Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimal Graph Cutting: An Adaptive and Extensible Framework for Multi-Omics Cancer Subtyping

Sreekumar Ranganathan

sreekumarx@yahoo.com

National Institute of Technology Srinagar

Research Article

Keywords: Multi-omics integration, cancer subtyping, graph optimization, adaptive data fusion, spectral clustering

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11506680/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Optimal Graph Cutting: An Adaptive and Extensible Framework for Multi-Omics Cancer Subtyping

Sreekumar R, NIT, Srinagar,

Abstract

Integrating heterogeneous molecular datasets is essential for characterizing cancer heterogeneity, yet differences in data reliability and the availability of new molecular information challenge conventional integration methods. We present Optimal Graph Cutting (OGC), an adaptive graph-based framework for multi-omics integration and cancer subtyping. OGC constructs modality-specific patient similarity graphs and employs a neighbourhood inconsistency measure (Dst) to estimate reliability at both modality and individual edge levels. These estimates guide adaptive weighting and graph refinement, followed by spectral optimization to identify molecularly distinct patient groups. A key feature of OGC is its extensibility: additional molecular modalities can be represented as new similarity graphs and incorporated without redesigning the entire integration framework, with their contributions determined by estimated neighbourhood reliability. OGC was applied to The Cancer Genome Atlas datasets integrating gene expression, DNA methylation and microRNA expression. Survival analyses reported significant differences between selected OGC-derived subgroups, including those in glioblastoma (log-rank $P < 0.0001$), kidney renal clear cell carcinoma ($P = 0.0031$) and lung adenocarcinoma ($P = 0.041$). These findings illustrate the potential of reliability-aware graph integration to support cancer subtype discovery while accommodating evolving multi-omics datasets.

Keywords: Multi-omics integration; cancer subtyping; graph optimization; adaptive data fusion; spectral clustering.

Introduction

A cancer-subtyping system is usually built from information available at a particular point in time. However, cancer biology is not static. New omics measurements, additional molecular layers, longitudinal samples, clinical variables, imaging information, or improved measurements can become available during the course of time. When such information is

introduced a rigid integration method may need to be completely redesigned or retrained. In contrast, a system that can naturally incorporate new information into its existing representation can update the patient similarity structure without discarding the information already captured.

Adaptability to Evolving Information

An important characteristic of a multi-omics cancer-subtyping framework is its ability to accommodate information that becomes available as datasets evolve. Cancer research increasingly generates heterogeneous molecular and multimodal measurements, including transcriptomic, epigenomic, miRNA, proteomic, imaging and clinical data. Consequently, a subtyping framework designed around a fixed set of input modalities may have limited applicability when additional information becomes available. Network-based approaches such as Similarity Network Fusion demonstrated that complementary molecular measurements can be represented as patient-specific similarity networks and integrated to identify clinically meaningful cancer subtypes [1]. Similarly, NEMO demonstrated that multi-omics clustering can be performed even when only a subset of omics measurements is available for individual patients, thereby reducing the need to discard patients or perform imputation when the data are incomplete [2,3]. More recent approaches such as MOSD have further demonstrated the value of adaptive weighting of molecular data types for improving survival discrimination of cancer subtypes [4,7].

The ability to accommodate evolving information does not imply that adding more data will invariably improve performance. Comprehensive benchmarking studies have shown that incorporating additional omics layers can sometimes reduce cancer-subtyping performance because of redundancy, noise or conflicting biological signals [5]. Therefore, the desirable property is not simply the capacity to accept additional information, but the ability to determine whether and to what extent the newly available information contributes useful signal. An adaptive integration strategy can retain previously learned molecular relationships while incorporating new complementary evidence and adjusting the relative contribution of the available modalities according to their reliability. This provides a more flexible representation of tumour heterogeneity and creates the potential for improved clustering accuracy, clinical discrimination and biological interpretability.

This consideration motivates the adaptive graph formulation adopted in the proposed Optimal Graph Cutting framework. Instead of requiring a fixed and equal contribution from all molecular modalities, OGC estimates the reliability of the available patient-similarity structures and incorporates them into the consensus graph. Consequently, when additional molecular information becomes available, it can be incorporated as an additional graph and its contribution can be determined according to its estimated reliability. Thus, extensibility and adaptive weighting together provide a mechanism for allowing the integrated representation to evolve with the available biological information rather than requiring the entire subtyping framework to be redesigned.

One approach to representing such relationships is through a **patient similarity graph**. In this representation, patients are represented as vertices and the similarity between a pair of patients is represented by the weight of the edge connecting them. Network-based integration methods have demonstrated the usefulness of such representations for combining multiple molecular data types and identifying clinically meaningful cancer subtypes [6]. In particular, Similarity Network Fusion (SNF) constructs a patient similarity network for each data type and fuses the individual networks into a single representation. SNF was demonstrated using mRNA expression, DNA methylation and miRNA expression data across multiple cancer datasets and showed the potential of network-based multi-omics integration for cancer subtyping and survival prediction [7,8].

However, the quality and reliability of the patient similarity relationships may differ among molecular modalities. A molecular layer with a more coherent neighbourhood structure may provide stronger evidence for patient similarity than a modality with a noisier or more heterogeneous representation. Furthermore, even within a single molecular modality, some individual patient-to-patient relationships may be reliable while others may be less reliable.

The **Optimal Graph Cutting (OGC)** framework was therefore developed in this research to address this problem. Rather than treating all molecular graphs as equally informative, OGC seeks to construct a consensus representation while incorporating information concerning the reliability of different modalities and individual graph edges.

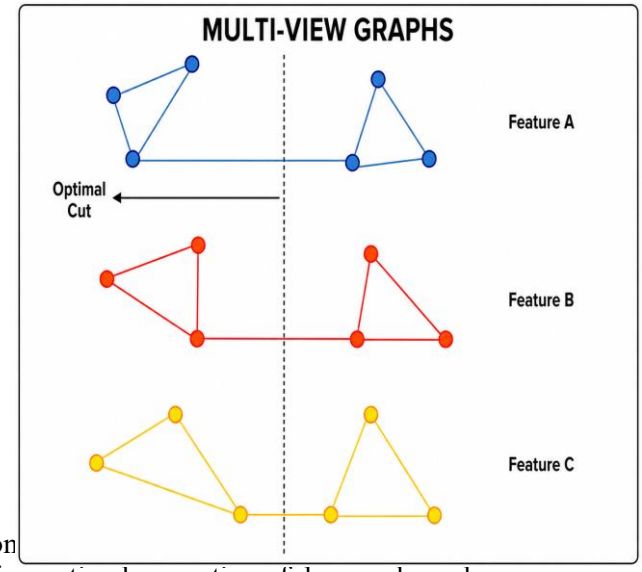


Figure1: Illustration ecific patient similarity graphs, highlighting optimal separation of dense subgraphs.

The fundamental objective of OGC is to learn **one consensus graph that explains the available modalities**, optimize the total weighted graph cut, and provide a framework capable of incorporating additional molecular features or modalities.

The central idea of this chapter is therefore:

Multi-Omics Graphs→Reliability Estimation→Weighted Consensus Graph→ Optimal Graph Cut→Cancer Subtypes

Method

Let the multi-omics dataset contain n patients and M molecular modalities. The data corresponding to modality m can be represented as. $X^{(m)} = \{x_1^{(m)}, x_2^{(m)}, \ldots\ldots x_n^{(m)}\}$.

For each modality, a patient similarity matrix is constructed: $W^{(m)} = \left[w_{ij}^{(m)}\right]$,

where $w_{ij}^{(m)}$ represents the similarity between patients i and j according to modality m.

This matrix can be interpreted as a weighted graph $G^{(m)} = \left(V, E^{(m)}, W^{(m)}\right)$,where V represents the common set of patients.

The use of similarity graphs is particularly suitable for multi-omics analysis because the different molecular measurements can be converted into a common patient-level representation without requiring the raw features from different modalities to be directly

concatenated. Network fusion approaches have demonstrated that patient similarity structures can capture complementary information from different omics layers [9].

Nevertheless, the individual graphs need not have identical structures. For example, two patients may have highly similar gene-expression profiles but substantially different methylation profiles. Another pair may show strong methylation similarity but relatively weak transcriptional similarity.

Consequently, an important question arises: How should the contribution of each molecular graph be determined when their neighbourhood structures differ in reliability?

The OGC framework addresses this problem through two levels of reliability estimation:

1. **Global modality reliability**, which determines the contribution of an entire omics graph.
2. **Local edge reliability**, which determines the contribution of individual patient-to-patient relationships.

This distinction forms the basis of the adaptive graph construction proposed in this research.

Graph Representation of Multi-Omics Data

For each molecular modality m, the patient similarity matrix is denoted by .$W^{(m)} = \left[w_{ij}^{(m)}\right]$.

The degree matrix corresponding to the graph is defined as $D_{ii}^{(m)} = \sum w_{ij}^{(m)}$.

The graph Laplacian is then defined as $L^{(m)} = D^{(m)} - W^{(m)}$.

The Laplacian provides a mathematical representation of the connectivity structure of the graph and is central to spectral graph-based clustering [10]. If two patients are strongly connected in the similarity graph, a desirable partition should tend to place them in the same group. Conversely, weakly connected patients are more likely to be separated into different groups. This makes graph partitioning particularly suitable for cancer subtyping, where a molecular subtype can be interpreted as a group of patients having relatively strong internal molecular similarity and weaker similarity to patients outside the group.

Graph Cut and RatioCut Formulation

Consider a graph G=(V,E) that is partitioned into K groups: $A_1, A_2, \ldots\ldots, A_K$.The graph-cut objective measures the total weight of edges connecting different groups.

For a group A_k, the cut between A_k and its complement $\overline{A_k}$ is represented as $cut\left(A_k, \overline{A_k}\right)$.

A RatioCut formulation can then be written as

$$RatioCut(A_1, \ldots, A_K) = \sum \frac{cut\left(A_k, \overline{A_k}\right)}{|A_k|}. \quad (1)$$

The normalization by the size of each group prevents the optimization from favouring trivial partitions consisting of extremely small groups. Graph-partitioning objectives such as normalized and ratio-based cuts have been widely studied as a means of separating groups with strong internal connectivity while minimizing connectivity between groups [86,87]. For a two-way partition, the discrete group membership can be represented by an indicator vector f. The corresponding graph objective can be expressed in quadratic form as $.f^T Lf$.

For a symmetric similarity matrix, this quadratic form is related to the pairwise graph structure through

$$.f^T Lf = \frac{1}{2} \sum W_{ij} \left(f_i - f_j\right)^{2.} \quad (2)$$

Thus, minimizing the quadratic form encourages connected samples to have similar values in the relaxed representation.

Spectral Relaxation of the Graph-Cutting Problem

The direct optimization of discrete cluster membership is computationally difficult. Spectral clustering provides a continuous relaxation of the graph-partitioning problem by expressing the objective using the eigenstructure of the graph Laplacian [86,87].For K clusters, the relaxed representation is written as $H = [h_1, h_2, \ldots, h_K]$,where $H \in R^{n \times K}$.

The corresponding optimization problem is $min_H\ Tr(H^T LH)$ subject to $.H^T H = I$.

The orthogonality constraint ensures that the spectral directions are linearly independent.

According to the spectral formulation, the solution is obtained from the eigenvectors associated with the smallest eigenvalues of the graph Laplacian [11,12]. These eigenvectors

provide a low-dimensional representation of the graph in which patients with similar connectivity patterns tend to be located close to one another. The resulting matrix H therefore provides a spectral embedding of the patient population. The rows of H can subsequently be clustered using k-means or another clustering algorithm.The complete process can therefore be summarized as

Similarity Matrix→Graph Laplacian→Eigenvectors→Spectral Embedding→Patient Clusters

Spectral clustering has become an established graph-based clustering approach because it transforms a graph-partitioning problem into an eigenvalue problem that can be efficiently solved using standard linear algebra techniques [11].

Extension of Graph Cutting to Multiple Omics

For M molecular modalities, each modality produces its own **Figure 1** Illustration of the ratio-cut principle applied to modality-specific patient similarity graphs, highlighting optimal separation of dense subgraphs. Laplacian:

$$L^{(1)}, L^{(2)}, \ldots, L^{(M)}.$$

The multi-omics graph can be represented by a weighted combination of these modality-specific Laplacians: $L_{fusion} = \sum k_m L^{(m)}$, (3)

where k_m represents the contribution of modality m.The corresponding optimization becomes $min_H Tr\left(H^T L_{fusion} H\right)$ subject to $H^T H = I$.

This formulation allows the molecular modalities to contribute jointly to the spectral representation.

The concept of combining patient similarity networks is related to existing multi-omics network-fusion approaches such as SNF, which constructs separate patient networks for each molecular data type and integrates them into a fused network [1]. The OGC framework differs in its formulation by explicitly incorporating graph-cut optimization and adaptive reliability weighting into the consensus graph construction.

Therefore, OGC can be viewed as a graph-partitioning framework in which the integrated graph is optimized according to the estimated reliability of the molecular representations.

Need for Adaptive Weighting

A simple multi-omics fusion strategy could assign equal weights to all modalities:

$$k_1 = k_2 = \ldots\ldots = k_M.$$

However, this assumes that all molecular layers provide equally reliable information about patient similarity. Such an assumption may not always be justified. The information content and noise characteristics of molecular modalities can vary across cancer types and patient cohorts. Multi-omics integration studies have consequently explored approaches that account for the relative contribution or complementary information of different molecular data types [13].

In the present work, an adaptive weighting mechanism was introduced based on the **Dst measure** developed as part of the proposed framework. Dst is used to assess the consistency of patient neighbourhood relationships. The basic principle is:

Low Dst⇒High neighbourhood consistency

whereas

High Dst⇒Low neighbourhood consistency.

The average Dst of an omics modality therefore provides an indication of the overall consistency of its patient neighbourhood structure.

Modality-Level Confidence Estimation

For modality m, let the average neighbourhood inconsistency be $\overline{Dst^{(m)}}$.

The modality-level confidence is calculated using an exponential transformation: $C_m = exp^{\left(-\beta\overline{Dst^{(m)}}\right)}$, where β controls the sensitivity of the confidence to the Dst value.

The exponential transformation has an important property: confidence decreases monotonically as Dst increases. The resulting confidence scores are normalized to obtain modality-specific weights:

$$k_m = \frac{C_m}{\sum C_r}. \quad (4)$$

Consequently, $\sum k_m = 1$.

The interpretation is:

Lower average Dst → Higher confidence → Higher modality weight,

while,

Higher average Dst → Lower confidence → Lower modality weight.

This adaptive weighting mechanism is one of the principal methodological components of OGC. It allows the relative importance of an omics layer to be determined from its observed neighbourhood structure rather than imposed arbitrarily.

Edge-Level Confidence Estimation

Modality-level weighting provides a global estimate of reliability. However, it does not account for the fact that individual patient relationships may have different reliability values within the same modality.

Therefore, OGC introduces a second level of confidence estimation.

For the edge connecting patients i and j in modality m, the edge confidence is defined as

$$c_{ij}^{(m)} = exp^{\left(-\alpha Dst_{ij}^{(m)}\right)}, \tag{5}$$

where α controls the rate of attenuation.

The original similarity value is then adjusted according to the edge confidence:

$$\tilde{W}_{ij}^{(m)} = W_{ij}^{(m)} c_{ij}^{(m)}. \tag{6}$$

Consequently, a reliable edge with low Dst retains a large proportion of its original similarity, whereas an unreliable edge with high Dst receives a substantially lower effective weight.

The two levels of weighting can therefore be summarized as follows:

Average Dst→Global Modality Weight

and

Individual Dst→Local Edge Confidence.

This provides OGC with both **global adaptivity** and **local adaptivity**.

Dst-Based Edge Refinement

The edge-level confidence function also provides a mechanism for suppressing unreliable relationships. In the analysis presented in this study, Dst values approximately ranged from 0.3 to 4.8. For a low Dst value such as 0.3, the corresponding exponential confidence remains close to one. As Dst increases, the confidence decreases progressively.

Table 1: Dst Vs Confidence

Dst	Confidence
0.3	Near to 1
2	0.264
3	0.135
4	0.069
4.8	0.041

This transformation ensures that highly inconsistent relationships have little influence on the final graph.

The analysis used approximately 30 nearest neighbours and considered a Dst value of approximately 2.7 as an acceptable neighbourhood separation threshold under the adopted similarity-difference criterion. Thus, the Dst mechanism provides a basis for retaining reliable graph edges while strongly attenuating relationships that are inconsistent with the local neighbourhood structure.

Algorithm for the construction of the Adaptive Consensus Graph

The final OGC graph is constructed by integrating the modality-level and edge-level reliability information. For each modality, the procedure consists of the following steps.

Step 1: Similarity graph construction

A patient similarity matrix is constructed from the corresponding omics data.

Step 2: Neighbourhood identification

The local neighbourhood of each patient is determined.

Step 3: Dst calculation

Dst values are calculated to assess neighbourhood consistency.

Step 4: Modality-level confidence

The average Dst is transformed into a confidence score.

Step 5: Modality weighting

The confidence scores are normalized to obtain kmk_m.

Step 6: Edge-level confidence

Individual Dst values are transformed into edge confidence values.

Step 7: Similarity refinement

The original similarity matrix is adjusted: $W_{ij}^{(m)} = W_{ij}^{(m)} c_{ij}^{(m)}$.

Step 8: Laplacian construction

The adjusted similarity matrix is converted into a Laplacian: $L^{(m)} = D^{(m)} - W^{(m)}$.

Step 9: Weighted fusion

The modality-specific Laplacians are integrated:$L_{fusion} = \sum k_m L^{(m)}$.

Step 10: Spectral optimization

The eigenvectors corresponding to the smallest eigenvalues of the fused Laplacian are obtained.

Step 11: Spectral representation

The selected eigenvectors form the matrix H.

Step 12: Cancer subtyping

The rows of H are clustered to obtain the final patient subtypes.

Thus, the complete OGC framework is:

Multi-
Omics Data→Similarity Graphs→Dst Estimation→Confidence Estimation→Adaptive Grap
h Refinement→Laplacian Fusion→Spectral Optimization→Cancer Subtypes

4.12 GBM Case Study: Modality Priority

The adaptive weighting mechanism was demonstrated using the glioblastoma multiforme (GBM) cohort. The analysis produced substantially different average Dst values for the three molecular modalities.

Table 2: Average Dst values of GBM Omics

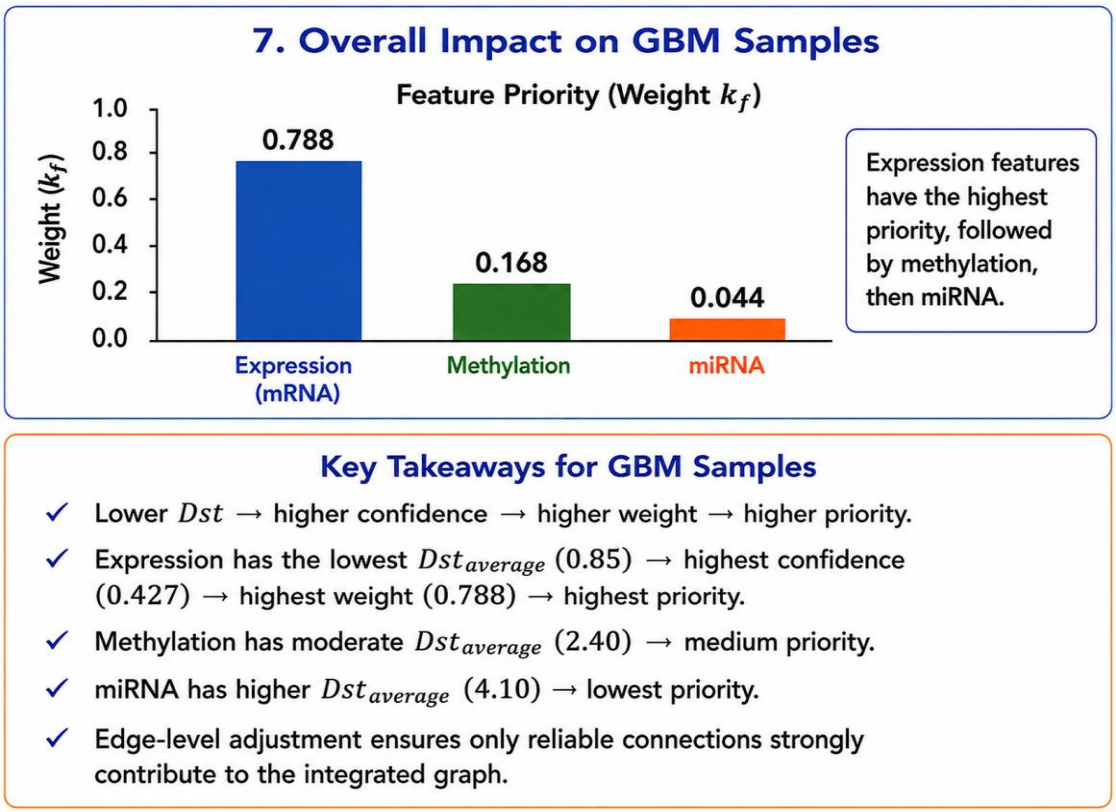
Molecular modality	Average Dst	Interpretation
Gene expression	0.85	Highest neighbourhood consistency
DNA methylation	2.40	Intermediate consistency
miRNA	4.10	Lowest neighbourhood consistency

The resulting normalized weights were approximately:

Table 3: Calculated weights of different Omics of GBM samples

Molecular modality	Normalized weight
Gene expression	0.788
DNA methylation	0.168
miRNA	0.044

Figure 2: Feature



Prioritization in GBM features using OGC methods

These values demonstrate the adaptive behaviour of the proposed method.

Gene expression obtained the largest weight because its average Dst was the lowest. DNA methylation obtained an intermediate weight, while miRNA received the smallest weight because it had the highest average Dst. The resulting weights do not imply that methylation or miRNA information is discarded. Instead, these modalities contribute to the consensus graph with reduced influence according to their estimated neighbourhood reliability.

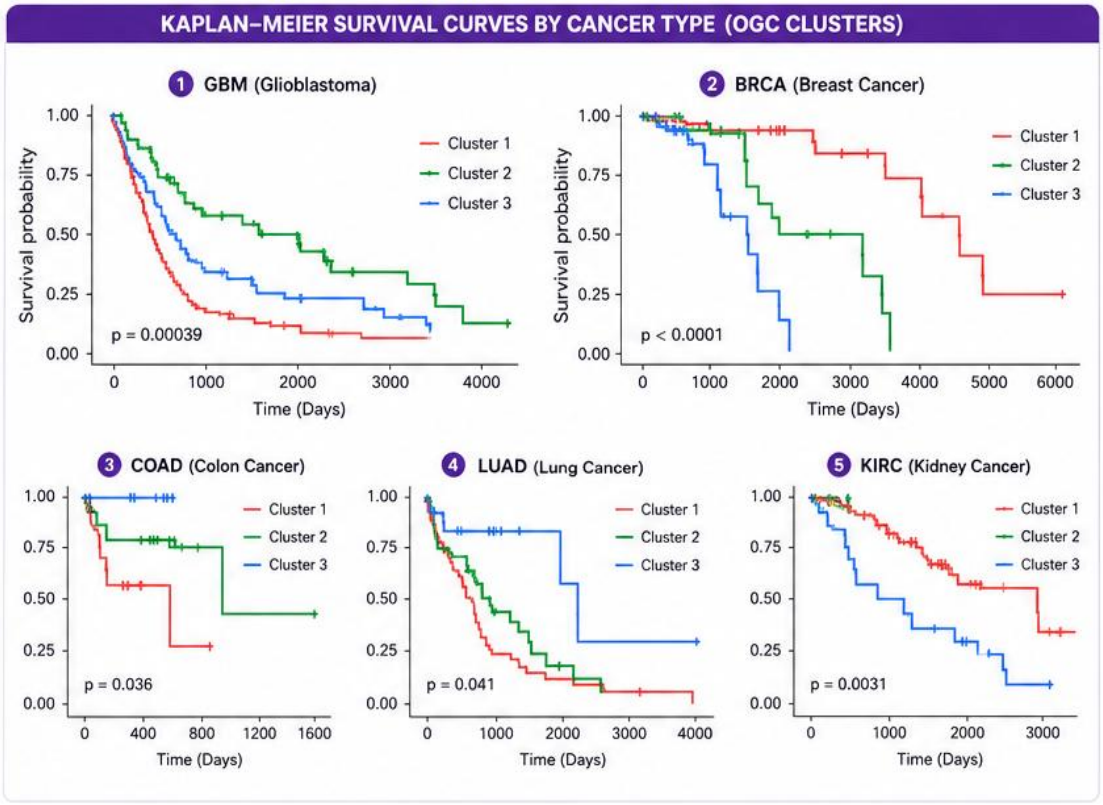
The GBM analysis therefore illustrates the principle:

Neighbourhood consistency→Confidence→Modality priority.

Survival-Based Validation of OGC Subtypes

The identification of mathematically distinct clusters does not, by itself, establish that the clusters have clinical significance [19]. Therefore, survival analysis was used to evaluate whether OGC-derived patient groups exhibited different survival distributions. The Kaplan–Meier estimator provides a non-parametric estimate of the survival function and is particularly

suitable for clinical datasets containing censored observations [15].Kaplan–Meier survival curves were generated for the OGC-derived clusters across the investigated cancer cohorts.



Optimal Graph

The analysis reported the following survival p-values:

Table 4: P-values of different cancer distributions created using 3 feature integration

Cancer type	Reported p-value
GBM	<0.0001
COAD	0.086
LUAD	0.041
KIRC	0.0031

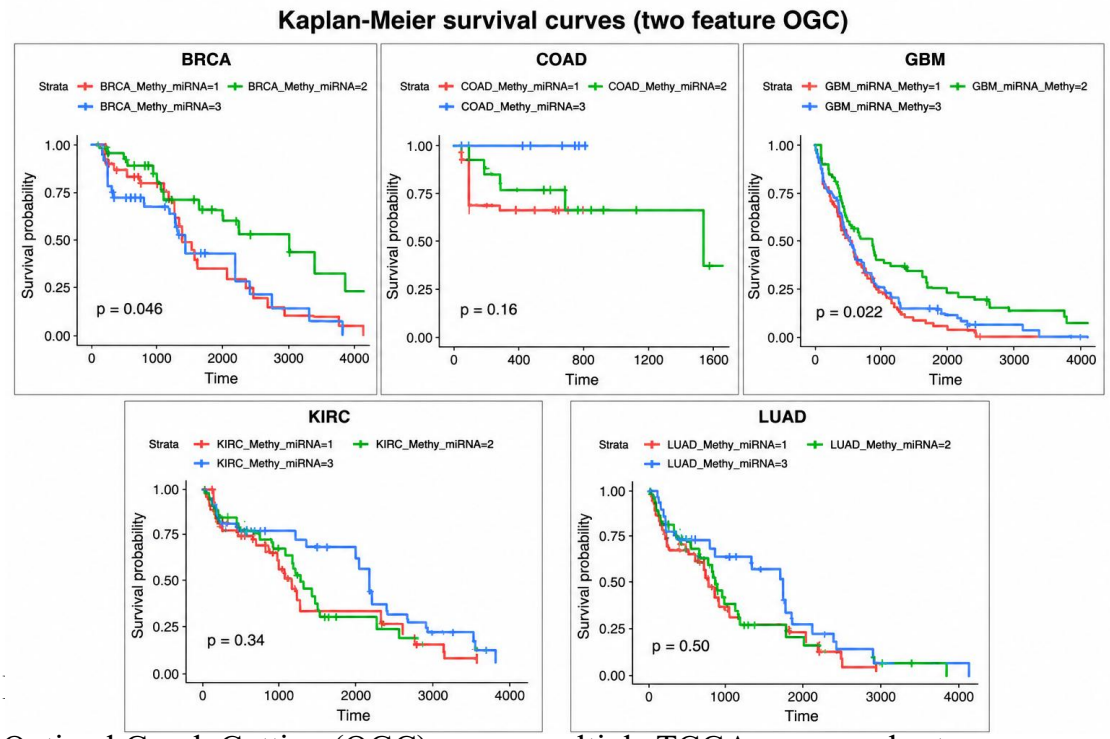
The OGC clusters therefore showed particularly strong survival separation in GBM and KIRC, while the separation was weaker in COAD.

The significance of the survival differences was assessed using the log-rank framework.The results demonstrate that the graph-derived patient groups can capture molecular differences

that are associated with clinically observable differences in survival. The survival analysis is therefore an important component of the validation process because it evaluates the clinical relevance of the computationally derived subtypes rather than relying solely on internal clustering measures [20].

Two-Feature OGC Analysis

The OGC framework was additionally evaluated using combinations of two molecular modalities. In particular, methylation and miRNA information were integrated to investigate whether complementary molecular representations could improve patient stratification.



Two feature Optimal Graph Cutting (OGC) across multiple TCGA cancer cohorts.

The survival analysis produced the following reported p-values:

Table 5: P-values of different cancer distributions created using 2 feature integration

Cancer type	P-value
BRCA	0.046
COAD	0.160
GBM	0.022

The results indicate that the effectiveness of a particular combination of molecular modalities can vary among cancer types. The statistically significant results observed for BRCA and GBM suggest that the combined molecular representation can provide clinically meaningful stratification in selected cohorts [21]. In contrast, the absence of statistically significant separation in COAD indicates that the same molecular combination does not necessarily provide equivalent prognostic discrimination across all cancer types [22]. This observation reinforces the rationale for adaptive weighting. Rather than assuming that the contribution of an omics modality is constant across diseases, OGC allows the molecular contribution to depend on the observed similarity structure.

Two-Feature Weighting in GBM

For the GBM analysis involving DNA methylation and miRNA, the normalized weights obtained from the OGC weighting mechanism were approximately

$k_{methylation} = 0.659$

and

$k_{miRNA} = 0.341.$

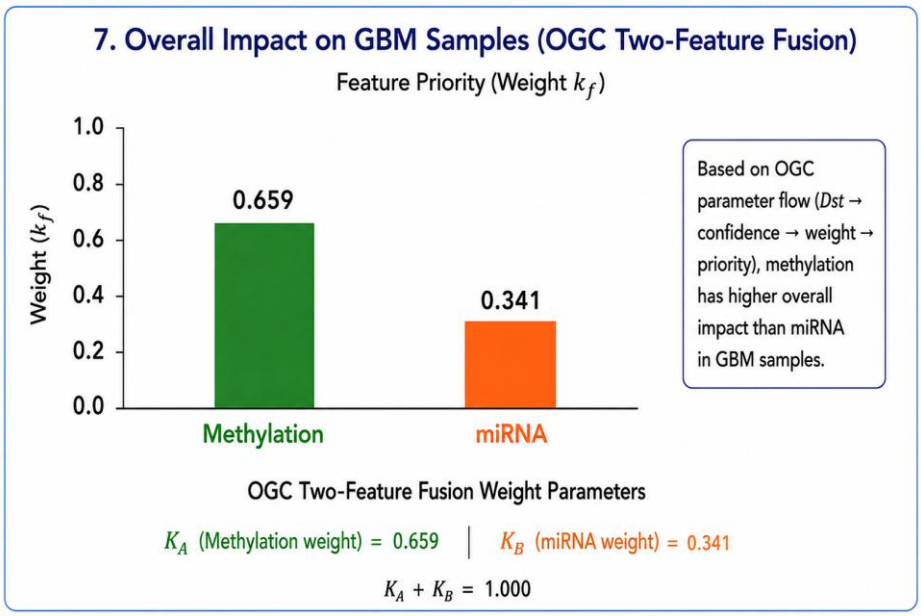
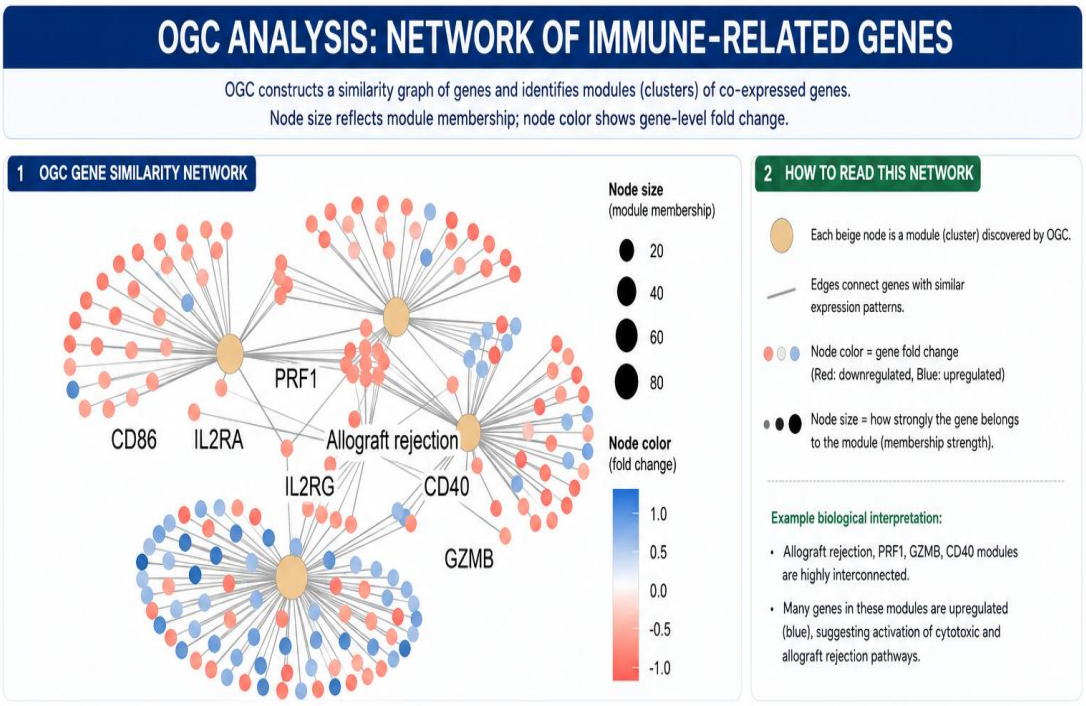


Figure 5: Adaptive weights for two feature clustering

Therefore, methylation contributed approximately twice as much as miRNA to the integrated two-feature graph.This result demonstrates that the adaptive weighting mechanism operates consistently when a subset of molecular modalities is integrated. The weights are not manually imposed but are derived from the estimated reliability of the molecular representations.

Biological Interpretation of OGC-Derived Networks



Figu ıd cytokine on analysis

The evaluation of OGC was extended from patient-level clustering to the interpretation of molecular relationships. A gene similarity network was constructed to examine immune-related genes. In this network, genes are represented as nodes and similarity relationships are represented by edges. The network structure therefore provides information about groups of genes exhibiting coordinated molecular behaviour. The analysis highlighted immune-related genes including **PRF1, GZMB, and CD40**, with modules associated with biological processes including allograft rejection. These pathways indicate a higher immune response in the cells, which predicts a better survival prognosis [23]. As a result, this group has the best survival properties among the clusters created by OGC.

Such graph-based representation is useful because it allows groups of genes to be considered collectively rather than as isolated features [24,25].The identification of coordinated gene modules provides a biological interpretation of the molecular structures underlying the patient groups [26].

Thus, the OGC framework provides a connection between:

Patient Similarity→Patient Clusters→Gene Similarity→Biological Modules.

Biological Characterization of Breast Cancer Groups

The biological interpretation was further examined for breast cancer.In this case two of the representative patient groups displayed distinct molecular characteristics.

Immune-Activated Group

One group was characterized as an **immune-activated cluster**. In this group, the enriched pathways included:

> Th1 cell differentiation, Th2 cell differentiation,Th17 cell differentiation,immune-associated processes,endocytosis, and proteoglycans in cancer.[27]

The enrichment of immune-associated processes indicates that this patient group (Group 1) is characterized by a strong immune-related molecular phenotype [28]. This is plotted as high surviving samples in the KM curves shown in the figure(4). This biological signature provides an explanation for the molecular separation observed through graph-based clustering.

Proliferative Group

Another group was characterized as a **proliferative cluster**.

The dominant biological characteristics involved:

> cell-cycle progression, cellular proliferation, mitotic activity, and proliferation-associated molecular processes [29].

The distinction between an immune-associated group and a proliferative group demonstrates that the OGC clusters can correspond to biologically interpretable tumour states rather than representing purely mathematical partitions.[30]

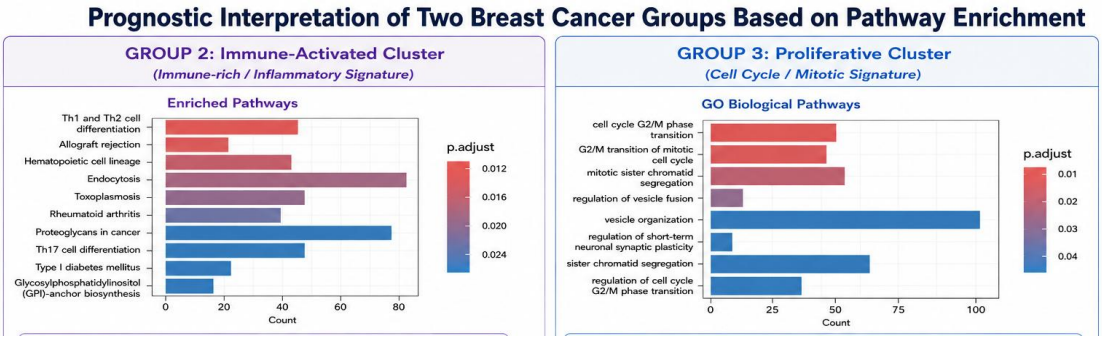


Figure 7: Gene Ontology (GO) biological process enrichment analysis of OGC-derived clusters, emphasizing cell-cycle and vesicle organization processes.

Clinical and Biological Significance

The results obtained through OGC can be considered at three complementary levels.

Computational significance: At the computational level, OGC constructs an integrated patient graph and performs spectral graph partitioning using an adaptive weighted Laplacian.

Clinical significance: At the clinical level, the resulting clusters can be evaluated using survival analysis. According to this study, the observed survival differences indicate that selected OGC-derived molecular groups are associated with clinically distinct outcomes.

Biological significance : At the biological level, the patient groups can be interpreted using gene similarity networks and pathway enrichment. The identification of immune-activated

and proliferative groups demonstrates that the computational clusters can correspond to recognizable molecular phenotypes.

Therefore, the overall interpretation follows the sequence:

Graph Structure→Cancer Subtype→Survival Difference→Biological Interpretation

This is particularly important in cancer informatics because the ultimate objective of cancer subtyping is not merely to generate mathematically separated groups. The groups should ideally provide molecular and clinical information that can contribute to understanding disease heterogeneity. [31,32]

Comparison with Conventional Multi-Omics Graph Integration

Existing multi-omics network approaches have demonstrated the value of constructing and integrating patient similarity networks. SNF, for example, constructs one network per data type and fuses the networks into a common patient representation [1].The OGC framework extends this general network-based philosophy in several important ways.

First, OGC explicitly formulates the integration problem in terms of graph-cut optimization.

Second, it introduces a modality-level confidence mechanism based on Dst.

Third, it introduces edge-level confidence so that unreliable patient relationships can be attenuated even within an otherwise informative modality.

Fourth, the resulting weighted graphs are integrated through a Laplacian formulation and optimized using spectral methods.

The conceptual difference can therefore be represented as:

Conventional network fusion→Network integration→Clustering

whereas OGC follows

Network construction→Reliability estimation→Adaptive graph refinement→Weighte d graph-cut optimization→Clustering.

This distinction is important because the objective is not merely to combine networks but to determine an integrated graph that is appropriate for the graph-partitioning problem.

Discussion

The OGC framework addresses several challenges associated with multi-omics cancer subtyping. First, different molecular modalities may provide different levels of information concerning patient similarity [33]. OGC addresses this issue through modality-level adaptive weighting. Second, reliability may vary within an individual modality. Edge-level Dst confidence addresses this local variability. Third, the reliability estimates are incorporated directly into the graph representation used for clustering rather than being treated as an unrelated preprocessing operation. Fourth, spectral optimization provides a mathematically established mechanism for converting the integrated graph into a low-dimensional patient representation [16,17]. Finally, survival and biological analyses provide complementary evidence concerning the relevance of the resulting clusters [34].

This study demonstrates that the optimal contribution of an omics layer cannot necessarily be assumed to be constant across different integration settings.The adaptive nature of OGC therefore provides a mechanism for allowing the observed molecular structure to determine the relative contribution of the available modalities.

Limitations

Although OGC provides an adaptive mechanism for multi-omics graph integration, several limitations should be considered.

Dependence on neighbourhood construction

Dst estimation depends on the neighbourhood structure used for each patient. Therefore, changes in the number of neighbours can influence the estimated reliability.

Parameter dependence

The edge-confidence function depends on the parameter α, while the modality-level confidence depends on β. Different parameter choices can change the degree to which unreliable relationships are attenuated.

Similarity construction

The quality of the resulting graph depends on the similarity measure used to construct the individual molecular graphs. Consequently, systematic differences in scaling or preprocessing can influence the graph structure.

Biological interpretation

Graph similarity and pathway enrichment provide associations rather than causal explanations. Therefore, the biological interpretation of the identified modules should be considered hypothesis-generating and should ideally be complemented by experimental or independent validation.

Cohort-specific weighting

Because the modality weights are estimated from the observed neighbourhood structure, the resulting weights may be cohort-specific. A modality receiving a high weight in one cancer type should not automatically be assumed to receive the same weight in another cohort.

These limitations are important when interpreting the OGC results and when considering application of the framework to independent datasets.

Summary of the Research Attainment

The OGC framework developed in this study provides an adaptive graph-based approach for multi-omics cancer subtyping.

The principal contributions are summarized below.

1. Consensus graph construction

OGC integrates multiple omics-specific patient similarity graphs into a common graph representation.

2. Graph-cut optimization

The integrated graph is partitioned through a spectral relaxation of the graph-cut objective [18,17].

3. Adaptive modality weighting

The Dst-based confidence mechanism determines the relative contribution of different omics modalities.

4. Edge-level reliability

Individual patient relationships are weighted according to their local neighbourhood consistency.

5. Cancer-specific adaptation

The contribution of the different molecular modalities can vary according to the characteristics of the cancer cohort.

6. Clinical validation

Kaplan–Meier and survival-based analyses were used to evaluate whether the resulting patient groups exhibit different clinical outcomes [15].

7. Biological interpretation

Gene similarity networks and pathway-level analyses were used to interpret the molecular characteristics of selected patient groups.

The overall methodological contribution can be summarized as:

Multi-Omics Feature Graphs→Dst-
Based Reliability→Adaptive Consensus Graph→Optimal Graph Cut→Cancer Subtypes

Conclusion

This chapter presented the **Optimal Graph Cutting (OGC)** framework developed for multi-omics cancer subtyping.

The approach begins by representing each molecular modality as a patient similarity graph. These graphs are then evaluated using the Dst measure to estimate neighbourhood consistency. The resulting reliability information is incorporated at two levels: modality-level confidence determines the global contribution of each molecular layer, while edge-level confidence determines the reliability of individual patient relationships.

The adjusted molecular graphs are represented using graph Laplacians and integrated into a weighted consensus Laplacian. Spectral optimization is subsequently used to obtain a low-dimensional patient representation, from which cancer subtypes are derived.

The analyses presented in this study demonstrate that the resulting clusters can exhibit clinically meaningful survival differences in selected cancer cohorts. The subsequent gene-network and pathway analyses further indicate that the identified patient groups can possess distinct biological characteristics, including immune activation and proliferative activity.

The principal contribution of OGC is therefore the integration of **multi-omics graph representation, adaptive reliability estimation, and graph-cut optimization** into a single framework.

Unlike a fixed-weight integration strategy, the proposed approach allows the molecular data to determine the relative contribution of the available modalities through their neighbourhood consistency. At the same time, the edge-level confidence mechanism allows unreliable patient relationships to be attenuated locally.

Consequently, OGC establishes an adaptive graph-based approach in which:

Molecular Data→Patient Similarity→Reliability→Adaptive Graph→Optimal Partition→Clinical/Biological Subtypes

This research attainment forms an important step in the overall thesis framework by moving from confidence-based multi-omics integration toward **adaptive graph-based cancer subtyping**, and provides the graph-based foundation for the subsequent progression-aware multi-omics analysis.

Robustness

Cross-validation experiments further demonstrate the robustness and reproducibility of the OGC framework. Across repeated data splits, OGC achieved an **Adjusted Rand Index (ARI) of at least 0.72**, indicating strong agreement between cluster assignments obtained from independent training folds. These values are comparable to, and in several cases exceed, those obtained using CBI[35], highlighting that the gains in separability introduced by ratio-cut pruning do not come at the expense of clustering consistency.

Importantly, the convergence of high ARI scores demonstrates that OGC improves both **discrete clustering outcomes** and **continuous geometric structure**. This dual stability is critical for downstream analyses, including gated pseudotime inference [36], where consistent eigen-subspaces are required to infer reliable progression trajectories.

Collectively, these cross-validation results corroborate the spectral stability of OGC and reinforce its suitability as a foundation for reproducible, progression-aware multi-omics modeling.

Table 6: Robustness summary

Metric	Result (OGC)	Comparator note
ARI (10-fold CV)	≥ 0.72	Comparable to or exceeding CBI

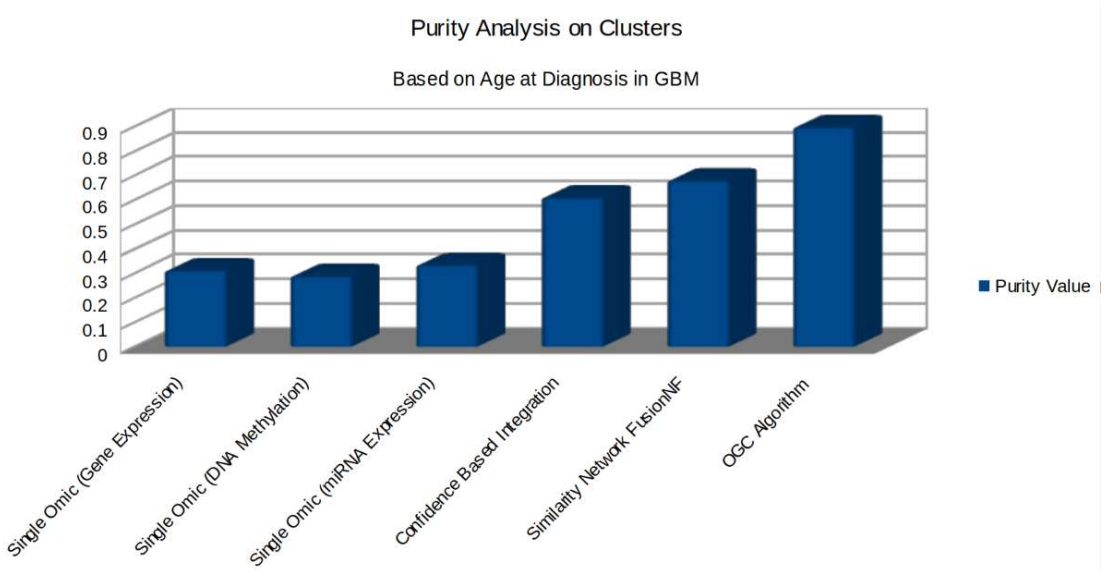


Figure 8: Comparison of cluster purity (based on age-at-diagnosis in GBM) across integration strategies. OGC achieves the highest purity, surpassing Confidence-Based Integration (CBI) and Similarity Network Fusion (SNF), validating its superior boundary sharpness.

Purity analysis (Figure 8) further demonstrates OGC's discriminative power. The gain over SNF and CBI confirms that ratio-cut pruning yields more coherent, biologically interpretable clusters.

Acknowledgement

The author acknowledges The Cancer Genome Atlas (TCGA) research program for providing access to the multi-omics datasets used in this study. These publicly available data resources enabled the investigation of graph-based integration methods for cancer subtyping [37]. The author also acknowledges the developers and maintainers of the open-source software packages and computational tools used in the implementation and evaluation of the proposed Optimal Graph Cutting (OGC) framework [38.39].

Data Availability Statement

The data underlying this article will be shared on reasonable request to the corresponding author.

References

[1] Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*. 2014;11:333–337. https://doi.org/10.1038/nmeth.2810

[2] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*. 2018;46(20):10546–10562. https://doi.org/10.1093/nar/gky889

[3] Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*. 2019;35(18):3348–3356. https://doi.org/10.1093/bioinformatics/btz058

[4] Duan X, Ding X, Zhao Z. Multi-omics integration with weighted affinity and self-diffusion applied for cancer subtypes identification. J Transl Med. 2024 Jan 19;22(1):79. doi: 10.1186/s12967-024-04864-x. PMID: 38243340; PMCID: PMC10799401.

[5] Duan R, Gao L, Gao Y, Hu Y, Xu H, Huang M, Song K, Wang H, Dong Y, Jiang C, Zhang C, Jia S. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLoS Comput Biol. 2021 Aug 12;17(8):e1009224. doi: 10.1371/journal.pcbi.1009224. PMID: 34383739; PMCID: PMC8384175.

[6] Liou K, Wang JP. Integrating genetic and gene expression data in network-based stratification analysis of cancers. BMC Bioinformatics. 2025 May 13;26(1):126. doi: 10.1186/s12859-025-06143-y. PMID: 40360993; PMCID: PMC12070578.

[7] Duan X, Ding X, Zhao Z. Multi-omics integration with weighted affinity and self-diffusion applied for cancer subtypes identification. J Transl Med. 2024 Jan 19;22(1):79. doi: 10.1186/s12967-024-04864-x. PMID: 38243340; PMCID: PMC10799401.

[8] Liu J, Xue X, Wen P, Song Q, Yao J, Ge S. Multi-fusion strategy network-guided cancer subtypes discovering based on multi-omics data. Front Genet. 2024 Nov 14;15:1466825. doi: 10.3389/fgene.2024.1466825. PMID: 39610828; PMCID: PMC11602503.

[9] Chierici M, Bussola N, Marcolini A, Francescatto M, Zandonà A, Trastulla L, Agostinelli C, Jurman G, Furlanello C. Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. Front Oncol. 2020 Jun 30;10:1065. doi: 10.3389/fonc.2020.01065. PMID: 32714870; PMCID: PMC7340129.

[10] Fiedler, M. (1973). "Algebraic connectivity of graphs." *Czechoslovak Mathematical Journal*, 23(2), 298-305.

[11] von Luxburg U. A tutorial on spectral clustering. *Statistics and Computing*. 2007;17:395–416. https://doi.org/10.1007/s11222-007-9033-z

[12] Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems 14*. MIT Press; 2002:849–856.

[13] Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J. 2022 Dec 1;21:134-149. doi: 10.1016/j.csbj.2022.11.050. PMID: 36544480; PMCID: PMC9747357.

[14] Kannan R, Vempala S, Vetta A. On clusterings: good, bad and spectral. *Journal of the ACM*. 2004;51(3):497–515. https://doi.org/10.1145/990308.990313

[15] Kaplan, E. L., & Meier, P. (1958). "Nonparametric estimation from incomplete observations." *Journal of the American Statistical Association*, 53(282), 457–481. DOI: 10.2307/2281868

[16] von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. *The Annals of Statistics*. 2008;36(2):555–586. https://doi.org/10.1214/009053607000000640

[17] Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. In: *Methods in Molecular Biology*. 2018;1831:141–154. https://doi.org/10.1007/978-1-4939-8649-6_10

[18] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543

[19] Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*. 2013;110(11):4245–4250. https://doi.org/10.1073/pnas.1208949110

[20] Poulakis K, Westman E. Clustering and disease subtyping in Neuroscience, toward better methodological adaptations. Front Comput Neurosci. 2023 Oct 19;17:1243092. doi: 10.3389/fncom.2023.1243092. PMID: 37927546; PMCID: PMC10620518.

[21] Harbi E, Aschner M. Role of BRCA1 in glioblastoma etiology. Cell Oncol (Dordr). 2024 Dec;47(6):2091-2098. doi: 10.1007/s13402-024-01024-7. Epub 2024 Dec 10. PMID: 39656422; PMCID: PMC12974051.

[22] Gu L, Jiang C, Xu C, Liu Y, Zhou H. Based on Molecular Subtypes, Immune Characteristics and Genomic Variation to Constructing and Verifying Multi-Gene Prognostic Characteristics of Colorectal Cancer. Front Cell Dev Biol. 2022 Feb 23;10:828415. doi: 10.3389/fcell.2022.828415. PMID: 35281077; PMCID: PMC8905350.

[23] Oshi M, Patel A, Wu R, Le L, Tokumaru Y, Yamada A, Yan L, Matsuyama R, Ishikawa T, Endo I, Takabe K. Enhanced immune response outperform aggressive cancer biology and is associated with better survival in triple-negative breast cancer. NPJ Breast Cancer. 2022 Aug 9;8(1):92. doi: 10.1038/s41523-022-00466-2. PMID: 35945417; PMCID: PMC9363489.

[24] Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinformatics. 2004 Jul 23;5:100. doi: 10.1186/1471-2105-5-100. PMID: 15272936; PMCID: PMC509016.

[25] Liu Y, Huse J, Kannan K. Expression graph network framework for biomarker discovery. Brief Bioinform. 2025 Aug 31;26(5):bbaf559. doi: 10.1093/bib/bbaf559. PMID: 41139924; PMCID: PMC12554635.

[26] Zhang Y, Shi W, Sun Y. A functional gene module identification algorithm in gene expression data based on genetic algorithm and gene ontology. BMC Genomics. 2023 Feb 17;24(1):76. doi: 10.1186/s12864-023-09157-z. PMID: 36797662; PMCID: PMC9936134.

[27] Huang G, Wang Y, Chi H. Regulation of TH17 cell differentiation by innate immune signals. Cell Mol Immunol. 2012 Jul;9(4):287-95. doi: 10.1038/cmi.2012.10. Epub 2012 Apr 16. PMID: 22504954; PMCID: PMC3423893.

[28] Li Z, Wang L, Yang S, Luo B, Liu Y, Chen M, Wang C. Immune-associated molecular classification and prognosis signature of sepsis. PLoS One. 2025 Jun 12;20(6):e0326083. doi: 10.1371/journal.pone.0326083. PMID: 40504881; PMCID: PMC12161593.

[29] Matson JP, Cook JG. Cell cycle proliferation decisions: the impact of single cell analyses. FEBS J. 2017 Feb;284(3):362-375. doi: 10.1111/febs.13898. Epub 2016 Oct 5. PMID: 27634578; PMCID: PMC5296213.

[30] Feng P, Li Z, Li Y, Zhang Y, Miao X. Characterization of Different Subtypes of Immune Cell Infiltration in Glioblastoma to Aid Immunotherapy. Front Immunol. 2022 Jun 21;13:799509. doi: 10.3389/fimmu.2022.799509. PMID: 35799789; PMCID: PMC9254719.

[31] Zhao L, Lee VHF, Ng MK, Yan H, Bijlsma MF. Molecular subtyping of cancer: current status and moving toward clinical applications. Brief Bioinform. 2019 Mar 25;20(2):572-584. doi: 10.1093/bib/bby026. PMID: 29659698.

[32] Shen JP. Artificial intelligence, molecular subtyping, biomarkers, and precision oncology. Emerg Top Life Sci. 2021 Dec 21;5(6):747-756. doi: 10.1042/ETLS20210212. PMID: 34881776; PMCID: PMC8786277.

[33] Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol. 2018 Sep 14;430(18 Pt A):2924-2938. doi: 10.1016/j.jmb.2018.05.037. Epub 2018 Jun 1. PMID: 29860027; PMCID: PMC6097926.

[34] Zhang JZ, Wang C. A comparative study of clustering methods on gene expression data for lung cancer prognosis. BMC Res Notes. 2023 Nov 8;16(1):319. doi: 10.1186/s13104-023-06604-8. PMID: 37941025; PMCID: PMC10630994.

[35] R. Sreekumar, Farida Khursheed, Identifying cancer sub-types from genomic scale data sets using confidence based integration (CBI), Journal of Biomedical Informatics,Volume 126, 2022,103997 ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2022.103997.

(https://www.sciencedirect.com/science/article/pii/S1532046422000132)

[36] Sreekumar, R., and Farida Khursheed. "Gated Pseudotime for Medical Informatics: Robust Multi-Omic Trajectories and Survival Stratification." *Intelligence-Based Medicine* (2026): 100441.

[37] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490:61–70. https://doi.org/10.1038/nature11412

[38] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–287. https://doi.org/10.1371/journal.pcbi.1005752

[39] Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*. 2017;13(11):e1005752.

https://doi.org/10.1089/omi.2011.0118