevertheless, it is common in biological studies to normalize the data, which suggests

that if we can shift the mean during normalization, the mentioned data shifting is likely acceptable. Furthermore, the clustering results remained unaffected because the shift doesn't alter the underlying variation in the data [46].

2.2.2.1 Integrative Clustering based on NMF

Integrative Nonnegative Matrix Factorization (intNMF) is an extended clustering method that builds upon NMF, allowing the classification of patients based on multiple biological sources of data. The intNMF method avoids assuming any specific distributional form of the data [11].

The goal of intNMF is to minimize the objective function defined as the weighted Frobenius norm :

$$\min_{U, V_1, \dots, V_N} \sum_{i=1}^N \theta_i \|X_i - UV_i\|_F. \quad (2.6)$$

Here, the parameter θ_i is introduced to control the weight of each view in the analysis. Users can specify this parameter, with the default value being 1. Another option is:

$$\theta_i = \frac{\max_{i=1, \dots, N} \text{mean}\|X_i\|_F}{\text{mean}\|X_i\|_F}. \quad (2.7)$$

This implies that the dataset with the largest average values has a weight of 1, and a dataset with the smallest values has a weight greater than 1.

In the intNMF algorithm, the estimation of U and V_i is done alternatively using least squares, an approach designated nonnegative Alternating Least Squares that allows the transformation of the objective function from nonconvex to convex U given V_i and otherwise. Therefore, only the U matrix needs initialization [11].

Due to the sensitivity of NMF methods to initialization, two methods are considered in the intNMF algorithm. First, U is randomly initialized from a uniform distribution ($U[0, 1]$), which is common for NMF-based algorithms. Second, nonnegative double singular value decomposition (NNDSVD) is employed, and then the factorization that minimizes the objective function is used. NNDSVD uses two singular value decomposition processes to obtain nonnegative approximations of the data matrix and its singular factors, and, then, uses them to initialize the NMF factors [53].

After the initialization of U , the objective function is solved for each $V_i, i = 1, \dots, N$:

$$\begin{aligned} \arg \min_{V_i} \|X_i - UV_i\|_F, \quad i = 1, \dots, N \text{ such that } V_i \geq 0 &\implies \\ V_i = (U^T U)^{-1} U^T X_i, \quad i = 1, \dots, N \text{ such that } V_i \geq 0 \end{aligned}$$

Afterwards, U is solved using X_i and the current values for $V_i, i = 1, \dots, N$:

$$\arg \min_U \sum_{i=1}^N \theta_i \|X_i - U V_i\|_F \text{ such that } U \geq 0 \implies$$

$$U^T = \left(\left(\sum_{i=1}^N \sqrt{\theta_i} V_i V_i^T \right)^T \right)^{-1} \left(\left(\sum_{i=1}^N \sqrt{\theta_i} X_i V_i^T \right)^T \right)$$

The steps are repeated until convergence: the method defines a connectivity binary matrix, $C_{n \times n}$. It uses its stability as a stopping algorithm, meaning the algorithm stops if there are no alterations in the entries for a pre-set number of consecutive iterations. In matrix C , $c_{ij} = 1$ if patient i and j are in the same cluster and 0 otherwise. Finally, the cluster membership for each sample is determined by the highest entry in each column of U .

Furthermore, the authors of the intNMF algorithm have introduced a new concept to find the optimum number of clusters, the Cluster Prediction Index (CPI). This index leverages a resampling-based cross-validation technique, aiming to identify the ideal cluster count, denoted as k . The primary objective is to maximize the agreement between predicted clusters within the training and testing datasets. The data matrix $X_i, i = 1, \dots, N$ (with dimensions $n \times J_i$) is divided repeatedly into training ($X_{i \text{ (train)}}$ with dimensions $n_1 \times J_i$) and testing ($X_{i \text{ (test)}}$ with size $n_2 \times J_i$) sets. During each iteration, the intNMF is applied to the training data, estimating the coefficient matrix $V_{i \text{ (train)}}$, which is then used to compute the cluster memberships U (with dimensions $N_2 \times k$) for the testing data. These memberships are labeled as "predicted". Simultaneously, the intNMF is applied to the test data, $V_{i \text{ (test)}}$ is estimated and then used to find the cluster memberships, designated "observed". Subsequently, the Adjusted Rand Index is calculated between the "predicted" and "observed" cluster results. It's essential to emphasize that this metric evaluates the agreement between the clusters "predicted" with solely the test data and the "observed" clusters with the influence of the training data (not the external ground truth labels). This process is then repeated multiple times and the CPI is computed as the average of the Adjusted Rand Indices (see [subsubsection 2.4.2.2](#)). Therefore, the optimal number of clusters k is the one that maximizes the CPI.

The intNMF algorithm has been successfully applied to different multi-view datasets, such as breast cancer [11], murine liver, liver cancer, and obesity datasets [41]. More relevantly, in 2017, the intNMF method was applied to GBM data sourced from TCGA [11] which is not in line with the current classification. In a more recent use of the intNMF method [46], the authors applied the method to LGG from the TCGA dataset, a dataset that also maintains the 2016 WHO Classification of Tumors of CNS. The clustering results were interpreted based on key molecular parameters, namely the IDH status (wildtype or mutant) and 1p/19q Co-deletion or Non-Codeletion, and not on the cancer type. However, this categorization does not align with the latest WHO CNS classification, which defines glioma types as astrocytoma, oligodendroglioma, and GBM according to several molecular biomarkers. This dissertation aims to address this limitation and fully explore the method

to meet the current glioma classification. TCGA LGG and GBM datasets updated by the MONET consortium to meet the 2021 WHO CNS guidelines [5] are used for that purpose.

2.2.3 Graph-based Methods

As previously mentioned, various types of data can be transformed into similarity graphs using similarity (or dissimilarity) functions, which are then represented as similarity (or dissimilarity) matrices. When applied to multi-omics data, similarities for patients are computed separately for each omic and then integrated through different methods [56]. Graph-based multi-view clustering methods offer the advantages of simplicity and efficiency, particularly in handling nonlinear data, and have consequently garnered increased attention recently [57].

2.2.3.1 Similarity Network Fusion

The SNF method starts by constructing a graph of patient similarity network for each omic X_i , $i = 1, \dots, N$: $G_i = (V, E_i)$ where the vertices, V , represent the patients and the edges, E_i , are weighted by the patient similarity. The weights are represented in a similarity matrix for each data set, W_i , where the scaled exponential similarity kernel is used and defined in the following equation:

$$W_i(i, j) = \exp \left(-\frac{\rho^2(X_i(i, \cdot), X_i(j, \cdot))}{\alpha \epsilon_{i,j}} \right) \quad (2.8)$$

where $\rho(\cdot)$ is the Euclidean distance, α is a hyper parameter between 0 and 1 and ϵ is added to fix a scaling issue.

Then, the networks are integrated into a single similarity network using a nonlinear combination method [45]. Thus, the method defines a normalized matrix P_i that contains the information on the similarity of each patient to every other, and a matrix S_i which only has the similarity to the T most similar patients for each patient (T is a constant parameter defined by the user). The definition of the mentioned matrices is present in Equation 2.9, where N_i is the set of neighbors of i of each omic graph, including the i point itself.

$$P_i(i, j) = \begin{cases} \frac{W_i(i, j)}{2 \sum_{i \neq j} W_i(i, j)} & , i \neq j \\ 1/2 & , i = j \end{cases} \quad S_i(i, j) = \begin{cases} \frac{W_i(i, j)}{\sum_{j \in N_i} W_i(i, j)} & , j \in N_i \\ 0 & , \text{otherwise} \end{cases} \quad (2.9)$$

The mentioned matrices are then combined and normalized. The latter guarantees that a patient is most similar to himself and the authors found that this normalization induced a quicker convergence. The normalized matrix sum is defined by Equation 2.10:

$$\tilde{P}_i = S_i \times \frac{\sum_{o' \neq o} P_{o'}}{N-1} \times S_i^T, i = 1, \dots, N. \quad (2.10)$$

Then, the method performs spectral clustering (subsubsection 2.1.2.3) in the normalized matrix sum.

Since the similarity information is only diffused in the common neighborhood, SNF is robust to biological noise. Furthermore, optimizing the number of neighbors (T) and the hyperparameter α is necessary.

2.2.3.2 Cancer Integration via Multikernel Learning

The CIMLR method is based on an algorithm developed for the analysis of single-cell RNA-Seq data, the single-cell interpretation via multikernel learning (SIMLR) method [58]. CIMLR uses the Gaussian kernel for each block, X_i , independently to obtain 55 Gaussian kernels with different variances for each block and each kernel is a measure of distance between patients. Initially, N Gaussian kernels were computed however this increases the computation time and 55 kernels were derived empirically as a plateau of information given. Considering x_i the i -th row (in this case, patient i) of the input data and ϵ_{ij}^2 the variance, the Gaussian kernel is calculated as:

$$K(x_i, x_j) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\epsilon_{ij}^2}\right). \quad (2.11)$$

For the CIMLR method, the Gaussian kernel is modified:

$$K_s^\sigma(i, j) = \frac{1}{\sigma(\mu_i + \mu_j)\sqrt{2\pi}} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2(\mu_i + \mu_j)^2}\right), \text{ with } \mu_i = \frac{\sum_{j \in KNN(i)} \|x_i - x_j\|}{s}, \quad (2.12)$$

where s and σ are hyperparameters defined as $s = 10, 12, 12, \dots, 30$ and $\sigma = 1, 1.25, 1.5, 1.75, 2$ and $KNN(i)$ denotes the top T neighbors of the i -th cell.

To determine the optimum number of clusters, the eigenvectors of the similarity Laplacian are studied. For a given k , the goal is to find the indication matrix $Z(R) = XR$, with X being the matrix of the top k eigenvectors and R a rotation matrix. The cost function is defined as:

$$J(R) = \sum_{i,j} \frac{[Z(R)]_{i,j}^2}{[M(R)]_{i,j}^2}$$

where $[M(R)]_{i,j} = \max_j [Z(R)]_{i,j}$ [40]. The cost function, denominated separation cost, allows us to recover the rotation R for which each row in Z there will be at most one non-zero entry and the minimization will provide the best alignment with the canonical coordinate system [59]. This means that in an ideal scenario where each cluster is distinct and well-separated, the eigenvectors of the Laplacian matrix (or the transformed matrix Z) would align closely with the canonical coordinate system, which is a common assumption in spectral studies.

To find the optimal number of clusters k , originally the minimum value of the previous cost function was chosen [59], yet the CIMLR authors [40] define k as the largest drop in the value of $J(R)$. While the authors don't theoretically justify this change, intuitively it allows

the determination of the point at which adding another cluster does not significantly reduce the cost function, which might help prevent overfitting [60].

Then, the multiple kernels are combined into a symmetric similarity matrix with k blocks, where each block is constituted by similar patients. The main idea behind the CIMLR is that if there are k clusters in the original data, then the similarity matrix should have an approximate block-diagonal structure with k blocks [40, 58].

The similarity matrix and the weight of each Gaussian kernel are obtained through the following optimization:

$$\begin{aligned} \min_{S, L, w} & - \sum_{i, j, l} w_l K_l(x_i, x_j) S_{ij} + \beta \|S\|_F^2 + \gamma \text{tr}(L^T (I_N - S) L) + \rho \sum_l w_l \log(w_l) \quad (2.13) \\ \text{subject to } & L^T L = I_k, \sum_l w_l = 1, w_l \geq 0, \sum_j S_{ij} = 1, S_{ij} \geq 0. \end{aligned}$$

The matrix $L_{N \times k}$ is an auxiliary matrix that transforms the original data into a lower-dimensional space and enforces a low rank on S . The w_l is the weight of each Gaussian kernel, I_N and I_k are $N \times N$ and $k \times k$ identity matrices, β and γ are nonnegative tuning parameters [40, 58], and l is the kernel index that ranges from 1 to $55 \times N$.

The first term of the objective function enforces that the learned similarity between two patients should be high if the Gaussian kernel value is high. The second term is a regularization term that prevents the learned similarity matrix from becoming similar to an identity matrix [58] or from having high similarity values, overall discouraging the learned similarity from being too extreme. The third term enforces a low-rank structure on the transformed matrix, $L^T (I_N - S) L$. This operation transforms the similarity matrix S into this lower-dimensional space and captures how the patients relate to each other in the reduced space. The matrix $(I_N - S)$ is essentially the graph Laplacian and, by minimizing the trace, a sparse pattern in the diagonal elements of the transformed matrix is induced. This can be interpreted as encouraging approximately k connected components (since L is a $N \times k$ matrix) in a similarity graph, and creating k connected components is conceptually similar to forming k clusters. The fourth term imposes constraints on the kernel weights to avoid having a single kernel while trying to assign appropriate importance to different similarity measures.

The method then performs dimension reduction on the similarity matrix obtained using the t-distributed stochastic neighbor embedding (t-SNE) [61] and K-means clustering.

The CIMLR algorithm has been applied to different multi-omics datasets such as liver hepatocellular carcinoma, lung adenocarcinoma, and, most relevantly, to LGG [40]. The latter study was developed in 2018, using the TCGA dataset with the previous WHO CNS classification, accounting for 282 patients and it showed that 3 clusters were the better fit, followed by 7 and 13 clusters. The 3 clusters were concordant with the contemporary gliomas classification (oligoastrocytoma, astrocytoma, and oligodendroglioma) and showed a clear separation.

Once more, while CIMLR has been applied to multi-omics glioma data, it only contemplated patients from the LGG type and the algorithm results were based on the outdated WHO Classification of Tumors of CNS regarding gliomas, which is a noteworthy limitation since some LGG in 2016 are now considered GBM, the most aggressive glioma type. These are drawbacks that the present thesis seeks to explore while implementing the CIMLR methodology.

Table 2.2: A summary of multi-view clustering methods.

Method	Type	Data Types	Variable Selection	Preprocessing	Multi-omics Implementations
iCluster [42]	MF with Gaussian latent model	Continuous (Gaussian distribution)	×	Row centering [42]	GBM and Breast cancer [11], Uveal melanoma cancer [46], Colorectal cancer [44], Breast cancer [48]
iCluster+ [43]	MF with Generalized linear regression	Continuous, binary, categorical	✓	Row centering [42]	Colorectal cancer [44], Breast cancer [48], Murine liver, Liver cancer and Obesity [41]
moCluster [44]	MF with Modified consensus PCA	Continuous	✓	Centering, scaling, and block scaling (weight each data matrix by the reverse of their first eigenvalue [44])	Colorectal cancer [44], Breast cancer [48], Murine liver, Liver cancer and Obesity [41]
MOFA [49]	Matrix Factorization	Continuous, binary, and count	✓	Normalization [49]	Chronic lymphocytic leukaemia [49]
intNMF [11]	Nonnegative Matrix Factorization	Continuous	✓	Block scaling (dividing each data matrix by the maximum value) and if data entries are negative, data shifting (add the absolute value of the smallest negative number).	GBM and Breast cancer [11], LGG [46], Murine liver, Liver cancer and Obesity [41]
SNF [45]	Similarity Matrix	Continuous, binary, categorical	×	Normalization	GBM, Breast invasive carcinoma, Kidney renal clear cell carcinoma, Lung squamous cell Carcinoma, and Colon adenocarcinoma [45]
CIMLR [40]	Similarity Matrix	Continuous, binary	✓	Normalization so that values ranged between 0 and 1	LGG, Kidney cancer (clear cell renal cell carcinoma) Cutaneous melanoma, Prostate, and Breast Cancer [40] Murine liver, Liver cancer and Obesity and diabetes datasets [41]

A summary and comparison of the exposed multi-view clustering methods can be consulted in Table 2.2 where the type of method, the input data type, if variable selection is available, the preprocessing steps, and various multi-omics implementations are present.

2.3 Consensus Clustering Methods

Consensus clustering (CC), also known as unsupervised ensemble learning or aggregation of clustering, gains relevance when clustering results vary due to local maximum or minimum, the specific algorithm applied, or the data (different data or feature spaces) used [62, 63]. The primary objective of CC is to combine these various partitions into a single robust clustering that offers improved quality and stability. Ensemble clustering aims to produce a more reliable and consistent grouping of data, since the different methods have their strengths and weaknesses, by integrating multiple clustering results it is expected their joint contribution to have a compensatory outcome [62, 64, 65].

Ensemble clustering can be divided into two stages: generation mechanisms and consensus function [66, 67]. The first which can also be designated as diversity [68] generates base clusterings and the latter combines the base clusterings into a final clustering [66–68]. For example, think of multiple runs of K-means as the generation step and a consensus function to combine them. In this thesis, the focus for this section is the choice of the appropriate consensus function since the generation step is obtained through multi-view clustering, as detailed previously. Furthermore, criteria exist to choose the base clusterings incorporated into the final clustering, mostly using cluster validation measures or covariance, yet these are dependent on the generation method and therefore not explored in this context. A table comparing the criteria can be consulted in [68].

Consensus functions fall broadly into two classes: median partition-based and object co-occurrence [63, 69]. The objective of the former is to maximize the similarity between the different partitions using a distance measure, while the latter identifies a consensus clustering based on the co-occurrence of objects within the same clusters. Available median partition-based methods include genetic algorithms, NMF methods, kernel methods, and Mirkin distance-based [62, 66]; and, for the object co-occurrence class, co-association based matrix, graph, and hypergraph partitioning, relabeling and voting approach, and finite mixture models can be highlighted [66, 68]. Additionally, locally adaptive clustering algorithms can also be considered relevant in the former category but these are dependent on the generation mechanism or in the dataset [66], therefore are not specified.

Throughout the previous sections clustering and partitions were used interchangeably, however for the current section consider $\mathbb{P} = \{P_1, P_2, \dots, P_m\}$ a set of partitions such that each $P_j = \{C_1^j, C_2^j, \dots, C_{d_j}^j\}$ is a partition of the data points with d_j clusters, and C_k^j is the k -th cluster from the j -th partition.

2.3.1 Median Partition

Consider a distance measure $d(\cdot, \cdot)$, and the main goal of the median-partition problem is to find a P' such that

$$P' = \arg \min_P \sum_{j=1}^m w_j d(P, P_j) \quad (2.14)$$

Note that in Equation 2.14 the parameter w_j is the weight for each clustering result, with $w_j \geq 0$ and $\sum_{j=1}^m w_j = 1$ [70], and most median partition-based methods set them with equal weight [62], meaning $w_j = 1, \forall j = 1, \dots, m$. Another possible way to write the Equation 2.14 is to maximize the similarity between P and P_j .

The main theoretical results obtained for the present category are only applicable when considering a symmetric difference distance or the Mirkin distance [66]. In this context, the Mirkin distance for two partitions, P_a, P_b , is defined as the sum of the number of pairs of objects clustered together in P_a , but separately in P_b and the number of pairs of data points clustered separately in P_a , but together in P_b [66]. Meaning that it accounts for the number of disagreements between the two partitions. In particular, using the Mirkin distance with a variable number of partitions was proven to be NP-hard [62, 66, 71] and caused the emergence of several heuristics to optimize the problem. For example, Best-of-k adopts a simplistic approach by choosing the partition in the clustering ensemble that minimizes the distance from it to all the other partitions [66].

Yet the complexity of the problem lies in the measure chosen with some consensus partition obtained in polynomial time. The most common similarity or dissimilarity measures are based on:

- Counting the pair of objects two partitions have in common or difference, for example using the Rand Index, the Mirkin distance, etc [66].
- Set cardinality comparisons, with Purity, the F measure, etc [66].
- Information theory that quantifies the information shared between two partitions, like Normalized Mutual Information, etc [66].
- Positive semidefinite kernels, such as Graph kernel and Subset significance [66].

2.3.2 Object Co-occurrence

Overall, most of the object co-occurrence algorithms assume that each clustering produced is independent evidence for the data organization and the “natural” cluster. The main idea is that when two points are clustered together in any given clustering method, it is considered a vote for their association, which is a concept introduced by the Evidence Accumulation Theory (EAC) [72]. This association is then mapped into an intermediate space, designated co-association matrix (*co_assoc*). Each entry ij in the co-association matrix represents the normalized number of votes for the association between data points i and j . The normalization is achieved by dividing the number of times i and j were clustered together by the total number of different clusterings. The co-association matrix thus provides a robust measure of the likelihood that two points belong to the same cluster, integrating information from multiple clustering solutions, and is then interpreted as a similarity matrix, where higher values in the matrix indicate a stronger likelihood that two points belong to the same cluster.

To determine the final clusters, different methods have various approaches, such as setting a threshold of 0.5 and objects with larger entries are in the same clusters (similar to fuzzy clustering); applying a clustering algorithm to the co-association matrix, with common choices including agglomerative hierarchical clustering [72] and K-means [64]; viewing the matrix as an adjacency matrix of a graph and apply cuts to define the final clustering [73].

The consensus clustering method introduced by Monti [74] can be considered one of the first and most influential CC methods, falling under the object co-occurrence branch. It is mainly employed to evaluate the stability of clustering results through multiple runs of a clustering method. For this CC method, an algorithm and resampling scheme should be chosen beforehand. To quantify the agreement between the m partitions from the algorithm, a consensus matrix is created by averaging the co-association matrix of every resampled (or perturbed) dataset created in each run. Then, after determining the number of clusters, the chosen clustering method is executed on the consensus matrix, originating the final partition, and the stability of the clusters can be assessed through the consensus cumulative distribution function.

However, it is important to highlight that this technique isn't viable for a multi-view dataset, particularly the implementation of the resampling scheme, meaning prevalence is given to consensus functions that only require the base clusterings. Yet, theoretically, Monti's CC method can be extended to compare the results without using the data itself. This is where the method Cluster of Clusters Analysis (COCA or C of C) gains relevancy. COCA has been applied frequently on tumor subtyping and was first used in [75], but only formalized in [76]. While the main goal behind Monti's CC is to assess cluster stability in a single dataset, COCA's objective is to summarize the partitions found in different omic datasets into a final global clustering across the datasets, meaning a different partition for each dataset. Similarly to the connectivity matrix, a "matrix of clusters" is defined which is then used as input as a similarity matrix in a hierarchical clustering method (or any other based on distance). Some additional considerations are:

- Number of clusters as a parameter, advised to be set as the value that maximizes the average silhouette value, as this tends to yield more sensible results compared to the original method, choosing the "best" matrix, described in [75].
- COCA only needs data partition allocations, the data is not an input.
- Sensitive to distance metric chosen and gains the disadvantages of the single-view clustering method chosen.
- All partitions results have the same weight.

The last point is highlighted in [76], where the authors propose an alternative based on kernel methods, making it a median partition method named Kernel Learning Integrative Clustering. Theoretically, this method builds upon COCA intending to improve

it. However, since it extends from COCA, it still focuses on combining the clustering of different omics — a single-view to multi-view problem — which is not the focus here. Nevertheless, COCA remains a frequently used method for cancer subtyping, hence its relevance in the literature review. Another single-view to multi-view model is ensemble based multi-view archived multi-objective simulated annealing [77] which obtains the final clustering through a Pareto optimal front.

Furthermore, a criticism of the co-association matrix methods is that as a latent representation, it should have more information about the real representation of the data, making the approach too simplistic [66]. With this in mind, variations of the co-association matrix have been presented, including a weighted and a probabilistic variant [78, 79].

2.3.2.1 Graph and Hypergraph

The graph and hypergraph methods transform this problem into a graph and hypergraph, respectively, and the main difference lies in how to build this representation. Note that the edges are the main difference between the graph and the hypergraph. In the former the edge connects exactly two nodes and a hyperedge in the hypergraph can connect any set of nodes, generalizing the concept of an edge.

One of the most used approaches is to use normalized mutual information (NMI) to define the consensus partition as the one that shares the most information with all base clusterings and it is introduced by Strehl and Ghosh [73]. This can be expressed as:

$$P' = \arg \max_P \sum_{j=1}^m NMI(P, P_j) \quad (2.15)$$

Yet this optimization problem is computationally not feasible [73], and NP-hard [67]. Therefore, the authors suggest three heuristics: Cluster-based Similarity Partitioning Algorithm (CSPA), HyperGraph Partitioning Algorithm (HGPA), and Meta-CLustering Algorithm (MCLA). Note that with Equation 2.15 some questions might be raised as to why this method is not median partition-based, but the following heuristics definition should clarify this.

CSPA's first step is to build a similarity matrix (or co-association matrix) which can also be considered an adjacency matrix of a graph with the objects (patients) as nodes and the weight of each edge as the fraction of clustering two nodes in the same cluster. Then, the graph partitioning algorithm METIS is applied to the adjacency matrix, and the results are the final partition. The heuristic considers all base clusterings with the same weight and has a simplistic approach overall. Still, the computational time is $O(kn^2m)$, meaning quadratic to the number of patients.

HGPA starts by building the hypergraph by mapping the data points and the cluster allocations like in Table 2.3, and then searches for a cut that minimizes the number of hyperedges to eliminate by partitioning the hypergraph in k connected components of approximately the same size. In this heuristic, all hyperedges, nodes, and partitions have the same weight and the final partition has balanced clusters [67, 68, 73].

Table 2.3: Example for building the hypergraph given in [68]: a) Cluster allocations of 5 objects with two partitions, one with 3 clusters and another with 2. b) Incidence matrix of the hypergraph with 5 hyperedges (3 from H_1 and 2 from H_2) and 5 nodes (each data point).

(a)			(b)					
	P_1	P_2	H_1			H_2		
			h_1	h_2	h_3	h_4	h_5	
x_1	1	1	n_1	1	0	0	1	0
x_2	1	1	n_2	1	0	0	1	0
x_3	2	2	n_3	0	1	0	0	1
x_4	2	2	n_4	0	1	0	0	1
x_5	3	1	n_5	0	0	1	1	0

The MCLA heuristic is based on clustering clusters [73], providing a different type of approach and visualization. After building the hypergraph as in Table 2.3, the similarity of any two clusters is computed, for example, using the Jaccard Index [66]:

$$JI = \frac{|C_a^1 \cap C_b^2|}{|C_a^1 \cup C_b^2|}$$

This makes it possible to build a meta-graph where each node is a cluster and the edge weight is the similarity between the clusters. Then, hyperedges are clustered by partitioning the meta-graph in k balanced meta-clusters, using the METIS algorithm [73]. Each meta-cluster consists of clusters that belong to different partitions and were originally represented as hyperedges in the original hypergraph.

The following step is to collapse the hyperedges in a single meta-hyperedge for each of the k meta-clusters. Each meta-hyperedge has an association vector detailing for each data point the level of association (computed as the average of indicator vectors) with the respective meta-cluster, and the association vector can take values between 0 and 1, the former indicating the weakest association and the latter the strongest [73].

Let's consider a trivial example where a meta-cluster has 3 clusters, C_1 , C_2 , C_3 , for 4 data points. Each cluster has an association vector with the data points allocation, $C_1 = [1, 1, 1, 0]$, $C_2 = [1, 1, 0, 1]$, $C_3 = [0, 1, 1, 0]$. Thus, the meta-hyperedge is the average of the vectors, $[0.67, 1, 0.33]$, showing the level of association between the meta-clusters and the original objects.

After obtaining a single meta-hyperedge for each meta-cluster, the objects are assigned to their most associated meta-cluster, meaning the one with the highest association vector entry. This signifies that not all meta-clusters will have data points. Furthermore, it is possible to obtain confidence estimations of cluster memberships for each data point [73].

Another relevant method is ClustOmics [63], a graph-based that calculates how often pairs of samples appear together in the input clusters and uses this frequency as a similarity metric to restructure the data into consensus clusters. Its implementation allows

a single-to-multi-view and a multi-view scenario, the latter being relevant in our context. The method allows for patient clinical and survival data to be considered besides the clustering obtained in single-view clustering algorithms, but for the multi-view data only the predicted clusters are first considered to get the consensus clustering. The first step of the method is creating a support graph, with each patient a node and a support edge exists between nodes if there is at least one clustering result where the patients were grouped. Then, since this graph tends to be densely connected, the graph is filtered automatically to remove non-consistent integration edges with a weighted modularization strategy, aiming to find the one that does not induce too much sparsity to be informative. With this filtered graph, two state-of-the-art graph clustering strategies (Louvain community detection and Markov clustering) are employed, and the one that maximizes a metric proposed by the authors, i.e. weighted Modularization quality, is chosen. Finally, an optional step is presented: reintroducing the filtered-out patients to the consensus cluster with the mean number of supports shared with patients from consensus clusters.

Table 2.4: A summary of consensus clustering methods.

Method	Type	Consensus function	Dataset necessary
COCA [75, 76]	Object co-occurrence	Co-association matrix	×
CSPA [73]	Object co-occurrence	Graph	×
HGPA [73]	Object co-occurrence	Hypergraph	×
MCLA [73]	Object co-occurrence	Hypergraph	×
ClustOmics [63]	Object co-occurrence	Co-association matrix, Graph	✓

2.4 Clustering Validation Measures

Clustering validation measures can be classified into internal and external validation [80]. Different articles propose various divisions of classes, sometimes adding or altering groups. For example, considering external, internal, and relative measures [81] or grouping into pair-counting, set-matching, and information-theoretic measures [82]. Despite these variations, the aforementioned classification is used since it seems to be the most encompassing. Let's explore each of these categories in detail:

2.4.1 Internal Validation

Internal validation metrics are calculated based on properties of the obtained clusters, such as compactness, separation, and roundness, without using additional information about the data [80, 81]. This group of methods operates on the premise that the algorithms should identify clusters where the members are near one another while being distant from the members of other clusters [81].

For this subsection, consider a single partition P_1 such that $P_1 = \{C_1^1, \dots, C_{d_1}^1\} = \{C_1, \dots, C_d\}$ with $d_1 = d$ the number of clusters and C_i the i -th cluster.

2.4.1.1 Dunn Index

The Dunn index is the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance:

$$\text{Dunn}(P_1) = \frac{\min_{i,j=1,\dots,d} d(C_i, C_j)}{\max_{i=1,\dots,d} \delta C_i} \quad (2.16)$$

where $d(C_i, C_j)$ is the distance between the two clusters, meaning the minimum distance between any point in cluster i and any point in cluster j , and δC_i is the diameter of cluster i , defined as the maximum distance between any two points within C_i [83, 84]. Note that there are different ways to calculate this index, in particular, the denominator can be set as the size of cluster i [81], meaning that every combination of distance measure and denominator definition (and cluster-size) defines a different Dunn's index.

The Dunn index ranges from 0 to infinity, and a high Dunn index means that the distance between the clusters is large and the diameter of the cluster is small, indicating compact and well-separated clusters [83, 84].

2.4.1.2 Calinski-Harabasz Index

The Calinski-Harabasz index is defined as:

$$\text{Calinski-Harabasz}(P_1) = \frac{\frac{BCSS}{k-1}}{\frac{WCSS}{n-k}} \quad (2.17)$$

where $BCSS$ and $WCSS$ are the averages between- and the within-cluster sum of squares, respectively, n is the number of data points, and k is the number of clusters [80, 85]. This means that the Calinski-Harabasz index shows the quality of the predicted clusters based on the average sum of squares between and within the cluster.

The average between-cluster sum of squares is computed as:

$$WCSS = \sum_{i=1}^k n_i d(c_i, c)^2$$

with c_i the centroid of cluster i , c the mean of all data points, and $d()$ the Euclidean distance [85, 86].

The average within-cluster sum of squares is calculated as follows:

$$WCSS = \sum_{i=1}^n \sum_{x \in C_i} d(x, c_i)^2$$

with k the number of cluster, x a sample, C_i the i -th cluster, c_i the centroid of cluster C_i , and $d()$ the Euclidean distance [85, 86].

The previous formulations are the original based on the Euclidean distance yet the index was generalized for dissimilarities [87]:

$$WCSS = \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x_j, x_h} d(x_j, x_h)^2 BCSS = \frac{1}{n} \sum_{j,h} d(w_i, w_j)^2 - WCSS$$

This formulation makes it possible to use a distance matrix instead of the original data to compute the index.

The Calinski-Harabasz metric ranges from 0 to infinity and higher values reflect a better result, meaning that a high BCSS and low WCSS are preferable.

2.4.1.3 Average Silhouette Width

The silhouette is the average of the silhouette width of each point of all clusters. Considering x a point in cluster C_j and n_j the number of points in C_j , the silhouette width of x is defined as:

$$S(x) = \frac{1}{k} \sum_{i=1}^k \frac{b(x) - a(x)}{\max[a(x), b(x)]} \quad (2.18)$$

with $a(x)$ the average pairwise distance between object (patient) x and all other objects in the same cluster C_j ,

$$a(x) = \frac{1}{n_j - 1} \sum_{y \in C_j, y \neq x} d(x, y) \quad (2.19)$$

and $b(x)$ the minimum of the average distances between x and the points in the other clusters,

$$b(x) = \min_{h=1, \dots, k, h \neq j} \left[\frac{1}{n_h} \sum_{y \in C_h} d(x, y) \right]. \quad (2.20)$$

Thus, the average silhouette width, also known as silhouette, index is defined as

$$S = \frac{1}{k} \sum_{j=1}^k \left[\frac{1}{n_j} \sum_{x \in C_j} S(x) \right] \quad (2.21)$$

which ranges from -1 to 1 . If the value is near -1 , it indicates that the point is on average closer to a different cluster than the assigned cluster. If the value is near 1 , it means the point's average distance to its cluster is much smaller compared to its distance to other clusters. A higher silhouette value signifies that the clusters are more distinct and well-separated from each other [81].

2.4.2 External Validation

Much like the name suggests, external validation indexes use external information to evaluate the structure of the predicted clusters, usually the ground truth labels [80, 81].

2.4.2.1 Normalized Mutual Information

NMI is used based on the Mutual Information concept from Information Theory, and it is a similarity measure of the information shared between two partitions. Let $P_a = \{C_1^a, C_2^a, \dots, C_A^a\}$ and $P_b = \{C_1^b, C_2^b, \dots, C_B^b\}$ be two partitions, A being the number of clusters in P_a and B the number of clusters in P_b . Let n_{ia} be the number of objects in the i -th cluster of the partition P_a , n_{bj} the number of objects in the j -th cluster of the partition P_b and n_{ij} the number of objects which are together in the i -th cluster of the partition P_a and in the j -th cluster of the partition P_b .

$$NMI(P_a, P_b) = \frac{-2 \sum_{i=1}^A \sum_{j=1}^B \frac{n_{ij}}{n} \log \left(\frac{n_{ij}n}{n_{ia}n_{bj}} \right)}{\sum_{i=1}^A n_{ia} \log \left(\frac{n_{ia}}{n} \right) + \sum_{j=1}^B n_{bj} \log \left(\frac{n_{bj}}{n} \right)}$$

NMI varies between 0 and 1, and, in a supervised context, it is usual to have one of the partitions as the ground truth labels and the other as the predicted partitions. Therefore, a higher NMI means a higher similarity between the ground truth and the prediction, meaning an accurate assignment.

2.4.2.2 Adjusted Rand Index

To calculate the Adjusted Rand Index (ARI), consider the following contingency matrix [88]:

Table 2.5: Contingency matrix of two partitions: a represents the number of pairs that are clustered together in both partitions; b the number of pairs that are in the same group in P_b but in different groups in P_a ; c the number of pairs that are in the same group in P_a but in different groups in P_b ; and d the number of pairs that are in different groups in both partitions.

P_b	P_a	
	Pair in the same group	Pair in different groups
Pair in the same group	a	b
Pair in different groups	c	d

From this definition, the ARI created by Hubert and Arabie [89] is calculated as shown in Equation 2.22:

$$ARI = \frac{\binom{n}{2}(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{\binom{n}{2}^2 - [(a + b)(a + c) + (c + d)(b + d)]} \quad (2.22)$$

where n is the number of samples, $\binom{n}{2}$ is the total number of pairs, and a , b , c , and d are defined as in Table 2.5.

The ARI ranges from -1 to 1 . It has an expected value of zero in the case of random partitions, values approach one as the two partitions become more similar to each other (with one being the perfect match of the classification) [89], and values below 0 are considered worse than a random partition.

METHODOLOGY

The present chapter displays the thesis pipeline and details the process of data collection and preprocessing, as well as the reasoning behind the implemented multi-view and consensus clustering methods. Furthermore, implementation details, such as programming language, version, and metric used, are presented, aiming at reproducibility and transparency.

3.1 Pipeline

Figure ?? shows the pipeline adapted in the present dissertation, which is detailed and justified in the following sections.

3.2 Data Collection and Preprocessing

The multi-omics glioma data used was obtained through TCGA, a publically accessible repository with various multi-omics datasets of multiple cancers [90]. The TCGA aims to catalog and discover major cancer-causing genome alterations and, at present time, has molecularly characterized over 20,000 primary cancers [90, 91]. The extensive datasets have enhanced cancer diagnosis, treatment, and prevention and are of public access to further the knowledge of the disease [90].

For the present thesis, two omics datasets were studied: transcriptomics and methylomics. The former is the set of all molecules of RNA in a cell and provides information about the genes that are being transcribed [92]. The latter includes methylation levels for each CpG site across all patient samples and it is considered to have a fundamental role in transcriptional regulation, abnormalities that may lead to a multitude of diseases, including cancer (with both hypo- and hypermethylation observed in cancer patients). Additionally, epigenetic events, including DNA methylation, are reversible, which makes it an interesting cell level to study and then develop potential new approaches to therapy [93].

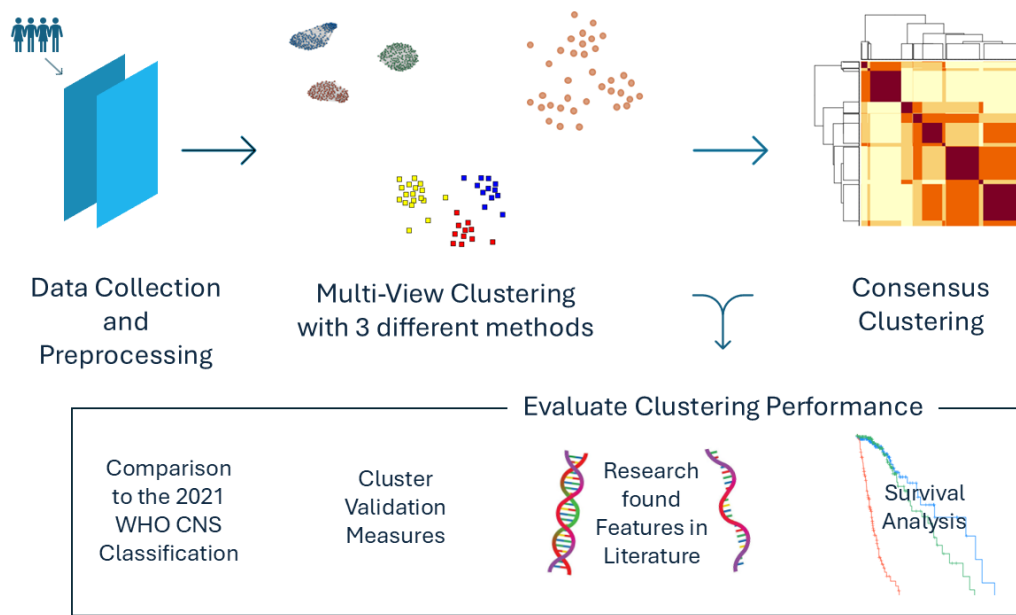


Figure 3.1: Thesis pipeline: Data is collected and preprocessed, followed by the application of the multi-view clustering methods chosen. These clusters are then evaluated by comparison to the 2021 WHO CNS classification, computing cluster validation measures, identifying key features (and if they are known or potential biomarkers for gliomas), and performing a survival analysis to evaluate the clusters’ predictive value. With this information in mind, the best result for each MVC method is selected, and consensus clustering is performed and evaluated in an identical process.

The dataset patient labels are in accordance with the current 2021 WHO guidelines for the classification of CNS tumors, following an update by the “MONET- Multi-omic networks in gliomas” project consortium [5]. Only the observations (i.e., patients) shared by the methylomics and transcriptomics datasets were considered. This means a total of 527 patients (249 astrocytomas, 112 GBM, and 166 oligodendrogliomas), 20,201 genes, and 301,541 CpG sites were considered, after the removable of constant variables.

To apply each method the data is normalized, which is important to maintain similar scales of the multiple datasets (methylomics ranges from 0 to 1, while transcriptomics varies from 0 to 2,554,530, approximately). Normalization ensures that the results are not skewed towards the data with a wider range [46]. Specifically, for the CIMLR and intNMF methods, the transcriptomic data was normalized so that it ranged from 0 to 1, by dividing by the maximum value, and the methylation data didn’t need this transformation since it naturally is a ratio between 0 and 1. To apply the moCluster method, both datasets were standardized and then divided by the first singular value, as recommended.

3.3 Multi-view Clustering Methods

Maintaining the thesis goals present, the methods outlined in Table 2.2 were considered for implementation. Identifying the most important features (such as genes and CpG sites) is interesting in this context, therefore methods that do not perform any kind of variable selection were disregarded.

Based on the literature review performed, iCluster+ has not consistently outperformed alternative methods of clustering and latent variable recovery [41, 44]. Additionally, the method is a computationally intensive algorithm and is considered especially demanding for nondeterministic algorithms, making it less efficient for large datasets.

On the other hand, the intNMF showed the most promising result in clustering a simulated multi-omics dataset [12]. In this article, MOFA was also implemented and demonstrated strong performance. However, intNMF outperformed it overall and is recommended as it is specifically designed for clustering tasks, unlike MOFA, which is not inherently intended for clustering. However, it should be noted that intNMF does not always yield the best results when investigating factor-level information, such as clinical annotations or survival data. In these cases, MOFA tends to outperform, particularly in its ability to integrate and interpret latent factors.

Although MOFA is highly regarded in the literature, the authors in [41] compared 13 different methods including MOFA on the ability to cluster and recover latent variables. They found that the MOFA method had difficulties converging on all the (simulated) benchmarks and could not find the structures, even in relatively simple scenarios. Furthermore, difficulty in tuning the sparsity parameters was highlighted. Considering the goals of this thesis, particularly the need for robust latent variable recovery and efficient clustering, MOFA will not be employed.

In contrast, CIMLR excelled in both simulated and real-world multi-omics datasets. For example, when tested on murine liver data, CIMLR achieved perfect F1 and ARI scores, demonstrating its strong clustering performance [41].

Furthermore, the moCluster algorithm was compared to different clustering algorithms in [41, 48], and both highlighted the overall performance of the method.

Therefore, the intNMF, moCluster, and CIMLR are the methods chosen for the present thesis, which represents a variety of multi-view clustering methods having dimensionality reduction (intNMF and moCluster), similarity/graph-based (CIMLR) and statistical-based (moCluster) methods.

3.4 Consensus Clustering Methods

The chosen clustering ensemble algorithm should be tailored to meet the specific objectives established. Namely, in the context of multi-omics with different clustering methods, there is an argument favoring the relevance of comparing object co-occurrence over maximizing cluster similarity (the median partition problem) [63]. The reason behind this statement,

while not explicitly stated, likely reflects differing objectives. In the study of multi-omics, particularly in patient stratification, understanding which patients are consistently clustered together may be more pertinent than identifying clusters that are closely located in space. With this idea in mind, the object co-occurrence becomes particularly relevant in the present thesis.

While researching these methods, it became apparent that there is a lack of current comprehensive practical and experimental studies on consensus methods based on object co-occurrence. Most reviews tend to focus on theoretical analyses [66–68, 94]. Kuncheva et al [95] conducted an experimental study, but it had some limitations: all clusters were created using the same method (i.e., no different methods were used for the same consensus function), the original co-association matrix was used without considering other variations (likely due to the article’s release in 2006), and the study focused only on traditional clustering methods and graph-based consensus functions. It was concluded that better global clustering is achieved when treating the co-association matrix as data rather than as a similarity matrix, using average-linkage followed by K-means clustering. Therefore, this information will be considered, and methods mentioned in surveys are given relevancy as a form of peer validation.

Fred and Jain [72] explored the EAC and different clustering methods applied to the co-association matrix. It was clear that, for most of the datasets, EAC with average or single-linkage had a lower error rate than the hypergraph methods, CSPA, HGPA, and MCLA. However, the choice between the hierarchical methods is not evident and context context-dependent. Overall the average-linkage is more robust and produces better results (lower error rate), especially in situations of touching clusters. However, when the clusters are well-separated, single-linkage shows a better performance.

Moreover, while obtaining a clear visualization of patient-to-patient relationships can enhance the understanding of the complex interactions within the data, the main hypergraph and graph methods tend to generate balanced and equal-sized clusters which is not desirable. Additionally, the MCLA method graphs another type of association than object to object, making them not as relevant to fulfill the mentioned visualization. Furthermore, the methods are easily understood and most have low computational complexity [66, 67], except for CSPA, which has computational time quadratic to the number of samples. This is a considerable limitation but in this context does not seem alarming since the multi-omics datasets under study are not high dimensional in the sample space. Yet, the 527 patients generate a co-association matrix with 277,729 entries of which 138,751 are unique (disregarding the diagonal).

It is noteworthy to highlight the surge of new consensus clustering methods after 2010 [96]. Despite this progress, there is a notable absence of comprehensive reviews that focus on these novel methodologies, with most existing reviews being pre-2010 or centered on classic methods. The survey by Boongoen et al [96] theoretically reviews several current methods for different data, objectives, and concerns, thereby aiming to fill this gap in the literature.

In light of this, although ClustOmics is a relevant state-of-the-art method, only Song et al [97] reviews and employs the method, but not for consensus clustering purposes, to compare their method's ability to determine the number of ideal clusters with ClustOmics's. Therefore, while the method shows great potential and seems promising, using ClustOmics specifically for consensus clustering purposes necessitates further investigation and empirical validation to establish its efficacy in this context.

Additionally, Xiuge Wu et al's study [98] is one of the few articles that provides an experimental comparison of consensus functions, including voting approach, single-linkage, complete-linkage, and mean-linkage. The study concludes that the best performance is achieved using K-means with different initializations as a generative mechanism and average-linkage agglomerative clustering as a consensus function. This makes it the second article to reach this conclusion regarding consensus functions. It is important to note that, while the article was released in 2018, the consensus functions discussed are classic methods. The article also addresses contemporary topics such as the comparison of clustering ensembles and selective clustering ensembles.

Taking into account the previous reflection, the best result from each multi-view clustering method is considered, and the consensus function used is the average-linkage agglomerative clustering with the traditional co-association matrix.

3.5 Evaluate Clustering Performance

To evaluate the clustering performance from both the MVC and the CC, 4 main steps are considered: comparison of predicted labels to the 2021 WHO CNS classification; cluster validation measures; study of relevant features found and research in existing literature; and survival analysis.

The latter is employed to evaluate the prognostic value of the clusters. To do so Kaplan-Meier curves were employed and they represent the probability of survival on the vertical axis and the time on the horizontal axis. To define the Kaplan-Meier curves, each patient needs 3 personal characterizations [99]:

- Serial time, meaning the length of time they have been observed. Can be days until death or days until the last check-up.
- Status at the end of serial time, indicating if the event occurred or if it is censored (the total survival time for that subject cannot be accurately determined). This second variable is binary.
- Study group they are inserted, which are the clusters predicted.

Consider that the n patients are followed-up at distinct times $t_1 < t_2 < t_3 < \dots < t_n$. The conditional probability of surviving past time t_i is $1 - d_i/n_i$, with d_i the number of events that occur at time t_i and n_i the number of patients just alive before t_i . Since the events are independent, the survival probability from one interval to the next can be

multiplied to give the cumulative survival probability. Meaning that the probability of being alive at t_i , $S(t_i)$, is:

$$S(t_i) = S(t_{i-1}) \left(1 - \frac{d_i}{n_i}\right) \quad (3.1)$$

where $t_0 = 0$ and $S(0) = 1$ [100]. The survival probability is constant between events, resulting in a step function, noncontinuous, where the estimated probability decreases vertically when at least one event occurs [99–101]. It is also important to highlight that once the first patient is censored the curve becomes an estimate, making it relevant to show censoring in the plot [99].

For this context, the Kaplan-Meier curves represent the survival probability of the predicted clusters, allowing a visual comparison. Furthermore, the curves can be statistically compared through a log-rank test, a nonparametric test. The null hypothesis is that the survival curves are equal over time [101]. The log-rank test calculates the χ^2 for each event time for each group and sums the results, which are then added to find the global χ^2 and this is compared to the curves of each group.

It is important to note that the clinical information is not available for 5 out of 527 patients, so only these were contemplated in the survival analysis.

3.6 Implementation Details

The Master thesis analysis was conducted in RStudio using R versions 4.4.1 and 4.0.5. Version 4.0.5 was used specifically on Kaggle due to the platform’s fixed version and the necessity for its CPU and RAM resources. Version 4.4.1 was utilized locally to leverage the latest features and improvements in the newest release. Key packages utilized in the multi-view clustering implementation include ‘CIMLR’ for the CIMLR method [40], ‘intNMF’ for the intNMF method [102], and ‘mogsas’ [103] for the moCluster method. For reproducibility, a random seed of 12345 was set.

The Euclidean norm was always employed throughout the calculations to measure distances between data points.

Furthermore, when given the choice of the single-view clustering method to apply, K-means was privileged since the CIMLR algorithm implementation relies on this clustering method, and maintaining this consistency helps in comparing the results more faithfully.

RESULTS AND DISCUSSION

The present section exposes the results obtained through the present thesis, including data visualization, the multi-view clustering methods (CIMLR, intNMF, and moCluster), and the consensus clustering method chosen. The partitions obtained were then evaluated by comparing them to the 2021 WHO CNS classification, calculating cluster validation measures, and a survival analysis was employed. Furthermore, an interpretation of the findings and a comparison of the identified features with existing literature are presented.

4.1 Data Visualization

As mentioned previously, the transcriptomics and the methylomics datasets were studied. In [Figure 4.1](#), a visualization of the patients with a t-SNE 2D projection of both omics can be observed: for the transcriptomics data, while overlap can be observed between glioma types, there is still a clear group structure across the gliomas; in the methylomics data, each glioma type is evident and distinct, even if some points are overlapping, in particular the astrocytoma and oligodendroglioma which is expected.

4.2 Multi-view Clustering

4.2.1 CIMLR

4.2.1.1 3 Clusters

The CIMLR method was implemented first considering 3 clusters to evaluate the method's ability to cluster the glioma types according to the 2021 WHO tumors CNS classification. Due to the large dataset, the default number of neighbors to be considered needs some tuning. The authors suggest setting this parameter between 30 to 50 when working with large datasets and setting the number of dimensions as the number of clusters [58]. Taking into consideration the cluster validation measures obtained ([Table A.1](#)), especially the internal validation ones, it seems appropriate to consider the top 50 neighbors. The confusion matrix between the predicted and the official labels can be consulted in [Table 4.1a](#), and the cluster validation metrics are in [Table 4.1b](#). The a 2D-visualization using t-SNE

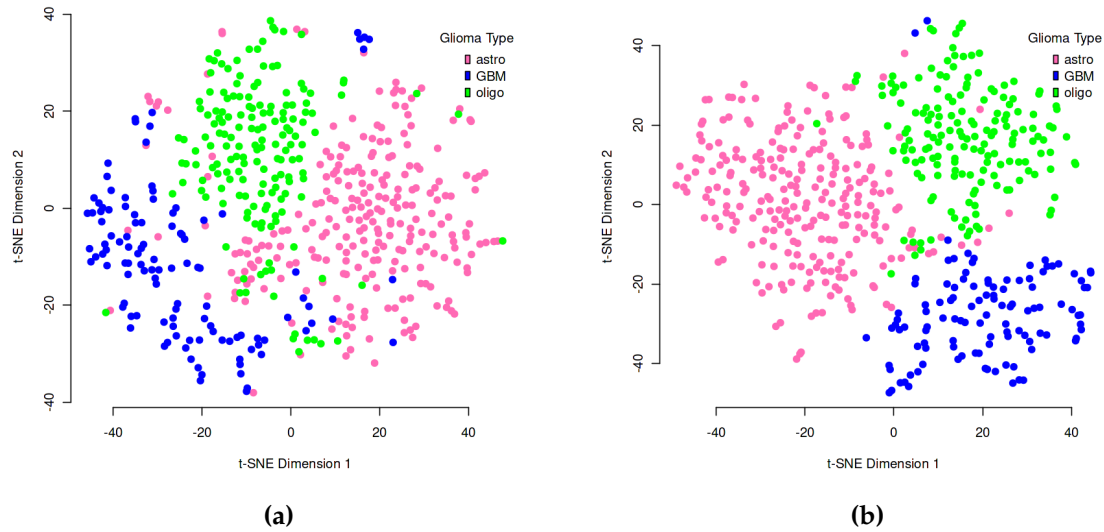


Figure 4.1: t-SNE 2D Projection of: a) Transcriptomics. b) Methyloomics.

of the similarity matrix with each cluster identified is in [Figure 4.2a](#)), the similarity matrix between the patients with the predicted and official assignments identified are in [Figure 4.2c](#), and the similarity graph for the obtained clusters is presented in [Figure 4.2b](#).

Table 4.1: CIMLR result: a) Confusion matrix for the 3 obtained clusters versus the official glioma types. b) Cluster validation measures for the 3 clusters output. c) Confusion matrix for the 4 obtained clusters versus the official glioma types. d) Cluster validation measures for the 4 clusters output.

(a)					(b)	
CIMLR	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	30	1	157	188	Adjusted Rand Index	0.61
Cluster 2	35	111	9	155	Normalized Mutual Information	0.604
Cluster 3	184	0	0	184	Calinski-Harabasz Index	64,540
					Dunn Index	8.37
					Average Silhouette Width	0.954
(c)					(d)	
CIMLR	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	36	0	137	173	Calinski-Harabasz Index	38,544
Cluster 2	182	0	0	182	Dunn Index	5.075
Cluster 3	28	1	27	56	Average Silhouette Width	0.94
Cluster 4	3	111	2	116		

From the 3 clusters, the first cluster has a total of 188 data points, the second 155, and the third 184, and the confusion matrix shows low error in contrast to the known glioma types ([Table 4.1a](#)), with GBM clearly distinguished from the other types, and each cluster mostly characterized by one glioma type. Some overlap of astrocytoma and GBM and

astrocytoma and oligodendroglioma, an expected result due to the molecular similarities known in each pair, especially in LGG. Figure 4.2a shows a clear separation between the predicted clusters in the 2D plane.

Regarding the measures in Table 4.1b, the external validation measures show a decent clustering result since they are above 0.5. The Calinski-Harabasz and the Dunn Index gain more relevance only when comparing clusters. The silhouette plot (Figure 4.2d) shows that the partitions are cohesive and the Average Silhouette Width is 0.954, demonstrating a strong clustering structure obtained [104].

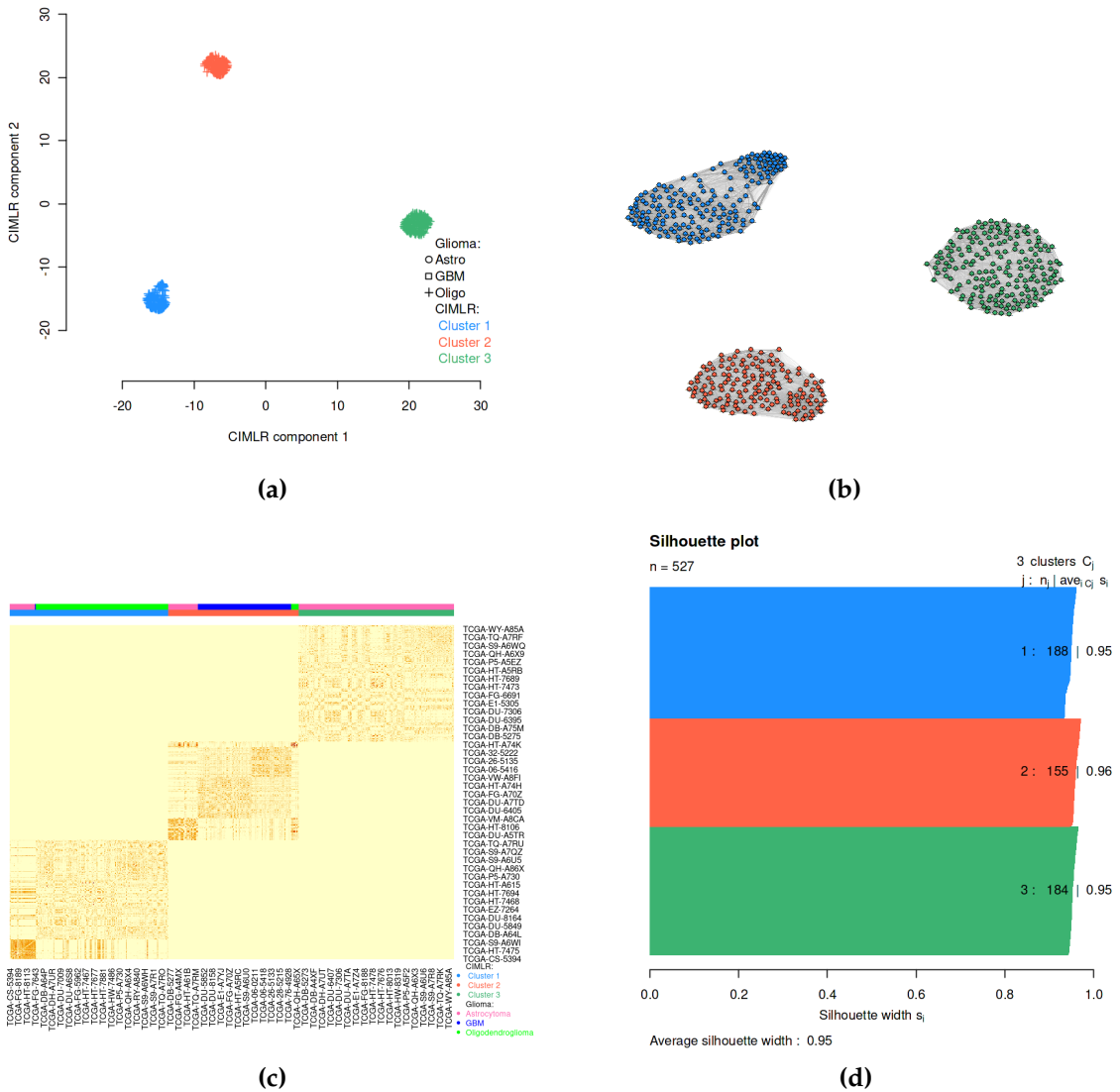


Figure 4.2: CIMLR result with 3 clusters: a) t-SNE 2D visualization of the CIMLR clusters. b) Similarity graph of the CIMLR clusters. c) Similarity matrix with official glioma types and CIMLR clusters identified. d) Silhouette plot.

Since the Calinski-Harabasz and the Dunn Index do not have a range, 1000 random

simulated partitions were generated, by attributing to each patient the label 1,2 or 3, randomly. Then, the indexes were calculated: the Calinski-Harabasz average is 1.004 (standard deviation is 0.690) and the average of the Dunn Index is 1.422×10^{-05} (standard deviation of 8.068×10^{-06}). Furthermore, the simulated and actual indexes presented have a high difference, so the simulated partitions were also calculated taking into consideration the glioma-type distribution in the dataset, yet the random metrics values were not affected. While this does not necessarily reflect a high quality for the obtained results, by comparison, it does mean the metrics values are better than a random partition one's.

In the similarity matrix (Figure 4.2c), 3 major blocks are distinguishable, which is the desirable outcome. In the first predicted cluster (in blue), the similarity between the patients is perceptible and intuitively it is clear why the patients are clustered together, yet 2 blocks can still be observed inside of this cluster. A similar point can be made with cluster 2 (orange) with 3 visible blocks, which consolidates the need to subtype the glioma types further. Analogously, in the similarity graph shown in Figure 4.2b, cluster 3 (green) seems cohesive, and in clusters 1 (blue) and 2 (orange) two subgroups can be identified, yet there are still enough edges between the groups for the clustering to tangibly make sense.

4.2.1.2 Optimum k

Subsequently, the separation cost was computed, considering up to 10 clusters, and it can be consulted in Figure 4.3. As mentioned previously, the function favors low values and the ideal number of clusters is the largest drop, which in this case means, 4 clusters should be considered, followed by 10 and 6. Notably, considering 2 or 3 clusters in the range for the optimum number of clusters might not seem relevant since the aim is to subtype glioma patients further. Yet in the analysis, they were contemplated to appraise the clustering methods chosen. Additionally, only 3 clusters were implemented to evaluate if the techniques are extracting information compatible with the current WHO guidelines.

4.2.1.3 4 Clusters

Thus, the CIMLR algorithm was applied for 4 clusters, the number of neighbors were tuned, and considering the metrics (see Table A.2), the top 42 neighbors were considered.

The confusion matrix is present in Table 4.1c and the cluster validation measures in Table 4.1d. The former shows that cluster 1 is mostly formed by oligodendroglioma (with 173 patients overall), cluster 2 is characterized by astrocytoma (182 patients), in cluster 3 oligodendroglioma and astrocytoma are mostly present with only a patient of GBM (56 patients in total), and cluster 4 has mostly GBM patients, 3 astrocytoma patients, and 2 oligodendroglioma patients. Cluster 4 in particular stands out since GBM is known to be the more aggressive glioma type, and the grouping of these LGGs with GBM might mean the identification of LGG patients with worse prognosis.

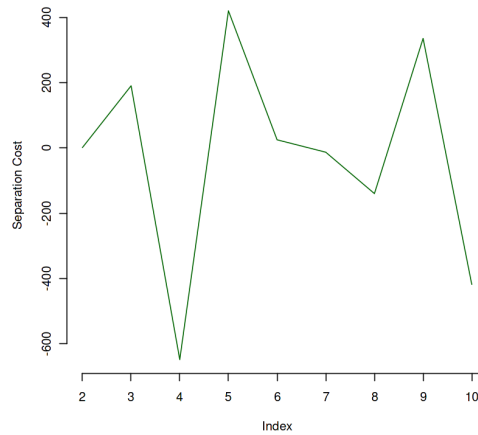


Figure 4.3: Separation cost (y-axis) for different numbers of clusters (x-axis).

The Calinski-Harabasz and the Dunn Index are lower than the 3 clusters' metrics but still have high values. The average silhouette is also lower but indicates that a strong structure was discovered [104]. Notably, cluster 3 has a near-perfect silhouette value, 0.99.

The t-SNE 2D visualization of the similarity matrix (Figure 4.4a) shows a clear separation of the CIMLR clusters. Specifically, cluster 3 (green) stands out in regards to tightness and compactness, indicating a strong cohesion among the data points, and this is supported by the silhouette value of 0.99 for this cluster Figure 4.4d, a near perfect value [104].

The similarity matrix (Figure 4.4c) shows a 4-block structure, as desired. In particular, cluster 3 in green (previously highlighted) has high similarity, and inside clusters 2 and 4 two other blocks can be identified. Yet, the connection between the blocks is apparent and intuitively justifies their grouping.

Figure 4.4b shows the similarity graph with the predicted clusters' colors. The previous analysis of cluster 3 is once again supported with a compact visualization. Clusters 2 (orange) and 4 (purple) also show a relatively close relationship between the objects. In contrast, cluster 1 is notable for two distinct high-density areas, prompting consideration as to whether further division is warranted, however, the decision remains unclear due to the notable presence of patients and edges bridging these areas.

Notably, cluster 3 (green) has a high silhouette value [104], accompanied by a prominent block in the similarity matrix and a compact 2D visualization. The fact that this cluster is predominately astrocytoma or oligodendroglioma means that molecular information similar to both glioma types is being detected, furthering our knowledge of the disease, especially regarding LGG where overlap is expected.

Furthermore, the consistency of the method's assignment was assessed by comparing the partitioning results from the 3 cluster and 4 cluster outputs. Notably, the 111 GBM

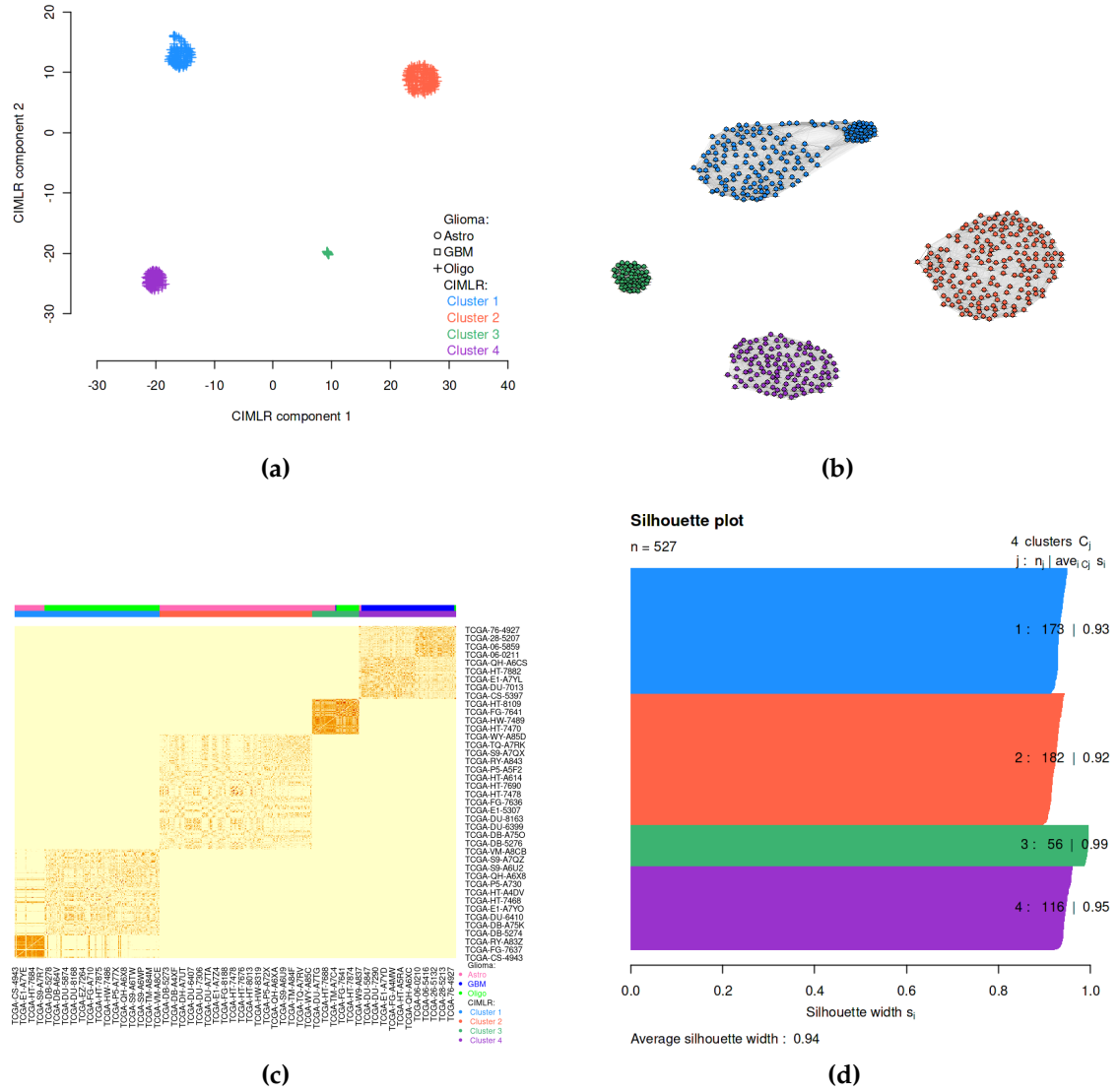


Figure 4.4: CIMLR result with 4 clusters: a) t-SNE 2D visualization of the CIMLR clusters. b) Similarity graph of the CIMLR clusters. c) Similarity matrix with official glioma types and CIMLR clusters identified. d) Silhouette plot.

patients grouped are the same in both results and the GBM patient, whose ID is TCGA-FG-7643, grouped with other glioma types is consistent. Similarly, the 182 astrocytoma patients present in cluster 3 from the 4 partitions are also present in cluster 3 from the 3 partitions. The only exceptions were two patients, TCGA-IK-7675 and TCGA-P5-A72W, who were reassigned to cluster 1, which predominantly consists of oligodendroglioma patients. Further comparison of cluster 1 across both results revealed a high degree of overlap, with most patients remaining consistent. The few reassigned cases from the 3 partition output were moved to cluster 3 in the 4 partition, where they formed a balance of astrocytoma and oligodendroglioma patients.

4.2.1.4 Relevant Features

Furthermore, the CIMLR method allows the determination of feature relevancy during the clustering method. With a significance of 5%, for the 3 clusters result, 188,583 (58.61 %) variables are significant, and for the 4 clusters, 183,887 (57.15%) features. Notably, most of the higher relevant variables' origin is from the methylomics data with the first RNA variable in the 75th (*SNAP91*) and 991st (*KIAA0495*) positions for 3 and 4 clusters, respectively. Additionally, there are 179,215 common significant features between the 3 and 4 clusters' result.

The CpG sites were also mapped to genes: for the 3 clusters, 17,553 genes are significant with 9.31% unique gene entries in the dataset, and 38,119 (20.21%) CpG sites did not have an equivalence in gene; and for the 4 clusters, this resulted in a total of 17,302 different genes, meaning 9.56% of the dataset is composed of unique genes (37,021 CpG sites did not have a gene correspondence, generating missing values in 20.45% of the entries).

The top 50 most frequently mapped genes for each result are presented in [Figure 4.5](#) in a word cloud. This is a visual representation where the frequency of a word is depicted by its size, meaning that the most frequent words are in bigger fonts, and the least frequent in smaller ones. Regarding the 3 clusters' output, it is clear that genes such as *PTPRN2*, *PRDM16*, *TNXB*, and *MAD1L1*, have high frequency, with a lot of the significant CpG sites mapped to these genes. For the 4 clusters, the top 4 remain the same, meaning these variables were detected in both results as high-frequency genes.

The effect of the *PTPRN2* gene in cancer has limited and contradicting data [105]. The study [105] concluded that *PTPRN2* is one of the seven *PTP* genes that was differentially expressed between grade 2 to 3 gliomas and grade 4 GBM, and Eun Joo Lee et al [106] identified that the gene has hypermethylation (more methylation than normal) in its promoter region which usually reduces the gene's activity, and hypomethylation (less methylation) in other parts, suggesting a tumor suppressor role [105]. However, a high *PTPRN2* expression has been correlated with poor clinical outcomes in breast cancer cases and the potential to promote the development of the tumor [105]. The authors [105] suggest that the biological impact of this gene in cancer may be context-dependent, yet it seems pertinent its identification in another study regarding gliomas.

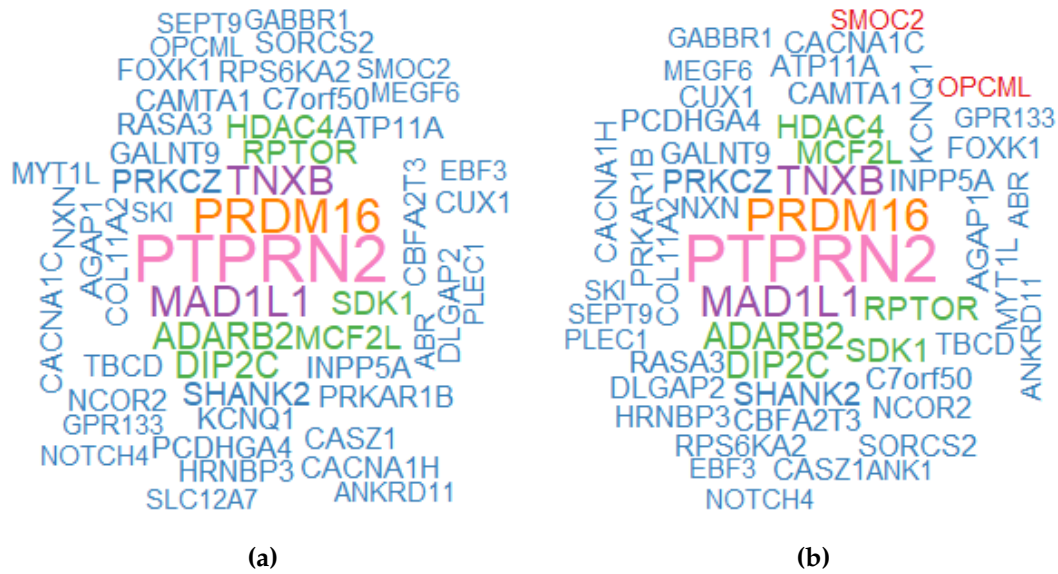


Figure 4.5: Wordcloud with the genes mapped from the significant CpG sites with: a) 3 clusters. b) 4 clusters.

Furthermore, in [106], *PRDM16* was identified as a differentially methylated gene (more common than expected) in GBM in all four established molecular functions, meaning that changes in this gene are likely to have an impact on the functions that are relevant to the biological process of gliomas. Overall, *PRDM16* has been mentioned as a gene of interest in different cancers assuming opposite functions, such as a tumor suppressor in lung cancer but a tumor promoter in brain tumors. It has been also associated with poor prognosis [107], and the CIMLR method was able to identify it as a significant feature for both clustering results.

The *TNXB* mRNA expression has been demonstrated to be downregulated in several types of cancers, yet for the glioma types, the concentration of *TNXB* was upregulated by 7.5% when compared to healthy controls [108]. The present identification of *TNXB* as a significant feature further supports its potential role in glioma biology. Overall, the identified genes through the CIMLR method validate the previous importance assigned to the mentioned biomarkers.

Notably, the gene *MAD1L1* and the WHO grades of glioma patients have been shown to have a significant positive correlation. In a 2011 study [109], *MAD1L1*, along with three other genes, was utilized to model the WHO grades of glioma. At the time, the WHO grade played a more prominent role in the diagnostic process.

Table 4.2 shows the top 10 most significant variables along with the obtained *p*-value and the gene the variables are mapped to. For the 3 clusters' top 10 significant variables (Table 4.2a), the first two variables are mapped into the same gene, *LOC100130987*, and the 8 and 9-th variables do not have a correspondence. Furthermore, *TRAF5* has been researched in the glioma context, namely, it was proven that a pathway containing *TRAF5* plays a crucial part in glioma cell migration and invasion [110]. The *CSF1R* gene has

Table 4.2: Mapping to genes the significant CpG sites obtained through CIMLR for: a) 3 clusters. b) 4 clusters.

(a)			(b)		
CpG site	<i>p</i> -value	Gene	CpG site	<i>p</i> -value	Gene
cg12472597	1.490×10^{-120}	<i>LOC100130987</i>	cg18222083	2.598×10^{-159}	<i>TMEM106A</i>
cg18476530	2.450×10^{-117}	<i>LOC100130987</i>	cg05866411	1.403×10^{-134}	<i>FGFRL1</i>
cg20912172	4.496×10^{-117}	<i>TRAF5</i>	cg17105609	5.394×10^{-134}	<i>FGFRL1</i>
cg16802592	1.501×10^{-115}	<i>LOC145845</i>	cg17701886	8.684×10^{-133}	<i>MCF2L2</i>
cg20294319	1.671×10^{-114}	<i>LRCH1</i>	cg05211768	1.405×10^{-132}	<i>FES</i>
cg05675514	7.738×10^{-114}	<i>FLJ22536</i>	cg15603424	9.219×10^{-127}	<i>ARNTL</i>
cg23505299	1.405×10^{-112}	<i>CSF1R</i>	cg00976453	1.294×10^{-126}	<i>KCNB1</i>
cg20955836	4.556×10^{-111}	<i>BMP7</i>	cg24940138	6.479×10^{-126}	<i>TMEM106A</i>
cg09790325	1.199×10^{-109}		cg10504751	4.340×10^{-124}	<i>GNAO1</i>
cg19414967	2.974×10^{-109}		cg05163329	5.193×10^{-124}	

been highlighted as a point of interest in gliomas, specifically as potentially improving treatment outcomes (by targeting *M-MDSCs*) [111].

Regarding the 4 clusters (Table 4.2b), the most significant CpG site is mapped onto *TMEM106A*, the second and third to *FGFRL1*, and the tenth site does not have a correspondence. Notably, none of the top 10 mapped genes are present in the other results of the top 10. Literature research yielded:

- *TMEM106A* has been mostly studied in other cancer types, proven to decrease cell proliferation and migration in kidney and lung cancer [112], yet it was recently highlighted in a glioma classification problem by influencing the distinction of GBM and identified as a potential novel biomarker [113].
- *FGFRL1* was recognized in the same context [113], yet this gene role in gliomas is well-documented: in GBM, the gene expression increases as a benign tumor progresses to malignant; in pediatric low-grade astrocytoma, tumor cell viability was reduced by *FGFR1* inhibition [114].
- *MCF2L2* was mentioned in [113] in a highly correlated block, and it is considered as possibly one of the most important marker genes in type 1 diabetes and polycystic ovary syndrome [115]. While its role in gliomas is not currently clear, its significance obtained in the CIMLR method might be indicative of a promising new biomarker.
- *FES* has an essential role in several cancers, being a cancer-promoting gene in breast cancer, and prostate cancer, but a tumor suppressor role in colorectal cancer [116], suggesting a context-dependent effect. In gliomas, the study [116] found a low expression in non-cancerous cells but high levels in GBM, indicating that *FES* might be contributing to the development of GBM. Yet the authors highlight that more analysis needs to be employed to further understand its role in glioma and if it can

be targeted in treatments. When testing *FES* gene expression and patient survival, significant differences in survival were obtained in patients with glioblastoma and LGG, a similar conclusion [117].

- The gene *ARNTL* is considered to promote LGG [118], a low expression of *ARNTL* is associated with high overall survival and disease-free survival [118, 119], meaning that *ARNTL* is highly correlated with glioma prognosis [119].
- *KCNB1* has been identified as a prognostic indicator in gliomas, with a low expression of *KCNB1* indicating a more aggressive glioma [120].
- Likewise, *GNAO1* has been recognized as a potential prognosis biomarker for gliomas [121, 122].

Therefore, the CIMLR method was able to identify several established key biomarkers in the glioma context, such as *TRAF5*, *CSF1R*, *KCNB1*, and *GNAO1*, validating previous knowledge. Furthermore, the genes *TMEM106A* and *MCF2L2* have recently been pointed out as a point of interest in glioma, and its obtained significance in this context further justifies its future research.

4.2.1.5 Survival Analysis

The characteristics were defined for each patient with available clinical data, and the survival curves were plotted with the CIMLR predicted partitions. Figure 4.6a and Figure 4.6b show these for the 3 and 4 clusters' output, respectively. In the figures, the events are represented with a vertical line in each curve, and for both the first event happens fairly earlier and overall frequently. Furthermore, both plots seem to have one cluster survival line visually distinctive from the others.

Regarding the 3 clusters, the survival distributions for the 3 clusters were statistically different, $\chi^2(2) = 183$, $p < 2 \times 10^{-16}$, for the log-rank test. Pairwise log-rank tests were performed to assess differences in survival distributions between clusters, and statistical significance was determined using the Benjamini-Hochberg (BH) correction, with adjusted p -values considered significant at a threshold of $p < 0.05$. The pairwise log-rank distributions revealed that clusters 1 and 2 and clusters 2 and 3 originated statistically different survival curves ($p < 2 \times 10^{-16}$). However, clusters 1 and 3 survival distributions were not statistically different, $p = 0.29$, a result expected from Figure 4.6a.

Notably, the conclusion that clusters 1 and 3 survival curves are not different could indicate that the clustering result does not have prognostic value. However, since the clusters are characterized by an LGG, it could also indicate that the data with the WHO classification does not have a distinct survival likelihood. To test this hypothesis, the Kaplan-Meier curve with the WHO glioma types was computed along with the log-rank test. Indeed, visually the GBM curve is distinct from the LGG but these do not seem to differ in Figure 4.7. The log-rank test, with $\chi^2(2) = 316$, $p < 2 \times 10^{-16}$, supported that

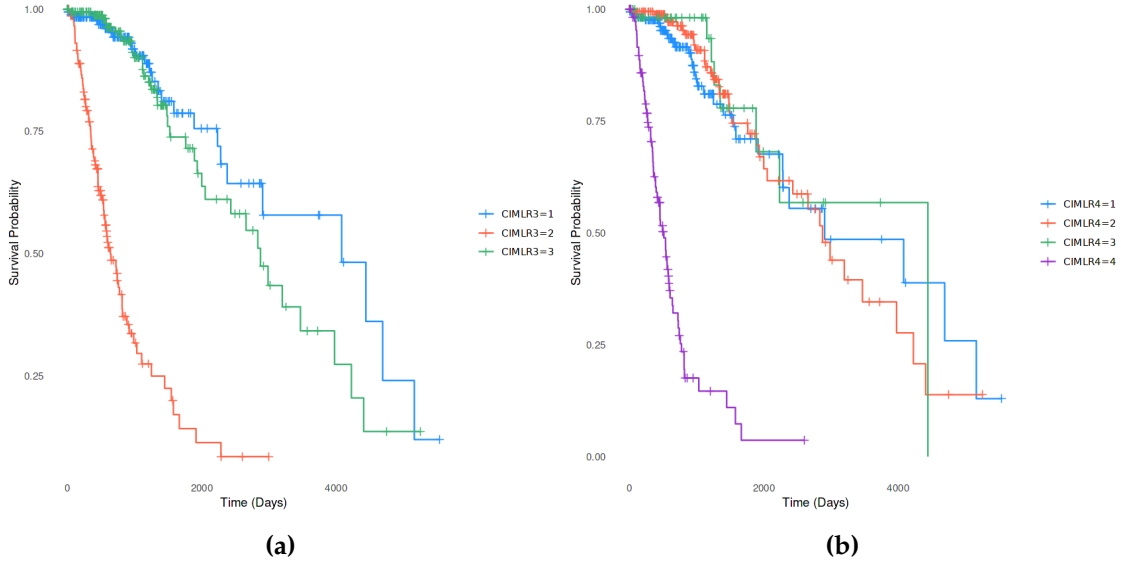


Figure 4.6: Kaplan-Meier curves for the CIMLR method with: a) 3 clusters. b) 4 clusters.

the 3 survival distributions are not equal, and the pair-wise comparisons sustained the previous hypothesis. Namely, the p -value for the pair GBM and astrocytoma, and GBM and oligodendroglioma is less than 2×10^{-16} , and for astrocytoma and oligodendroglioma is 0.1. Therefore, not obtaining different survival curves in the predicted clusters is not automatically a bad result, especially regarding the survival of the LGG which is known to be similar. Note that the clustering method does not necessarily aim to distinguish between survival outcomes, and can identify molecular characteristics that are not directly linked to survival but are still meaningful to understand glioma heterogeneity and personalize cancer treatment. Specifically, the separation of the LGG in the clustering is relevant for treatment stratification.

Cluster 2 from the 3 clusters output is constituted mostly by GBM patients and it has a significantly different survival curve and overall worse prognosis. This is coherent with previous knowledge of the GBM type, and the presence of LGG in this cluster means patients with a higher mortality rate than usual for these types might have been identified by the method. To verify this the clinical data was studied, focusing on patients who had passed away. The average of days from diagnosis until death was 437 (± 329) for GBM, 1,644 ($\pm 1,100$) for astrocytoma, and 1,578 ($\pm 1,513$) for oligodendroglioma. [Table 4.3](#) presents patient IDs for astrocytoma patients, along with their corresponding survival times (days from diagnosis to death). Of the 35 astrocytoma patients in cluster 2, 10 have passed away. Five of these deceased patients show survival times that are more consistent with the GBM average than with astrocytoma, suggesting an overlap in the aggressive nature of their disease. However, others exhibit survival times between 1,915 and 2,286 days, more typical of LGG. For the 9 oligodendroglioma, patients in this cluster, only one passed away after 919 days, which is closer to the GBM average. Taken together, these

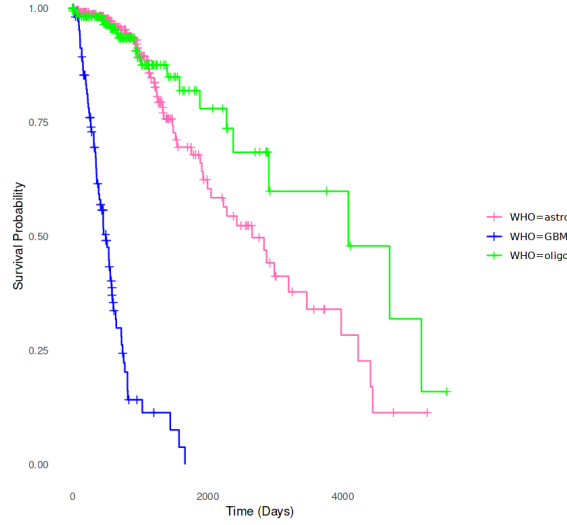


Figure 4.7: Kaplan-Meier curves with the official WHO glioma classification.

findings indicate that the CIMLR method successfully identified several LGG patients with shorter-than-expected survival times, resembling GBM characteristics.

Table 4.3: Days until death after diagnosis of the astrocytoma patients in cluster 2 from the CIMLR method with 3 clusters.

Patient ID	Days until death
TCGA-CS-4943	1,106
TCGA-DB-5277	1,547
TCGA-DU-6396	2,286
TCGA-DU-7007	1,915
TCGA-DU-7010	456
TCGA-E1-5304	1,251
TCGA-E1-A7YE	886
TCGA-E1-A7YI	111
TCGA-E1-A7YV	987
TCGA-S9-A7IS	241

For the CIMLR output with 4 clusters survival curve (Figure 4.6b), the log-rank test concluded that the four clusters survival estimation was significantly different, $\chi^2(3) = 279$, $p < 2 \times 10^{-16}$. Once again, the pairwise tests were performed, and the BH correction was employed. Clusters 1 and 4 ($p < 2 \times 10^{-16}$), 2 and 4 ($p < 2 \times 10^{-16}$), and 3 and 4 ($p < 7.5 \times 10^{-14}$) have statistically different survival curves, however clusters 1 and 2 ($p = 0.74$), 1 and 3 ($p = 0.68$), and 2 and 3 ($p = 0.68$) survival's distributions are equal.

Previously, in the 4 clusters' output, cluster 4 stood out due to its composition which is mostly GBM patients and some LGG. The survival distribution shows sharp drops and an overall low probability of survival, expected behavior for GBM. It was hypothesized that LGG patients with worse prognoses were identified. All patients in this cluster were alive

as of the last check-up, which makes the previous sentence inconclusive. Additionally, the GBM patient TCGA-FG-7643, who was placed in a different cluster than the other GBM patients in both the 3- and 4-cluster results, had a survival time of 611 days (as of the last follow-up). This timeframe, however, is inconclusive regarding the patient’s prognosis.

4.2.2 intNMF

Similar to the steps followed for CIMLR, the intNMF method was initially applied with 3 clusters to compare against the WHO classification of gliomas. Subsequently, various numbers of clusters were evaluated to determine the optimal cluster count. Specifically, 4 to 6 clusters were considered, with the results for 5 and 6 clusters presented in the [Appendix B](#).

4.2.2.1 3 Clusters

Due to the algorithm’s sensitivity to the initial matrices, 30 different initializations were considered, and the stability of the connectivity binary matrix $C_{n \times n}$ was used as a stopping criterion, which was defined as 20 and achieved at 112 computational iterations. With these parameters set, the output for the 3 clusters considering equal weights for the omics can be observed in [Figure 4.8](#). The implementation with different weights for each omic as defined in [Equation 2.7](#) was also implemented but yielded equal partitions.

By analyzing the confusion matrix in [Table 4.4a](#), it is possible to observe the following: Cluster 1 has 163 patients, mostly astrocytoma, followed by oligodendroglioma and GBM. Cluster 2 is composed of 137 astrocytoma and 117 oligodendroglioma patients (with a total of 254 patients). Cluster 3 is characterized by 93 GBM patients, with a small overlap of astrocytoma and oligodendroglioma (overall 110 patients).

Table 4.4: intNMF result with 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
intNMF	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	97	19	47	163	Adjusted Rand Index	0.235
Cluster 2	137	0	117	254	Normalized Mutual Information	0.314
Cluster 3	15	93	2	110	Calinski-Harabasz Index	8,179
					Dunn Index	0.231
					Average Silhouette Width	0.921

The consensus matrix can also be obtained and studied to determine the number of clusters and the stability of the predicted result. Analyzing [Figure 4.8a](#), the cluster pattern is clear, with three diagonal blocks. These blocks are dark and well-defined, indicating high consensus within each cluster, meaning that the patients in these clusters are consistently grouped in multiple runs, yet there are several lines and points with

varying intensities outside the main diagonal blocks. These represent patient pairs that are sometimes, but not always, clustered together, indicating some clustering instability.

In Table 4.4b, the cluster validation measures are shown, and in particular, the external validation measures have low values, below 0.5, which is expected from the study of the confusion matrix. The Calinski-Harabasz and the Dunn Index are both lower than the values obtained through the CIMLR method. The Average Silhouette Width value is high, indicating the clusters are well-separated [104]. However, in the silhouette plot Figure 4.8b, some data points have a negative silhouette value, which might suggest that a substantial structure was not found for these [104]. This could indicate the potential for refinement of the clustering results.

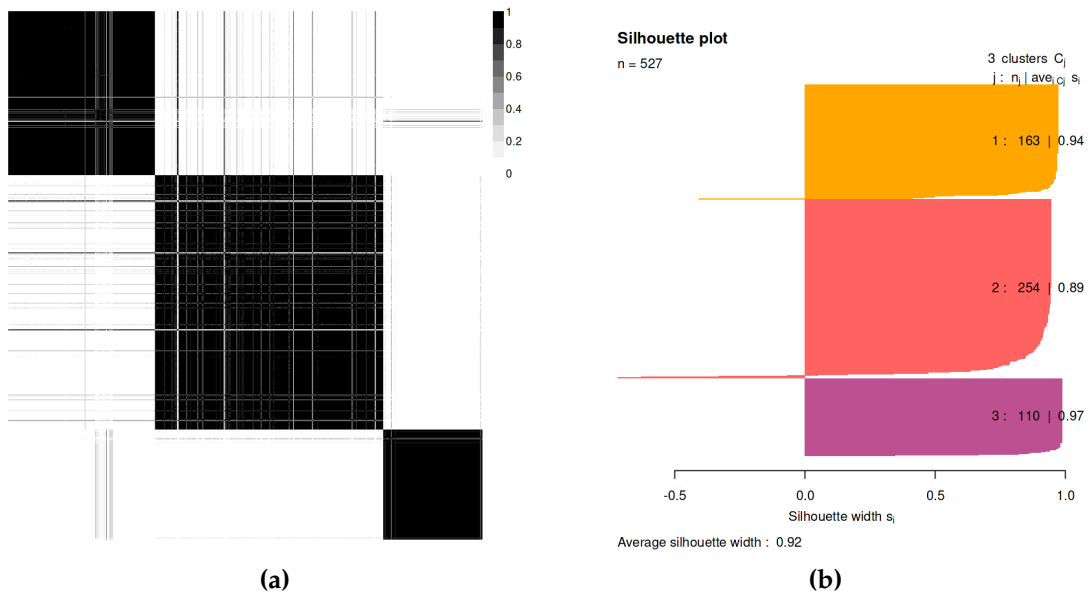


Figure 4.8: intNMF result with 3 clusters: a) Consensus matrix plot. b) Silhouette plot.

Relevant Features

Each omic's specific coefficient matrix was analyzed for sparsity. For transcriptomics data, cluster 1 exhibited 1,314 variables with zero coefficients, while cluster 2 had 4,360, and cluster 3 had 1,281. In the methylomics matrix, cluster 1 had 10,489 variables with zero coefficients, cluster 2 had 3,564, and cluster 3 had 5,747. Furthermore, studying cluster-specific variables, meaning variables that have zero coefficients in all clusters but one, is pertinent. Cluster 1 comprised 413 and 1 cluster-specific variables originating from the RNA and DNA omics, respectively. Cluster 2 exhibited 73 cluster-specific features from transcriptomics and 671 from methylomics. Cluster 3 has 227 cluster-specific variables rooting from the transcriptomics data and 136 from the DNA omic.

Subsequently, to interpret the relevance of each variable, the specific coefficient matrix for each omic were normalized by dividing the maximum value. The relative importance of each variable for cluster 1 can be found at Figure 4.9a and the normalized coefficient of the

DNA cluster-specific variable is not shown, since it only exists one CpG site (cg14149007) with value 0.069. The relative importance of each variable for cluster 2 is shown at Figure 4.9b and Figure 4.9c for the transcriptomics and methylomics, respectively. For cluster 3, Figure 4.9d and Figure 4.9e illustrate the relative importance of the transcriptomics and methylomics, respectively. Note that for the transcriptomics omic the maximum normalized coefficients range from 0.008 to 0.02, approximately, implicating a low relative importance. While in the normalized coefficients from the methylomics data, higher values are present, particularly in cluster 2 with values above 0.8, indicating the presence of variables with high relevance. Additionally, the specific coefficient matrices were combined and then normalized to compare the coefficients of the omics, but the conclusions drawn are similar.

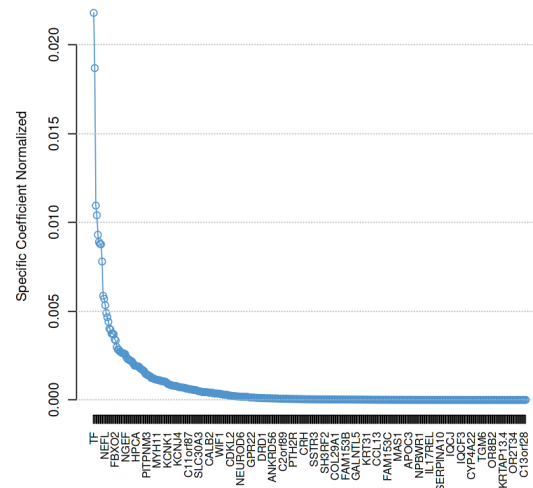
Only from cluster 3 existed cluster-specific variables with a normalized coefficient above 0.5, with a total of 244 CpG sites. Note that all 244 were significant variables for the CIMLR method with 3 and 4 clusters. In Table 4.5, the top 10 CpG sites (ordered by coefficient), their normalized coefficient, and the gene they are mapped into are shown.

Table 4.5: Mapping to genes the CpG sites specific to cluster 2 (from the 3 cluster output), with a normalized coefficient on the specific coefficient matrix bigger than 0.5.

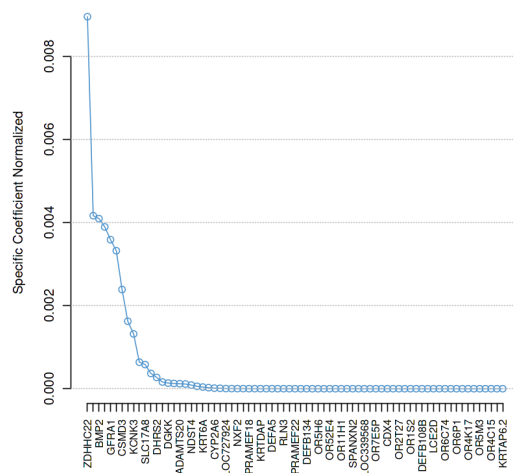
CpG site	Coefficient	Gene
cg00681103	0.885	<i>CFLAR</i>
cg17130504	0.883	<i>LZTS2</i>
cg20452714	0.876	<i>ARTN</i>
cg17759443	0.874	<i>ADAM33</i>
cg15472784	0.863	<i>KIAA0495</i>
cg07201456	0.857	<i>CHST6</i>
cg21819468	0.848	<i>PDPN</i>
cg20944305	0.832	<i>P2RY1</i>
cg05784314	0.830	<i>SLC27A3</i>
cg08971281	0.826	<i>PTRF</i>

Some considerations for the genes identified in Table 4.5:

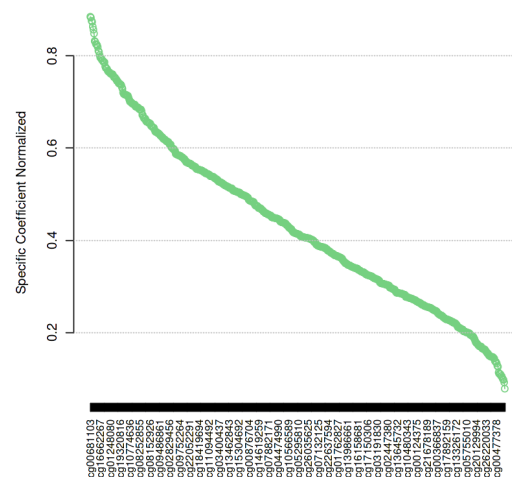
- The gene *CFLAR* has been studied in the glioma context, namely, it correlates significantly with overall survival [123], it has decreased expression in LGG compared to non-tumor [124], and it played a risk role in the survival of LGG [123].
- *LZTS2* is up-regulated in LGG but down-regulated in GBM [125].
- On the other hand, *ARTN*'s role in glioma does not seem to have been studied, yet this gene is linked to increased tumor size, quick relapse, and shorter survival in hepatocellular carcinoma (primary liver cancer), as well as promoting drug resistance [126], and the identification of the gene in this context makes it a potential biomarker for glioma.



(a)



(b)



(c)

Figure 4.9: Normalized coefficients for the intNMF method with 3 clusters (part 1): a) RNA cluster-specific variables in cluster 1. b) RNA cluster-specific variables in cluster 2. c) DNA cluster-specific variables in cluster 2.

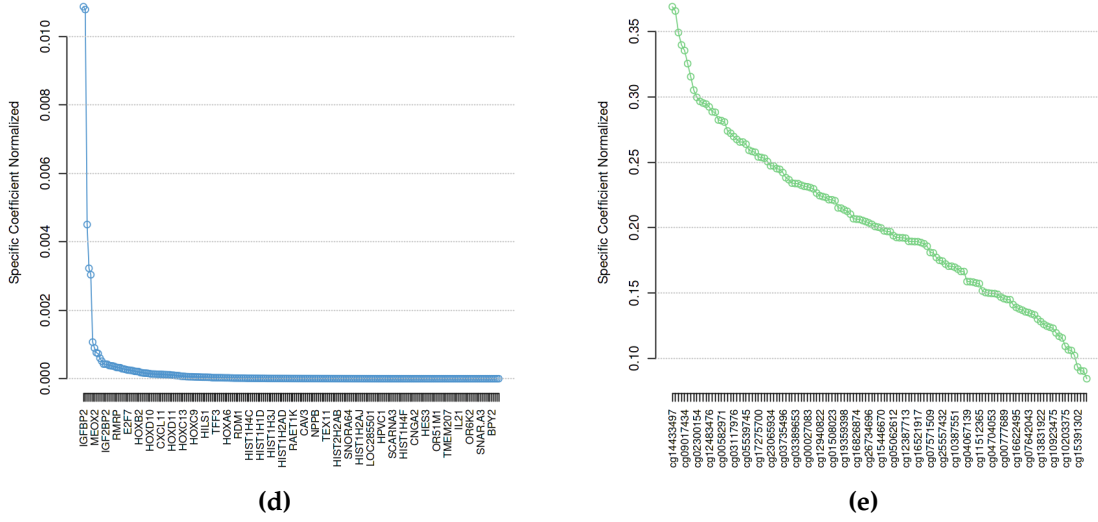


Figure 4.9: Normalized coefficients for the intNMF method with 3 clusters (part 2): d) RNA cluster-specific variables in cluster 3. **e)** DNA cluster-specific variables in cluster 3.

- Similarly, *ADAM33* is a known biomarker for different types of cancer, such as breast cancer [127] and gastric cancer [128], but its expression in gliomas seems unstudied.
- The gene *KIAA0495* was highlighted in article [113] as having the greatest influence in distinguishing GBM, and was found down-regulated in astrocytoma.
- *CHST6* is also a known molecular factor having a significant negative correlation to the prognosis and survival of glioma patients, meaning that higher expression of *CHST6* indicates a worse prognosis [129].
- A high expression of *PDPN* was thought to be correlated negatively with prognosis, however, recent findings suggest that the gene itself did not have a direct factor in response to treatment, and it is hypothesized that the previous correlation was due to other underlying pathways [130].
- *P2RY1* was found to be up-regulated in high-grade gliomas [131].
- *SLC27A3*'s high expression has been connected to worse outcomes in GBM, and its expression is higher in men than women, and in the latter, it was associated with the patient's age and Body Mass Index[132]. (cg05784314)
- In gliomas, a high *PTRF* mRNA expression is linked with increased malignancy, particularly in high-grade gliomas and IDH wildtype tumors, indicating a link to poor prognosis, which contrasts with its downregulation in other cancers [133].

Thus, several of the relevant features identified have been associated with gliomas, like *KIAA0495*, *P2RY1*, and *PTRF*. On the other hand, known biomarkers in other cancers were also highlighted and could be a point of interest in future glioma research.

4.2.2.2 4 Clusters

The method was implemented considering 4 clusters, omics with equal weights, a maximum of 1,000 iterations, 30 different initializations, and a stability count of the connectivity matrix of 15 (reached at the 190th iteration). Once again, considering different weight for the omics resulted in the same partitions. [Table 4.6a](#) shows the confusion matrix, [Figure 4.10a](#) the consensus matrix plot, [Table 4.6b](#) the metrics obtained, and [Figure 4.10b](#) the silhouette plot.

Table 4.6: intNMF method with 4 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
intNMF	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	2	102	1	105	Calinski-Harabasz Index	1,301
Cluster 2	163	0	7	170	Dunn Index	0.098
Cluster 3	4	0	120	124	Average Silhouette Width	0.734
Cluster 4	80	10	38	128		

Analysis of the confusion matrix ([Table 4.6a](#)) indicates that cluster 1 has a total of 105 patients, mostly from the GBM type, cluster 2 is characterized by 163 astrocytoma and 7 oligodendroglioma patients, cluster 3 is composed by mostly oligodendroglioma with 120 patients and some astrocytoma patients are also present, and cluster 4 is mostly astrocytoma patients, followed by oligodendroglioma and GBM. Each glioma type has a characteristic cluster, an improvement from the intNMF result with 3 clusters ([Table 4.4a](#)).

In [Table 4.4b](#), the internal validation measures can be observed, in particular, Calinski-Harabasz is 1,301 and the Dunn Index is 0.098, which is notably lower than metrics obtained for 3 clusters using the same method. However, the silhouette value is high (still lower than the 3 clusters output), with an average of 0.734 which signifies a strong structure with the obtained clusters well-separated [104].

The consensus plot ([Figure 4.10a](#)) displays 4 blocks in the diagonal, indicating the distinct clusters. However, there are a considerable number of points outside these blocks, which are not confined to any specific clusters. This suggests that while some clusters are defined (like cluster 4), there is a notable overlap and uncertainty in cluster membership for many data points. This diffuse pattern outside the main blocks indicates instability in the clustering results, with some clusters not being as well-separated as others, and it shows a more unstable result than the 3 clusters consensus matrix.

In the silhouette plot ([Figure 4.10b](#)), all clusters have data points with negative silhouette values with a minimum of -0.7404 which indicates that a point is on average closer to a different cluster than the assigned cluster [81]. This shows an aggravation of the problem detected in the 3 clusters' results. On the other side, cluster 4 has a high silhouette value of 0.914 indicating a strong structure [104]. Both of these results are compatible with the analysis of the consensus matrix plot.

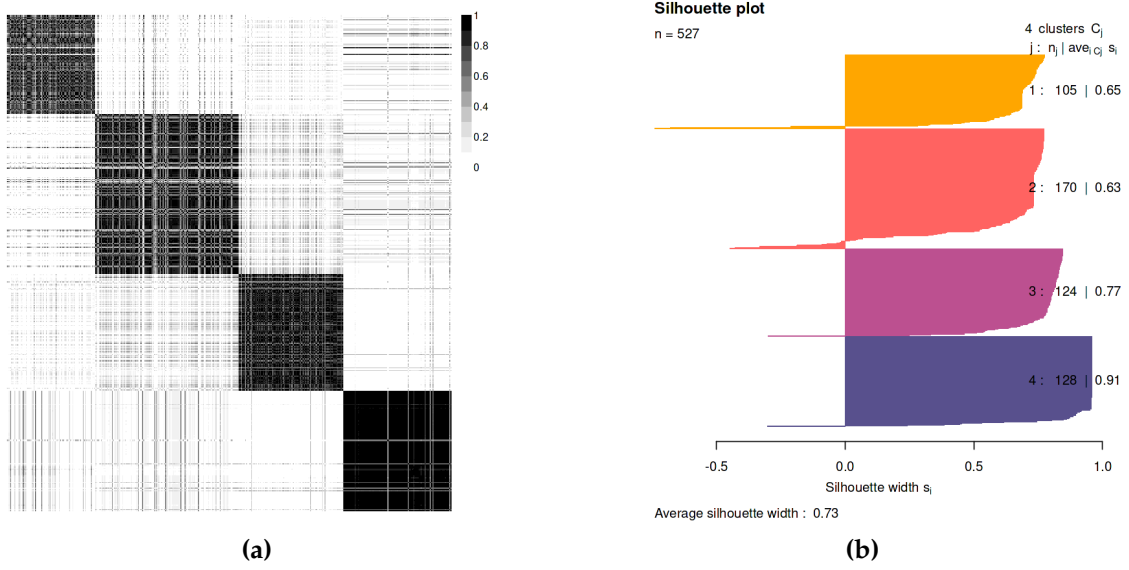


Figure 4.10: intNMF result with 4 clusters: a) Consensus matrix plot. b) Silhouette plot.

Relevant Features

Moreover, the sparsity of the obtained specific coefficient matrices is analyzed. For the transcriptomics, the specific matrix:

- **Cluster 1:** 984 variables have a coefficient of zero, out of which 205 are cluster-specific.
- **Cluster 2:** 4,827 variables have a coefficient of zero, with 32 being specific to this cluster.
- **Cluster 3:** 7,517 variables have a coefficient of zero, and 40 are cluster-specific.
- **Cluster 4:** 1,854 variables have a coefficient of zero, out of which 19 are specific to this cluster.

The specific coefficient matrix for the methylomics has the following characteristics:

- **Cluster 1:** 8,858 variables have a coefficient of zero, with 168 specific to this cluster.
- **Cluster 2:** 9,767 variables have a coefficient of zero, with 39 cluster-specific.
- **Cluster 3:** 4,275 variables have a coefficient of zero, and 246 are specific to this cluster.
- **Cluster 4:** 10,997 variables have a coefficient of zero, and no variables are specific to this cluster.

The specific coefficient matrices were normalized, and then the cluster-specific variable's coefficient was plotted. The transcriptomics variables coefficients show low values with the maximum value in clusters 1, 2, 3, and 4 approximately between 0.004 and 0.012 (Figure 4.11a, Figure 4.11c, Figure 4.11e, and Figure 4.11g, respectively). While for the

DNA omic these values in clusters 1 to 3 have higher values, especially in cluster 2 where they range to 0.8 (Figure 4.11b, Figure 4.11d, and Figure 4.11f, respectively).

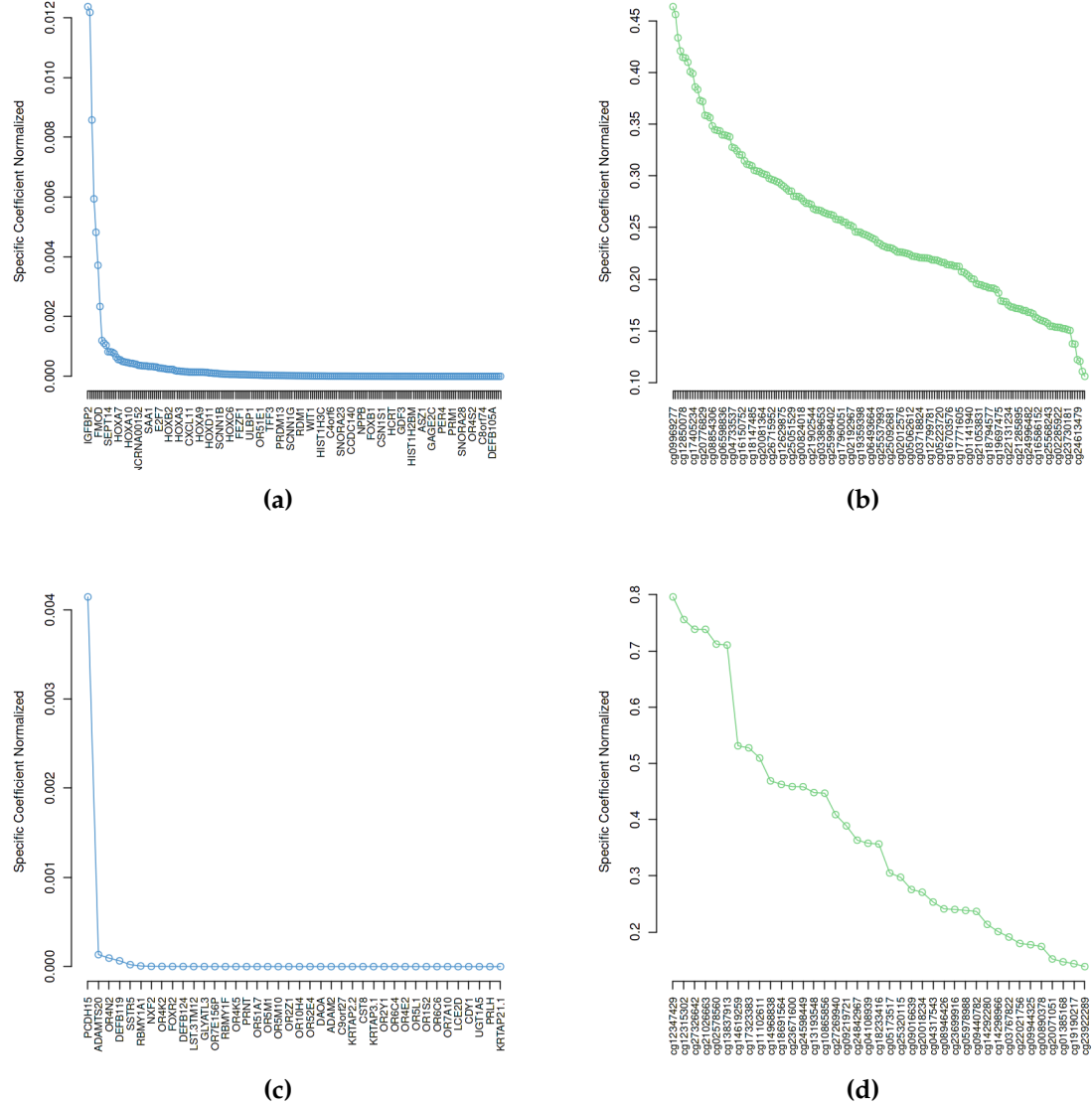


Figure 4.11: Normalized coefficients for the intNMF method with 4 clusters (part 1): a) RNA cluster-specific variables in cluster 1. b) DNA cluster-specific variables in cluster 1. c) RNA cluster-specific variables in cluster 2. d) DNA cluster-specific variables in cluster 2.

The CpG sites with a coefficient bigger than 0.5 are present in Table 4.7, along with the gene they map into. First, note that the cluster-specific variables are all from cluster 2 and only 9 sites satisfy the condition. These 9 CpG sites were also obtained as significant variables for the CIMLR method with 3 and 4 clusters.

Regarding the mapping, 4 sites are mapped into the same gene, *HIST1H4D*, and 2 sites are mapped onto *MCF2L2*. The former gene's, *HIST1H4D*, role in gliomas seems to not have been studied, yet *HIST1H4D* was found upregulated in gynecological cancers and

Table 4.7: Mapping to genes the CpG sites specific to cluster 2 (from the 4 cluster output), with a normalized coefficient on the specific coefficient matrix bigger than 0.5.

CpG site	Coefficient	Gene
cg12347429	0.796	<i>FLJ22536</i>
cg12315302	0.756	<i>HIST1H4D</i>
cg27326642	0.738	<i>HIST1H4D</i>
cg21026663	0.738	<i>HIST1H4D</i>
cg02578560	0.712	<i>TTBK1</i>
cg13837913	0.710	<i>HIST1H4D</i>
cg14619259	0.531	<i>SPRY4</i>
cg17323383	0.528	<i>MCF2L2</i>
cg11102611	0.509	<i>MCF2L2</i>

associated with poorer relapse-free survival and metastasis-free survival in breast cancer [134]. The latter gene is also present in the top 10 mapped genes in the CIMLR method with 4 clusters, and its relevancy in literature was addressed.

The first gene *FLJ22536*, alternatively known as *CASC15*, in high levels has been shown to be significantly associated with poor survival outcomes in glioma patients, as well as linked to more aggressive clinicopathological features, suggesting a potential role for *CASC15* in glioma progression and prognosis [135]. On the other hand, in the consulted literature, the gene *TTBK1* does not seem to have a strong association with cancer [136], but it plays an important part in neurodegenerative diseases, such as Alzheimer's and Parkinson's disease [137]. Recently, a study in primary and recurrent GBM identified the *TTBK1* gene as downregulated in primary GBM compared to normal brain tissues, suggesting a potential role for *TTBK1* in GBM, yet it is highlighted that further studies are necessary to determine if the low expression of the gene is related to primary GBM tumor growth [138]. Finally, the gene *SPRY4* is downregulated in glioma patients, and these have shorter overall and disease-free survival, highlighting the gene's tumor suppressor role [139].

4.2.2.3 Survival Analysis

To evaluate the prognostic ability of the molecular characteristic of the obtained clusters, the Kaplan-Meier curves for each partition were computed (Figure 4.12), and the log-rank test was employed to determine with a confidence of 95% if the survival curves between clusters are equal.

In Figure 4.12a, the 3 clusters' survival distribution is shown, and the cluster 3 curve stands out with a lower probability. The log-rank test resulted in $\chi^2(2) = 239$, $p < 2 \times 10^{-16}$, and at a significance level of 5%, it is possible to conclude that the 3 survival curves are statistically different. The pair-wise comparison wielded that all the clusters' survival distributions are statistically distinct ($p_{12} = 0.016$, $p_{13}, p_{23} < 2 \times 10^{-16}$), meaning that the identified clusters have a predictive quality. In light of the cluster compositions

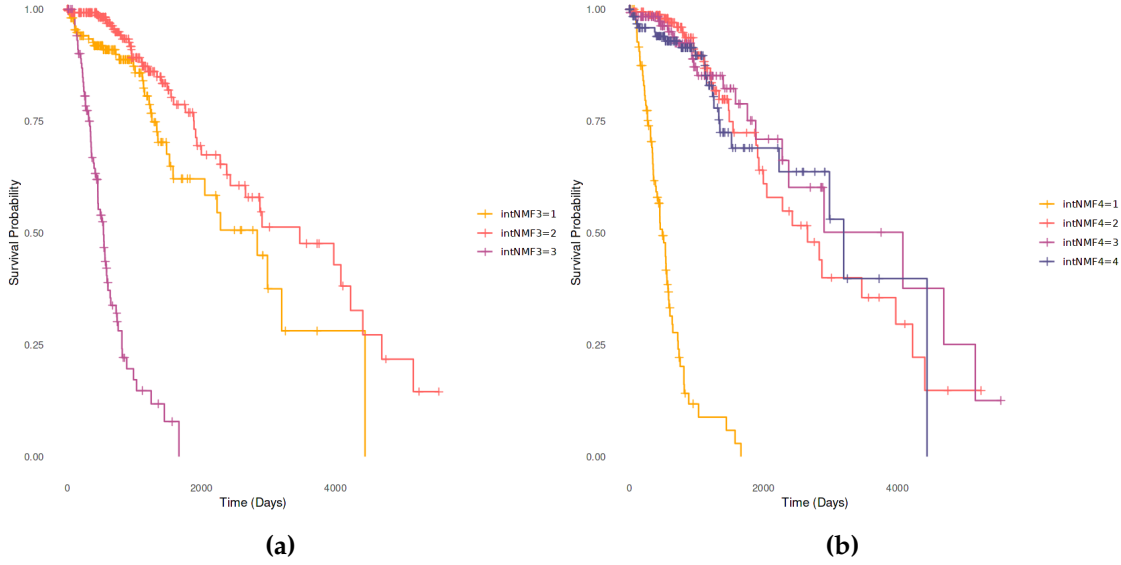


Figure 4.12: Kaplan-Meier curves for the intNMF method with: a) 3 clusters. b) 4 clusters.

(Table 4.4a), cluster 3, which predominantly consists of GBM patients, shows a worse survival likelihood, as expected.

Of 15 astrocytoma patients who were grouped with GBM in cluster 3, 6 have passed away and 5 of these have days until death closer to the GBM average. Of the 2 oligodendroglioma patients in the same cluster, one passed away 547 days after diagnosis, once again closer to GBM, and the other is alive and 1,354 days have passed since the last check-up.

Furthermore, of the 19 GBM patients in cluster 1 (not grouped with the other GBM patients), 12 have passed away within a few days after diagnosis (29 to 775) and the other patient died 1,578 days after diagnosis which is closer to the LGG average.

Interestingly, for the 3 clusters, while there is not a cluster characterized by either astrocytoma or oligodendroglioma, clusters 1 and 2 present a mix of both. Despite this, the survival rates across the clusters are not equal. This indicates that although the method does not perfectly align with the WHO glioma classifications, it is capable of capturing molecular differences that correlate with varying survival probabilities, which is a meaningful result that could suggest a reorganization of the WHO classification. This result contrasts with CIMLR’s 3-cluster output, where each cluster corresponded to a specific glioma type, but the survival rates within the LGG group were similar. Thus, we have identified two distinct outcomes: one aligns with the 2021 WHO CNS classification, validating it with molecular features; the other diverges from this classification, revealing three new groups with distinct survival distributions, but still supports the current separation of GBM and LGG.

For the 4 clusters’ output, the survival distributions are present in Figure 4.12b, and, once again, a cluster survival likelihood is distinct, cluster 1. The log-rank test

with the BH correction obtained $\chi^2(3) = 309$, $p < 2 \times 10^{-16}$, which means that the four survival curves are statistically different. The pair-wise comparisons reveal that cluster 1 survival distribution is different from all others, with $p < 2 \times 10^{-16}$, yet all others differ: $p_{23} = p_{34} = 0.67$, $p_{24} = 0.87$. Once more, the method is capable of distinguishing between GBM and LGG since cluster 1 has the most GBM patients, and potentially identifies three cases of LGG with a worse survival rate. This result is pertinent since the 4 clusters confusion matrix (Table 4.6a), where clusters 1 to 3 were comprised of a glioma type and cluster 4 was a mixture of all types. This showed improved performance compared to the 3 clusters, which lacked specificity for glioma types. However, the 4 clusters' result still has limitations. Notably, cluster 4's survival distribution does not differ from the other clusters, and the survival curves for clusters 2 (astrocytoma) and 3 (oligodendroglioma) are equal. Although the latter similarity might be explained by the WHO classification survival curve, it does not account for the lack of distinction in cluster 4.

In cluster 1 from the 4 clusters' output, 2 patients with astrocytoma were grouped with most of the GBM patients. They have both passed away, 111 or 886 days after diagnosis, values closer to the GBM average. One oligodendroglioma patient was also assigned to this cluster, and they passed away 547 days after diagnosis. Thus, the clinical data corroborates the hypothesis that LGG patients with lower survival rates were clustered together with the majority of the GBM patients.

In cluster 4, there are present 10 patients diagnosed with GBM, so it was theorized that this grouped had higher survival rates. However, 6 of these have passed away very closely to the diagnosis date (29 to 775).

4.2.3 moCluster

The first step in the moCluster method was to apply consensus principal component analysis, without inducing sparsity and find the number of latent variables to keep. It is considered important not to add sparsity immediately to evaluate the relative contribution of variance from each omic and to have an accurate permutation test evaluating the latent variables to keep, which represent the concordant structures in different data [103].

First, 15 CPCs were considered, and, in Figure 4.13a, it is possible to observe the relative contribution of variance from each data set: in the first component, the variance seems similar for both omics, but, in the following, the transcriptomics contributes relatively more than the methylomics. The explained variance is present in Figure 4.13b, the first consensus component explains 0.274 and the second 0.122.

Next, as described in section 2.2, a permutation step was used to assess the significance of each principal component, and the number of JLVs to keep can be determined through the "elbow" point in a scree plot. In Figure 4.14, the scree plot with permutation step is shown, with 20 permutations performed and the confidence interval calculated as permutation average plus or minus 1.96 times the standard deviation. In the plot, the elbow spot is present in the second component, therefore, 2 latent variables (with a

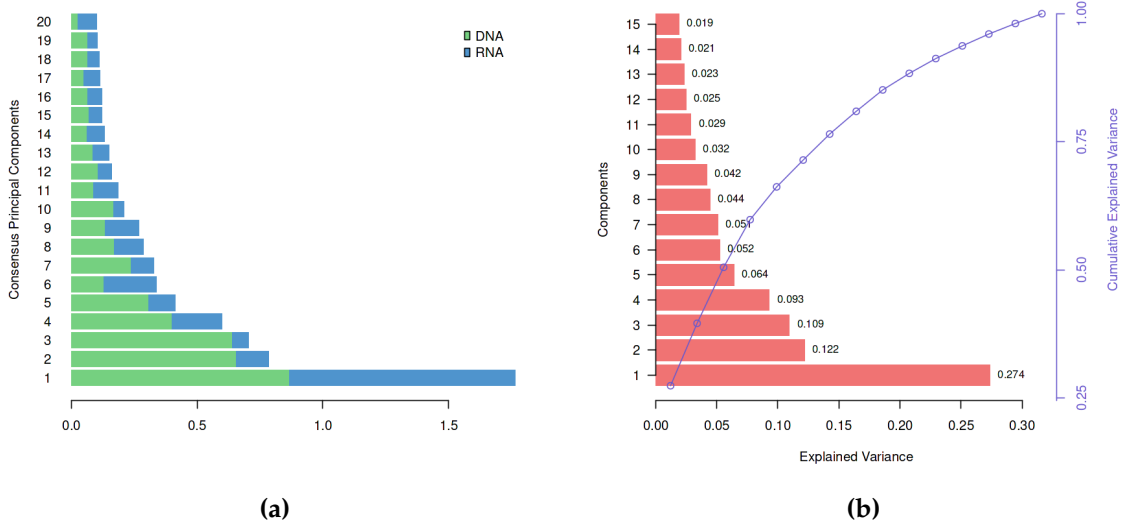


Figure 4.13: Consensus principal components: a) Relative contribution of variance from each data set. b) Explained and cumulative variance per principal component.

cumulative variance explained of 0.395) should be used in the analysis. Yet a second elbow can also be observed in 6 (0.714 cumulative variance) so this analysis is also contemplated. Note also that the first component is the only one above the confidence interval, meaning that we only obtained one concordant structure.

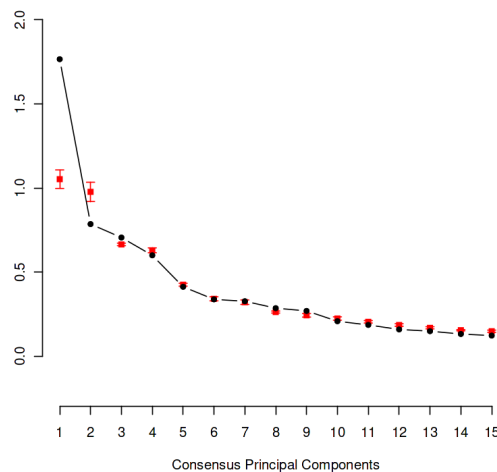


Figure 4.14: Elbow plot and permutation step of the consensus principal components. Horizontal axis represent the 15 PCs, the vertical axis the eigenvalues of the principal components, and the error bars exhibiting the 95% confidence intervals for the resampled eigenvalues.

4.2.3.1 2 Joint Latent Variables

After obtaining the 2 latent variables, K-means clustering was performed with 3 clusters and 200 random starts (analogous to the settings in the CIMLR function [40]).

Table 4.8a shows the confusion matrix between the glioma types and the obtained clusters, where cluster 1 is characterized by 209 astrocytoma and 136 oligodendroglioma patients with only one GBM patient, cluster 2 is balanced between astrocytoma and oligodendroglioma types, with 28 of each and 3 GBM patients, and cluster 3 is defined by mostly GBM with 108 patients, 12 astrocytoma patients and 2 oligodendroglioma. The method effectively distinguishes between GBM and LGG, though it encounters challenges in differentiating between LGG subtypes. However, this may not necessarily be a limitation, as the method could be identifying molecular characteristics that support this grouping, which may differ from the WHO classification.

Table 4.8: moCluster method with 2 JLV and 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	209	1	136	346	Adjusted Rand Index	0.356
Cluster 2	28	3	28	59	Normalized Mutual Information	0.387
Cluster 3	12	108	2	122	Calinski-Harabasz Index	680.893
					Dunn Index	0.021
					Average Silhouette Width	0.55

Table 4.8b displays the cluster validation measures. Notably, the internal values are below 0.5, an unfavorable result. Both the Calinski-Harabasz Index and the Dunn Index have the lowest value obtained until this point. The Average Silhouette Width is 0.55 which indicates that a reasonable structure was found [104]. The silhouette plot (Figure 4.15) shows that cluster 1 and 3 have similar silhouette values with reasonable structures, but cluster 2 has a value below 0.5, meaning a weak structure was found and could be artificial [104].

Moreover, for the K-means algorithm, the gap statistic was calculated to determine the optimum number of clusters, ranging from 1 to 20, and is present in Figure 4.16. Since the heuristic favors high values (until stabilization), the ideal number of clusters is clearly 2 followed by 3.

The sparsity was introduced, and the consensus principal components were recalculated. The sparsity parameters range from 0.05 to 1 with a step of 0.05, and note that, for example, setting sparsity as 0.1 means that only 10% of the variables with non-zero coefficients were considered for clustering.

The correlation of the original and sparse latent variables was compared, and if a very high correlation was obtained, the sparse and non-sparse components should account for the same variance or biological effects [103]. In Figure 4.17, heatmaps are present with

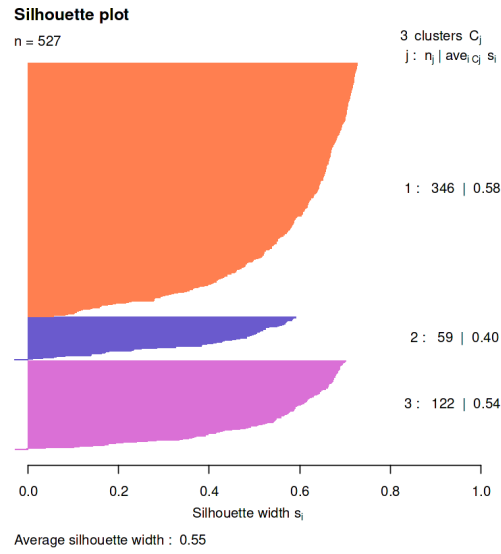


Figure 4.15: moCluster method with 2 JLV and 3 clusters: Silhouette plot.

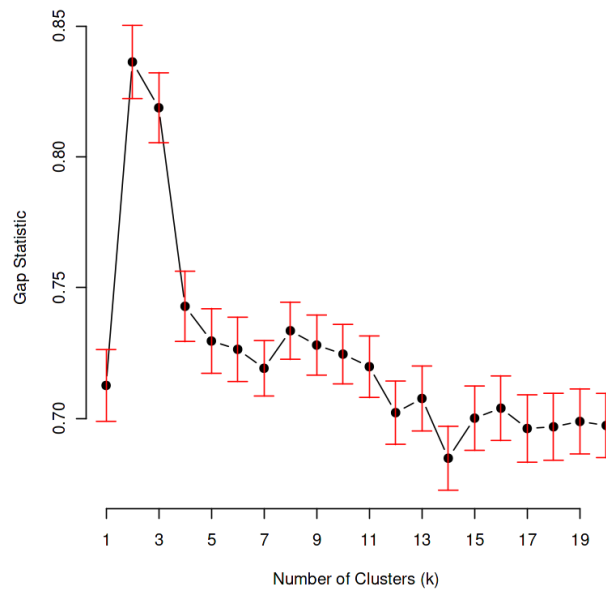


Figure 4.16: moCluster method with 2 JLV and 3 clusters: Gap statistic with x -axis as the range for the number of clusters, the y -axis the gap statistic values, and the error bars around each point show the standard error of the gap statistic, which indicate the variability or precision of the gap statistic estimate based on the simulation with 100 bootstraps.

the correlation of the first non-sparse component with the first sparse component in the first column and the second component's correlation in the second column. Consider a very high correlation ranging from 0.90 to 1 and a high correlation from 0.70 to 0.89 [140]. Except for sparsity 1, all correlations have a very high absolute value ranging from 0.901 to 0.999, indicating that it is reasonable to use these sparse components.



Figure 4.17: Correlation heatmap between non-sparse and sparse 2 consensus principal components.

Then, the gap statistic was calculated for the relevant highly correlated sparse components:

- **0.05 to 0.30:** the highest value was obtained with $k = 1$ (Figure C.1), meaning the data should not be clustered.
- **0.35 to 0.65:** yields a clear maximum for 3 clusters (Figure C.2).
- **0.70 to 0.95:** definitive peak at 2 clusters (Figure C.3).

The K-means algorithm for 3 clusters was also applied for the relevant parameters and, in Figure 4.18, a comparison of the metrics of non-sparse and sparse metrics for 3-means is present. Notably, for the Adjusted Rand Index, Normalized Mutual Information, and Average Silhouette Width, all sparse components perform better. However, for the Dunn Index, only a sparsity of 0.45 performs better than the non-sparse 3-means result. For the Calinski-Harabasz Index, a sparsity level between 0.50 and 0.65 provides higher values than the non-sparse 3-means result. Since the Dunn Index values are generally low and sparsity levels between 0.35 and 0.45 worsen the Calinski-Harabasz Index, a sparsity of 0.50 is preferable. This level of sparsity maximizes the Adjusted Rand Index, Normalized Mutual Information, and Average Silhouette Width while balancing the overall cluster validation metrics. Note that the gap statistic obtained for the moCluster output considering 2 sparse joint latent variables validates the current glioma classification, indicating no need to subtype further the glioma types. This might be derived from only

considering 2 latent variables, which is typically considered a low number. Hence the relevancy of considering 6 in [subsubsection 4.2.3.2](#).

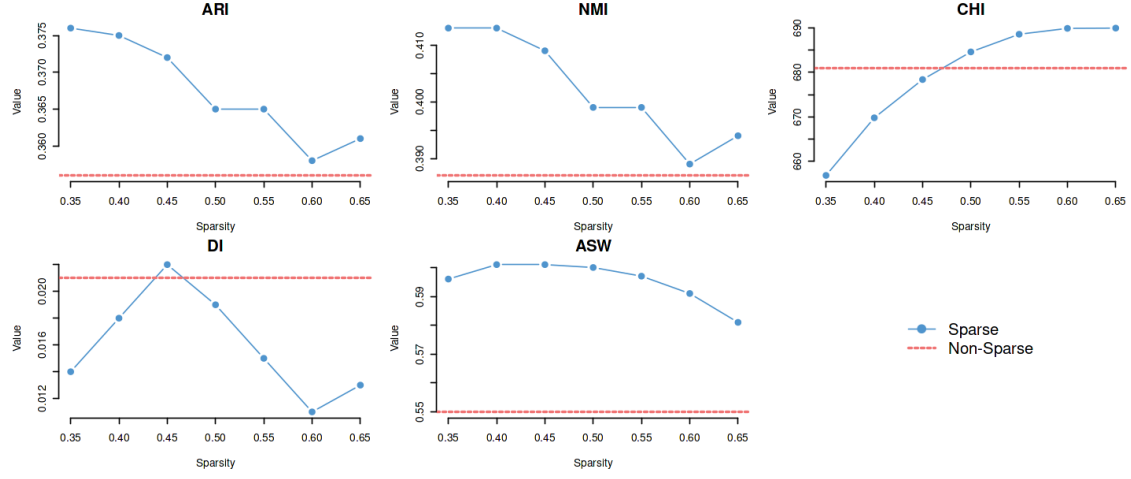


Figure 4.18: Metrics for 3-means with non-sparse and sparse consensus principal components.

Therefore, the result of the 3-means clustering with 0.50 sparsity is presented. The confusion matrix ([Table 4.9a](#)) shows that cluster 1 is mostly constituted by astrocytoma (212) and oligodendroglioma (141) patients with one GBM patient, cluster 2 has a similar structure with 27 astrocytoma and 23 oligodendroglioma patients and two GBM patients, and cluster 3 is characterized by 109 GBM, 10 astrocytoma, and 2 oligodendroglioma samples.

Table 4.9: moCluster method with 2 sparse JLV and 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	212	1	141	354	Adjusted Rand Index	0.365
Cluster 2	27	2	23	52	Normalized Mutual Information	0.399
Cluster 3	10	109	2	121	Calinski-Harabasz Index	684.603
					Dunn Index	0.019
					Average Silhouette Width	0.600

The cluster validation measures are in [Table 4.9b](#). The internal validation and the Dunn Indexes are low, with the former being lower than 0.5. The Calinski-Harabasz Index is 684, one of the lowest values obtained. The average silhouette value is 0.6 meaning that a reasonable structure was found. The silhouette plot in [Figure 4.19](#) shows that cluster 1 has the highest Average Silhouette Width, followed by cluster 3, and cluster 2. Notably, only in cluster 3 are some data points with negative values (the minimum being -0.034, which is still far from -1), indicating some objects are in a boundary area between two clusters.

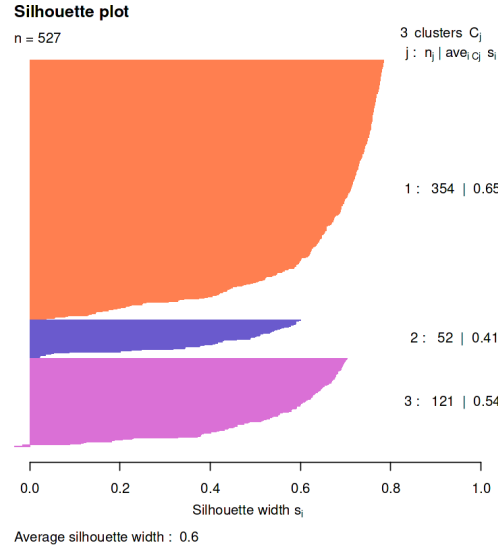


Figure 4.19: moCluster method with 2 sparse JLV and 3 clusters: Silhouette plot.

Relevant Features

The principal component scores represent the projection of the samples onto the latent variables axes, with the size of the number of objects by the number of JLVs. The minimum and maximum values of these scores are -0.043 and 0.166 , respectively, and the first latent variable has an average of 8.989×10^{-18} (variance of 8.456×10^{-4}), the second a mean of 1.230×10^{-19} (variance of 6.601×10^{-4}), overall low values.

Furthermore, since sparsity was induced in [Figure 4.20](#) the number of variables and their signal coefficients in the sparse JLVs for each omic is shown. In particular, 150,770 and 10,100 variables have non-null coefficients for the methylomics and transcriptomics, respectively. The transcriptomics coefficients' have a similar distribution of positive and negative for the second PC, and mostly a positive contribution for the first. The methylomics has more negative coefficients for both latent variables, indicating a negative contribution.

Then, principal component-specific variables, meaning features with a zero coefficient for one of the components and non-zero for the other, were found. Component one has 57,316 specific variables, of which 52,395 and 50,341 are common to the significant variables from the CIMLR method with 3 and 4 clusters, respectively. Compared to the intNMF variables with a normalized coefficient above 0.5, 33 are common to the intNMF with 3 clusters. Component 2 also has 57,316 specific variables with 23,448 and 23,448 common to the significant variables from the CIMLR method with 3 and 4 clusters, respectively.

The absolute value was ordered and the top 10 variables for principal components 1 and 2 are present in [Table 4.10a](#) and [Table 4.10b](#), respectively. Note that all shown features are from the transcriptomics data, and the first variable originating from the methylomics only shows up in the 2,738 and 952-th positions for PCs one and two, respectively.

For the first JLV's identified variables, note that:

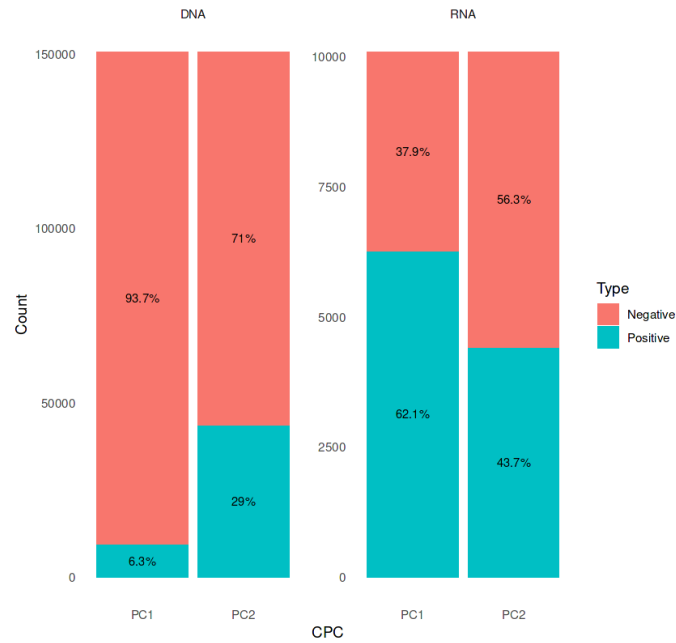


Figure 4.20: Stacked bar chart of RNA and DNA counts and type by principal component with sparsity set as 0.50.

Table 4.10: Top 10 of the 2 principal component-specific variables ordered by absolute coefficient score: a) First sparse JLV. b) Second sparse JLV.

(a)		(b)	
Variable	Coefficient	Variable	Coefficient
<i>MSN</i>	0.0144	<i>SAMD12</i>	0.02014
<i>DCTD</i>	0.0140	<i>DGKE</i>	0.02014
<i>SLC43A3</i>	0.0140	<i>KIAA1045</i>	0.02007
<i>TAGLN2</i>	0.0139	<i>GLT1D1</i>	0.02006
<i>KIAA0495</i>	0.0137	<i>NCRNA00087</i>	0.02005
<i>EMP3</i>	0.0136	<i>AK5</i>	0.02005
<i>CLIC1</i>	0.0136	<i>TMEM130</i>	0.01998
<i>CD58</i>	0.0135	<i>MAP7D2</i>	0.01997
<i>PSMC2</i>	0.0134	<i>CHN1</i>	0.01997
<i>PRICKLE3</i>	0.0133	<i>PAK1</i>	0.01997

- A recent study demonstrated improved targeting of glioma cells by using *MSN* (or *MSN*) to load the chemotherapeutic drugs, reducing the overall drug dosage needed, minimizing toxic potency, and reducing chemotherapy resistance [141]. While this study represents a novel direction for treatment, the authors highlight that larger samples and deeper mechanisms should be analyzed, yet the potential given to *MSN* in glioma treatment is clear.
- The gene *DCTD* was introduced in a study by Huimin Hu et al [142] as having a potential role in the invasion capacity of glioma cells and as a therapeutic target, since a high level of *DCTD* (mRNA) showed shorter survival of patients in different databases. This does seem to be the only study directly linking the gene to gliomas.
- Elevated levels of *SLC43A3* have been found to have a strong association with higher risk levels in GBM, and the gene could serve as a potential target for GBM treatment [143, 144].
- *TAGLN2* was found to promote cell invasion and proliferation in gliomas, as well as highly expressed in GBMs compared to LGGs thus linked to poor prognosis [145].
- The gene *KIAA0495* has been previously identified in the intNMF 3 clusters result after being mapped from the CpG site cg15472784, therefore its literature review has been contemplated. Additionally, it was the first significant variable from transcriptomics in the CIMLR method with 4 clusters.
- *EMP3* expression in brain tumors, particularly in GBM, has been well established as upregulated, and the gene is considered a predictor for clinical progression and outcomes in gliomas [146–148].
- In gliomas, elevated levels of *CLIC1* have been linked with worse survival outcomes and the suppression of *CLIC1* has been found to reduce proliferation, making it an important prognostic factor [149].
- Zhi Tian et al [150] studied the connection between *CD58* and gliomas from three views, bioinformatics databases, clinical samples, and cell models, and consistently concluded that higher levels of *CD58* meant more aggressive glioma subtypes and poorer prognosis. While a low level correlated with decreased cell proliferation, increased apoptosis, and diminished glioma cell metastasis.
- *PSMC2* was found significantly upregulated in gliomas, divided into GBM and LGG, and its expression elevated with the increase of tumor malignancy. Likewise, patients with high *PSMC2* expression had a substantially poorer survival rate than those with low expression. Thus, *PSMC2* was identified as a potential diagnostic marker and a risk factor for gliomas [151, 152].

- *PRICKLE3*'s role in gliomas seems mostly unstudied yet Wen-Jing Zeng et al [153] detected a significant link with overall patient survival in LGG and the gene expression, however, it was not included in the final model to predict prognosis. Radaszkiewicz et al [154] analyzed the expression of *PRICKLE3* on several cancers: most mutations to the gene happened in ovarian and esophageal cancers, and mutations had a significant negative prognostic impact within the TCGA samples from the various cancers.

For the second JLV's variables, we have

- *SAMD12* has been identified in a breast cancer study [155] as the most target gene in a regulatory network highlighting its importance in molecular mechanisms. However, this gene does not seem investigated overall.
- In non-small cell lung cancer, *DGKE* expression is linked to sensitivity to erlotinib (medication) and radiotherapy and a significant association with responses to both systemic therapies and radiotherapy [156].
- *KIAA1045* was recently identified as a hub for astrocytoma patients (according to the 2021 WHO CNS classification) [157].
- The expression of *GLT1D1* is hypermethylated in gliomas. Furthermore, the gene has a negative correlation with immune-stromal cells and immune checkpoints and might have a strong antagonistic effect on the antitumor drugs, thus *GLT1D1* is closely associated with anti-tumor and pro-tumor immunity [158].
- mRNA *NCRNA00087* was found significantly underexpressed in prostate cancer [159].
- *AK5* plays a dual role in cancer, promoting survival in prostate cancer but serving as an unfavorable marker in gastric and colorectal cancers [160].
- *TMEM130* expression in GBM is reported as down-regulated [161], yet no additional studies connecting gliomas and *TMEM130* are reported, since this gene could be associated with GBM.
- *MAP7D2* is considered a biomarker for various cancers, namely, a risk gene in carcinoma and a prognostic biomarker of gastric cancer [162]. Additionally, it has an increased expression in non-small cell lung cancer tissues [163].
- Statistically significant changes in the gene *CHN1* have been detected for GBM [164].
- In gliomas, *PAK1* is abnormally expressed and glioma cell invasion, migration and proliferation are all inhibited by the gene [165].

Thus, several genes' roles have been established in glioma literature, such as *MSN*, *SLC43A3*, and *TAGLN2*. Additionally, emerging evidence highlights the potential importance of genes such as *GLT1D1*, which has been linked to immune response and drug resistance in gliomas, *PAK1*, known for inhibiting glioma cell invasion and proliferation, and *PSMC2*, identified as a diagnostic marker associated with poor prognosis. Moreover, while genes like *SAMD12* and *TMEM130* show promise in cancer research, their roles in gliomas remain largely unexplored.

4.2.3.2 6 Joint Latent Variables

Now 6 JLV were considered and K-means clustering was executed with 3 clusters and 200 random starts. Table 4.11a shows the confusion matrix between the glioma types and the obtained clusters, Table 4.11b the cluster validation measures and Figure 4.21 the silhouette value for each predicted cluster.

Table 4.11: moCluster method with 6 JLV and 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	209	1	136	346	Adjusted Rand Index	0.362
Cluster 2	11	110	2	123	Normalized Mutual Information	0.402
Cluster 3	29	1	28	58	Calinski-Harabasz Index	181.718
					Dunn Index	0.046
					Average Silhouette Width	0.29

In the confusion matrix, cluster 1 is mostly constituted by astrocytoma patients (209) followed closely by 136 oligodendroglioma patients and 1 GBM patient, cluster 2 contains nearly all the GBM patients (110), some astrocytoma patients (11), and two oligodendroglioma patients, and cluster 3 is balanced with astrocytoma and oligodendroglioma patients, having 29 and 28 respectively, and one GBM patient. Once again, the confusion matrix indicates that the method separated GBM and LGG, but did not divide the latter into astrocytoma and oligodendroglioma.

It is important to highlight that cluster 1 from the 6 JLV partitions has the same glioma type distribution as cluster 1 from the 2 JLV, but 5 patients differ in the clusters. Their ID, glioma type, and cluster allocations for both partitions are presented in Table 4.12, and note that they are not consistently clustered together.

In Table 4.11b, the internal validation measures are below 0.5. The Calinski-Harabasz Index and the Dunn Index are also relatively low compared to other results, but the Dunn Index is higher than the K-means with 2 latent variables (Table 4.8b). The Average Silhouette Width has the lowest value obtained so far (0.29) indicating non-cohesive clusters with considerable overlap and a weak structure that could be artificial [104]. The silhouette plot (Figure 4.22) shows that all clusters have a similar silhouette value.

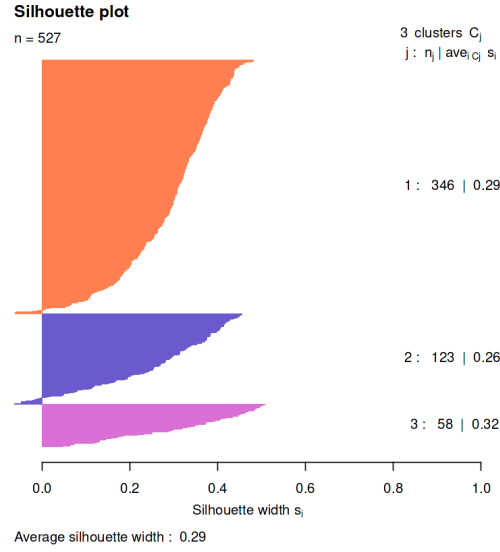


Figure 4.21: moCluster method with 6 JLV and 3 clusters: Silhouette plot.

Table 4.12: Comparison of cluster 1 for 2 and 6 latent variables partitions.

Patient	Glioma type	Cluster in 2 CPC	Cluster in 6 CPC
TCGA-DU-7010	Astrocytoma	1	2
TCGA-DU-A7TC	Astrocytoma	1	3
TCGA-TM-A84I	Astrocytoma	1	2
TCGA-CS-6186	GBM	1	2
TCGA-HT-A4DV	oligodendroglioma	1	3
TCGA-S9-A6WG	Astrocytoma	3	1
TCGA-S9-A7R7	Astrocytoma	3	1
TCGA-TM-A84F	Astrocytoma	3	1
TCGA-HT-7854	GBM	2	1
TCGA-HT-7877	oligodendroglioma	2	1

The gap statistic was employed to evaluate the ideal number of clusters considering 6 latent variables, 20 clusters as the maximum, and 200 random initializations for K-means. The obtained values and error bars can be observed in [Figure 4.22](#), and the largest increase happens from $k = 1$ to 2, and the second largest is from 4 to 5, with the subsequent increases smaller indicating stabilization. Therefore, the highest value of the gap statistic until stabilization is 5, which should be considered the number of clusters.

Thus, K-means with 5 clusters were computed [Table 4.13a](#) shows the confusion matrix between the predicted cluster and the glioma type, and it is clear there is a cluster for each glioma type: clusters 1, 4, and 5 are characterized by astrocytoma, GBM, and oligodendroglioma respectively. Clusters 2 and 3 seem to further divide the LGG, with each having only a GBM patient.

[Table 4.13b](#) has the external cluster validation measures. A low Average Silhouette Width of 0.228 with the highest value obtained in cluster 2 of 0.33, and the lowest in cluster

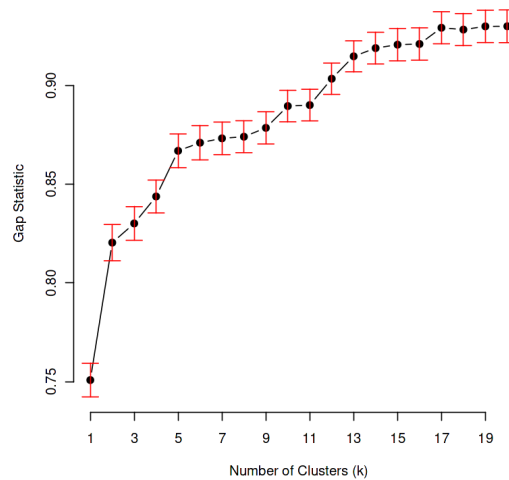


Figure 4.22: Gap statistic for the 6 latent variables obtained through moCluster.

Table 4.13: moCluster method with 6 JLTV and 5 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	162	2	13	177	Calinski-Harabasz Index	152.721
Cluster 2	25	1	19	45	Dunn Index	0.034
Cluster 3	47	1	28	76	Average Silhouette Width	0.228
Cluster 4	7	108	2	117		
Cluster 5	8	0	104	112		

3 of 0.10 (Figure 4.23).

The next step was inducing sparsity and checking if the correlation values between the non-sparse and sparse latent variables was high, which can be done in Figure 4.24. For this correlation between non sparse and sparse CPC, the range is bigger with a lot of the sparsity levels having one or more latent variables not with low correlation. A sparsity of 0.80 to 0.95 is required to ensure that all sparse latent variables have a very high correlation. However, if more leniency is allowed—accepting that 5 out of the 6 CPCs have a very high correlation while the remaining one has a high correlation—then sparsity levels of 0.35 to 0.75 can also be considered.

In light of this, 15 sparse components with the mentioned sparsities and the explained and cumulative variance were calculated. In Table 4.14, it is possible to see the cumulative variance, with a lower sparsity parameter in the range of 60% and a higher with 71%.

For the relevant sparse CPC, the gap statistic was calculated and it is present in Figure C.4. From this, it is possible to determine that:

- **0.30:** The first high is $k = 1$ followed by a drop and a rise until 12 where the value

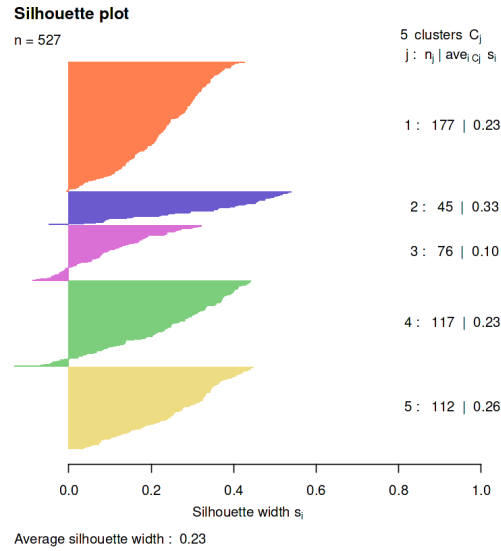


Figure 4.23: moCluster method with 6 JLV and 5 clusters: Silhouette plot.



Figure 4.24: Correlation heatmap between non-sparse and sparse 6 consensus principal components.

Table 4.14: Cumulative explained variance for 6 consensus principal components out of 15 with sparsity parameters ranging from 0.40 to 0.75.

Sparsity	Cumulative Explained Variance
0.40	0.633
0.45	0.657
0.50	0.676
0.55	0.690
0.60	0.701
0.65	0.708
0.70	0.711
0.75	0.714

stabilizes until 13 and rises again, which makes it a delicate case where 12 can be considered a good value but in practicality setting it as 12 makes interpretability unfeasible.

- **0.35:** The first maximum is at 1 cluster then drops and rises again steadily until 20 clusters, not stabilizing, making it hard to determine the number of clusters to set and other techniques should be used to assess.
- **0.65 to 0.95:** 5 clusters are the ideal, analogous to the non-sparse latent variables.

However, 0.30 and 0.35 sparsity values were not considered for clustering due to the low correlation mentioned previously with the 6th not sparse and sparse components and the lower cumulative variance comparatively.

Thus, for sparsity 0.65 to 0.95, the K-means algorithm with 5 centers and 200 random starts is computed and the comparison of metrics between the 5-means with sparse and non-sparse CPCs are shown in Figure 4.25. From the plot analysis, comes that:

- **Calinski-Harabasz Index:** a sparsity of 0.80 to 0.95 shows higher values than the non-sparse metrics, and other sparsity parameters have a lower metric.
- **Dunn Index:** 0.65 and 0.70 originate a better value than the non-sparse, 0.80 and 0.85 the same value, and others lower.
- **Average Silhouette Width:** 0.85 and 0.95 generate the same value as the sparse, and others have a worse value.

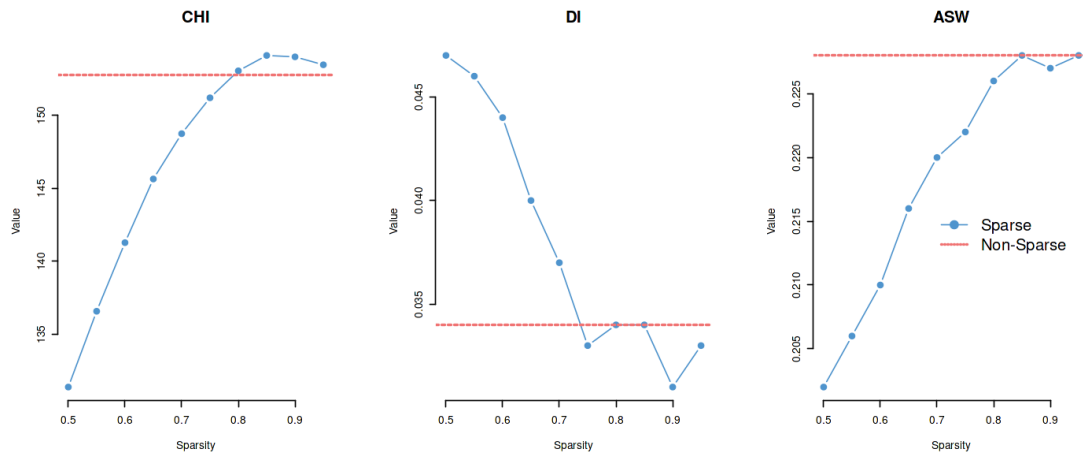


Figure 4.25: Metrics for 5-means with non-sparse and sparse consensus principal components.

Based on the comparison of metrics, the sparsity value of 0.85 is identified as the optimal choice. Specifically, it improves the Calinski-Harabasz Index while maintaining the Dunn Index and the Average Silhouette Width at levels comparable to non-sparse configurations.

The 5-means confusion matrix and cluster validation measures for the sparsity of 0.85 are detailed in Table 4.15a and Table 4.15b, respectively. Cluster 1 consists mainly of astrocytoma patients, with 25 individuals, along with 18 oligodendroglioma patients and only 1 GBM patient. Cluster 2 is similarly diverse but has a higher number of astrocytoma patients (48), followed by 29 oligodendroglioma patients and again only 1 GBM patient. Cluster 3 is mostly composed of GBM patients, with 108 out of 117 total, indicating that this cluster is primarily associated with GBM. Cluster 4, like cluster 1, is dominated by astrocytoma patients (161), with 14 oligodendroglioma patients and just 2 GBM patients, showing another astrocytoma-heavy group. Finally, cluster 5 is composed mostly of oligodendroglioma patients (103), with only a few astrocytoma (8) and no GBM patients, making it the most oligodendroglioma-dominated cluster.

Table 4.15: moCluster method with 6 sparse JLV and 5 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	25	1	18	44	Calinski-Harabasz Index	154.070
Cluster 2	48	1	29	78	Dunn Index	0.034
Cluster 3	7	108	2	117	Average Silhouette Width	0.228
Cluster 4	161	2	14	177		
Cluster 5	8	0	103	111		

The silhouette plot in Figure 4.26 reflects the low average width with all clusters showing a weak structure that could be artificial.

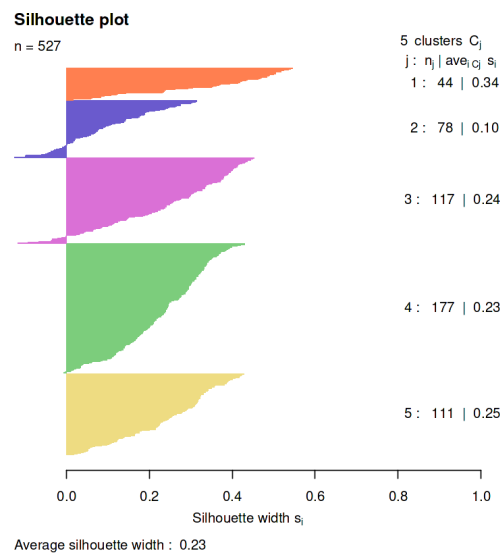


Figure 4.26: moCluster method with 6 sparse JLV and 5 clusters: Silhouette plot.

Relevant Features

The principal component scores represent the projection of the samples onto the latent variables axes, with the size of the number of objects by the number of JLV. The maximum and minimum values of these scores are 0.218 and -0.091 , respectively. The mean and variance of each latent variable have low values (around 10^{-18}), with the first CPC being the only positive mean.

Furthermore, in Figure 4.27 the number of variables and their signal coefficients in the sparse JLVs for each omic is shown. In particular, 256,310 and 17,171 variables have non-null coefficients for the methylomics and transcriptomics, respectively. The transcriptomics coefficients' have a similar distribution of positive and negative, while the methylomics has mostly negative coefficients, indicating a negative contribution.

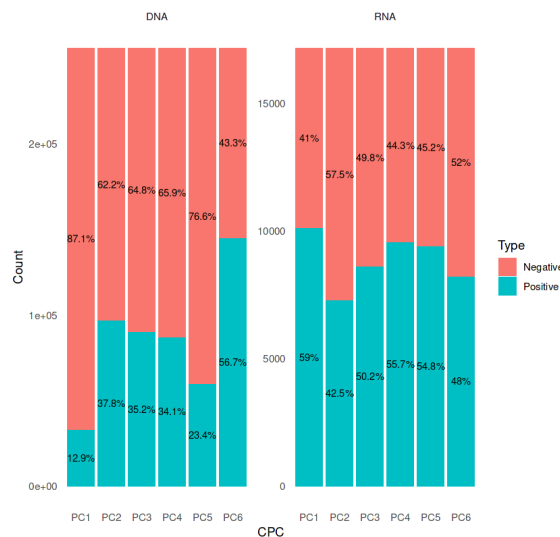


Figure 4.27: Stacked bar chart of RNA and DNA counts and type by principal component with sparsity set as 0.85.

Then, principal component-specific variables, meaning features with a non-zero coefficient for one of the components and zero for all the others, were found. For principal component 1, 36 variables are exclusive, PC2 has 60 specific variables, PC3 has 55, PC4 28, PC5 47, and PC6 80. From the features specific for the first component, 6 were also significant for the CIMLR method with 3 clusters (cg25954539, cg15213225, cg02932689, cg20088847, cg18557684, and cg04297264), 5 features with 4 clusters (cg25954539, cg15213225, cg20088847, cg02932689, and cg04297264). 33 are common to the variables with normalized coefficients bigger than 0.5 in the intNMF method with 3 clusters, and 8 features to the first JLV from 2 sparse JLV of the moCluster method. The JLV-specific variables for the second component only have 18 features in common with the second component of the 2 sparse JLV of the moCluster method. For the variables specific to the fifth JLV, 2 are common with the features identified by the CIMLR method with 3 clusters (cg13552712, and SDR9C7), and 1 is common with the 4 clusters'

(cg13552712).

Due to the poor performance of moCluster method with 6 sparse JLV and 5 clusters, the features were not cross-referenced with literature, as we considered that the clustering results did not provide meaningful or reliable biological insights.

4.2.3.3 Survival Analysis

The Kaplan-Meier curves were plotted for the moCluster predicted cluster with 2 JLVs with sparsity parameter of 0.5 and 6 JLVs with sparsity set as 0.85, and these can be consulted in Figure 4.28.

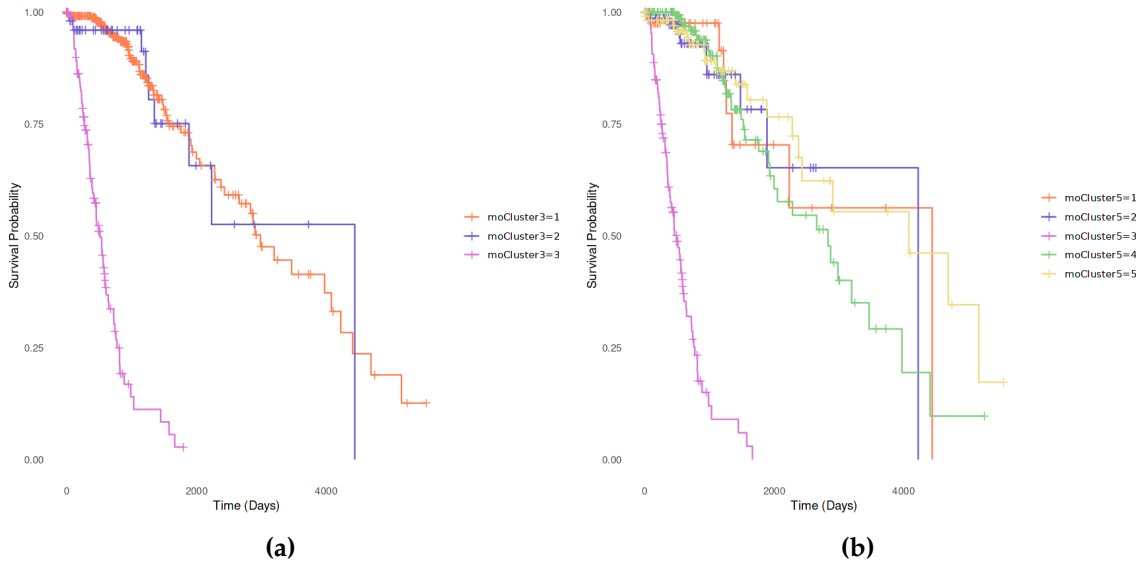


Figure 4.28: Kaplan-Meier curves for the moCluster method with: a) 2 JLVs with sparsity of 0.5 and 3 clusters. b) 6 JLVs with sparsity of 0.85 and 5 clusters.

Regarding the 3 clusters with 2 sparse JLVs, visually the survival curve in Figure 4.28a of cluster 3 seems to be different from the other clusters, with an overall worse survival likelihood. The log-rank test was employed and, with a significance of 5%, the survival distributions for the 3 clusters were statistically different, $\chi^2(2) = 301$, $p < 2 \times 10^{-16}$. Afterward, the pairwise comparison tests were employed and these revealed that clusters 1 and 3 ($p < 2 \times 10^{-16}$) and clusters 2 and 3 ($p < 4.2 \times 10^{-13}$) survival curves are statistically different. On the other hand, clusters 2 and 3 survival distributions are not statistically different ($p = 0.93$). In light of this, cluster 3 which is mostly constituted by the GBM type shows a worse prognosis than the other clusters, which is compatible with previous knowledge of the disease, and likely indicates that the astrocytoma and oligodendroglioma patients in this cluster have a worse survival likelihood.

To verify the previous statement, the clinical data was analyzed. Of the 10 astrocytoma patients in the cluster, 4 have passed away and the days to death after the diagnosis are low, ranging from 111 to 987, similar to the GBM average of 437 days. One of astrocytoma

patient does not have clinical data. The 2 oligodendroglioma patients have 243 or 805 days from the last follow-up appointment. This means that the majority of the patients in the cluster that passed away has a closer average to the GBM than the LGG.

Furthermore, the survival of the GBM patients not in cluster 3 was also studied. For the 2 patients in cluster 2, one passed away 36 days after diagnosis, which is expected from GBM, and the other patient last check-up happened 611 days, which is ambiguous. The GBM patient in cluster 1 has 1,201 since their last follow-up, which is higher than the type average and closer to oligodendroglioma.

For the 5 clusters with 6 sparse JLVs, cluster 3 stands out in [Figure 4.28b](#) with a distinguishable curve from the other clusters, and a rapid descending survival likelihood. The log-rank test concluded that the five clusters survival estimation was significantly different, $\chi^2(4) = 333$, $p < 2 \times 10^{-16}$. Once again, the pairwise tests were performed, and the BH correction was employed. Cluster 3 is statistically different from all other clusters ($p_{12} < 2 \times 10^{-16}$, $p_{13} = 3.5 \times 10^{-12}$, $p_{14} < 2 \times 10^{-16}$, $p_{15} < 2 \times 10^{-16}$). However, clusters 4 and 5 ($p = 0.43$) and all other combinations ($p = 0.88$) have statistically equal distribution curves. The conclusion drawn from this analysis is similar: moCluster separates the LGG and GBM and identifies patients with a low survival probability.

Once again, the clinical data of the LGG patients grouped with GBM was analyzed. Of the 7 astrocytoma patients, 4 have passed away 111 to 987 days after diagnosis. Once again, the astrocytoma patients have a closer average to GBM. The GBM patients not in cluster 3 were also studied: of the 2 in cluster 4, one passed away 538 days after diagnosis, which is consistent for GBM.

4.2.4 Comparison

In [Table 4.16](#), the cluster validation measures are present with different cluster configurations for each multi-view clustering method employed. For the CIMLR method, the 3 and 4 clusters result are present with the 3 clusters having better external validation measures than the 4 clusters. Regarding the intNMF method, 3 to 6 clusters were considered and the 3 clusters' output had higher measures overall, in particular, the Calinski-Harabasz Index is 8,179 while the others are around 1,000. In the moCluster method, multiple JLVs and combinations of sparsity were considered, and the best sparsity parameter was selected for 2 and 6 PC with the corresponding ideal number of clusters. Thus, for 2 JLVs sparsity of 0.5 was considered with 3 clusters, and for the 6 JLVs sparsity of 0.85 with 5 clusters. Comparing the external validation measures, the 2 sparse JLVs with 3 clusters shows a better performance than the 6 sparse JLVs with 5 clusters, for the Calinski-Harabasz Index and the Average Silhouette Width. The Dunn Index shows a higher value for the 5 clusters, yet both values are low. In light of this, the 2 sparse JLVs with 3 clusters was considered for consensus clustering.

Some additional considerations regarding the omics: In the CIMLR feature selection, in both the 3 and 4 clusters output, the most relevant variables originated from the

Table 4.16: Comparison of the multi-view clustering methods across various cluster validation measures: a) CIMLR method with 3 and 4 clusters. b) intNMF method with 3 to 6 clusters. c) moCluster method with 3 clusters and 2 JLVs with sparsity of 0.5 and 4 clusters with 6 JLVs with sparsity of 0.85.

(a)				
Cluster Validation Measures	CIMLR 3	CIMLR 4		
Adjusted Rand Index	0.610			
Normalized Mutual Information	0.604			
Calinski-Harabasz Index	64,540	38,544		
Dunn Index	8.37	5.074		
Average Silhouette Width	0.954	0.94		
(b)				
Cluster Validation Measures	intNMF 3	intNMF 4	intNMF 5	intNMF 6
Adjusted Rand Index	0.235			
Normalized Mutual Information	0.314			
Calinski-Harabasz Index	8,179	1,301	1,188	1,072
Dunn Index	0.231	0.098	0.062	0.042
Average Silhouette Width	0.921	0.734	0.792	0.794
(c)				
Cluster Validation Measures	moCluster 3	moCluster 5		
Adjusted Rand Index	0.365			
Normalized Mutual Information	0.399			
Calinski-Harabasz Index	684	154		
Dunn Index	0.019	0.034		
Average Silhouette Width	0.600	0.228		

methyloomics and the first transcriptomics variable is not in a high position. While this might be due to there being more CpG sites than genes, it can also mean that the methyloomics are more relevant than transcriptomics to group the glioma patients, which intuitively is corroborated by the 2D visualization of the omics (Figure 4.1). Furthermore, in the cluster-specific variables for the intNMF method with 3, 4, and 5 clusters, the transcriptomic data had a low normalized coefficient in the relevant matrix. However, when considering methyloomics, there were some features with a normalized coefficient higher or equal to 0.5 suggesting a notable contribution from this omic. Contrary to this tendency, in the moCluster latent variables, aside from the first principal component, the transcriptomic shows a higher relative contribution than the transcriptomics (Figure 4.13a), which can be theorized as one of the possible reasons to a poor predictive outcome (low overall metrics).

Furthermore, for the CIMLR and intNMF methods with 3 clusters, which yielded better results, a different seed was employed. The results are shown in section A.2 and section B.4, respectively, and the partitions originated from the different seeds for each

method were identical.

4.3 Consensus Clustering

4.3.1 3 Methods

The initial step in consensus clustering involved constructing a co-association matrix using the outputs from three multi-view clustering methods. This matrix is required to be symmetric with ones along the diagonal, conditions that have been confirmed. In this matrix, the possible values for the entries are 0 when the objects were not clustered together in a single method, $1/3$ if the patients are in the same cluster in one method, $2/3$ when they are grouped in 2 methods, and 1 if they were clustered together in all outputs.

After constructing the co-association matrix, it was converted into a dissimilarity matrix by computing one minus the co-association matrix. To meet the requirements of the hierarchical clustering function, this dissimilarity matrix is then transformed into a distance matrix using the “as.dist()” function, which converts it into the appropriate format. Finally, average-link hierarchical clustering is applied to the resulting distance matrix.

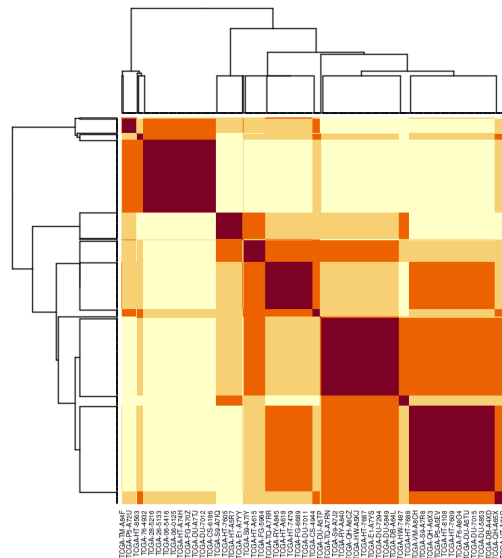


Figure 4.29: Heatmap of the co-association matrix from the 3 MVC methods with dendrogram.

In Figure 4.29, the heatmap of the co-association matrix and the corresponding dendrogram are presented. The heatmap reveals several prominent dark blocks along the diagonal, with four major blocks being particularly noticeable, alongside smaller blocks. These indicate groups of patients that were consistently clustered together across the three different methods. A key observation is the block in the upper left corner, which displays a variety of colors representing different degrees of connection. This block has relatively few and weak connections to patients outside of it, as illustrated by the dendrogram above

it, which shows a notable height required to merge into a single agglomerative cluster. In contrast, outside of this upper left block, the patients appear to be more interconnected with each other, making it challenging to discern distinct clusters. Nevertheless, in the lower right corner of the heatmap, there is a potential cluster comprising two major blocks.

In [Figure 4.30](#), the dendrogram from the average-link hierarchical clustering is shown, with two red lines highlighting the cuts. The first jump between the height of the cluster happens around 0.2, originating 16 clusters.

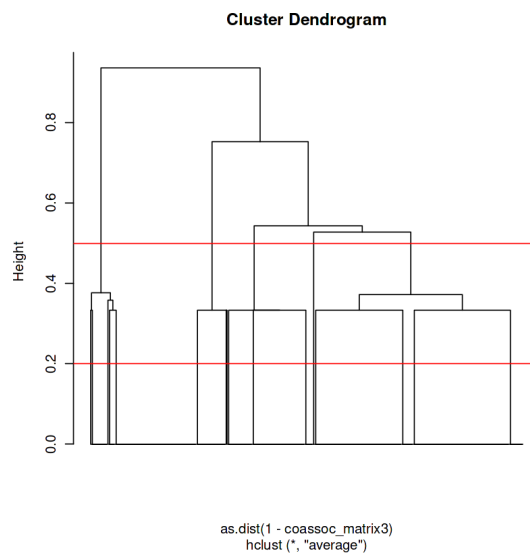


Figure 4.30: Dendrogram from the average-link hierarchical clustering of the co-association matrix with 3 methods.

The 16 clusters have various numbers of data points with the biggest group having 117 patients and 4 clusters are singletons, meaning they are characterized by a single object. With the predicted clusters, the cluster validation measures were computed, and the mean Average Silhouette Width is 0.992 (all clusters with more than one element have a perfect score and the single element a value of 0), and the Dunn Index and the Calinski-Harabasz Index are returning an infinite value. For the Dunn Index, the phenomenon is explained as the maximum distance between any two points in each cluster is 0, meaning that the patients in each consensus cluster were grouped in the 3 MVC methods. The Calinski-Harabasz value explanation is analogous. While the three cluster validation measures have high values, the interpretation of the 16 clusters is difficult and since the distance matrix is derived from the co-association matrix the values of the Dunn and Calinski-Harabasz Index can be considered artificially high.

An alternative dendrogram cut can be considered at a height of 0.5 originating 5 clusters, a more interpretable result. In [Table 4.17a](#), the confusion matrix is present: cluster 1 is characterized by 137 astrocytoma and 117 oligodendroglioma patients, clusters 2 and 4 have mostly astrocytoma patients, followed by oligodendroglioma and few GBM patients, cluster 3 is constituted by 109 GBM patients and some astrocytoma and oligodendroglioma

types, and cluster 5 has only two oligodendroglioma patients. The cluster validation measures are shown in Table 4.17b where the average silhouette of 0.589 means a reasonable structure was found [104]. The Calinski-Harabasz Index shows an improvement from the moCluster results but is lower than obtained in the other methods, while the Dunn Index of 0.333 is higher than the moCluster and intNMF but lower than the CIMLR results.

Table 4.17: Consensus clustering result with 3 methods and 5 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	137	0	117	254	Calinski-Harabasz Index	798.524
Cluster 2	73	1	30	104	Dunn Index	0.333
Cluster 3	18	109	2	129	Average Silhouette Width	0.589
Cluster 4	21	2	15	38		
Cluster 5	0	0	2	2		

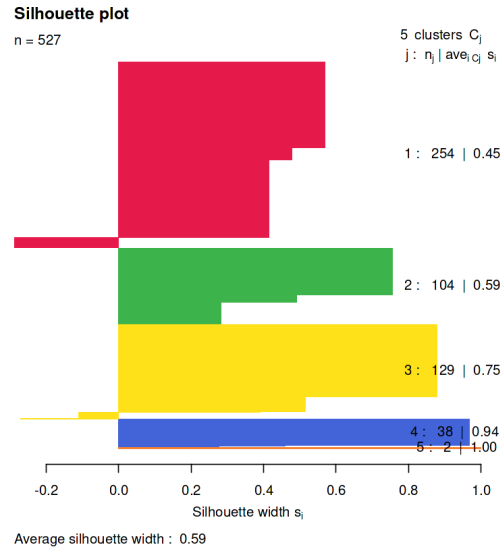


Figure 4.31: Consensus clustering result with 3 methods and 5 clusters: Silhouette plot.

4.3.1.1 Survival Analysis

The Kaplan-Meier curves for the 5 predicted clusters are present in Figure 4.32. The survival curve of cluster 3 (in yellow) seems to be distinct with an accentuated drop in the survival probability, compatible with the cluster composition of mostly GBM patients. Cluster 5 (in orange) also shows a stiff drop in the estimation, explained by it only having two data points.

The log-rank test was employed and resulted in $\chi^2(4) = 268$, $p < 2 \times 10^{-16}$, meaning that with a significance of 5% the survival distributions of the predicted clusters are not

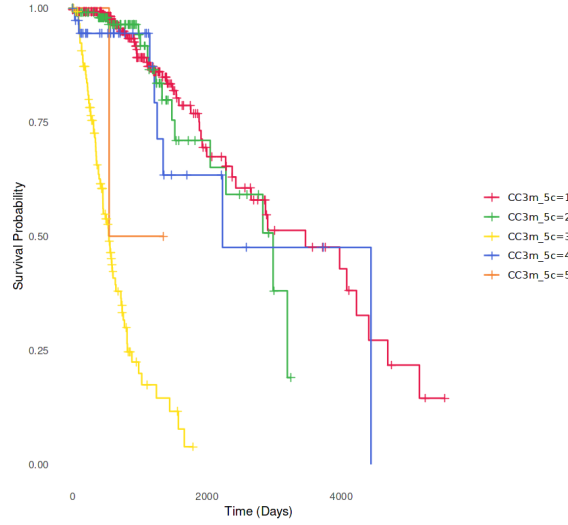


Figure 4.32: Kaplan-Meier curves for the 5 predicted consensus clusters.

equal. Furthermore, the pair-wise comparisons with the BH correction were performed, and the survival functions from clusters 1 and 3, 2 and 3, and 3 and 4 are all not equal, $p_{13}, p_{23} < 2 \times 10^{-16}$, $p_{34} = 9.9 \times 10^{-8}$. Yet the other comparisons yielded that the survival distributions are equal: $p_{12} = 0.56$, $p_{14} = 0.47$, $p_{15} = 0.16$, $p_{24} = 0.8$, $p_{25} = 0.33$, $p_{35} = 0.47$, $p_{45} = 0.56$. This is in line with the results obtained in the multi-view clustering methods, GBM is separated from the LGG but the method did not separated the latter.

4.3.2 2 Methods

Moreover, as the moCluster cluster validation measures are notably lower, the consensus clustering process was additionally employed only considering the 3 clusters' output from CIMLR and intNMF. The new co-association matrix was computed, followed by the dissimilarity matrix and the average-link hierarchical clustering. Note that the entries of the co-association matrix are now 0, 1/2 or 1.

The heatmap of the co-association matrix and the corresponding dendrogram are presented in [Figure 4.33](#). The heatmap reveals several prominent dark blocks along the diagonal, with five major or medium blocks and two smaller blocks, and overall it is easier to analyze the connections. Once again the block in the upper left corner shows minor connections with patients outside of it, yet the dendrogram's height seems smaller to merge all the clusters. This might indicate that the moCluster assignment is not as coherent with the other two methods. Outside of this upper left block, the patients still appear to be more interconnected with each other, making it challenging to discern distinct clusters. However, there seems to have been an improvement in readability.

The dendrogram obtained from the average-link clustering to the co-association matrix of the two methods can be observed in more detail in [Figure 4.34](#). The first cut of the dendrogram can be around height 0.2 to represent the highest jump, generating 9 clusters.

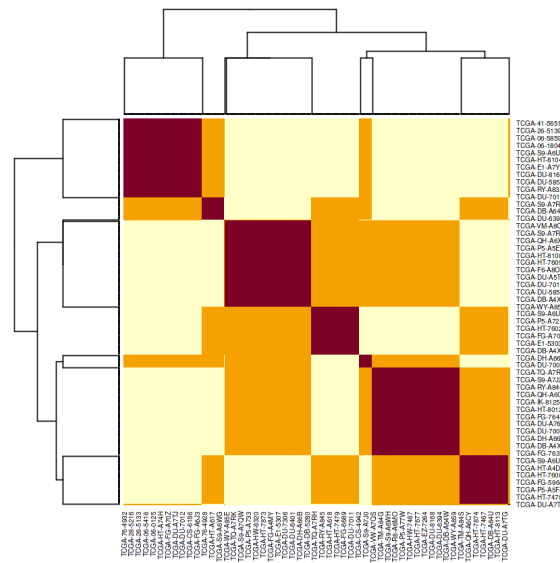


Figure 4.33: Heatmap of the co-association matrix from the 2 MVC methods with dendrogram.

Another potential cut is at a height of 0.6.

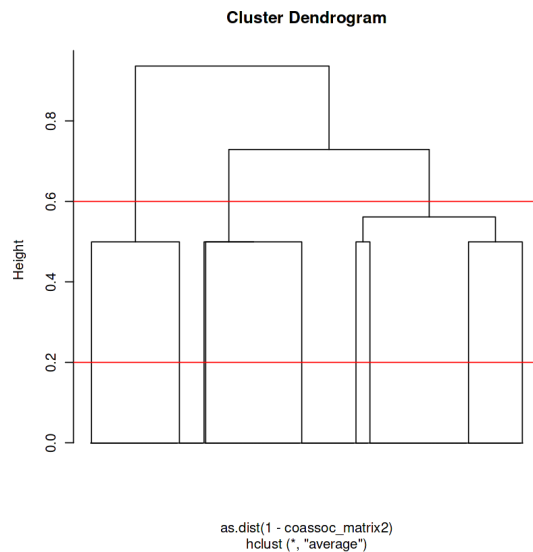


Figure 4.34: Dendrogram from the average-link hierarchical clustering of the co-association matrix with 2 methods.

The cut at height 0.2 originates 9 clusters, and the confusion matrix is present in [Table 4.18](#). Notably, clusters 1, 2, and 8 are constituted solely by astrocytoma patients, and cluster 9 only by oligodendroglioma. Clusters 3 and 4 have astrocytoma and oligodendroglioma patients, cluster 5 by astrocytoma and GBM, while 6 and 7 have all types of gliomas. The cluster validation measures were calculated and the Average Silhouette Width is 0.998, since all clusters except 8 have a perfect value, and cluster 8 has an average of 0. Once again, the Dunn and the Calinski-Harabasz Index have an infinite value.

Table 4.18: Consensus clustering result with 2 methods and 9 clusters: Confusion matrix for the obtained clusters versus the official glioma types.

moCluster	Astro	GBM	Oligo	Total
Cluster 1	117	0	0	117
Cluster 2	66	0	0	66
Cluster 3	11	0	6	17
Cluster 4	9	0	111	120
Cluster 5	14	93	0	107
Cluster 6	21	1	44	66
Cluster 7	10	18	3	31
Cluster 8	1	0	0	1
Cluster 9	0	0	2	2

Subsequently, the cut that originates 3 clusters is considered. In Table 4.19a, the confusion matrix is present, and a glioma type, with some overlap between the LGG and the astrocytoma and GBM type, characterize each cluster. Table 4.19b shows the cluster validation indexes: the internal measures are above 0.5, and the Average Silhouette Width of 0.666 indicates a reasonable structure was found [104]. It is worth highlighting that the Adjusted Rand Index and the Normalized Mutual Information of the consensus clustering are higher than the 3 clusters output from CIMLR and intNMF. However, both cases have higher Average Silhouette Width, and the Calinski-Harabasz Index. The Dunn Index for the consensus clustering assignment is higher than the intNMF prediction but lower than the CIMLR result.

Table 4.19: Consensus clustering result with 2 methods and 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
moCluster	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	184	0	0	184	Adjusted Rand Index	0.629
Cluster 2	41	1	163	205	Normalized Mutual Information	0.639
Cluster 3	24	111	3	138	Calinski-Harabasz Index	840.436
					Dunn Index	0.5
					Average Silhouette Width	0.666

In the context of CC, it is important to note that the CIMLR and intNMF methods have 244 CpG sites as relevant features in common. While cross-referencing all of these sites with existing literature is not practical due to the sheer number, the top 10 most relevant variables, based on their contribution to clustering performance, were thoroughly investigated and highlighted.

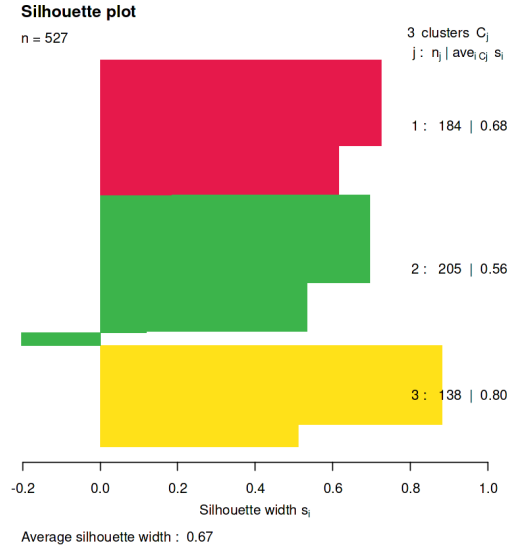


Figure 4.35: Consensus clustering result with 2 methods and 3 clusters: Silhouette plot.

4.3.2.1 Survival Analysis

Figure 4.36 shows the Kaplan-Meier curves for the 3 predicted clusters, building the co-association matrix with 2 MVC methods. Visibly, the survival line of cluster 3 (in yellow) is very distinctive from the other, and there seems to be a notable overlap of the other 2 curves. This is expected with the previous knowledge obtained through the Kaplan-Meier curves of the official WHO glioma classification (Figure 4.7).

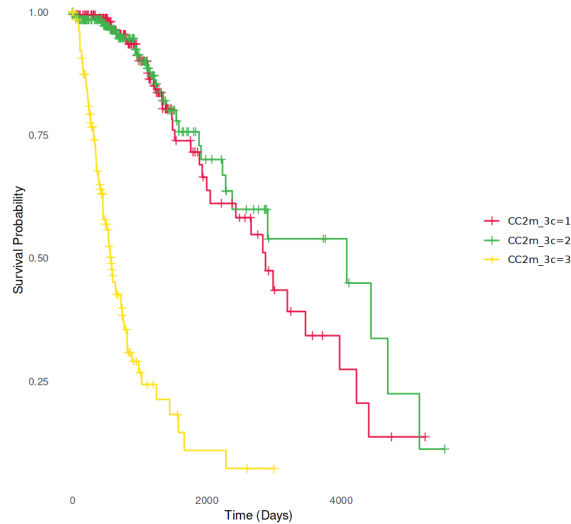


Figure 4.36: Kaplan-Meier curves for the 3 predicted consensus clusters.

Next, to test if the survival functions are equal for the cluster, the log-rank test was applied and resulted in $\chi^2(2) = 211$, $p < 2 \times 10^{-16}$. This means that the survival curves are not equal, and the pair-wise comparisons become relevant. From these, it is possible

to conclude that the survival distributions from clusters 1 and 2 are equal ($p_{12} = 0.47$) and clusters 1 and 3 and 2 and 3 are distinct ($p_{13}, p_{23} < 2 \times 10^{-16}$). Once again, this is coherent with [Figure 4.7](#).

Of the 24 astrocytoma patients grouped with GBM in cluster 3, 7 have passed away, 5 with low days after diagnosis (11-987) but 2 passed away 1,251 and 2,286 days after diagnosis. The oligodendroglioma patient in this clusters was alive at the last follow-up. This means that most of the LGG patients in this cluster had a lower survival than expected.

CONCLUSION

The present thesis explored the application of state-of-the-art multi-view clustering methods to glioma patients by integrating two different omics datasets: transcriptomics and methylomics. To evaluate the performance of the clustering methods, various metrics were computed, key features were analyzed, and survival analysis was conducted to evaluate the predictive value of the clusters. Furthermore, a consensus clustering approach was implemented using the co-association matrix and average-linkage hierarchical clustering, and the results were evaluated using the same performance metrics and survival analyses.

Specifically, the CIMLR algorithm produced clusters with a similar structure to the 2021 WHO CNS classification of gliomas, enabling the identification of several significant features and the analysis of the top 10 most significant features allowed further validation of known biomarkers. CIMLR had strong performance metrics with well-defined and cohesive clusters, however, its prognostic value in survival analysis was limited. This is in line with similar observations in the 2021 WHO classification survival analysis, suggesting that the challenge may be inherent to the available data to the current classification or the omic datasets used. Expanding the study to include additional omics data could offer more comprehensive insights and potentially improve the survival prediction capabilities of the clustering models.

In contrast, the intNMF technique displayed lower overall performance metrics and was not in line with the current glioma types, except for the separation of GBM and LGG. The method yielded a surprising result in the survival analysis, where the three identified clusters showed distinct survival patterns, which is an important and unexpected discovery that justifies further investigation. The comparison to the 2021 WHO CNS classification was highlighted in the results of all methods and this is relevant for two reasons which were exemplified by the CIMLR and the intNMF methods. The former supports the 2021 WHO classification of gliomas and this is beneficial to validate the current classification, to add biomarkers that deepen our knowledge, and allow for more accurate diagnosis and personalized treatment. The latter still supports a 3-type division of the glioma classification which is not in line with the 2021 WHO CNS classification, and the new grouping is still advantageous since it is supported statistically with different survival

curves and might provide more information about the underlying molecular information of the disease.

The moCluster method, when optimized with two latent variables, supported a three-cluster solution that is coherent with the current glioma classification. With six latent joint variables, a five-cluster implementation was found to be ideal. Despite these findings, moCluster achieved the weakest metrics overall and failed to distinguish the official glioma subtypes effectively.

An important consideration for future research is the impact of random seeds on the results of non-deterministic algorithms. Since these methods can produce varying outcomes depending on the seed, a systematic evaluation using multiple seeds is essential to assess the robustness of the clustering results. Preliminary analysis using two different seeds provided insight into this variability, but a more extensive study is necessary to draw definitive conclusions.

Consensus clustering was initially attempted by combining all three methods, but the results did not show substantial improvement. Due to the poor performance of moCluster, a second consensus clustering was performed without it. The optimal solution resulted in nine clusters, but these were difficult to interpret, with performance metrics that seemed artificially inflated. As a result, the three-cluster solution was selected as the second-best choice, offering more reliable results and metrics. The clusters closely mirrored the 2021 WHO glioma classification, further validating the current classification and contributing to a deeper understanding of glioma biology.

Future research could benefit from exploring different consensus clustering techniques that leverage recent advancements in the field. While the methods used in this thesis were chosen for their peer-reviewed, well-established frameworks, other approaches may offer enhanced performance and more accurate clustering.

Many of the biomarkers identified in this dissertation have been previously reported in glioma studies, and their presence in the literature adds trustworthiness to the findings, further validating their established roles in glioma biology. On the other hand, newly discovered biological features, such as *TMEM106A*, *MCF2L2*, *FES*, *TTBK1*, *ADAM33*, and *ARTN*, which have either an important role in other cancers or only have one other study supporting its relevancy in glioma, necessitate further biological validation. This analysis is contemplated in the MONET project consortium where future efforts will be dedicated to experimental and clinical investigation of the relevance and therapeutic potential of these biomarkers.

In conclusion, addressing the limitations and exploring the areas for improvement discussed in this thesis will not only advance our understanding of the robustness and applicability of multi-view clustering but also enhance the integration of multi-omic data. This could lead to more significant biological discoveries and refined subtyping of gliomas, which is crucial given the aggressive nature of the disease. Continued efforts to subtype gliomas and deepen the knowledge of their underlying biology are essential for developing better treatment strategies and improving patient outcomes.

BIBLIOGRAPHY

- [1] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User's Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf>.
- [2] P. Wesseling and D. Capper. "WHO 2016 Classification of gliomas". In: *Neuropathology and Applied Neurobiology* 44.2 (2018), pp. 139–150. DOI: [10.1111/nan.12432](https://doi.org/10.1111/nan.12432).
- [3] Z. Zhang, Y. Xiao, S. Zhao, et al. "FAM109B plays a tumorigenic role in low-grade gliomas and is associated with tumor-associated macrophages (TAMs)". In: *Journal of Translational Medicine* 22 (2024), p. 833. DOI: [10.1186/s12967-024-05641-6](https://doi.org/10.1186/s12967-024-05641-6).
- [4] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary". In: *Acta Neuropathologica* 131.6 (2016), pp. 803–820. DOI: [10.1007/s00401-016-1545-1](https://doi.org/10.1007/s00401-016-1545-1).
- [5] M. L. Mendonça, R. Coletti, C. S. Gonçalves, E. P. Martins, B. M. Costa, S. Vinga, and M. B. Lopes. "Updating TCGA glioma classification through integration of molecular profiling data following the 2016 and 2021 WHO guidelines". In: *bioRxiv* (2023). DOI: [10.1101/2023.02.19.529134](https://doi.org/10.1101/2023.02.19.529134).
- [6] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, and D. W. Ellison. "The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary". In: *Neuro-Oncology* 23.8 (2021), pp. 1231–1251. DOI: [10.1093/neuonc/noab106](https://doi.org/10.1093/neuonc/noab106).
- [7] W. Aiman, D. Gasalberti, and A. Rayi. "Low-Grade Gliomas". In: *StatPearls*. Ed. by S. Publishing. Updated. Treasure Island, FL, 2023-05. URL: <https://www.ncbi.nlm.nih.gov/books/NBK560668/>.

-
- [8] A. M. Rulseh, J. Keller, J. Klener, J. Sroubek, V. Dbalý, M. Syrůček, F. Tovaryš, and J. Vymazal. "Long-term Survival of Patients Suffering from Glioblastoma Multiforme Treated with Tumor-treating Fields". In: *World Journal of Surgical Oncology* 10 (2012), p. 220. DOI: [10.1186/1477-7819-10-220](https://doi.org/10.1186/1477-7819-10-220).
 - [9] Z. Cai, R. C. Poulos, J. Liu, and Q. Zhong. "Machine learning for multi-omics data integration in cancer". In: *iScience* 25.2 (2022), p. 103798. DOI: [10.1016/j.isci.2022.103798](https://doi.org/10.1016/j.isci.2022.103798).
 - [10] L. Yang, Y. Yang, L. Huang, X. Cui, and Y. Liu. "From Single- to Multi-Omics: Future Research Trends in Medicinal Plants". In: *Briefings in Bioinformatics* 24.1 (2023), bbac485. DOI: [10.1093/bib/bbac485](https://doi.org/10.1093/bib/bbac485).
 - [11] P. Chalise and B. L. Fridley. "Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm". In: *PLOS ONE* (2017-05). DOI: [10.1371/journal.pone.0176278](https://doi.org/10.1371/journal.pone.0176278).
 - [12] L. Cantini, P. Zakeri, C. Hernandez, and et al. "Benchmarking Joint Multi-Omics Dimensionality Reduction Approaches for the Study of Cancer". In: *Nat Commun* 12 (2021), p. 124. DOI: [10.1038/s41467-020-20430-7](https://doi.org/10.1038/s41467-020-20430-7).
 - [13] A. Barzegar Behrooz, H. Latifi-Navid, S. C. da Silva Rosa, M. Swiat, E. Wiechec, C. Vitorino, R. Vitorino, Z. Jamalpoor, and S. Ghavami. "Integrating Multi-Omics Analysis for Enhanced Diagnosis and Treatment of Glioblastoma: A Comprehensive Data-Driven Approach". In: *Cancers* 15.12 (2023), p. 3158. DOI: [10.3390/cancers15123158](https://doi.org/10.3390/cancers15123158).
 - [14] W. F. C. Alvin C. Rencher. "Cluster Analysis". In: *Methods of Multivariate Analysis*. John Wiley & Sons, Ltd, 2012. Chap. 15, pp. 501–554. ISBN: 9781118391686. DOI: <https://doi.org/10.1002/9781118391686.ch15>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118391686.ch15>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118391686.ch15>.
 - [15] B. Reddy. "Literature Survey On Clustering Techniques". In: *IOSR Journal of Computer Engineering* 3 (2012-01), pp. 01–12. DOI: [10.9790/0661-0310112](https://doi.org/10.9790/0661-0310112).
 - [16] M. B. Lopes, E. P. Martins, S. Vinga, and B. M. Costa. "The Role of Network Science in Glioblastoma". In: *Cancers* 13.5 (2021), p. 1045. DOI: [10.3390/cancers13051045](https://doi.org/10.3390/cancers13051045).
 - [17] M. Serra-Burriel and C. Ames. "Machine Learning-Based Clustering Analysis: Foundational Concepts, Methods, and Applications". In: *Acta Neurochirurgica Supplement* 134 (2022), pp. 91–100. DOI: [10.1007/978-3-030-85292-4_12](https://doi.org/10.1007/978-3-030-85292-4_12).
 - [18] J. Hartigan. *Clustering Algorithms*. University Microfilms, 1988. URL: <https://books.google.pt/books?id=cl73oQEACAAJ>.

- [19] M. Bredel, D. M. Scholtens, G. R. Harsh, C. Bredel, J. P. Chandler, J. J. Renfrow, A. K. Yadav, H. Vogel, A. C. Scheck, R. Tibshirani, and B. I. Sikic. “A Network Model of a Cooperative Genetic Landscape in Brain Tumors”. In: *JAMA* 302.3 (2009-07), pp. 261–275. ISSN: 0098-7484. DOI: [10.1001/jama.2009.997](https://doi.org/10.1001/jama.2009.997). eprint: https://jamanetwork.com/journals/jama/articlepdf/184262/joc90050_261_275.pdf. URL: <https://doi.org/10.1001/jama.2009.997>.
- [20] M. Brendel, C. Su, Z. Bai, H. Zhang, O. Elemento, and F. Wang. “Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review”. In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022). AI in Omics, pp. 814–835. ISSN: 1672-0229. DOI: [10.1016/j.gpb.2022.11.011](https://doi.org/10.1016/j.gpb.2022.11.011). URL: <https://www.sciencedirect.com/science/article/pii/S1672022922001668>.
- [21] G. Riddick and H. A. Fine. “Integration and analysis of genome-scale data from gliomas”. In: *Nature Reviews Neurology* 7.8 (2011), pp. 439–450. ISSN: 1759-4766. DOI: [10.1038/nrneurol.2011.100](https://doi.org/10.1038/nrneurol.2011.100). URL: <https://doi.org/10.1038/nrneurol.2011.100>.
- [22] W.-L. Zhao, C.-H. Deng, and C.-W. Ngo. “k-means: A revisit”. In: *Neurocomputing* 291 (2018), pp. 195–206. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2018.02.072>. URL: <https://www.sciencedirect.com/science/article/pii/S092523121830239X>.
- [23] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. *Scalable K-Means++*. 2012. arXiv: [1203.6402](https://arxiv.org/abs/1203.6402) [cs.DB].
- [24] F. Li, Z. Sun, J.-X. Liu, J. Shang, L. Dai, X. Liu, and Y. Li. “NESM: a network embedding method for tumor stratification by integrating multi-omics data”. In: *G3 Genes | Genomes | Genetics* 12.11 (2022-09), jkac243. ISSN: 2160-1836. DOI: [10.1093/g3journal/jkac243](https://doi.org/10.1093/g3journal/jkac243). eprint: <https://academic.oup.com/g3journal/article-pdf/12/11/jkac243/46801325/jkac243.pdf>. URL: <https://doi.org/10.1093/g3journal/jkac243>.
- [25] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Knowledge Discovery and Data Mining*. 1996. URL: <https://api.semanticscholar.org/CorpusID:355163>.
- [26] W.-c. Wong and A. W.-c. Fu. “Incremental Document Clustering for Web Page Classification”. In: *Enabling Society with Information Technology*. Ed. by Q. Jin, J. Li, N. Zhang, J. Cheng, C. Yu, and S. Noguchi. Tokyo: Springer Japan, 2002, pp. 101–110. ISBN: 978-4-431-66979-1.
- [27] R. Guimerà, S. Mossa, A. Turtshi, and L. A. N. Amaral. “The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.22 (2005), pp. 7794–7799. DOI: [10.1073/pnas.0407994102](https://doi.org/10.1073/pnas.0407994102). URL: <https://doi.org/10.1073/pnas.0407994102>.

- [28] Y. Xu, V. Olman, and D. Xu. "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees". In: *Bioinformatics* 18.4 (2002-04), pp. 536–545. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/18.4.536](https://doi.org/10.1093/bioinformatics/18.4.536). eprint: <https://academic.oup.com/bioinformatics/article-pdf/18/4/536/48850282/bioinformatics\18\4\536.pdf>. URL: <https://doi.org/10.1093/bioinformatics/18.4.536>.
- [29] F. Boyer, A. Morgat, L. Labarre, J. Pothier, and A. Viari. "Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data". In: *Bioinformatics* 21.23 (2005-10), pp. 4209–4215. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bti711](https://doi.org/10.1093/bioinformatics/bti711). eprint: <https://academic.oup.com/bioinformatics/article-pdf/21/23/4209/50567317/bioinformatics\21\23\4209.pdf>. URL: <https://doi.org/10.1093/bioinformatics/bti711>.
- [30] Q. Yang and S. Lonardi. "A parallel algorithm for clustering protein-protein interaction networks". In: *2005 IEEE Computational Systems Bioinformatics Conference - Workshops (CSBW'05)*. 2005, pp. 174–177. DOI: [10.1109/CSBW.2005.13](https://doi.org/10.1109/CSBW.2005.13).
- [31] U. von Luxburg. "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4 (2007-12), pp. 395–416. ISSN: 1573-1375. DOI: [10.1007/s11222-007-9033-z](https://doi.org/10.1007/s11222-007-9033-z). URL: <https://doi.org/10.1007/s11222-007-9033-z>.
- [32] M. Saha. "A Graph Based Approach to Multiview Clustering". In: *Pattern Recognition and Machine Intelligence*. Ed. by P. Maji, A. Ghosh, M. N. Murty, K. Ghosh, and S. K. Pal. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 128–133. ISBN: 978-3-642-45062-4.
- [33] H. Wang, Y. Yang, B. Liu, and H. Fujita. "A study of graph-based system for multi-view clustering". In: *Knowledge-Based Systems* 163 (2019), pp. 1009–1019. ISSN: 0950-7051. DOI: <https://doi.org/10.1016/j.knosys.2018.10.022>. URL: <https://www.sciencedirect.com/science/article/pii/S0950705118305082>.
- [34] C. Hou, F. Nie, H. Tao, and D. Yi. "Multi-View Unsupervised Feature Selection with Adaptive Similarity and View Weight". In: *IEEE Transactions on Knowledge and Data Engineering* 29.9 (2017), pp. 1998–2011. DOI: [10.1109/TKDE.2017.2681670](https://doi.org/10.1109/TKDE.2017.2681670).
- [35] Y. Wang, L. Wu, X. Lin, and J. Gao. "Multiview Spectral Clustering via Structured Low-Rank Matrix Factorization". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.10 (2018), pp. 4833–4843. DOI: [10.1109/TNNLS.2017.2777489](https://doi.org/10.1109/TNNLS.2017.2777489).
- [36] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). URL: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

- [37] R. Tibshirani, G. Walther, and T. J. Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2000). URL: <https://api.semanticscholar.org/CorpusID:59738652>.
- [38] M. Hloch, M. M. Kubek, and H. Unger. "Graph-based Clustering Algorithms – A Review on Novel Approaches". In: *Advances in Science, Technology and Engineering Systems Journal* (2021). URL: <https://api.semanticscholar.org/CorpusID:238224369>.
- [39] G. Chao, S. Sun, and J. Bi. "A Survey on Multiview Clustering". In: *IEEE Transactions on Artificial Intelligence* 2.2 (2021), pp. 146–168. DOI: [10.1109/TAI.2021.3065894](https://doi.org/10.1109/TAI.2021.3065894).
- [40] D. Ramazzotti, A. Lal, B. Wang, et al. "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival". In: *Nature Communications* 9 (2018), p. 4453. DOI: [10.1038/s41467-018-06921-8](https://doi.org/10.1038/s41467-018-06921-8).
- [41] M. Pierre-Jean, J.-F. Deleuze, E. Le Floch, and F. Mauger. "Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration". In: *Briefings in Bioinformatics* 21.6 (2019-12), pp. 2011–2030. ISSN: 1477-4054. DOI: [10.1093/bib/bbz138](https://doi.org/10.1093/bib/bbz138). eprint: <https://academic.oup.com/bib/article-pdf/21/6/2011/34672159/bbz138.pdf>.
- [42] R. Shen, A. B. Olshen, and M. Ladanyi. "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22 (2009-09), pp. 2906–2912. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp543](https://doi.org/10.1093/bioinformatics/btp543). eprint: https://academic.oup.com/bioinformatics/article-pdf/25/22/2906/48997731/bioinformatics_25_22_2906.pdf. URL: <https://doi.org/10.1093/bioinformatics/btp543>.
- [43] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen. "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.11 (2013), pp. 4245–4250. DOI: [10.1073/pnas.1208949110](https://doi.org/10.1073/pnas.1208949110).
- [44] C. Meng, D. Helm, M. Frejno, and B. Kuster. "moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets". In: *Journal of Proteome Research* 15.3 (2016), pp. 755–765. DOI: [10.1021/acs.jproteome.5b00824](https://doi.org/10.1021/acs.jproteome.5b00824).
- [45] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11 (2014-03), pp. 333–337. ISSN: 1548-7105. DOI: [10.1038/nmeth.2810](https://doi.org/10.1038/nmeth.2810).

-
- [46] B. Fridley and X. Wang, eds. *Statistical Genomics*. ISBN: 978-1-0716-2986-4. New York, NY: Humana, 2023-03. DOI: [10.1007/978-1-0716-2986-4](https://doi.org/10.1007/978-1-0716-2986-4).
- [47] J. A. Westerhuis, T. Kourti, and J. F. MacGregor. "Analysis of multiblock and hierarchical PCA and PLS models". In: *Journal of Chemometrics* 12.5 (1998), pp. 301–321. DOI: [https://doi.org/10.1002/\(SICI\)1099-128X\(199809/10\)12:5<301::AID-CEM515>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S). eprint: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291099-128X%28199809/10%2912%3A5%3C301%3A%3AAID-CEM515%3E3.0.CO%3B2-S>. URL: <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-128X%28199809/10%2912%3A5%3C301%3A%3AAID-CEM515%3E3.0.CO%3B2-S>.
- [48] C. Chauvel, A. Novoloaca, P. Veyre, F. Reynier, and J. Becker. "Evaluation of integrative clustering methods for the analysis of multi-omics data". In: *Briefings in Bioinformatics* 21.2 (2019-02), pp. 541–552. ISSN: 1477-4054. DOI: [10.1093/bib/bbz015](https://doi.org/10.1093/bib/bbz015). eprint: https://academic.oup.com/bib/article-pdf/21/2/541/33582876/suppl_bbz015.pdf. URL: <https://doi.org/10.1093/bib/bbz015>.
- [49] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. "Multi-Omics Factor Analysis - A Framework for Unsupervised Integration of Multi-Omics Data Sets". In: *Molecular Systems Biology* 14.6 (2018), e8124. DOI: [10.15252/msb.20178124](https://doi.org/10.15252/msb.20178124).
- [50] M. E. Glickman and D. A. van Dyk. "Basic Bayesian Methods". In: *Topics in Biostatistics*. Ed. by W. T. Ambrosius. Totowa, NJ: Humana Press, 2007, pp. 319–338. ISBN: 978-1-59745-530-5. DOI: [10.1007/978-1-59745-530-5_16](https://doi.org/10.1007/978-1-59745-530-5_16). URL: https://doi.org/10.1007/978-1-59745-530-5_16.
- [51] R. Argelaguet, D. Arnol, D. Bredikhin, et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1 (2020), pp. 1–17. DOI: [10.1186/s13059-020-02015-1](https://doi.org/10.1186/s13059-020-02015-1).
- [52] P. Santamarina-Ojeda, J. R. Tejedor, R. F. Pérez, V. López, A. Roberti, C. Mangas, A. F. Fernández, and M. F. Fraga. "Multi-omic integration of DNA methylation and gene expression data reveals molecular vulnerabilities in glioblastoma". In: *Molecular Oncology* 17.9 (2023), pp. 1726–1743. DOI: <https://doi.org/10.1002/1878-0261.13479>. eprint: <https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1878-0261.13479>. URL: <https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1878-0261.13479>.
- [53] C. Boutsidis and E. Gallopoulos. "SVD based initialization: A head start for nonnegative matrix factorization". In: *Pattern Recognition* 41.4 (2008), pp. 1350–1362. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2007.09.010>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320307004359>.

- [54] J. A. Tropp. "Literature Survey: Non-Negative Matrix Factorization". In: 2003. URL: <https://api.semanticscholar.org/CorpusID:58269165>.
- [55] V. Raufaste-Cazavieille, R. Santiago, and A. Droit. "Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immunoncology". In: *Frontiers in Molecular Biosciences* 9 (2022). ISSN: 2296-889X. DOI: [10.3389/fmolb.2022.962743](https://doi.org/10.3389/fmolb.2022.962743). URL: <https://www.frontiersin.org/articles/10.3389/fmolb.2022.962743>.
- [56] N. Rappoport and R. Shamir. "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". In: *Nucleic Acids Research* 46.20 (2018-10), pp. 10546–10562. ISSN: 0305-1048. DOI: [10.1093/nar/gky889](https://doi.org/10.1093/nar/gky889). eprint: <https://academic.oup.com/nar/article-pdf/46/20/10546/26817354/gky889.pdf>. URL: <https://doi.org/10.1093/nar/gky889>.
- [57] S. Zhu, W. Wang, W. Fang, and M. Cui. "Autoencoder-assisted latent representation learning for survival prediction and multi-view clustering on multi-omics cancer subtyping". In: *Mathematical Biosciences and Engineering* 20.12 (2023), pp. 21098–21119. ISSN: 1551-0018. DOI: [10.3934/mbe.2023933](https://doi.org/10.3934/mbe.2023933). URL: <https://www.aimspress.com/article/doi/10.3934/mbe.2023933>.
- [58] B. Wang, J. Zhu, E. Pierson, et al. "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning". In: *Nature Methods* 14.4 (2017), pp. 414–416. DOI: [10.1038/nmeth.4207](https://doi.org/10.1038/nmeth.4207).
- [59] L. Zelnik-Manor and P. Perona. "Self-tuning spectral clustering". In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. NIPS'04. Vancouver, British Columbia, Canada: MIT Press, 2004, pp. 1601–1608.
- [60] R. T. McGibbon and V. S. Pande. "Variational cross-validation of slow dynamical modes in molecular kinetics". In: *The Journal of Chemical Physics* 142.12 (2015-03), p. 124105. ISSN: 0021-9606. DOI: [10.1063/1.4916292](https://doi.org/10.1063/1.4916292). eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4916292/14045907/124105_1_online.pdf. URL: <https://doi.org/10.1063/1.4916292>.
- [61] L. van der Maaten and G. E. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: <https://api.semanticscholar.org/CorpusID:5855042>.
- [62] P. Xanthopoulos. "A Review on Consensus Clustering Methods". In: *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*. 2014. URL: <https://api.semanticscholar.org/CorpusID:58626921>.
- [63] G. Brière, É. Darbo, P. Thébault, and R. Uricaru. "Consensus clustering applied to multi-omics disease subtyping". In: *BMC Bioinformatics* 22.1 (2021-07), p. 361. ISSN: 1471-2105. DOI: [10.1186/s12859-021-04279-1](https://doi.org/10.1186/s12859-021-04279-1). URL: <https://doi.org/10.1186/s12859-021-04279-1>.

- [64] A. Fred. "Finding Consistent Clusters in Data Partitions". In: *Multiple Classifier Systems*. Ed. by J. Kittler and F. Roli. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 309–318. ISBN: 978-3-540-48219-2.
- [65] Goder and Filkov. "Consensus Clustering Algorithms: Comparison and Refinement". In: Society for Industrial and Applied Mathematics, 2008-01. DOI: [10.1137/1.9781611972887.11](https://doi.org/10.1137/1.9781611972887.11).
- [66] S. Vega-Pons and J. Ruiz-Shulcloper. "A Survey of Clustering Ensemble Algorithms." In: *IJPRAI* 25 (2011-05), pp. 337–372. DOI: [10.1142/S0218001411008683](https://doi.org/10.1142/S0218001411008683).
- [67] R. Ghaemi, m. n. Sulaiman, and H. Ibrahim. "A Survey: Clustering Ensembles Techniques". In: vol. 50. 2009-02.
- [68] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar. "From clustering to clustering ensemble selection: A review". In: *Engineering Applications of Artificial Intelligence* 104 (2021), p. 104388. ISSN: 0952-1976. DOI: <https://doi.org/10.1016/j.engappai.2021.104388>. URL: <https://www.sciencedirect.com/science/article/pii/S0952197621002360>.
- [69] A. Lourenço, S. Rota Bulò, A. Fred, and M. Pelillo. "Consensus Clustering with Robust Evidence Accumulation". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Ed. by A. Heyden, F. Kahl, C. Olsson, M. Oskarsson, and X.-C. Tai. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 307–320. ISBN: 978-3-642-40395-8.
- [70] M. Zhang. "Weighted clustering ensemble: A review". In: *Pattern Recognition* 124 (2022-04), p. 108428. ISSN: 0031-3203. DOI: [10.1016/j.patcog.2021.108428](https://doi.org/10.1016/j.patcog.2021.108428). URL: <http://dx.doi.org/10.1016/j.patcog.2021.108428>.
- [71] M. Křivánek and J. Morávek. "NP-hard problems in hierarchical-tree clustering". In: *Acta Informatica* 23.3 (1986-06-01), pp. 311–323. ISSN: 1432-0525. DOI: [10.1007/BF00289116](https://doi.org/10.1007/BF00289116). URL: <https://doi.org/10.1007/BF00289116>.
- [72] A. L. Fred and A. K. Jain. "Combining multiple clusterings using evidence accumulation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.6 (2005), pp. 835–850. DOI: [10.1109/TPAMI.2005.113](https://doi.org/10.1109/TPAMI.2005.113).
- [73] A. Strehl and J. Ghosh. "Cluster ensembles — a knowledge reuse framework for combining multiple partitions". In: *J. Mach. Learn. Res.* 3.null (2003-03), pp. 583–617. ISSN: 1532-4435. DOI: [10.1162/153244303321897735](https://doi.org/10.1162/153244303321897735). URL: <https://doi.org/10.1162/153244303321897735>.
- [74] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". In: *Machine Learning* 52.1 (2003-07), pp. 91–118. ISSN: 1573-0565. DOI: [10.1023/A:1023949509487](https://doi.org/10.1023/A:1023949509487). URL: <https://doi.org/10.1023/A:1023949509487>.

- [75] Cancer Genome Atlas Network. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012-10), pp. 61–70. ISSN: 0028-0836. DOI: [10.1038/nature11412](https://doi.org/10.1038/nature11412). URL: <https://europepmc.org/articles/PMC3465532>.
- [76] A. Cabassi and P. D. W. Kirk. “Multiple kernel learning for integrative consensus clustering of omic datasets”. In: *Bioinformatics* 36.18 (2020-09), pp. 4789–4796. DOI: [10.1093/bioinformatics/btaa593](https://doi.org/10.1093/bioinformatics/btaa593).
- [77] S. Mitra and S. Saha. “A multiobjective multi-view cluster ensemble technique: Application in patient subclassification”. In: *PLOS ONE* 14.5 (2019-05), pp. 1–30. DOI: [10.1371/journal.pone.0216904](https://doi.org/10.1371/journal.pone.0216904). URL: <https://doi.org/10.1371/journal.pone.0216904>.
- [78] S. Vega-Pons, J. Correa-Morris, and J. Ruiz-Shulcloper. “Weighted Cluster Ensemble Using a Kernel Consensus Function”. In: 2008-09, pp. 195–202. ISBN: 978-3-540-85919-2. DOI: [10.1007/978-3-540-85920-8_24](https://doi.org/10.1007/978-3-540-85920-8_24).
- [79] A. Lourenço, S. Rota Buló, N. Rebagliati, et al. “Probabilistic consensus clustering using evidence accumulation”. In: *Machine Learning* 98 (2015), pp. 331–357. DOI: [10.1007/s10994-013-5339-6](https://doi.org/10.1007/s10994-013-5339-6).
- [80] H. Xiong and Z. Li. “Clustering Validation Measures”. In: *Data Clustering*. Ed. by C. C. Aggarwal and C. K. Reddy. Chapman and Hall/CRC, 2014, pp. 571–605. ISBN: 9781466558212.
- [81] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty. “Model-based evaluation of clustering validation measures”. In: *Pattern Recognition* 40.3 (2007), pp. 807–824. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2006.06.026>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320306003104>.
- [82] N. X. Vinh, J. Epps, and J. Bailey. “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”. In: *J. Mach. Learn. Res.* 11 (2010-12), pp. 2837–2854. ISSN: 1532-4435.
- [83] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. “On Clustering Validation Techniques”. In: *Journal of Intelligent Information Systems* 17.2 (2001-12), pp. 107–145. ISSN: 1573-7675. DOI: [10.1023/A:1012801612483](https://doi.org/10.1023/A:1012801612483). URL: <https://doi.org/10.1023/A:1012801612483>.
- [84] B. Tavakkol, M. K. Jeong, and S. L. Albin. “Validity indices for clusters of uncertain data objects”. In: *Annals of Operations Research* 303.1 (2021-08), pp. 321–357. ISSN: 1572-9338. DOI: [10.1007/s10479-018-3043-4](https://doi.org/10.1007/s10479-018-3043-4). URL: <https://doi.org/10.1007/s10479-018-3043-4>.
- [85] T. Caliński and H. JA. “A Dendrite Method for Cluster Analysis”. In: *Communications in Statistics - Theory and Methods* 3 (1974-01), pp. 1–27. DOI: [10.1080/03610927408827101](https://doi.org/10.1080/03610927408827101).

-
- [86] R. Ünlü and P. Xanthopoulos. "Estimating the number of clusters in a dataset via consensus clustering". In: *Expert Systems with Applications* 125 (2019), pp. 33–39. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2019.01.074>. URL: <https://www.sciencedirect.com/science/article/pii/S0957417419300892>.
 - [87] C. Hennig and T. F. Liao. "How to Find an Appropriate Clustering for Mixed-Type Variables with Application to Socio-Economic Stratification". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 62.3 (2013-04), pp. 309–369. ISSN: 0035-9254. DOI: [10.1111/j.1467-9876.2012.01066.x](https://doi.org/10.1111/j.1467-9876.2012.01066.x). eprint: https://academic.oup.com/jrssc/article-pdf/62/3/309/49562625/jrssc_62_3_309.pdf. URL: <https://doi.org/10.1111/j.1467-9876.2012.01066.x>.
 - [88] D. Steinley. "Properties of the Hubert-Arabie adjusted Rand index". In: *Psychological Methods* 9.3 (2004), pp. 386–396. DOI: [10.1037/1082-989X.9.3.386](https://doi.org/10.1037/1082-989X.9.3.386).
 - [89] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. ISSN: 1432-1343. DOI: [10.1007/BF01908075](https://doi.org/10.1007/BF01908075). URL: <https://doi.org/10.1007/BF01908075>.
 - [90] N. C. Institute. *The Cancer Genome Atlas (TCGA)*. Accessed: 2024-09-18. 2024. URL: <https://www.cancer.gov/ccg/research/genome-sequencing/tcga>.
 - [91] T. C. G. A. R. Network. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". In: *Nature* 455 (2008), pp. 1061–1068. DOI: [10.1038/nature07385](https://doi.org/10.1038/nature07385).
 - [92] L. P. Chaves, C. M. Melo, W. Lautert-Dutra, A. L. Caliari, and J. A. Squire. "Transcriptomics and Immune Response in Human Cancer". In: *Transcriptomics in Health and Disease*. Ed. by G. A. Passos. Cham: Springer International Publishing, 2022, pp. 299–315. ISBN: 978-3-030-87821-4. DOI: [10.1007/978-3-030-87821-4_13](https://doi.org/10.1007/978-3-030-87821-4_13). URL: https://doi.org/10.1007/978-3-030-87821-4_13.
 - [93] M. Kulis and M. Esteller. "2 - DNA Methylation and Cancer". In: *Epigenetics and Cancer, Part A*. Ed. by Z. Herceg and T. Ushijima. Vol. 70. Advances in Genetics. Academic Press, 2010, pp. 27–56. DOI: <https://doi.org/10.1016/B978-0-12-380866-0.60002-2>. URL: <https://www.sciencedirect.com/science/article/pii/B9780123808660600022>.
 - [94] S. Sarumathi, N. Shanthi, S. Vidhya, and M. Sharmila. "A Comprehensive Review on Different Mixed Data Clustering Ensemble Methods". In: *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 8 (2014), pp. 1456–1466. URL: <https://api.semanticscholar.org/CorpusID:27706879>.
 - [95] L. Kuncheva, S. Hadjitodorov, and L. Todorova. "Experimental Comparison of Cluster Ensemble Methods". In: *2006 9th International Conference on Information Fusion*. 2006, pp. 1–7. DOI: [10.1109/ICIF.2006.301614](https://doi.org/10.1109/ICIF.2006.301614).

- [96] T. Boongoen and N. Iam-On. "Cluster ensembles: A survey of approaches with recent extensions and applications". In: *Computer Science Review* 28 (2018), pp. 1–25. ISSN: 1574-0137. DOI: <https://doi.org/10.1016/j.cosrev.2018.01.003>. URL: <https://www.sciencedirect.com/science/article/pii/S1574013717300692>.
- [97] W. Song, W. Wang, and D.-Q. Dai. "Subtype-WESLR: identifying cancer subtype with weighted ensemble sparse latent representation of multi-view data". In: *Briefings in Bioinformatics* 23.1 (2021-10), bbab398. ISSN: 1477-4054. DOI: [10.1093/bib/bbab398](https://doi.org/10.1093/bib/bbab398). eprint: https://academic.oup.com/bib/article-pdf/23/1/bbab398/42230823/supplementary_material_bbab398.pdf. URL: <https://doi.org/10.1093/bib/bbab398>.
- [98] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim. "A comparative study of clustering ensemble algorithms". In: *Computers & Electrical Engineering* 68 (2018), pp. 603–615. ISSN: 0045-7906. DOI: <https://doi.org/10.1016/j.compeleceng.2018.05.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0045790617325417>.
- [99] M. Jason T. Rich, M. J. Gail Neely, M. Randal C. Paniello, M. D. Courtney C. J. Voelker, M. Brian Nussenbaum, and M. Eric W. Wang. "A practical guide to understanding Kaplan-Meier curves". In: *Otolaryngology–Head and Neck Surgery* 143.3 (2010). PMID: 20723767, pp. 331–336. DOI: [10.1016/j.otohns.2010.05.007](https://doi.org/10.1016/j.otohns.2010.05.007). eprint: <https://doi.org/10.1016/j.otohns.2010.05.007>. URL: <https://doi.org/10.1016/j.otohns.2010.05.007>.
- [100] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. "Survival Analysis Part I: Basic concepts and first analyses". In: *British Journal of Cancer* 89.2 (2003), pp. 232–238. ISSN: 1532-1827. DOI: [10.1038/sj.bjc.6601118](https://doi.org/10.1038/sj.bjc.6601118). URL: <https://doi.org/10.1038/sj.bjc.6601118>.
- [101] P. Schober and T. R. Vetter. "Kaplan-Meier Curves, Log-Rank Tests, and Cox Regression for Time-to-Event Data". In: *Anesthesia & Analgesia* 132.4 (2021), pp. 969–970. DOI: [10.1213/ANE.0000000000005358](https://doi.org/10.1213/ANE.0000000000005358).
- [102] P. Chalise. *intNMF*. R package version 1.2.0. 2018. URL: <https://www.rdocumentation.org/packages/IntNMF/versions/1.2.0>.
- [103] C. Meng. *mogsa: Multiple omics data integrative clustering and gene set analysis*. R package version 1.38.0. 2024. DOI: [10.18129/B9.bioc.mogsa](https://doi.org/10.18129/B9.bioc.mogsa). URL: <https://bioconductor.org/packages/mogsa>.
- [104] L. Kaufman and P. Rousseeuw. "Partitioning Around Medoids (Program PAM)". In: *Finding Groups in Data*. John Wiley & Sons, Ltd, 1990. Chap. 2, pp. 68–125. ISBN: 9780470316801. DOI: <https://doi.org/10.1002/9780470316801.ch2>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316801.ch2>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2>.

- [105] A. Bourgonje, K. Verrijp, J. Schepens, A. Navis, J. Piepers, C. Palmen, M. van den Eijnden, R. Hooft van Huijsduijnen, P. Wesseling, W. Leenders, and W. Hendriks. “Comprehensive protein tyrosine phosphatase mRNA profiling identifies new regulators in the progression of glioma”. In: *Acta Neuropathologica Communications* 4.1 (2016), p. 96. DOI: [10.1186/s40478-016-0372-x](https://doi.org/10.1186/s40478-016-0372-x).
- [106] E. J. Lee, P. Rath, J. Liu, D. Ryu, L. Pei, S. K. Noonepalle, A. Y. Shull, Q. Feng, N. S. Litofsky, D. C. Miller, D. C. Anthony, M. D. Kirk, J. Laterra, L. Deng, H.-B. Xin, X. Wang, J.-H. Choi, and H. Shi. “Identification of Global DNA Methylation Signatures in Glioblastoma-Derived Cancer Stem Cells”. In: *Journal of Genetics and Genomics* 42.7 (2015), pp. 355–371. DOI: [10.1016/j.jgg.2015.06.003](https://doi.org/10.1016/j.jgg.2015.06.003).
- [107] Q. Shi, G. Song, L. Song, Y. Wang, J. Ma, L. Zhang, and E. Yuan. “Unravelling the function of prdm16 in human tumours: A comparative analysis of haematologic and solid tumours”. In: *Biomedicine & Pharmacotherapy* 178 (2024), p. 117281. ISSN: 0753-3322. DOI: <https://doi.org/10.1016/j.biopha.2024.117281>. URL: <https://www.sciencedirect.com/science/article/pii/S075333222401165X>.
- [108] S. Liot, A. Aubert, V. Hervieu, N. E. Kholti, J. Schalkwijk, B. Verrier, U. Valcourt, and E. Lambert. “Loss of Tenascin-X expression during tumor progression: A new pan-cancer marker”. In: *Matrix Biology Plus* 6-7 (2020). Complexity of Matrix Phenotypes, p. 100021. ISSN: 2590-0285. DOI: <https://doi.org/10.1016/j.mbplus.2020.100021>. URL: <https://www.sciencedirect.com/science/article/pii/S2590028520300028>.
- [109] L. Bie, G. Zhao, P. Cheng, G. Rondeau, S. Porwollik, Y. Ju, X. Q. Xia, and M. McClelland. “The accuracy of survival time prediction for patients with glioma is improved by measuring mitotic spindle checkpoint gene expression”. In: *PLoS One* 6.10 (2011), e25631. DOI: [10.1371/journal.pone.0025631](https://doi.org/10.1371/journal.pone.0025631).
- [110] T. Tao, C. Cheng, Y. Ji, G. Xu, J. Zhang, L. Zhang, and A. Shen. “Numbl inhibits glioma cell migration and invasion by suppressing TRAF5-mediated NF-KB activation”. In: *Molecular Biology of the Cell* 23.14 (2012). PMID: 22593207, pp. 2635–2644. DOI: [10.1091/mbc.e11-09-0805](https://doi.org/10.1091/mbc.e11-09-0805). eprint: <https://doi.org/10.1091/mbc.e11-09-0805>. URL: <https://doi.org/10.1091/mbc.e11-09-0805>.
- [111] G. P. Takacs, J. S. Garcia, C. A. Hodges, C. J. Kreiger, A. Sherman, and J. K. Harrison. “CSF1R Ligands Expressed by Murine Gliomas Promote M-MDSCs to Suppress CD8+ T Cells in a NOS-Dependent Manner”. In: *Cancers* 16.17 (2024). ISSN: 2072-6694. DOI: [10.3390/cancers16173055](https://doi.org/10.3390/cancers16173055). URL: <https://www.mdpi.com/2072-6694/16/17/3055>.
- [112] G. A. Herrera-Quiterio and S. Encarnación-Guevara. “The transmembrane proteins (TMEM) and their role in cell proliferation, migration, invasion, and epithelial-mesenchymal transition in cancer”. In: *Frontiers in Oncology* 13 (2023). ISSN:

- 2234-943X. DOI: [10.3389/fonc.2023.1244740](https://doi.org/10.3389/fonc.2023.1244740). URL: <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1244740>.
- [113] F. G. Vieira, R. Bispo, and M. B. Lopes. "Integration of Multi-Omics Data for the Classification of Glioma Types and Identification of Novel Biomarkers". In: *Bioinformatics and Biology Insights* 18 (2024). PMID: 38737722, p. 11779322241249563. DOI: [10.1177/11779322241249563](https://doi.org/10.1177/11779322241249563). eprint: <https://doi.org/10.1177/11779322241249563>. URL: <https://doi.org/10.1177/11779322241249563>.
- [114] A. Ardizzone, S. A. Scuderi, D. Giuffrida, C. Colarossi, C. Puglisi, M. Campolo, S. Cuzzocrea, E. Esposito, and I. Paterniti. "Role of Fibroblast Growth Factors Receptors (FGFRs) in Brain Tumors, Focus on Astrocytoma and Glioblastoma". In: *Cancers* 12.12 (2020). ISSN: 2072-6694. DOI: [10.3390/cancers12123825](https://doi.org/10.3390/cancers12123825). URL: <https://www.mdpi.com/2072-6694/12/12/3825>.
- [115] J. Wang, J. Li, R. Chen, et al. "DNA methylation-based profiling reveals distinct clusters with survival heterogeneity in high-grade serous ovarian cancer". In: *Clinical Epigenetics* 13 (2021), p. 190. DOI: [10.1186/s13148-021-01178-3](https://doi.org/10.1186/s13148-021-01178-3).
- [116] K. Stoltz, M. Sinyuk, J. S. Hale, Q. Wu, B. Otvos, K. Walker, A. Vasanthi, J. N. Rich, A. B. Hjelmeland, and J. D. Lathia. "Development of a Sox2 reporter system modeling cellular heterogeneity in glioma". In: *Neuro-Oncology* 17.3 (2014-11), pp. 361–371. ISSN: 1522-8517. DOI: [10.1093/neuonc/nou320](https://doi.org/10.1093/neuonc/nou320). eprint: <https://academic.oup.com/neuro-oncology/article-pdf/17/3/361/17431103/nou320.pdf>. URL: <https://doi.org/10.1093/neuonc/nou320>.
- [117] B. H. Kim, Y. J. Kim, M.-H. Kim, Y. R. Na, D. Jung, S. H. Seok, J. Kim, and H. J. Kim. "Identification of FES as a Novel Radiosensitizing Target in Human Cancers". In: *Clinical Cancer Research* 26.1 (2020-01), pp. 265–273. ISSN: 1078-0432. DOI: [10.1158/1078-0432.CCR-19-0610](https://doi.org/10.1158/1078-0432.CCR-19-0610). eprint: <https://aacrjournals.org/clincancerres/article-pdf/26/1/265/2058971/265.pdf>. URL: <https://doi.org/10.1158/1078-0432.CCR-19-0610>.
- [118] W. Hou, W. Hou, and X. Zhao. "Exploring transcriptomic databases: unraveling circadian gene disruptions in lower grade glioma". In: *Scientific Reports* 14 (2024), p. 16960. DOI: [10.1038/s41598-024-67559-9](https://doi.org/10.1038/s41598-024-67559-9).
- [119] R. Chai, M. Liao, L. Ou, Q. Tang, Y. Liang, N. Li, W. Huang, X. Wang, K. Zheng, and S. Wang. "Circadian Clock Genes Act as Diagnostic and Prognostic Biomarkers of Glioma: Clinic Implications for Chronotherapy". In: *BioMed Research International* 2022.1 (2022), p. 9774879. DOI: <https://doi.org/10.1155/2022/9774879>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/9774879>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/9774879>.
- [120] H.-Y. Wang, W. Wang, Y.-W. Liu, et al. "Role of KCNB1 in the prognosis of gliomas and autophagy modulation". In: *Scientific Reports* 7 (2017), p. 14. DOI: [10.1038/s41598-017-00045-7](https://doi.org/10.1038/s41598-017-00045-7).

- [121] M. Tu, L. Ye, S. Hu, W. Wang, P. Zhu, X. Lu, and W. Zheng. "Identification of Glioma Specific Genes as Diagnostic and Prognostic Markers for Glioma". In: *Current Bioinformatics* 16.1 (2021), pp. 120–129. DOI: [10.2174/1574893615999200424090954](https://doi.org/10.2174/1574893615999200424090954).
- [122] H. Wang, Y. Zhang, C. Zhang, and X. Tong. "Screening prognosis-related genes in glioblastoma by microarray analysis". In: *Anti-Cancer Drugs* 31.2 (2020), pp. 150–157. DOI: [10.1097/CAD.0000000000000839](https://doi.org/10.1097/CAD.0000000000000839).
- [123] X. Li, Z. Huang, L. Zhu, F. Yu, M. Feng, A. Gu, and et al. "Prognostic Model and Nomogram Construction and Validation With an Autophagy-Related Gene Signature in Low-Grade Gliomas". In: *Frontiers in Genetics* 13 (2022). ISSN: 1664-8021. DOI: [10.3389/fgene.2022.905751](https://doi.org/10.3389/fgene.2022.905751). URL: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.905751>.
- [124] Q. Qiao, Y. Wang, R. Zhang, and Q. Pang. "Autophagy related DNA methylation signature predict clinical prognosis and immune microenvironment in low-grade glioma". In: *Translational Cancer Research* 11.7 (2022), pp. 2157–2174. DOI: [10.21037/tcr-22-310](https://doi.org/10.21037/tcr-22-310).
- [125] C. Yan, L. Su, X. Qi, Y. Yang, and J. Wu. "Oncogenic involvement of NF-κB signaling pathway in glioma". In: *Cancer Research* 80.16_Supplement (2020-08), pp. 4735–4735. DOI: [10.1158/1538-7445.AM2020-4735](https://doi.org/10.1158/1538-7445.AM2020-4735). URL: <https://doi.org/10.1158/1538-7445.AM2020-4735>.
- [126] K. Hezam, J. Jiang, F. Sun, X. Zhang, and J. Zhang. In: *Reviews in the Neurosciences* 29.1 (2018), pp. 93–98. DOI: [doi:10.1515/revneuro-2017-0029](https://doi.org/10.1515/revneuro-2017-0029). URL: <https://doi.org/10.1515/revneuro-2017-0029>.
- [127] G. Manica, C. Ribeiro, M. Oliveira, et al. "Down regulation of ADAM33 as a Predictive Biomarker of Aggressive Breast Cancer". In: *Scientific Reports* 7 (2017), p. 44414. DOI: [10.1038/srep44414](https://doi.org/10.1038/srep44414).
- [128] K.-E. Kim, H. Song, C. Hahm, S. Y. Yoon, S. Park, H.-r. Lee, and et al. "Expression of ADAM33 Is a Novel Regulatory Mechanism in IL-18-Secreted Process in Gastric Cancer1". In: *The Journal of Immunology* 182.6 (2009-03), pp. 3548–3555. ISSN: 0022-1767. DOI: [10.4049/jimmunol.0801695](https://doi.org/10.4049/jimmunol.0801695). eprint: <https://journals.aai.org/jimmunol/article-pdf/182/6/3548/1277951/zim00609003548.pdf>. URL: <https://doi.org/10.4049/jimmunol.0801695>.
- [129] B. Zhao, R. Xuan, G. Yang, et al. "A novel Golgi related genes based correlation prognostic index can better predict the prognosis of glioma and responses to immunotherapy". In: *Research Square Preprint* (2024). Version 1. DOI: [10.21203/rs.3.rs-3961611/v1](https://doi.org/10.21203/rs.3.rs-3961611/v1). URL: <https://doi.org/10.21203/rs.3.rs-3961611/v1>.
- [130] X. Wang, X. Wang, J. Li, et al. "PDPN contributes to constructing immunosuppressive microenvironment in IDH wildtype glioma". In: *Cancer Gene Therapy* 30 (2023), pp. 345–357. DOI: [10.1038/s41417-022-00550-6](https://doi.org/10.1038/s41417-022-00550-6).

- [131] E.-A. Ko and T. Zhou. "GPCR genes as a predictor of glioma severity and clinical outcome". In: *Journal of International Medical Research* 50.7 (2022). PMID: 35903880, p. 03000605221113911. DOI: [10.1177/03000605221113911](https://doi.org/10.1177/03000605221113911). eprint: <https://doi.org/10.1177/03000605221113911>. URL: <https://doi.org/10.1177/03000605221113911>.
- [132] J. Korbecki, K. Kojder, D. Jeżewski, D. Simińska, P. Tomasiak, and et al. "Reduced Expression of Very-Long-Chain Acyl-CoA Synthetases SLC27A4 and SLC27A6 in the Glioblastoma Tumor Compared to the Peritumoral Area". In: *Brain Sciences* 13.5 (2023). ISSN: 2076-3425. DOI: [10.3390/brainsci13050771](https://www.mdpi.com/2076-3425/13/5/771). URL: <https://www.mdpi.com/2076-3425/13/5/771>.
- [133] S. Sun, C. Yang, K. Wang, et al. "Molecular and clinical characterization of PTRF in glioma via 1,022 samples". In: *BMC Cancer* 23 (2023), p. 551. DOI: [10.1186/s12885-023-11001-2](https://doi.org/10.1186/s12885-023-11001-2).
- [134] W. Xie, J. Zhang, P. Zhong, et al. "Expression and potential prognostic value of histone family gene signature in breast cancer". In: *Experimental and Therapeutic Medicine* 18.6 (2019), pp. 4893–4903. DOI: [10.3892/etm.2019.8129](https://doi.org/10.3892/etm.2019.8129).
- [135] W. Chen, W. Qian, J. Nie, et al. "A study of the prognostic value of long non-coding RNA CASC15 in human solid tumors utilizing The Cancer Genome Atlas (TCGA) datasets and a meta-analysis". In: *Clinical and Experimental Medicine* 23 (2023), pp. 65–78. DOI: [10.1007/s10238-021-00789-7](https://doi.org/10.1007/s10238-021-00789-7).
- [136] J. Ji, R. Xu, K. Ding, G. Bao, X. Zhang, B. Huang, X. Wang, and e. a. Martinez. "Long Noncoding RNA SCHLAP1 Forms a Growth-Promoting Complex with HNRNPL in Human Glioblastoma through Stabilization of ACTN4 and Activation of NF-KB Signaling". In: *Clinical Cancer Research* 25.22 (2019-11), pp. 6868–6881. ISSN: 1078-0432. DOI: [10.1158/1078-0432.CCR-19-0747](https://doi.org/10.1158/1078-0432.CCR-19-0747). eprint: <https://aacrjournals.org/clincancerres/article-pdf/25/22/6868/2056606/6868.pdf>. URL: <https://doi.org/10.1158/1078-0432.CCR-19-0747>.
- [137] K. Purushothaman, E. Sivasankar, M. Krishnamoorthy, et al. "Computational identification of potential tau tubulin kinase 1 (TTBK1) inhibitors: a structural analog approach". In: *In Silico Pharmacology* 12 (2024), p. 66. DOI: [10.1007/s40203-024-00242-z](https://doi.org/10.1007/s40203-024-00242-z).
- [138] M. A. Bustos, N. Rahimzadeh, R. G. Suyeon Ryu, L. T. Tran, V. M. Renteria-Lopez, R. I. Ramos, and et al. "Cell-free plasma microRNAs that identify patients with glioblastoma". In: *Laboratory Investigation* 102.7 (2022), pp. 711–721. ISSN: 0023-6837. DOI: <https://doi.org/10.1038/s41374-021-00720-4>. URL: <https://www.sciencedirect.com/science/article/pii/S0023683722002069>.
- [139] Z. Chai, H. Fan, Y. Li, et al. "miR-1908 as a novel prognosis marker of glioma via promoting malignant phenotype and modulating SPRY4/RAF1 axis". In: *Oncology Reports* 38.5 (2017), pp. 2717–2726. DOI: [10.3892/or.2017.6003](https://doi.org/10.3892/or.2017.6003).

-
- [140] A. G. Asuero, A. Sayago, and A. G. González. “The Correlation Coefficient: An Overview”. In: *Critical Reviews in Analytical Chemistry* 36 (2006), pp. 41–59. URL: <https://api.semanticscholar.org/CorpusID:39213560>.
- [141] P. Zhang, F. Cao, J. Zhang, Y. Tan, and S. Yao. “Temozolomide and chloroquine co-loaded mesoporous silica nanoparticles are effective against glioma”. In: *Heliyon* 9.8 (2023). DOI: [10.1016/j.heliyon.2023.e18490](https://doi.org/10.1016/j.heliyon.2023.e18490). URL: <https://doi.org/10.1016/j.heliyon.2023.e18490>.
- [142] H. Hu, Z. Wang, M. Li, et al. “Gene Expression and Methylation Analyses Suggest DCTD as a Prognostic Factor in Malignant Glioma”. In: *Scientific Reports* 7 (2017), p. 11568. DOI: [10.1038/s41598-017-11962-y](https://doi.org/10.1038/s41598-017-11962-y).
- [143] W. Wu, C. Jiang, W. Zhu, and X. Jiang. “Multi-omics analysis reveals the association between specific solute carrier proteins gene expression patterns and the immune suppressive microenvironment in glioma”. In: *Journal of Cellular and Molecular Medicine* 28.9 (2024). e18339 JCOMM-01-2024-076.R1, e18339. DOI: <https://doi.org/10.1111/jcmm.18339>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcmm.18339>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jcmm.18339>.
- [144] M. Zhang, X. Lv, Y. Jiang, G. Li, and Q. Qiao. “Identification of aberrantly methylated differentially expressed genes in glioblastoma multiforme and their association with patient survival”. In: *Experimental and Therapeutic Medicine* 18.3 (2019), pp. 2140–2152. DOI: [10.3892/etm.2019.7807](https://doi.org/10.3892/etm.2019.7807).
- [145] M. Han, R. Xu, Y. Xu, et al. “TAGLN2 is a candidate prognostic biomarker promoting tumorigenesis in human gliomas”. In: *Journal of Experimental & Clinical Cancer Research* 36 (2017), p. 155. DOI: [10.1186/s13046-017-0619-9](https://doi.org/10.1186/s13046-017-0619-9).
- [146] Y.-F. Gao, T. Zhu, C.-X. Mao, Z.-X. Liu, Z.-B. Wang, X.-Y. Mao, L. Li, J.-Y. Yin, H.-H. Zhou, and Z.-Q. Liu. “PPIC, EMP3 and CHI3L1 Are Novel Prognostic Markers for High Grade Glioma”. In: *International Journal of Molecular Sciences* 17.11 (2016). ISSN: 1422-0067. DOI: [10.3390/ijms17111808](https://doi.org/10.3390/ijms17111808). URL: <https://www.mdpi.com/1422-0067/17/11/1808>.
- [147] L. Li, S. Xia, Z. Zhao, et al. “EMP3 as a prognostic biomarker correlates with EMT in GBM”. In: *BMC Cancer* 24 (2024), p. 89. DOI: [10.1186/s12885-023-11796-0](https://doi.org/10.1186/s12885-023-11796-0).
- [148] A. A. Martija and S. Pusch. “The Multifunctional Role of EMP3 in the Regulation of Membrane Receptors Associated with IDH-Wild-Type Glioblastoma”. In: *International Journal of Molecular Sciences* 22.10 (2021). ISSN: 1422-0067. DOI: [10.3390/ijms22105261](https://doi.org/10.3390/ijms22105261). URL: <https://www.mdpi.com/1422-0067/22/10/5261>.
- [149] C. Wang and Z. He. “Multi-omics analysis reveals CLIC1 as a therapeutic vulnerability of gliomas”. In: *Frontiers in Pharmacology* 14 (2023). ISSN: 1663-9812. DOI: [10.3389/fphar.2023.1279370](https://doi.org/10.3389/fphar.2023.1279370). URL: <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1279370>.

- [150] Z. Tian, W. Jia, Z. Wang, et al. "Clinical significance of immune-related antigen CD58 in gliomas and analysis of its potential core related gene clusters". In: *Heliyon* 10.8 (2024). DOI: [10.1016/j.heliyon.2024.e29275](https://doi.org/10.1016/j.heliyon.2024.e29275). URL: <https://doi.org/10.1016/j.heliyon.2024.e29275>.
- [151] X. Zheng, Y. Wang, D. Wang, et al. "PSMC2 is overexpressed in glioma and promotes proliferation and anti-apoptosis of glioma cells". In: *World Journal of Surgical Oncology* 20 (2022), p. 84. DOI: [10.1186/s12957-022-02533-1](https://doi.org/10.1186/s12957-022-02533-1).
- [152] Y. Wang, S. Zhang, Z. Zhao, Q. Jin, Z. Wang, Z. Song, L. Liu, and Z. Zhao. "PSMC2 promotes glioma progression by regulating immune microenvironment and PI3K/AKT/mTOR pathway". In: *Immunobiology* 229.3 (2024), p. 152802. ISSN: 0171-2985. DOI: <https://doi.org/10.1016/j.imbio.2024.152802>. URL: <https://www.sciencedirect.com/science/article/pii/S0171298524000202>.
- [153] W.-J. Zeng, Y.-L. Yang, Z.-Z. Liu, Z.-P. Wen, Y.-H. Chen, X.-L. Hu, Q. Cheng, J. Xiao, J. Zhao, and X.-P. Chen. "Integrative Analysis of DNA Methylation and Gene Expression Identify a Three-Gene Signature for Predicting Prognosis in Lower-Grade Gliomas". In: *Cellular Physiology and Biochemistry* 47.1 (2018-05), pp. 428–439. ISSN: 1015-8987. DOI: [10.1159/000489954](https://doi.org/10.1159/000489954). eprint: <https://karger.com/cpb/article-pdf/47/1/428/2452454/000489954.pdf>. URL: <https://doi.org/10.1159/000489954>.
- [154] K. A. Radaszkiewicz, M. Sulcova, E. Kohoutkova, et al. "The role of prickle proteins in vertebrate development and pathology". In: *Molecular and Cellular Biochemistry* 479 (2024), pp. 1199–1221. DOI: [10.1007/s11010-023-04787-z](https://doi.org/10.1007/s11010-023-04787-z).
- [155] E. E. Cilek, H. Ozturk, and B. G. Dedeoglu. "Construction of miRNA-miRNA networks revealing the complexity of miRNA-mediated mechanisms in trastuzumab treated breast cancer cell lines". In: *PLOS ONE* 12.10 (2017), e0185558. DOI: [10.1371/journal.pone.0185558](https://doi.org/10.1371/journal.pone.0185558).
- [156] Q. Ye, R. Raese, D. Luo, S. Cao, Y.-W. Wan, Y. Qian, and N. L. Guo. "MicroRNA, mRNA, and Proteomics Biomarkers and Therapeutic Targets for Improving Lung Cancer Treatment Outcomes". In: *Cancers* 15.8 (2023), p. 2294. DOI: [10.3390/cancers15082294](https://doi.org/10.3390/cancers15082294).
- [157] R. Coletti, M. L. de Mendonça, S. Vinga, and M. B. Lopes. "Inferring Diagnostic and Prognostic Gene Expression Signatures Across WHO Glioma Classifications: A Network-Based Approach". In: *Bioinformatics and Biology Insights* 18 (2024), p. 11779322241271535. DOI: [10.1177/11779322241271535](https://doi.org/10.1177/11779322241271535). eprint: <https://doi.org/10.1177/11779322241271535>. URL: <https://doi.org/10.1177/11779322241271535>.

- [158] C. Zhang and W. Zhou. "Machine learning-based identification of glycosyltransferase-related mRNAs for improving outcomes and the anti-tumor therapeutic response of gliomas". In: *Frontiers in Pharmacology* 14 (2023). ISSN: 1663-9812. DOI: [10.3389/fphar.2023.1200795](https://doi.org/10.3389/fphar.2023.1200795). URL: <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2023.1200795>.
- [159] R. Wang, Y. Wu, J. Yu, G. Yang, H. Yi, and B. Xu. "Plasma Messenger RNAs Identified Through Bioinformatics Analysis are Novel, Non-Invasive Prostate Cancer Biomarkers". In: *OncoTargets and Therapy* 13 (2020), pp. 541–548. DOI: [10.2147/OTT.S221276](https://doi.org/10.2147/OTT.S221276).
- [160] M. S. Siddiqui, M. H. Shahi, and J. S. Castresana. "The role of the adenylate kinase 5 gene in various diseases and cancer". In: *Journal of Clinical and Translational Science* 8.1 (2024), e96. DOI: [10.1017/cts.2024.536](https://doi.org/10.1017/cts.2024.536).
- [161] J. Tang, D. He, P. Yang, et al. "Genome-wide expression profiling of glioblastoma using a large combined cohort". In: *Scientific Reports* 8 (2018), p. 15104. DOI: [10.1038/s41598-018-33323-z](https://doi.org/10.1038/s41598-018-33323-z).
- [162] F. Xie, D. Zhang, X. Qian, H. Wei, L. Zhou, C. Ding, W. Pan, and Q. Ye. "Analysis of cancer-promoting genes related to chemotherapy resistance in esophageal squamous cell carcinoma". In: *Annals of Translational Medicine* 10.2 (2022), p. 92. DOI: [10.21037/atm-21-7032](https://doi.org/10.21037/atm-21-7032).
- [163] J. Luo, Q. Hu, M. Gou, X. Liu, Y. Qin, J. Zhu, C. Cai, T. Tian, Z. Tu, Y. Du, and H. Deng. "Expression of Microtubule-Associated Proteins in Relation to Prognosis and Efficacy of Immunotherapy in Non-Small Cell Lung Cancer". In: *Frontiers in Oncology* 11 (2021). ISSN: 2234-943X. DOI: [10.3389/fonc.2021.680402](https://doi.org/10.3389/fonc.2021.680402). URL: <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.680402>.
- [164] D. B. Hoelzinger, L. Mariani, J. Weis, T. Woyke, T. J. Berens, W. McDonough, A. Sloan, S. W. Coons, and M. E. Berens. "Gene Expression Profile of Glioblastoma Multiforme Invasive Phenotype Points to New Therapeutic Targets". In: *Neoplasia* 7.1 (2005), pp. 7–16. ISSN: 1476-5586. DOI: <https://doi.org/10.1593/neo.04535>. URL: <https://www.sciencedirect.com/science/article/pii/S1476558605800946>.
- [165] Q. Yi, T. Chen, K. Zhou, Q. Ma, Z. Sun, and H. Shi. "PAK1 Inhibition Suppresses the Proliferation, Migration and Invasion of Glioma Cells". In: *Current Protein and Peptide Science* 24.2 (2023), pp. 178–189. ISSN: 1389-2037. DOI: [doi:10.2174/1389203724666221226150329](https://doi.org/10.2174/1389203724666221226150329). URL: <https://www.ingentaconnect.com/content/ben/cpps/2023/00000024/00000002/art00006>.

A.1 Parameter Tuning

In this section, the number of neighbors rangin from 30 to 50 and the corresponding cluster validation measures for the CIMLR method with 3 and 4 clusters are present in [Table A.1](#) and [Table A.2](#), respectively.

Table A.1: Tuning number of neighbors for the CIMLR method with 3 clusters.

Tuning Parameter	ARI	NMI	CHI	DI	ASW
30	0.619	0.688	2,447	0.791	0.827
31	0.623	0.690	25,389	3.625	0.920
32	0.623	0.690	18,344	3.239	0.917
33	0.623	0.690	27,480	4.459	0.928
34	0.623	0.690	2,442	0.886	0.830
35	0.619	0.688	29,208	4.467	0.934
36	0.597	0.596	42,815	5.680	0.940
37	0.754	0.727	38,646	5.060	0.935
38	0.623	0.685	2,121	0.655	0.815
39	0.602	0.599	31,847	5.243	0.934
40	0.602	0.599	5,055	1.311	0.872
41	0.602	0.599	21,459	3.510	0.920
42	0.627	0.687	12,444	2.033	0.896
43	0.636	0.691	8,825	1.472	0.886
44	0.627	0.687	11,678	1.901	0.900
45	0.754	0.727	14,911	2.486	0.914
46	0.602	0.599	11,339	0.934	0.864
47	0.606	0.601	61,488	7.549	0.952
48	0.610	0.604	69,948	6.256	0.943
49	0.606	0.601	70,578	6.529	0.945
50	0.610	0.604	65,558	8.091	0.952

A.2 Different Seed - 3 Clusters

In the present section, seed 123 was considered for the CIMLR method with 3 clusters. The number of neighbors was tuned in the same way and the parameter that maximized the most measures was 50, like with the previous seed.

The result is present in [Table A.3](#): the confusion matrix between the two seeds show an equal decomposition of the patients, only the labels of clusters 1 and 2 are exchanged. The internal validation measures are also equal, which is in line with the previous statement. The external validation measures do show some variation, with the Calinski-Harabasz Index slightly higher and the Dunn Index, and Average Silhouette Width slightly lower. Upon further verification, it was confirmed that the patients in cluster 1 from the result with seed 12345 are the same as the patients in cluster 2 with seed 123, and that cluster 3 from both results also have the same patients. Meaning that the partition is equal only the denomination of the clusters is different.

Table A.2: Tuning number of neighbors for the CIMLR method with 4 clusters.

Tuning Parameter	CHI	DI	ASW
30	25,537	4.269	0.927
31	28,715	4.088	0.929
32	32,290	4.498	0.933
33	30,759	4.452	0.933
34	34,123	4.690	0.935
35	34,074	4.796	0.935
36	22,166	2.898	0.914
37	2,589	0.900	0.850
38	36,566	4.726	0.937
39	36,403	4.669	0.936
40	40,138	4.664	0.937
41	34,844	4.713	0.937
42	38,545	5.075	0.940
43	19,876	2.954	0.923
44	17,799	2.450	0.918
45	6,096	1.481	0.886
46	9,269	1.940	0.905
47	9,301	1.781	0.902
48	23,597	3.033	0.920
49	11,176	1.496	0.887
50	27,597	2.979	0.917

Table A.3: CIMLR result with 3 clusters for a different seed.

(a)					(b)	
CIMLR	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	35	111	9	155	Adjusted Rand Index	0.61
Cluster 2	30	1	157	188	Normalized Mutual Information	0.604
Cluster 3	184	0	0	184	Calinski-Harabasz Index	65,336
					Dunn Index	7.588
					Average Silhouette Width	0.952

Originally, the method to determine the number of clusters using the CPI method (see [section 2.2](#)) was going to be employed since it is provided by the intNMF package [102]. However, when applying it to the two datasets an unexpected error occurred with the message “dim = 0”, indicating an issue related to the dimensionality or the structure of the data. By rechecking the data matrices it was verified both were correct, with non-zero dimensions, and no missing or negative values were present.

Additionally, it could be theorized that the error might be linked to the presence of zeros or very small values in the data, however, upon further inspection, it became clear that was not the case. The example data provided by the package contains much smaller values with one of the datasets ranging from 1.487×10^{-17} to 1 whereas the data that provoked the error ranges from 0 to 1 (transcriptomic) or 0.004 to 0.996 (methyloomic), and the data structures were the same with patients in rows and molecular features in columns.

Given the large dimensionality of the data, particularly the methyloomics matrix with 301,541 features, it was hypothesized that the error might stem from internal operations within the NMF algorithm that inadvertently led to a zero-dimension matrix or had difficulty handling large data matrices.

In light of this, an alternative approach was considered: employing various numbers of clusters, evaluating metrics, and choosing the optimal k based on the performance. Thus, 3 to 6 clusters were considered.

B.1 5 Clusters

The intNMF method was applied for 5 clusters considering 1,000 maximum iterations, 29 different initializations, and a stability count of the connectivity matrix of 15, reached at the 276th iteration. Previously 30 initializations were considered, however, due to the time-out limit this was not feasible. Nevertheless, the reduction from 30 to 29 initializations was not anticipated to affect the stability of the results notably.

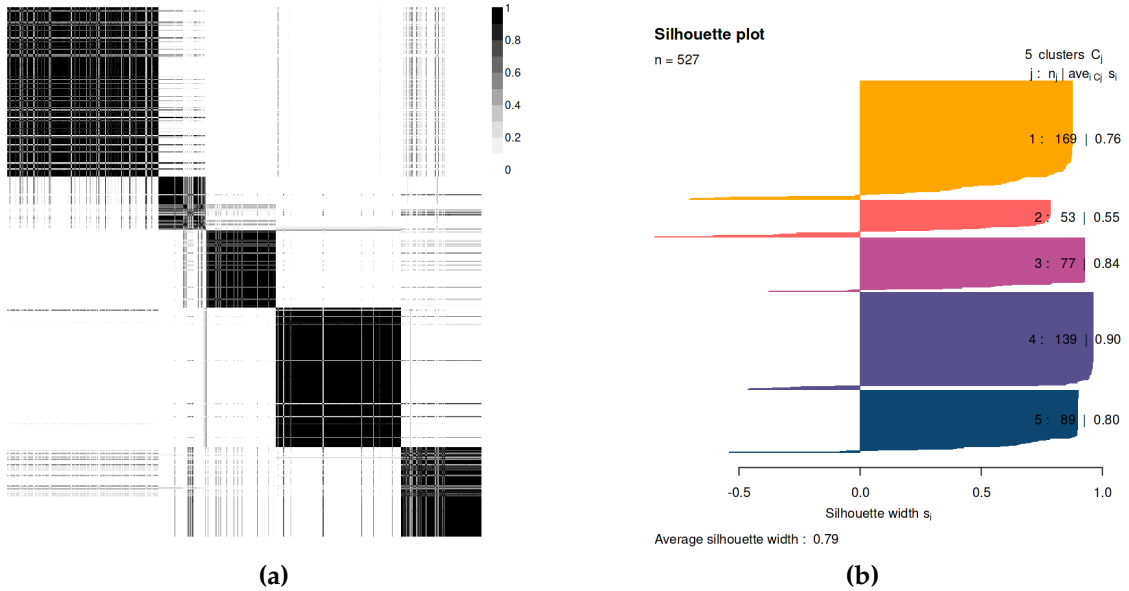
From [Table B.1a](#), the confusion matrix, it is possible to analyze the cluster composition:

Table B.1: intNMF result with 5 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
intNMF	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	169	0	0	169	Calinski-Harabasz Index	1,188
Cluster 2	27	22	4	53	Dunn Index	0.062
Cluster 3	0	77	0	77	Average Silhouette Width	0.792
Cluster 4	4	0	135	139		
Cluster 5	49	13	27	89		

cluster 1 is characterized solely by astrocytoma patients; cluster 2 has 27 astrocytoma patients followed closely by 22 GBM patients and 4 oligodendroglioma patients; cluster 3 only has GBM patients; cluster 4 is mostly oligodendroglioma patients and 4 astrocytoma; and cluster 5 has 49 from the astrocytoma type, 27 from oligodendroglioma, and 13 from GBM. Overall, each glioma type has one cluster of which they are the majority, and notably, GBM is not highly distinguishable from the LGG.

In Table 4.6b, the internal validation measures are present. In particular, the Calinski-Harabasz Index is 1,188 and the Dunn Index is 0.062, both lower than considering 3 and 4 clusters in the intNMF method. The average silhouette width is 0.792, higher than the 4 clusters' output but lower than the 3 clusters. This value means a strong structure was found and the clusters are well-separated, yet the low Dunn Index indicates the opposite.

**Figure B.1: intNMF result with 5 clusters:** a) Consensus matrix plot. b) Silhouette plot.

The consensus plot (Figure B.1a) reveals 5 prominent dark blocks along the diagonal, each representing a well-defined cluster. This pattern indicates strong intra-cluster consistency, suggesting that the clustering method effectively grouped similar data points, and overall suggests robust clustering results, particularly when contrasted with the consensus

matrix plot for 4 clusters (Figure 4.10a). Upon closer examination of the first two blocks, it becomes evident that there is partial overlap among patient labels. However, the fifth block representing cluster 5 displays multiple connections with all other clusters, indicating some instability in the clustering process for the patients grouped within this segment.

The silhouette plot shown in Figure B.1b shows a strong structure for all clusters except cluster 2, which has a silhouette value of 0.55 meaning a reasonable arrangement. It is also apparent the various objects have a negative silhouette, the minimum being -0.848 for a data point in cluster 2. This could explain the lower Dunn Index, as the presence of negative silhouette values indicates poor cluster separation and higher intra-cluster distances, which negatively impact the Dunn Index calculation.

B.1.1 Relevant Features

The specific coefficient matrices for each omic are analyzed and the cluster-specific variables are selected. Namely, the transcriptomics matrix:

- **Cluster 1:** 6,611 variables have a coefficient of zero, out of which 39 are cluster-specific.
- **Cluster 2:** 1,384 variables have a coefficient of zero, with 12 being specific to this cluster.
- **Cluster 3:** 2,975 variables have a coefficient of zero, and 13 are cluster-specific.
- **Cluster 4:** 5,874 variables have a coefficient of zero, out of which 182 are specific to this cluster.
- **Cluster 5:** 1,560 zero coefficient variables with 23 cluster-specific.

The specific coefficient matrix for the methylomicss has the following characteristics:

- **Cluster 1:** 7,127 variables with a coefficient of zero, without features specific to this cluster.
- **Cluster 2:** 19,955 variables have a coefficient of zero, with 13 cluster-specific.
- **Cluster 3:** 5,666 variables with a coefficient of zero, and 136 are specific to this cluster.
- **Cluster 4:** 3,952 variables have a coefficient of zero, and no variables are specific to this cluster.
- **Cluster 5:** 12,768 variables have zero coefficients and none are cluster specific.

Subsequently, the matrices are normalized and then the cluster-specific variables' coefficient is plotted, as a way to assess relative importance. For clusters 1, 2, and 5, the only cluster-specific variables are from the transcriptomics data with the highest normalized coefficient of approximately 0.0025 and 5×10^{-04} , indicating low relative importance. For cluster 3, both omics have cluster-specific variables with the highest for

RNA being 0.0012 and for DNA 0.45. For cluster 4, the values for both omics are similar to the previous cluster, 0.006 to RNA and around 0.5 to DNA. However, no cluster-specific variables have a normalized coefficient above 0.5.

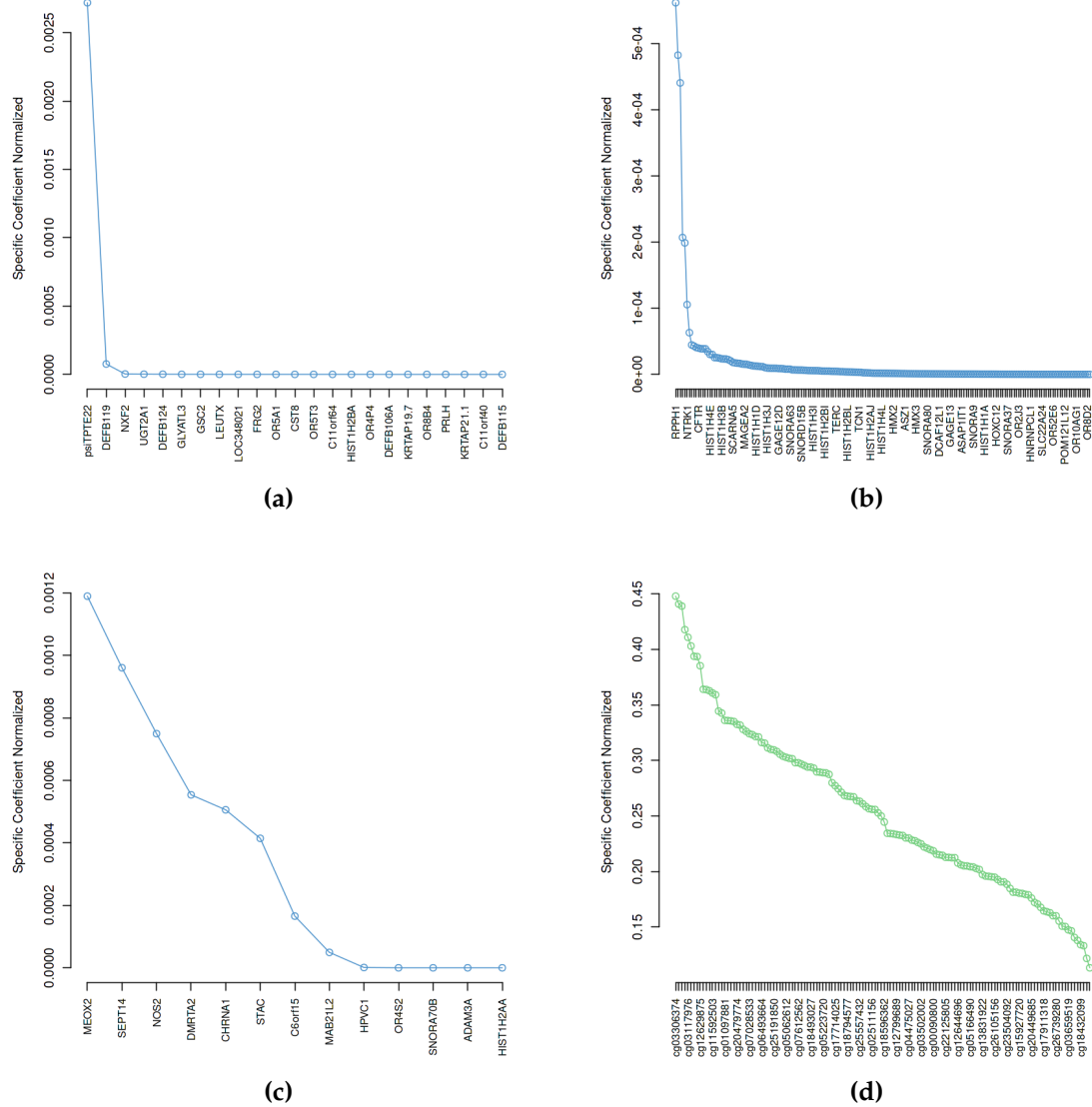
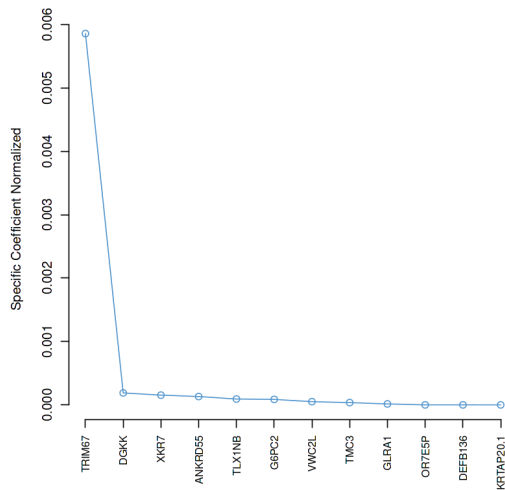


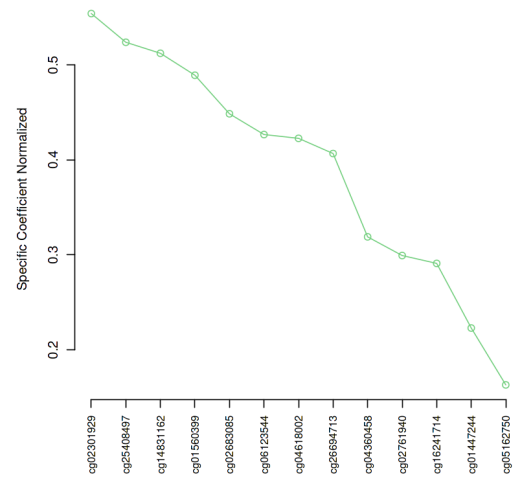
Figure B.2: Normalized coefficients for the intNMF method with 5 clusters: a) RNA cluster-specific variables in cluster 1. b) RNA cluster-specific variables in cluster 2. c) RNA cluster-specific variables in cluster 3. d) DNA cluster-specific variables in cluster 3.

B.1.2 Survival Analysis

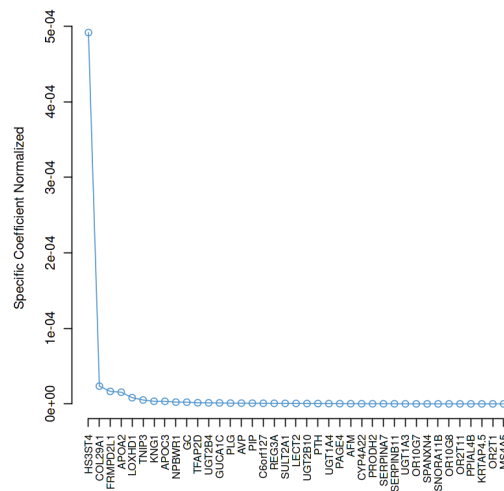
In Figure B.3, the survival function for the 5 clusters results are present, and clusters 2 and 3 seem to have the lower survival probability and similar curves. The log-rank test with the BH correction resulted in $\chi^2(4) = 236$, $p < 2 \times 10^{-16}$, meaning that the 5 survival distributions are significantly different. The pair-wise comparisons were employed, and



(e)



(f)



(g)

Figure B.2: Normalized coefficients for the intNMF method with 5 clusters: e) RNA cluster-specific variables in cluster 4. f) DNA cluster-specific variables in cluster 4. g) DNA cluster-specific variables in cluster 5.

the following clusters' survival distributions are statistically different: clusters 1 and 2 ($p = 5.8 \times 10^{-16}$), 1 and 3 ($p < 2 \times 10^{-16}$), 2 and 3 ($p = 0.0156$), 2 and 4 ($p = 4 \times 10^{-12}$), 2 and 5 ($p = 0.0001$), 3 and 4 ($p < 2 \times 10^{-16}$), 3 and 5 ($p = 3.1 \times 10^{-13}$), and 4 and 5 ($p = 0.0499$). Nevertheless, clusters 1 and 4 ($p = 0.6406$), and 1 and 5 ($p = 0.1175$) survival functions are statistically equal. The relatively high p-value for the comparison between clusters 4 and 5 suggests that these clusters may not be distinct, which aligns with the observation that clusters 1 and 4, as well as clusters 1 and 5, also show no significant differences. Note that clusters 1 and 4 are expected to have an equal survival curve since cluster 1 has astrocytoma patients and cluster 4 has mostly oligodendroglioma and some astrocytoma patients, which aligns with knowledge from the Kaplan-Meier curve of the WHO classification. It also does not seem alarming that cluster 5's survival is similar due to being mostly astrocytoma, followed by oligodendroglioma and some GBM patients. It does however highlight that cluster 2, which is a smaller cluster with mostly astrocytoma and GBM, and some oligodendroglioma, survival function is not equal, possibly due to the inclusion of more GBM patients. Again, GBM is distinguishable from the LGG with cluster 3 (solely GBM) having a not equal survival curve.

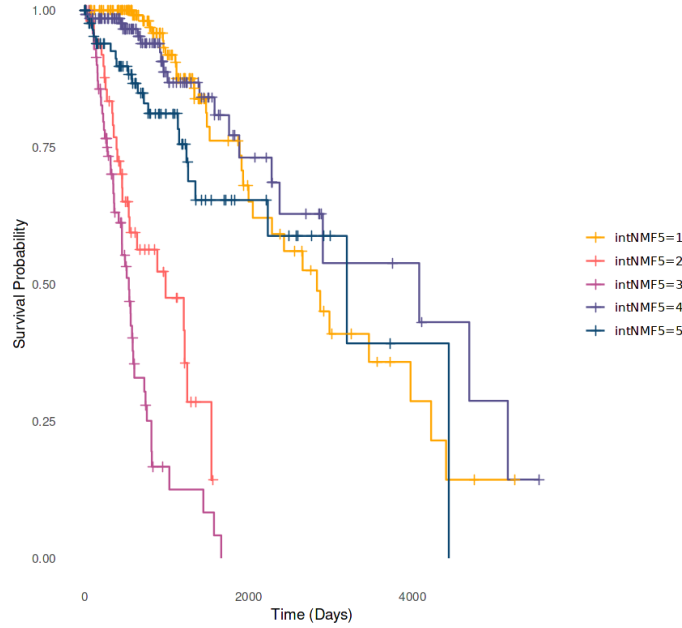


Figure B.3: intNMF method with 5 clusters.

B.2 6 Clusters

The intNMF method was applied for 6 clusters considering 1,000 maximum iterations, 24 different initializations, and a stability count of the connectivity matrix of 15, which was reached at the 399th iteration. Once again, the initializations were constrained by the run time out.

Table B.2: intNMF result with 6 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
intNMF	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	160	0	0	160	Calinski-Harabasz Index	1,072
Cluster 2	0	69	0	69	Dunn Index	0.042
Cluster 3	44	25	10	79	Average Silhouette Width	0.794
Cluster 4	15	0	63	78		
Cluster 5	27	18	4	49		
Cluster 6	3	0	89	92		

In the confusion matrix, present in [Table B.2a](#), cluster 1 is characterized strictly by astrocytoma patients; cluster 2 by GBM; clusters 3 and 5 have mostly astrocytoma patients, followed by GBM, and then oligodendroglioma; and clusters 4 and 6 are constituted predominantly by oligodendroglioma patients and some astrocytoma. Meaning that there are clusters constituted solely by astrocytoma and GBM, two clusters with all glioma types, and two clusters with only LGG.

In [Table B.2b](#), the cluster validations measures are present. The Calinski-Harabasz Index is 1,072 and the Dunn Index is 0.042, lower than previously. The average silhouette width is 0.794, slightly higher than the 5 clusters output, and evidences a strong structure with well-separated clusters. The silhouette plot ([Figure B.4b](#)) demonstrated this and corroborated the consensus plot analysis, with cluster 4 having a strong structure and cluster 5 a weak structure that could be artificial [104]. The minimum value obtained is -0.875 , and negative silhouette values are present in all clusters with higher values than before.

The consensus plot ([Figure B.4a](#)) exhibits 6 blocks in the diagonal, in particular, cluster 6 block is dark, suggesting a robust clustering. Cluster 1 also mostly seems robust except for some data points, however, cluster 5 stands out with many white lines within the diagonal block and various connections with the other clusters. Similar, not as grave, behavior is present for clusters 2 to 4. This indicates instability in the clustering result and points toward the need to adjust parameters, particularly the number of initializations to increase robustness.

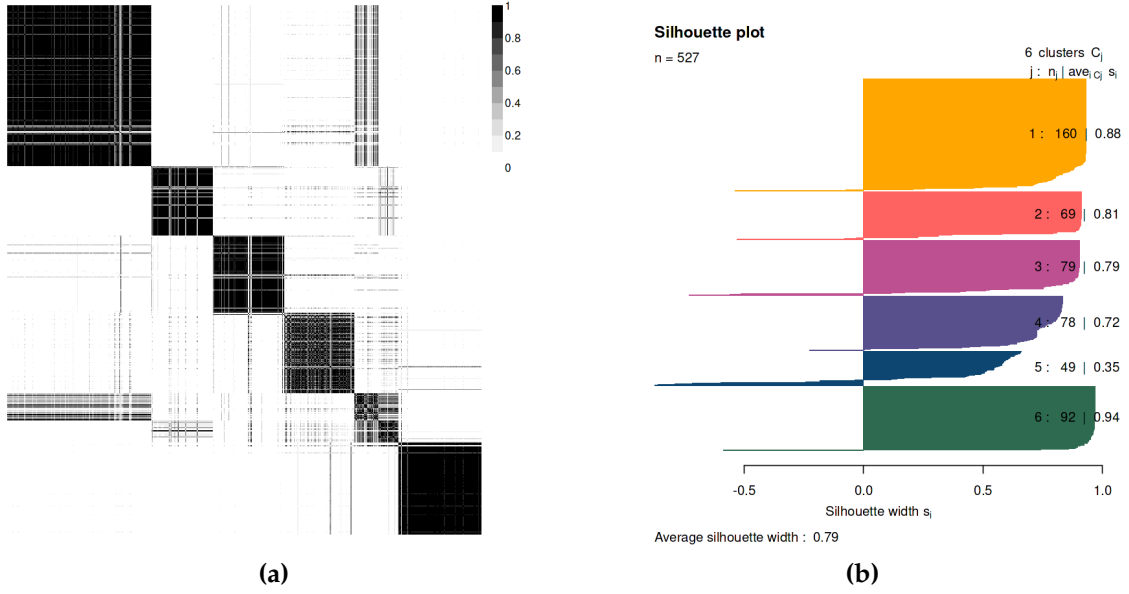


Figure B.4: intNMF result with 6 clusters: a) Consensus matrix plot. b) Silhouette plot.

B.3 Convergence Study

In the present section, the convergence of the intNMF method for 3 to 6 clusters is studied.

B.3.1 3 Clusters

Studying the algorithm's convergence with parameters set to perform adjustments is relevant. In the final iteration, the value of the objective function was 1,095.9, the absolute difference with the previous iteration is 1.519×10^{-5} .

The behavior of objective function and the absolute difference with the previous iteration throughout the iterations can be consulted in Figure B.5. It is noteworthy to mention that the objective function values and the absolute difference between them in iterations are both monotonously decreasing. Moreover, the objective function values stabilize and, since a monotonically decreasing series of positive values bounded below must converge, it is possible to affirm that the absolute difference converges to 0, the desirable outcome.

B.3.2 4 Clusters

Figure B.6 shows two plots to help evaluate the method's convergence as this is a key point in ensuring the stability of the result and judging the appropriateness of the parameters set.

In Figure B.6a, the value of the objective function and the absolute difference between iterations are shown, and both tend to stabilize. The Figure B.6b also presents the absolute difference but excludes the initial iterations. This emphasizes the non-monotone behavior of the convergence process, illustrating the fluctuations and the eventual trend toward

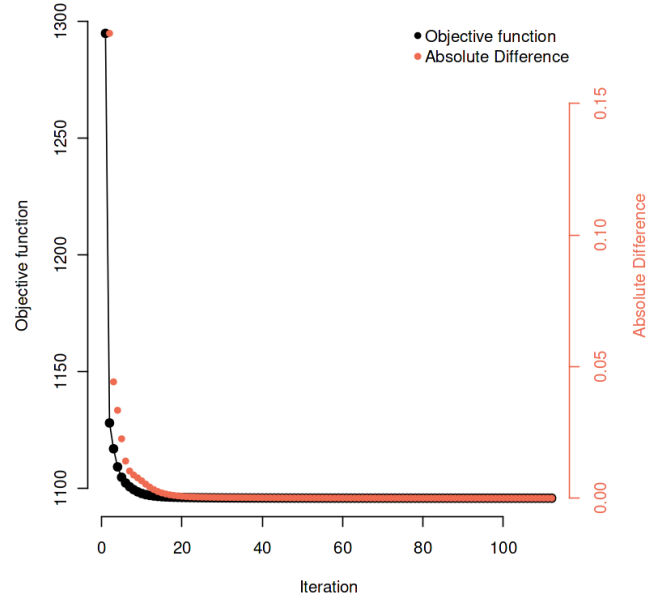


Figure B.5: Value of the objective function in each iteration (black) and the absolute difference between iterations (coral).

stabilization. It is worth highlighting that while a monotone decrease in the objective function value is a straightforward indicator of convergence, non-monotone sequences can still converge, in this case, towards zero.

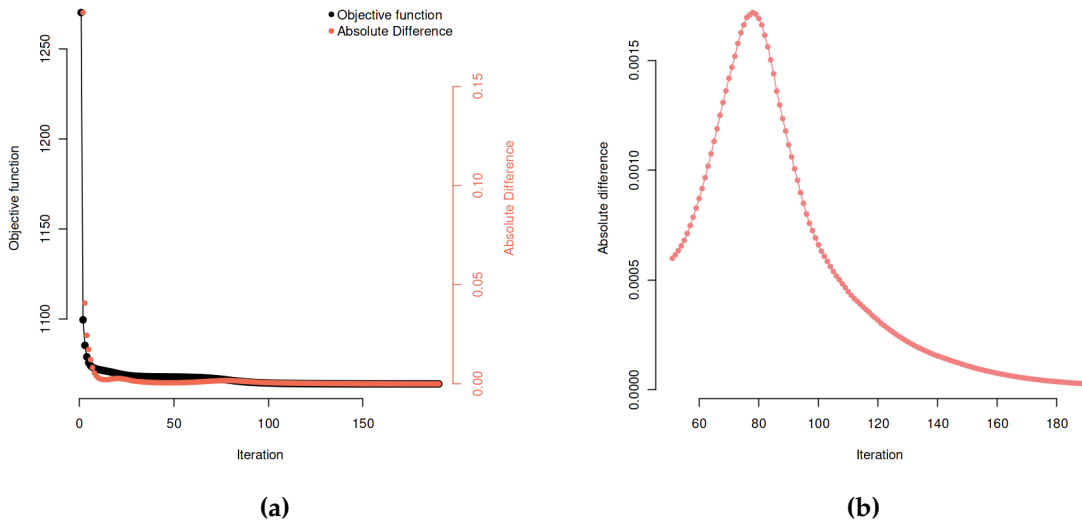


Figure B.6: Convergence study for intNMF with 4 clusters: a) Value of the objective function in each iteration (black) and the absolute difference between iterations (coral). b) Detailed view of the absolute differences in the objective function through iterations.

In the last (190) iteration, where the stability set for the connectivity matrix was reached,

the value of the objective function is 1,063.8 and the absolute difference between iterations is 2.695×10^{-5} .

B.3.3 5 Clusters

Next, the convergence of the matrices from the 5 clusters output is studied. The overall behavior of the objective function through the iterations is present in [Figure B.7](#). In particular, while the absolute difference between the values in iterations is not monotonously decreasing ([Figure B.7b](#)), it is clear it converges to 0 with its value in the last iteration being 2.614×10^{-05} . The objective function values are decreasing monotonously ([Figure B.7a](#)), reaching the value of 1,043.099 in the last iteration.

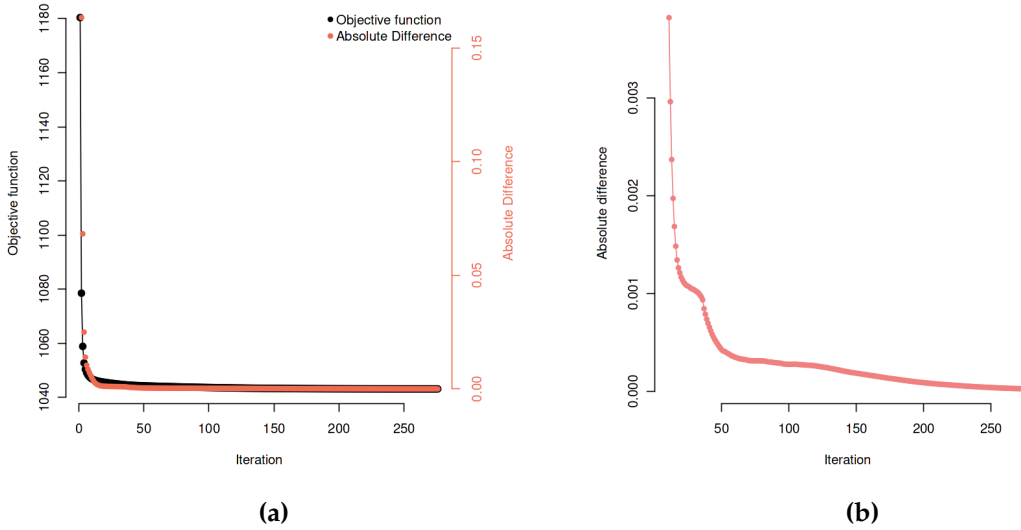


Figure B.7: Convergence study for intNMF with 5 clusters: a) Value of the objective function in each iteration (black) and the absolute difference between iterations (coral). b) Detailed view of the absolute differences in the objective function through iterations.

B.3.4 6 Clusters

Finally, a convergence analysis for the 6 clusters is expo in [Figure B.8](#). The objective function value in the last iteration is 1,026.423 with an absolute difference from the previous iteration of 5.554×10^{-05} , which is very low. In [Figure B.8a](#), the objective function values are monotonously decreasing, yet the same is not verified in the absolute difference as evidenced in [Figure B.8b](#). However given the value of the difference in the last iterations, close to zero, and the overall behavior after the 100 iterations, it is still possible to affirm that the sequence converges to 0.

Given the poor performance of the 6 clusters output exposed, the feature analysis is not fulfilled. Furthermore, no further number of clusters is considered since overall the cluster validation measures are decreasing with the increase in the number of clusters,

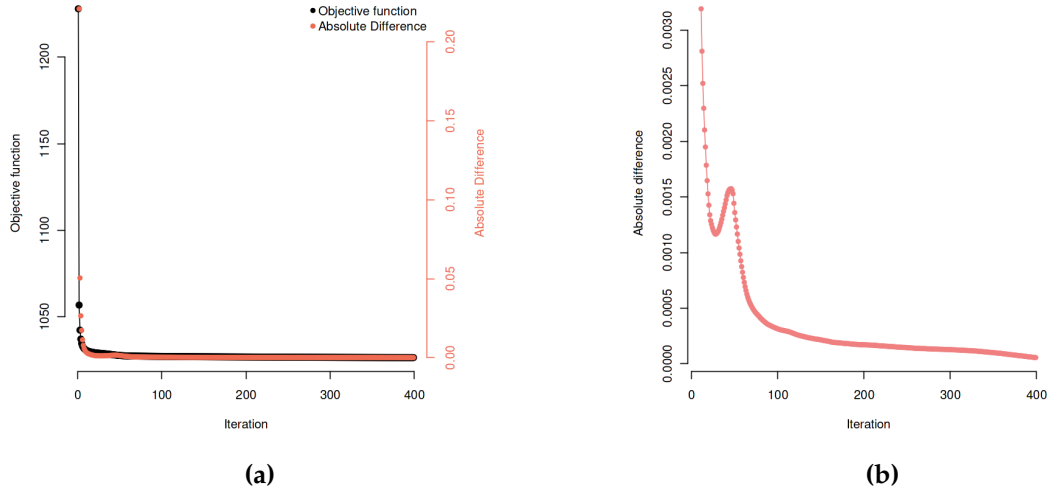


Figure B.8: Convergence study for intNMF with 6 clusters: a) Value of the objective function in each iteration (black) and the absolute difference between iterations (coral). b) Detailed view of the absolute differences in the objective function through iterations.

the consensus plot for 6 clusters shows instability, meaning the clustering is less robust, and there is difficulty in considering various initializations to increase robustness, and the interpretation of the confusion matrix is lower.

B.4 Different Seed - 3 Clusters

In the present section, seed 123 was considered for the intNMF method with 3 clusters, with the same parameters: 30 different initializations and the stability of the connectivity matrix set as 20,

The result is present in [Table B.3](#): the confusion matrix between the two seeds show an equal distribution of the patients through the glioma types and the cluster validation measures are all equal to the result with seed 12345. After verification, it was concluded that the partitions originated from different seeds are equal.

Table B.3: intNMF result with 3 clusters: a) Confusion matrix for the obtained clusters versus the official glioma types. b) Cluster validation measures.

(a)					(b)	
intNMF	Astro	GBM	Oligo	Total	Measures	Value
Cluster 1	97	19	47	163	Adjusted Rand Index	0.235
Cluster 2	137	0	117	254	Normalized Mutual Information	0.314
Cluster 3	15	93	2	110	Calinski-Harabasz Index	8,179
					Dunn Index	0.231
					Average Silhouette Width	0.921

C.1 Gap statistic

The present section shows the gap statistic for different sparsity parameters considering 2 and 6 joint latent variables.

C.1.1 2 Joint Latent Variables

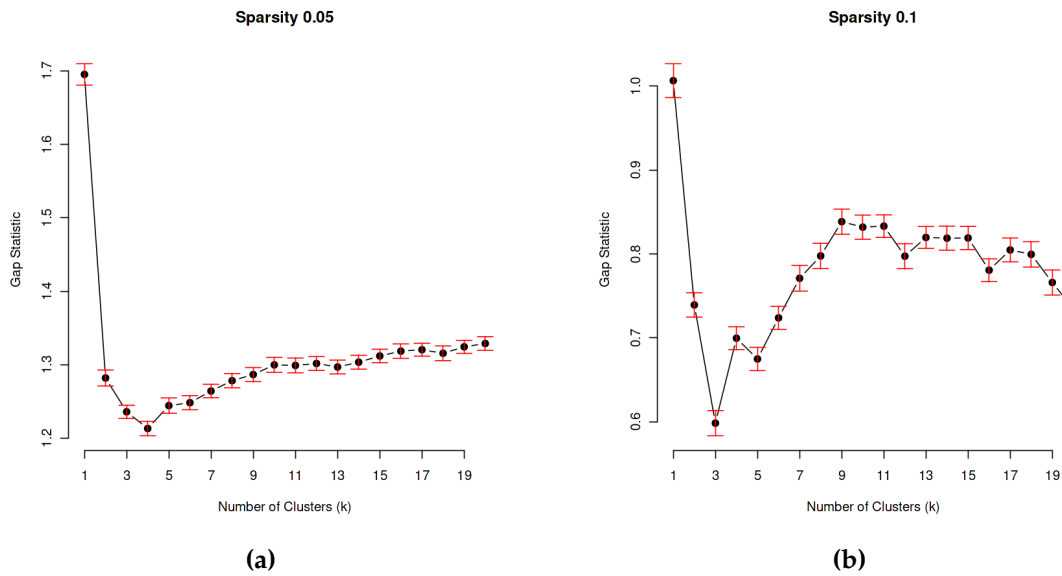


Figure C.1: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 1): a) 0.05. b) 0.10.

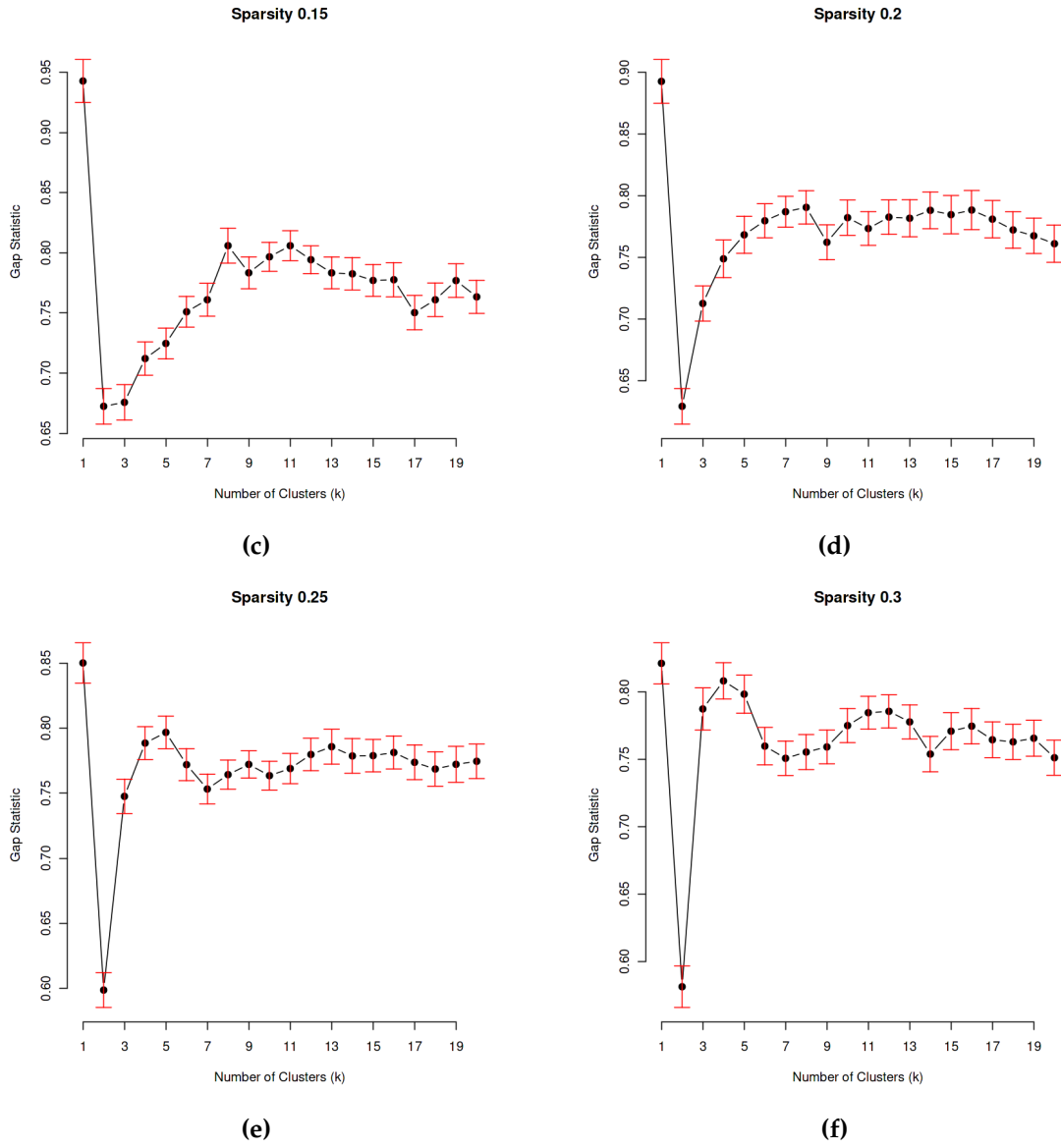


Figure C.1: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 2):
 c) 0.15. d) 0.20. e) 0.25. f) 0.30.

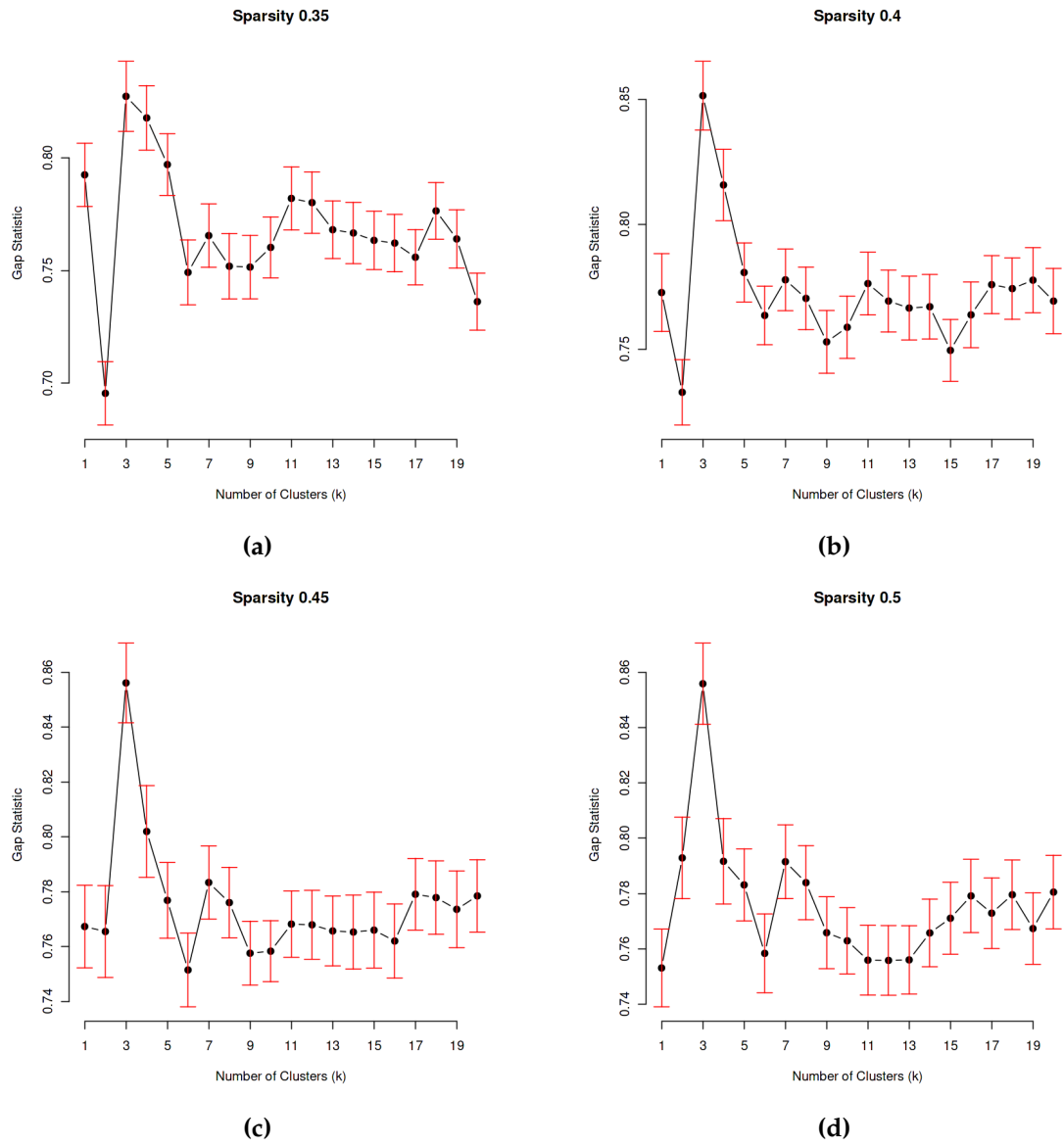


Figure C.2: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 1):
a) 0.35. b) 0.40. c) 0.45. d)

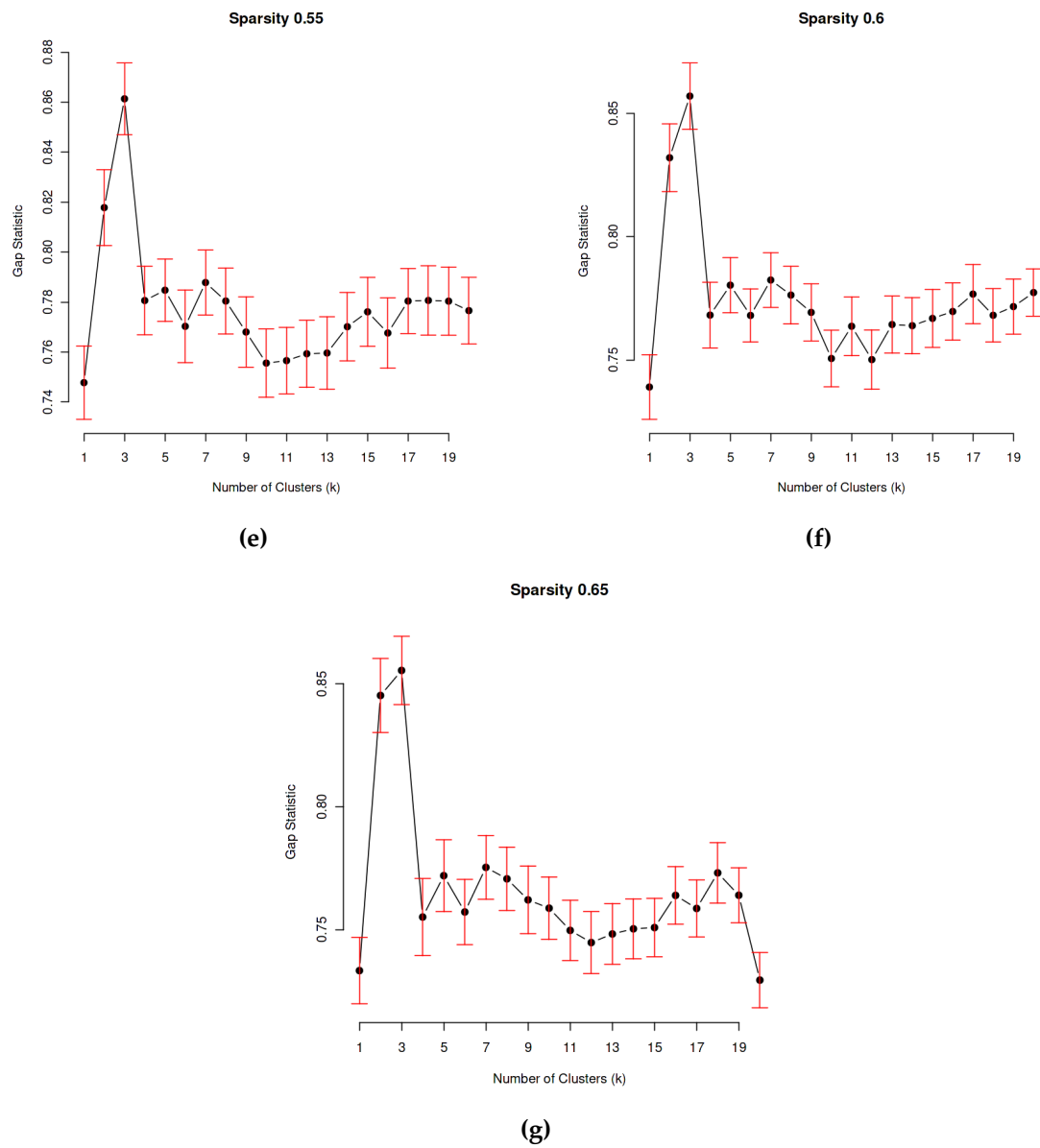


Figure C.2: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 2): e) 0.55. f) 0.60. g) 0.65.

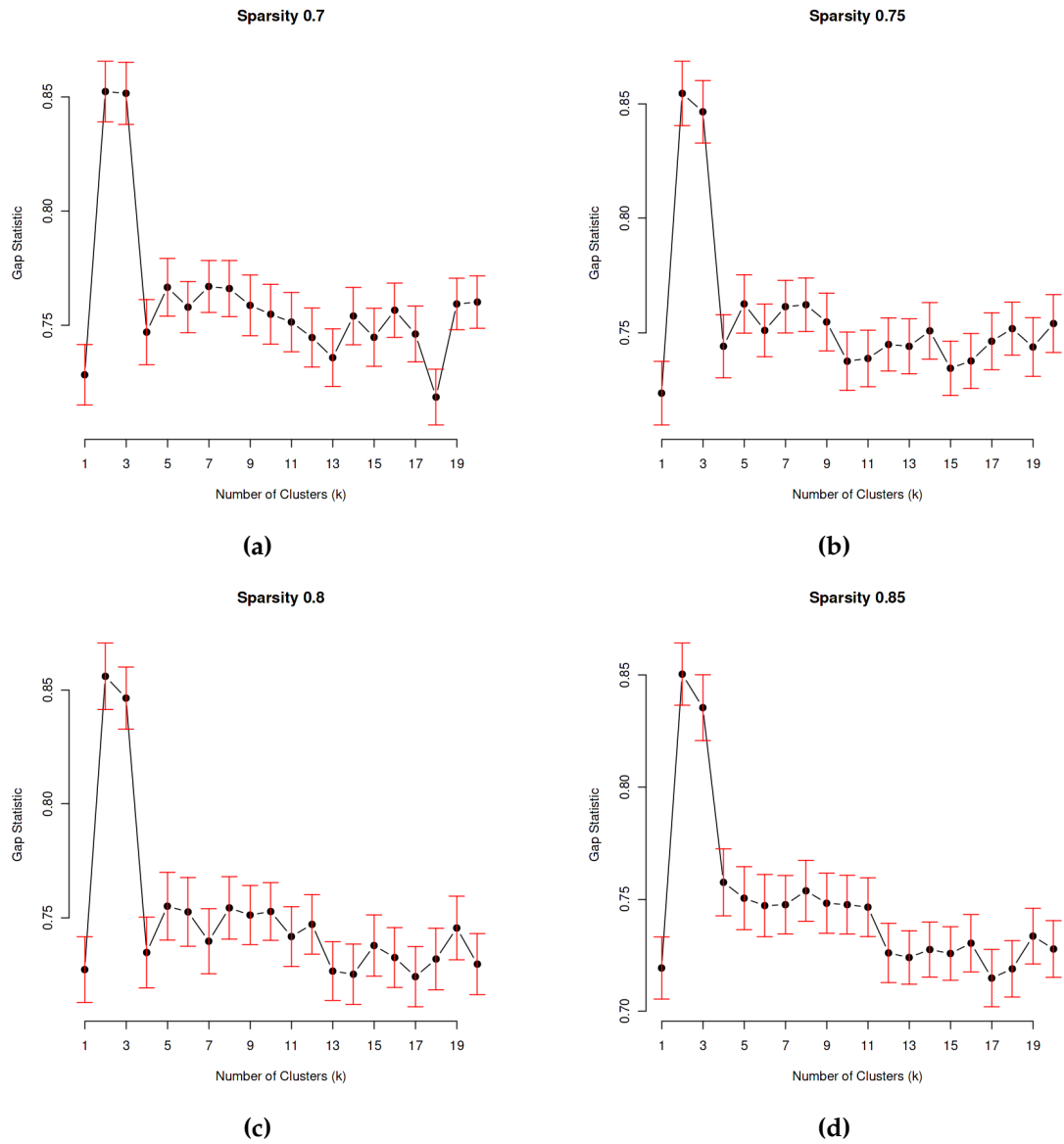


Figure C.3: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 1):
a) 0.70. b) 0.75. c) 0.80. d) 0.85

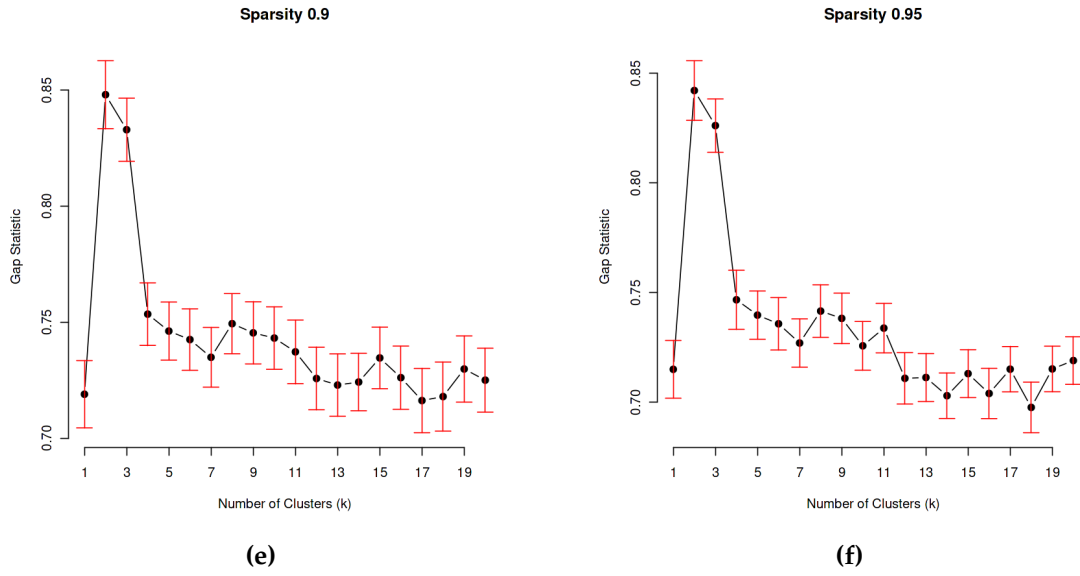


Figure C.3: Gap statistic for the moCluster method with 2 JLV and sparsity of (part 2): e) 0.90. f) 0.95.

C.1.2 6 Joint Latent Variables

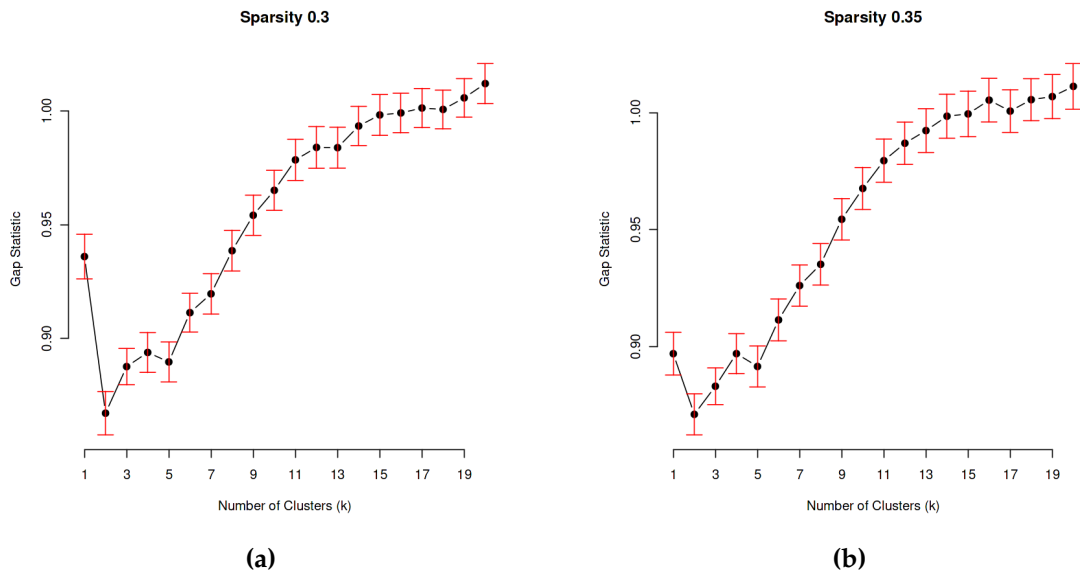


Figure C.4: Gap statistic for the moCluster method with 6 JLV and sparsity of (part 1): a) 0.30. b) 0.35. c) 0.40.

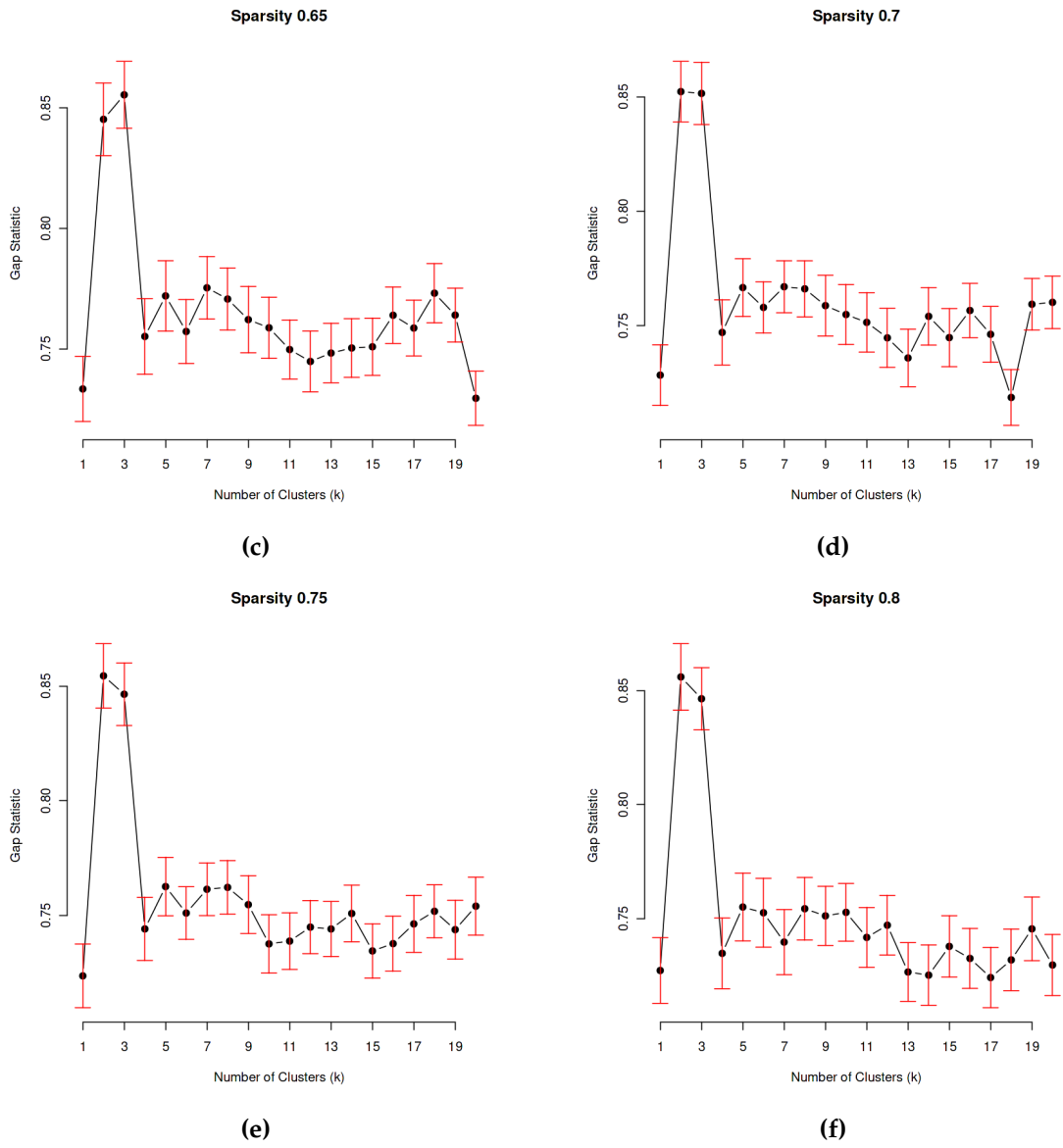


Figure C.4: Gap statistic for the moCluster method with 6 JLV and sparsity of (part 2):
c) 0.65. d) 0.70. e) 0.75. f) 0.80.

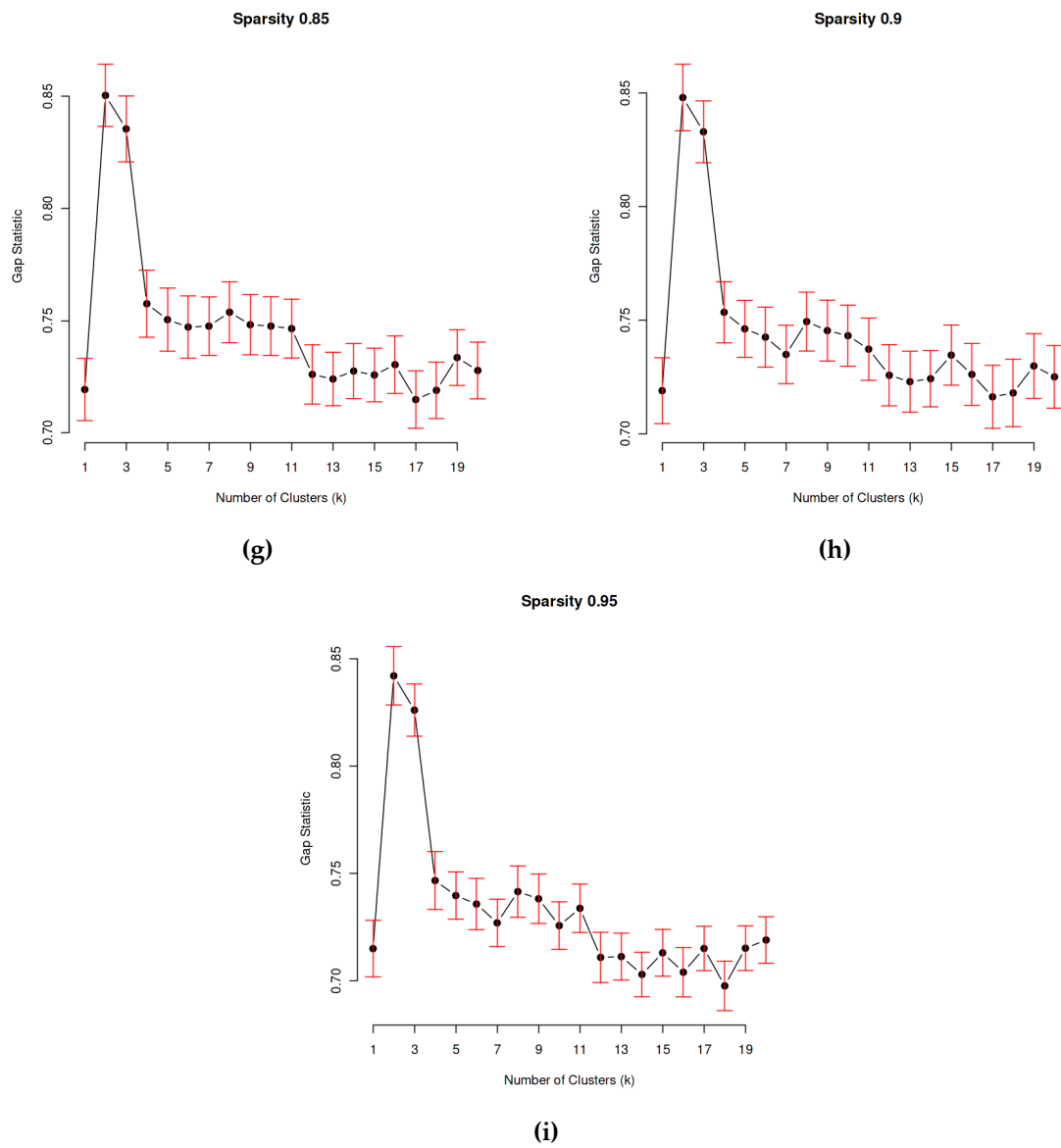


Figure C.4: Gap statistic for the moCluster method with 6 JLV and sparsity of (part 3): g) 0.85. h) 0.90. i) 0.95.



2024 DECRYPTING GLIOMA HETEROGENEITY THROUGH MULTI-VIEW CLUSTERING: Carla Pérez-Cortés



NOVA SCHOOL OF
SCIENCE & TECHNOLOGY