Multi-Omics Integration Data for PDAC Stratification

Valentina Debbia - 74280A

November 25, 2025

Abstract

Pancreatic adenocarcinoma (PDAC) exhibits high molecular heterogeneity, limiting the ability of conventional clinical classifications to capture the biological complexity of the disease. This study develops an integrative framework based on six omics levels - circRNA, miRNA, mRNA, phosphoproteome, proteome, and copy number alterations - for patient stratification using Similarity Network Fusion (SNF). The resulting integrated network is analyzed using spectral clustering to identify distinct molecular subtypes, while machine learning models applied to the derived embeddings are used to predict tumor stage and assess the informative contribution of each omics level.

1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal malignancies, characterized by late diagnosis, aggressive progression, and limited therapeutic options. The five-year survival rate stands at only 8% for pancreatic adenocarcinoma specifically, and 13% for all pancreatic cancers combined [1]. This dismal prognosis, combined with resistance to conventional therapies, underscores the urgent need for improved diagnostic and therapeutic approaches. Recent advances in high-throughput sequencing and proteomics technologies have enabled comprehensive molecular profiling of cancer patients, revealing extensive inter-patient heterogeneity that traditional histopathological classification fails to capture.

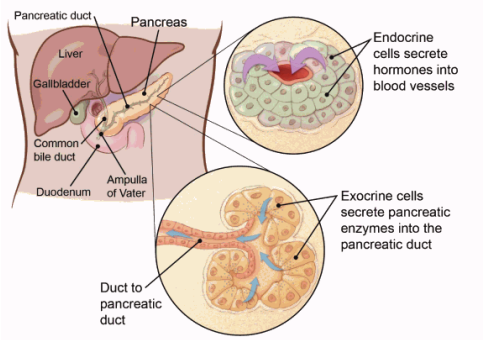

Figure 1: Pancreas anatomy

Multi-omics data integration has emerged as a powerful strategy for understanding cancer complexity. By combining information from different molecular layers - including genomics, transcriptomics, and proteomics - researchers can obtain a more holistic view of disease mechanisms and identify clinically relevant patient subtypes. However, integrating heterogeneous omics data presents significant computational challenges due to differences in data dimensionality, scale, and noise characteristics.

Similarity Network Fusion (SNF) has emerged as a powerful framework for multi-omics data integration. SNF constructs patient similarity networks from individual omics layers and iteratively fuses them through a nonlinear combination process, allowing complementary information from different data types to enhance overall network structure while preserving layer-specific patterns. This approach has demonstrated success in cancer subtyping and outcome prediction across various malignancies [2].

Building on these advances, this study aims to leverage Similarity Network Fusion to integrate six complementary omics layers - circRNA, miRNA, mRNA, phosphoproteome, proteome, and somatic copy number alterations - from a cohort of pancreatic adenocarcinoma patients. The objective is to construct a unified multi-layer patient similarity network that captures the molecular heterogeneity of PDAC. Specifically, the project seeks to (1) identify data-driven molecular subtypes emerging from the fused network, (2) characterize their clinical and pathological features, and (3) evaluate the extent to which integrated multi-omics profiles can support the prediction of tumor stage. This framework is designed to provide a comprehensive, systems-level perspective on PDAC biology that extends beyond traditional clinical classification.

2 Methods

This section provides a comprehensive overview of the methodological framework employed in this study, with each component described in sufficient detail to ensure full reproducibility of the analyses.

2.1 Dataset

The study utilized data from 140 pancreatic adenocarcinoma patients, with 137 patients having complete multi-omics profiles. The dataset comprises six omics layers:

- `circRNA`: 3,979 circular RNA features (0.00% missing)
- `miRNA`: 2,416 microRNA features (64.85% missing)
- `mRNA`: 28,057 messenger RNA features (0.00% missing)

- `Phosphoproteome`: 8,004 phosphorylated protein features (32.04% missing)
- `Proteome`: 11,662 protein features (24.56% missing)
- `SCNA`: 19,906 somatic copy number alteration features (0.08% missing)

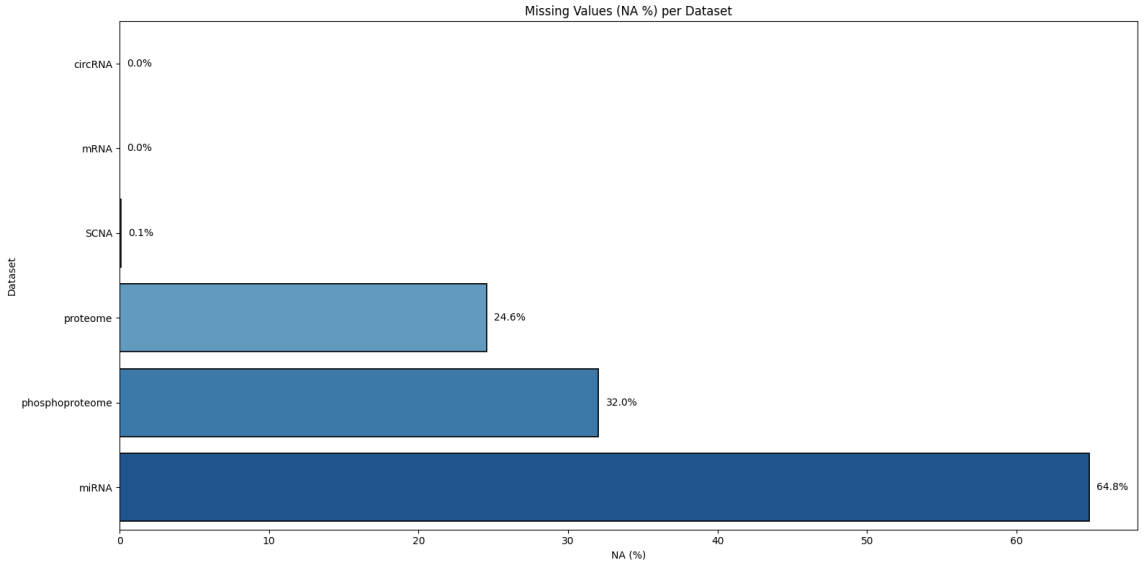

Figure 2: Missing Values per Dataset

Clinical data including demographic information, tumor characteristics, treatment history, and survival outcomes were also collected for all patients.

2.2 Data Preprocessing

To prepare all inputs for integrated analysis, a unified preprocessing framework was applied to ensure consistency, minimize technical noise, and preserve meaningful biological and clinical variation. The following sections outline the specific steps used for omics and clinical data.

2.2.1 Omics Data preprocessing

All six omics layers underwent a standardized preprocessing workflow designed to ensure comparability across data types and to minimize technical artifacts.

First, patient alignment was performed to retain only those individuals for whom complete profiles were available across all omics platforms and clinical records, resulting in a final cohort of 137 patients. Each dataset was then transposed so that patients corresponded to rows and molecular features to columns, providing a consistent structure for downstream analyses.

Missing values were addressed through median imputation applied independently within each feature, a strategy that preserves the central tendency of the data while avoiding assumptions about distributional shape. To eliminate non-informative variables, features exhibiting zero variance across patients were removed.

Finally, all remaining features were standardized using z-score normalization (mean = 0, standard deviation = 1), ensuring that differences in measurement scale or dynamic range across omics layers did not bias similarity calculations or clustering procedures. This preprocessing pipeline establishes a harmonized and analytically robust foundation for multi-omics integration.

After quality control and removal of zero-variance features, the final datasets contained 3,167 circRNA, 2,344 miRNA, 25,954 mRNA, 7,951 phosphoproteome, 11,631 proteome, and 19,892 SCNA features. Clinical data included demographic information, tumor characteristics, treatment history, and survival outcomes.

Omics Features	**Original**	**Final**
circRNA	3,979	3,167
miRNA	2,416	2,344
mRNA	28,057	25,954
Phosphoproteome	8,004	7,951
Proteome	11,662	11,631
SCNA	19,906	19,892

Table 1: Comparison before and after preprocessing

2.2.2 Clinical Data preprocessing

Clinical variables underwent careful encoding based on their data types:

- `Binary variables`: for example, sex, vital status, tumor necrosis. They were encoded as 0/1 numeric values
- `Ordinal variables`: for example, pathologic staging pN, pT, pM; tumor stage. They were mapped to ordered numeric scales preserving biological progression
- `Categorical variables`: for example,race, country, tumor site. They were one-hot encoded to create binary indicator variables for each category

Missing values in ordinal variables were explicitly encoded as NaN to distinguish true missingness from low-grade categories. The final preprocessed clinical dataset contained 67 numeric features for 137 patients.

2.3 Similarity network construction

For each omics layer, patient-to-patient similarity networks were constructed using a kernel-based approach with k-nearest neighbors (k-NN) adaptive scaling. Pairwise distances between patients were first computed, using Pearson correlation distance (1 - correlation) for expression-based datasets (circRNA, miRNA, mRNA) to capture co-expression structure, and Euclidean distance for phosphoproteomic, proteomic, and SCNA data. For each patient i, the local scaling parameter σ_i was then defined as the mean distance to its 20 nearest neighbors, enabling the kernel bandwidth to adapt to local density variations. These distances were subsequently transformed into similarities through a Gaussian kernel, $S(i,j) = exp(-d(i,j)^2/(2\sigma_i^2))$, such that closer patients received higher similarity scores. Finally, the resulting matrix was symmetrized by averaging S with its transpose, yielding a symmetric similarity matrix with unit diagonal suitable for downstream network-based integration.

This procedure generated six patient similarity matrices, one for each omics layer.

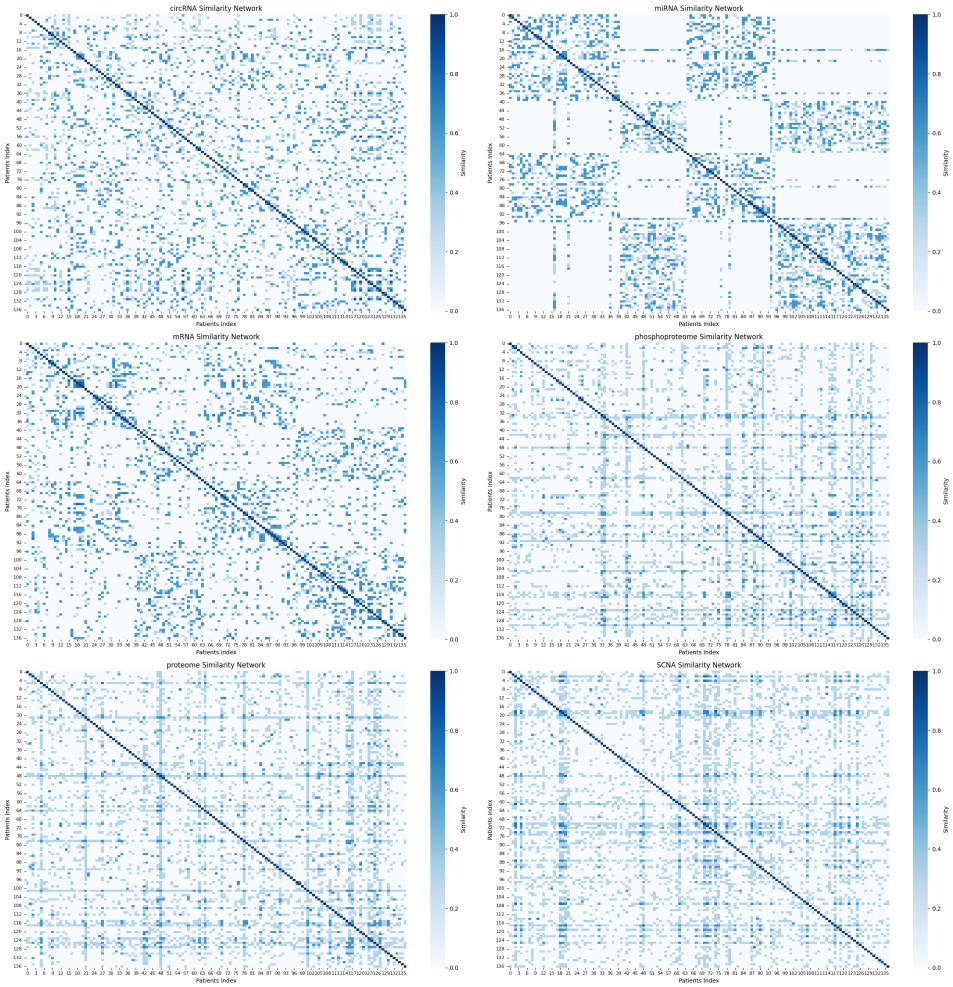

Figure 3: SNF for each omics

2.4 Similarity Network Fusion (SNF)

Similarity Network Fusion integrates multiple similarity networks through an iterative nonlinear message-passing process. The algorithm operates as follows:

1. **Initialization**. For each omics layer m, two matrices are maintained:
   - $P^{(m)}$: the similarity network for layer m
   - $S^{(m)}$: a normalized version capturing local neighborhood structure
2. **Iterative Fusion**. At each iteration t, networks are updated by $P^{(m)}(t+1) = S^{(m)} \times \left[\frac{\sum_{n \neq m} P^{(n)}(t)}{M-1}\right] \times \left(S^{(m)}\right)^T$ where M is the total number of omics layers. This formula allows each network to borrow information from other networks while preserving its local structure through $S^{(m)}$.
3. **Convergence**. The process iterates for T iterations ($T = 20$ in this study) until the fused network converges to a stable state that integrates complementary information from all omics layers.

After defining the iterative fusion procedure, three hyperparameters govern the behavior of SNF. The parameter K controls the construction of the local similarity structure used to derive $S^{(m)}$. The number of fusion iterations T determines how many times information is exchanged across networks before convergence. Finally, the weighting parameter α regulates the balance between local neighborhood preservation and global similarity during normalization. Together, these hyperparameters shape how rapidly and how strongly information propagates across omics layers throughout the fusion process. The output is a single fused similarity matrix representing integrated multi-omics patient relationships.

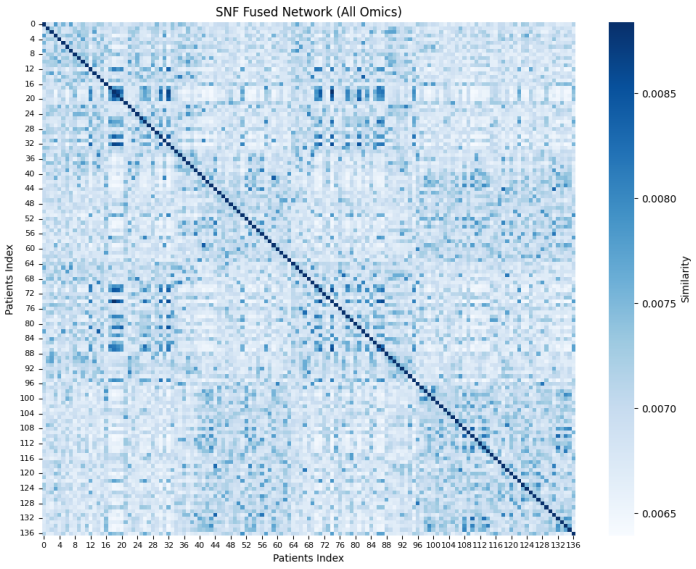

Figure 4: SNF combining all omics

3 Clustering

To enable visualization and support downstream clustering analyses, multiple dimensionality reduction techniques were applied to the fused similarity network.

a) Spectral Embedding was computed on the affinity matrix to extract 20 principal eigenvectors, capturing the dominant modes of variation while preserving local neighborhood structure.

b) Multidimensional Scaling was performed on the distance matrix (1−similarity) to obtain 20-dimensional representations optimized for global distance preservation.

c) UMAP was then used to generate 2 - dimensional embeddings from the distance matrix, employing 15 nearest neighbors and a minimum distance of 0.1 to balance local and global structure.

d) Finally, t-SNE produced additional 2 - dimensional embeddings using a perplexity of 30 and initialization based on the first two spectral components to enhance stability.

These reduced representations facilitated both the visualization of patient relationships and served as input features for subsequent clustering procedures.

3.1 Optimal cluster number determination

To identify the optimal number of patient clusters, clustering quality was evaluated for $K = 2$ to $K = 10$ on the fused similarity matrix. Three complementary metrics were computed. The `Silhouette Score` quantifies cluster cohesion and separation, ranging from -1 to 1, with higher values indicating better-defined clusters. The `Davies-Bouldin Index` measures the ratio between within-cluster and between-cluster distances, where lower values reflect more compact and well-separated clusters. The `Calinski-Harabasz Score` evaluates the ratio of between-cluster to within-cluster dispersion, with higher values indicating clearer and more distinct cluster structure.

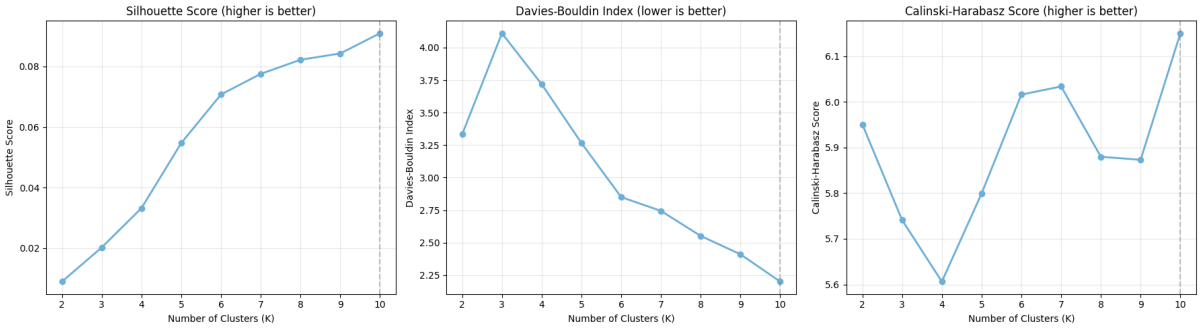

Figure 5: Identify the optimal K

Across all three indices, the evaluation consistently pointed to $K = 10$ as the most appropriate solution. The Silhouette Score reached its highest value at $K = 10$ (0.091), the Davies–Bouldin Index attained its lowest value at the same point (2.200), and the Calinski–Harabasz Score also peaked at $K = 10$ (6.2). This convergence of independent metrics provided robust evidence supporting a ten-cluster structure.

3.2 Clustering algorithm comparison

Four distinct clustering algorithms were applied to identify patient subtypes:

1. **Spectral Clustering**: applied directly to the fused similarity matrix (`affinity='precomputed'`) with k-means label assignment. This method is well-suited for graph-based data and can identify non-convex clusters.
2. **Gaussian Mixture Model (GMM)**: applied to the 20 - dimensional spectral embedding. GMM assumes data are generated from a mixture of Gaussian distributions and provides soft clustering with probabilistic cluster assignments. Full covariance matrices were used to capture cluster shape variations.
3. **HDBSCAN (Hierarchical Density-Based Spatial Clustering)**: applied to the distance matrix using hierarchical density estimation. This algorithm automatically identifies clusters of varying densities and flags outliers as noise points (`cluster label = -1`). Parameters: minimum cluster `size=5`, minimum `samples=2`.
4. **Hierarchical Clustering**: applied Ward linkage to the distance matrix and extracted $K = 10$ clusters. Ward's method minimizes within-cluster variance at each merge step.

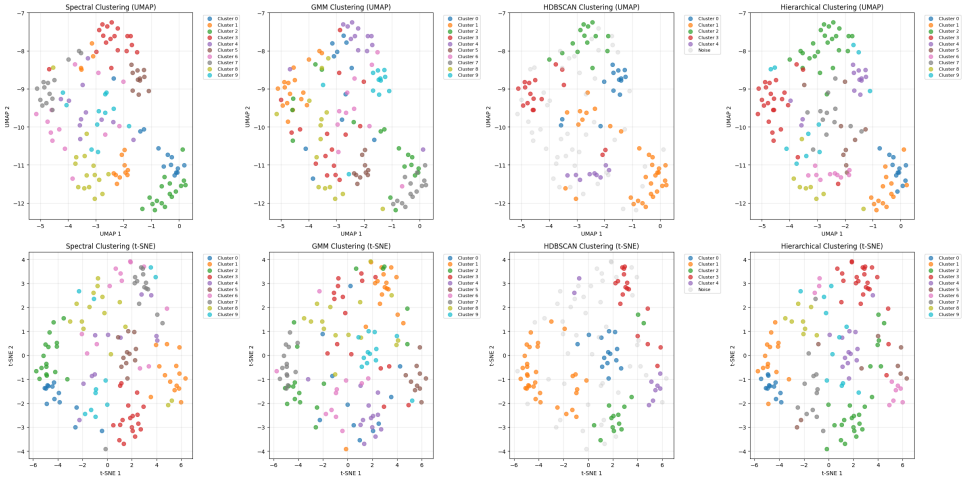

Figure 6: Comparison of four clustering methods in UMAP (top) and t-SNE (bottom)

Across these methods, Spectral Clustering produced the most balanced and well-separated partitions. To systematically compare performance, a composite score was computed by combining normalized clustering validity metrics: 30% Silhouette Score, 25% Davies-Bouldin Index, 25% Calinski-Harabasz Score, and 20% cluster balance.

Method	Silhouette	Davies-Bouldin	Calinski-Harabasz	Composite Score
Spectral	0.091	2.200	6.2	0.651
GMM	0.095	2.257	6.1	0.539
HDBSCAN	0.125	2.338	6.4	0.550
Hierarchical	0.077	2.294	5.9	0.131

Table 2: Comparison of clustering methods across validity metrics

This evaluation identified Spectral Clustering as the most reliable approach, achieving the highest composite score while maintaining well-proportioned cluster sizes suitable for downstream analyses.

Having identified Spectral Clustering as the most reliable method, the resulting ten clusters were examined to assess their size distribution and overall balance.

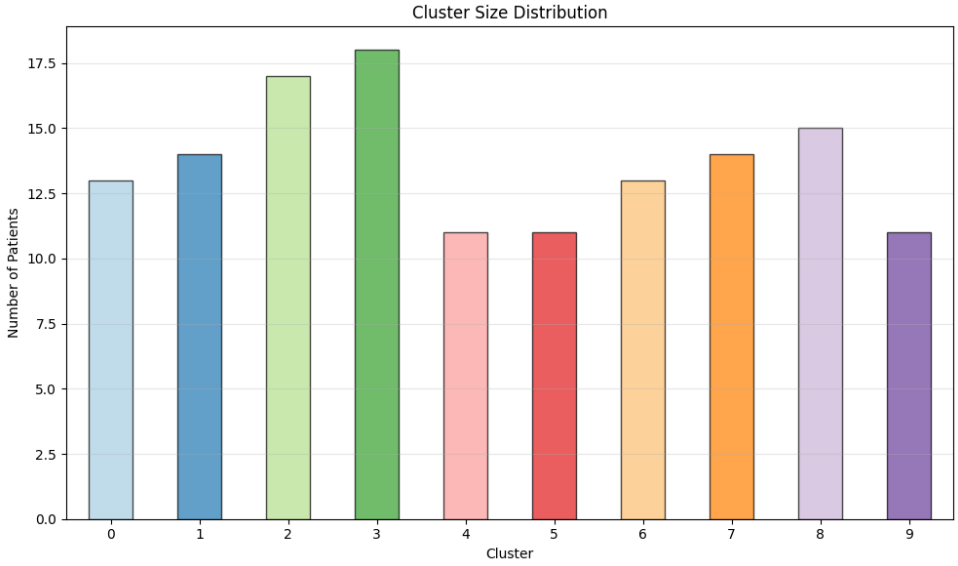

Figure 7: Distribution of clusters

The cluster membership counts showed a relatively even partitioning of patients across groups, with cluster sizes ranging from 11 to 18 individuals (8.0% – 13.1%). This distribution indicates that the algorithm did not collapse patients into a few dominant clusters nor produce excessively small or unstable groups. Instead, the resulting structure reflects a well-balanced segmentation of the cohort, consistent with the composite-score evaluation and supportive of the robustness of the selected clustering solution.

3.3 Clinical features

Following the assessment of cluster stability, the analysis systematically examined whether the ten molecular groups also differed in their clinical and demographic characteristics. To provide an integrated overview, categorical clinical features - including race, country of origin, tumor site, alcohol consumption patterns, and cause of death - were first evaluated.

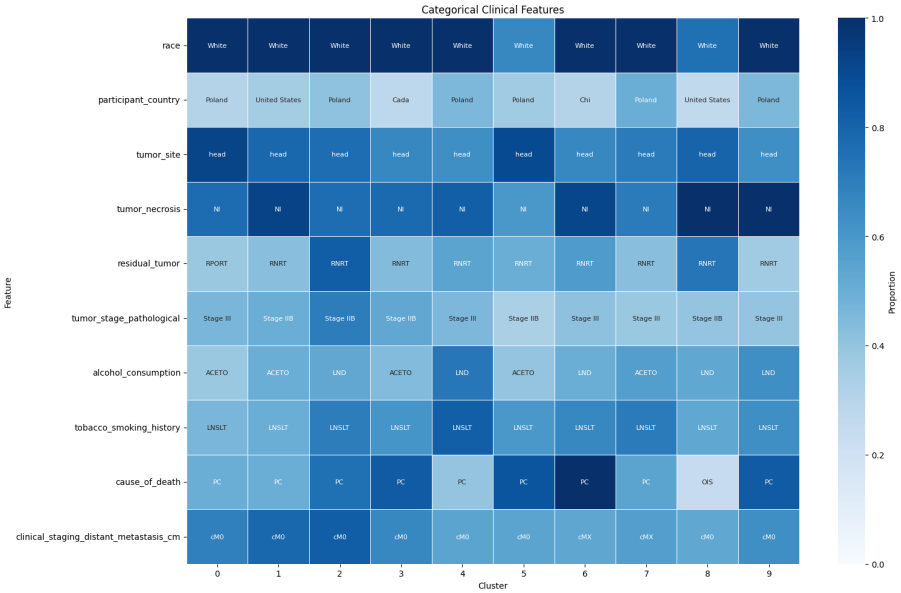

Figure 8: Categorical features across clusters

The resulting heatmap revealed a heterogeneous but largely unstructured distribution of these variables across clusters, with no single group showing a dominant enrichment for most categories. A similar pattern emerged when focusing on binary clinical attributes - including sex, BMI category, and vital status -whose heatmap, which likewise suggested only modest inter-cluster variation.

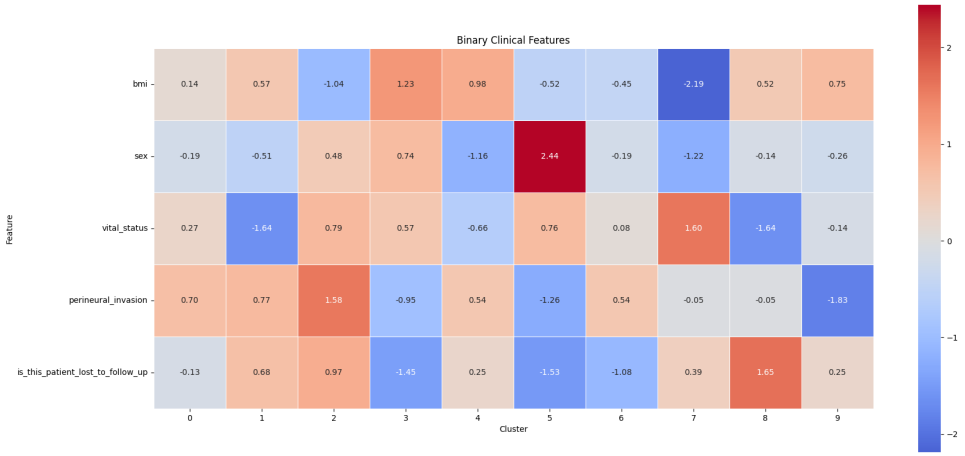

Figure 9: Binary features across clusters

To formally quantify these observations, one-way ANOVA was applied to continuous and ordinal variables. None of the tested features reached statistical significance. They all showed comparable distributions across clusters.

Variable	**F**	**p-value**
BMI	1,05	0,4010
Tumor stage	1,91	0,0562
Nodal involvement	1,07	0,3875
Primary tumor extent	1,69	0,0982
Residual disease	1,40	0,1971
Alcohol consumption	0,73	0,6848
Tobacco exposure	0,57	0,8184

Table 3: one-way ANOVA test

Binary variables were further evaluated using chi-square tests, again confirming the absence of significant associations. In fact, sex distribution (χ^2=4.94, p=0.8398) appeared broadly balanced across clusters, as illustrated below.

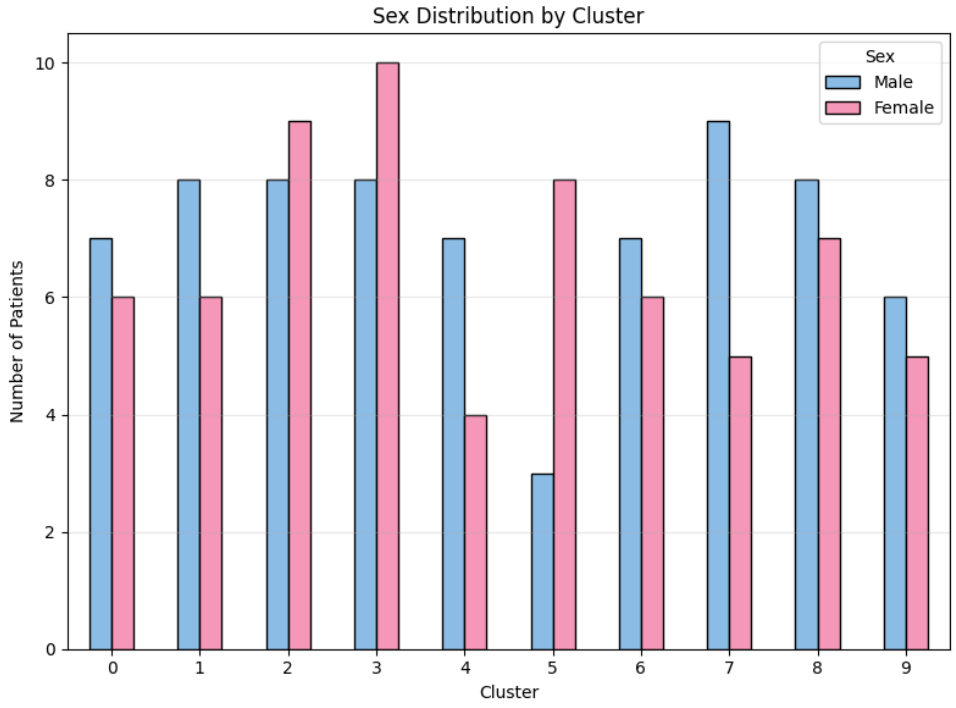

Figure 10: Sex distribution across clusters

Vital status showed a slightly more structured pattern (χ^2=23.19, p=0.1836), with three clusters (2, 3, and 7) exhibiting a higher proportion of deceased patients - slightly above ten individuals each - while clusters 1 and 8 contained an equivalent number of patients

who were predominantly alive. These tendencies are depicted in the graph below, although statistical evidence does not support strong differences between clusters.

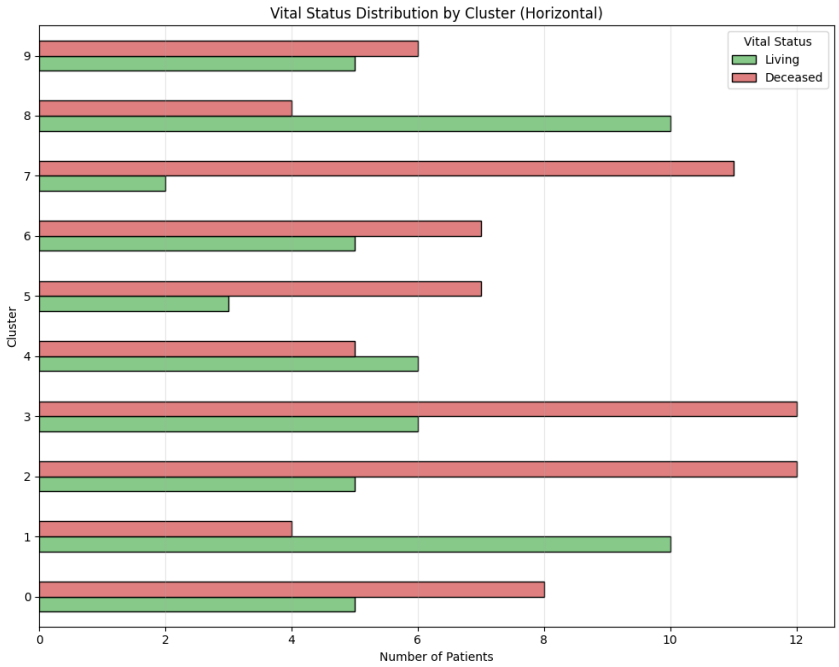

Figure 11: Vital status distribution across clusters

Geographical origin represented one of the few variables with a more recognizable structure. A dedicated heatmap highlighted the predominance of patients from Poland, the United States, and China, while the remaining countries appeared sparsely represented. Although these major groups showed some variation across clusters, their relative proportions differed only moderately, and no cluster exhibited a distinct geographical signature.

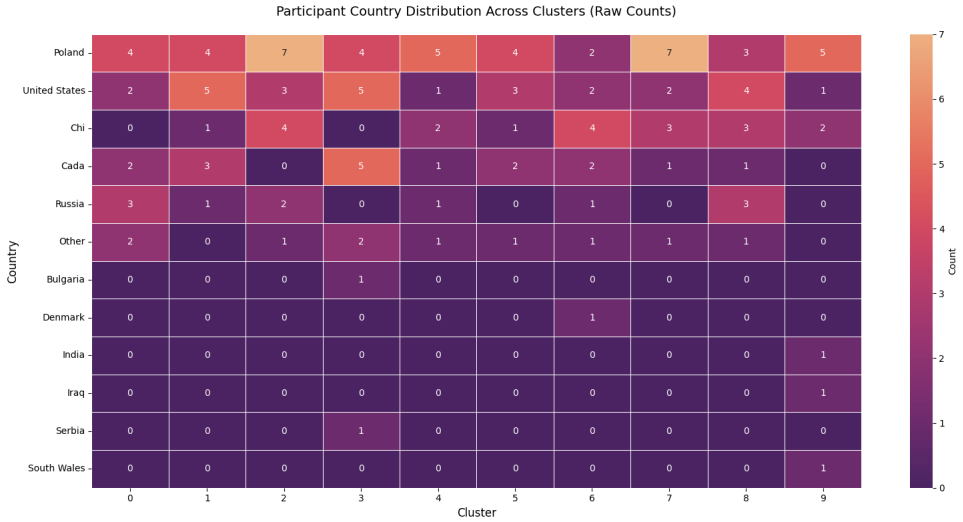

Figure 12: Country distribution across clusters

Lifestyle-related variables, such as smoking and alcohol consumption, were summarized using diverging bar charts. Overall, the cohort was characterized by a majority of non-smokers and non-drinkers; however, the number of smokers exceeded that of regular drinkers, while low-frequency drinkers remained the largest subgroup within alcohol-consumption categories.

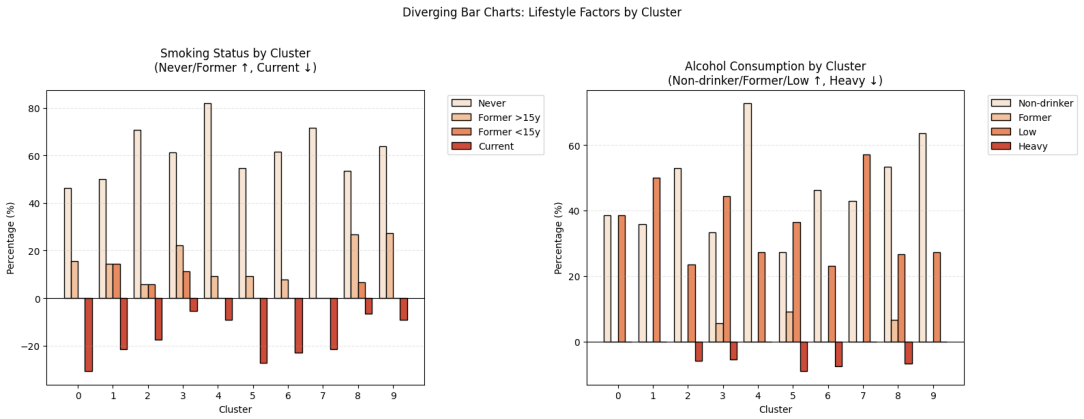

Figure 13: Lifestyle distribution across clusters

To complete the analysis, the BMI distribution across clusters was observed using boxplots. This showed substantial overlap and no evidence of cluster or specific shifts.

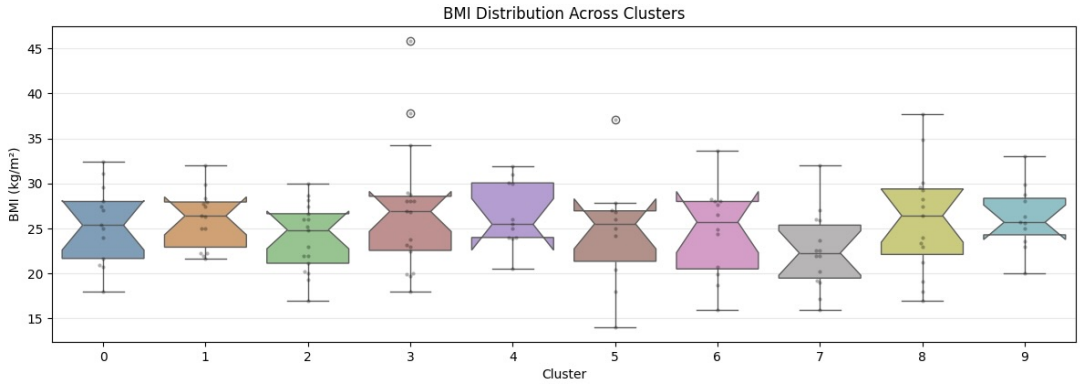

Figure 14: IBM distribution across clusters

Finally, the distribution of tumor stage was visualized using a dedicated heatmap. Stage II was significantly enriched in cluster 2, while stage III showed a relatively homogeneous presence across clusters, typically ranging from four to six patients. Stage I was consistently rare, and stage IV was almost absent in all groups.

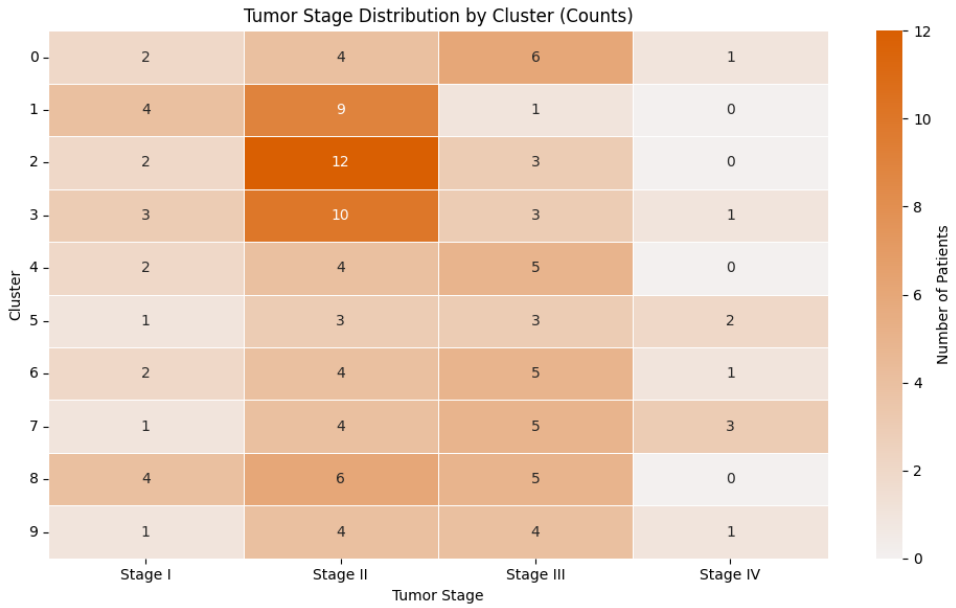

Figure 15: Tumor stage distribution across clusters

These patterns, although visually distinct, did not translate into statistically significant differences in the ANOVA framework. It is worth noting that the distribution of vital status across clusters highlighted clusters with higher mortality (2, 3, 7) and those with predominantly living patients (1, 8).

Taken together, these analyses indicate that the ten clusters, while molecularly distinct, do not exhibit strong or systematic divergence in most clinical and demographic variables. The observed patterns suggest that the clustering structure is primarily driven by molecular features rather than by baseline clinical characteristics.

4 Data Prediction

Predictive modeling refers to a broad class of computational approaches designed to estimate unknown variables or anticipate future outcomes based on patterns learned from existing data. These methods identify relationships, similarities, or latent structures within a dataset and use them to infer information that is not directly observable.

In oncology, such models are particularly valuable because they can support early risk stratification, guide treatment decisions, and help identify patients who may benefit from more intensive monitoring. By learning patterns embedded in clinical, molecular, or network-derived features, prediction algorithms attempt to estimate clinically relevant variables even when these are missing, uncertain, or difficult to assess.

In the context of this study, tumor stage prediction serves two purposes. First, it provides

a supervised learning framework to evaluate whether the fused similarity network and its spectral embedding encode clinically meaningful structure. Second, it allows us to test whether patient similarity patterns can be leveraged to infer disease severity, offering insight into the biological and clinical relevance of the integrated network representation.

4.1 Tumor stage imputation

The tumor stage pathological variable served as the prediction target. Six patients (4.4%) had missing tumor stage values. To enable supervised learning, these were imputed using **k-Nearest Neighbors imputation** (k=5, distance-weighted) based on all preprocessed clinical features. For each patient with missing stage:

1. Clinical features were used to identify 5 most similar patients with known stages
2. Stage was imputed as the rounded weighted average of neighbor stages
3. Imputation decisions were validated by examining neighbor stage distributions

Patient ID	**Neighbor Stages**	**Mean Neighbor**	**Imputed Stage**
C3L-02701	[2, 2, 2, 2, 1]	1.80	2
C3L-03635	[2, 1, 2, 1, 2]	1.60	2
C3N-00709	[2, 2, 2, 1, 2]	1.80	2
C3N-01715	[3, 2, 4, 2, 2]	2.60	3
C3N-01719	[3, 3, 2, 2, 2]	2.40	2
C3N-03039	[2, 2, 2, 2, 3]	2.20	2

Table 4: k-Nearest Neighbors imputation

After imputation, the dataset contained a complete distribution of tumor stages, with the following class frequencies:

- Stage I: 22 patients (16.1%),
- Stage II: 65 patients (47.4%),
- Stage III: 41 patients (29.9%),
- Stage IV: 9 patients (6.6%).

This resulted in a marked class imbalance, with a ratio of 7.22:1 between the most and least represented classes (Stage II vs. Stage IV), underscoring the need for imbalance-aware modeling strategies.

4.2 Train - Test split

After completing tumor stage imputation and establishing a fully annotated dataset, the next step was to design a robust evaluation framework for predictive modeling. To ensure that model performance could be assessed on unseen data while preserving the underlying class distribution, the cohort was partitioned into independent training and test subsets using a stratified splitting strategy. The final subdivision consisted of:

- Training set: 102 patients (74.5%)
- Testing set: 35 patients (25.5%)

Stratification ensured that all four tumor stages (I-IV) were proportionally represented in both subgroups, a critical requirement given the marked class imbalance observed after imputation. This approach ensured that the predictive models were trained on a representative sample of the population and evaluated with the most realistic class proportions possible.

4.3 Predictive models

Two complementary predictive approaches were implemented to assess the feasibility of tumor stage prediction from the fused similarity network and its spectral embedding. The first method operates directly on the patient–patient similarity structure, while the second evaluates whether low-dimensional spectral features extracted from the network encode clinically meaningful information.

4.3.1 k-Nearest Neighbors (k-NN)

k-NN classification was performed directly on the fused similarity network, using precomputed distances defined as $1-$similarity. This approach evaluates whether local neighborhood structure in the integrated network is informative for tumor stage discrimination. The number of neighbors was optimized through 5-fold stratified cross-validation on the training set, exploring values of k from 3 to 19 in steps of two. Balanced accuracy served as the optimization metric to mitigate the impact of class imbalance. This procedure identified $k = 17$ as the optimal value, corresponding to a cross-validated balanced accuracy of 0.287. The final model was trained exclusively on distances among training patients, and predictions for test patients were obtained through majority voting among their 17 nearest neighbors in the fused similarity space.

4.3.2 Machine Learning

To assess whether global network structure encodes clinically relevant information, two supervised learning models were trained on the 20 - dimensional spectral embedding derived from the fused similarity network.

- **Logistic Regression**. Multiclass logistic regression with L2 regularization, balanced class weights, and a maximum of 1000 iterations. This linear model serves as an interpretable baseline.
- **Random Forest**. An ensemble of 200 decision trees with maximum depth = 10 and balanced class weights. This nonlinear model captures higher-order interactions among spectral features.

Prior to training both models, all features were standardized using z-score normalization to ensure comparable scaling across dimensions.

Model performance was evaluated on the held-out test set using multiple complementary metrics. Accuracy quantified the overall proportion of correct predictions, while balanced accuracy provided a class-size-independent measure by averaging recall across tumor stages. The weighted F1-score summarized precision and recall while accounting for class frequencies, and the confusion matrix offered a detailed view of prediction errors across individual stages.

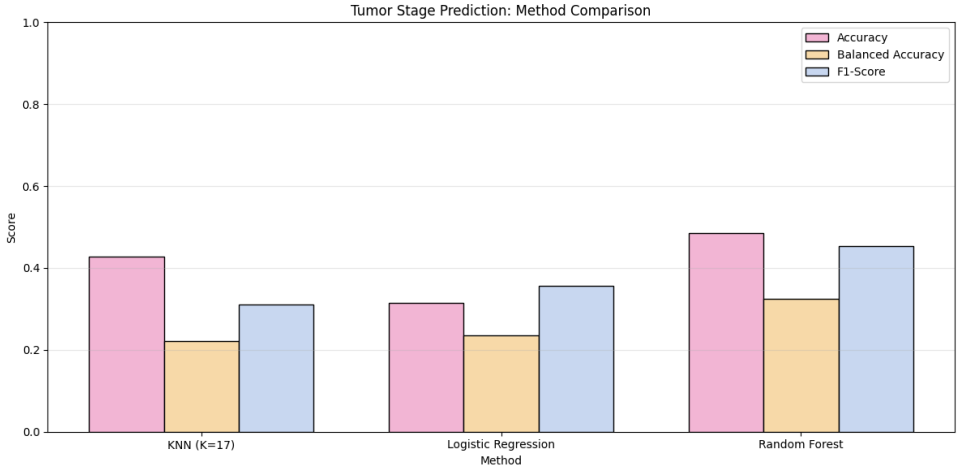

Figure 16: Metrics across predictive models

Based on the evaluation metrics obtained on the held-out test set, the Random Forest classifier trained on the spectral embedding emerged as the most effective predictive model in this study.

4.4 Omics layer contribution

Despite the pronounced class imbalance, Random Forest demonstrated a superior ability to capture nonlinear relationships and extract clinically relevant information from the fused network representation. For this reason, this model was selected as the reference model for downstream interpretability analyses.

To investigate which omics layers contributed most to successful tumor stage predictions, layer-specific similarity patterns among test patients were examined. For each omics layer, the mean similarity among patients whose tumor stage was correctly predicted by the Random Forest was compared with the mean similarity among those misclassified by the model. The difference between these two quantities defined a contribution score, calculated as $mean_{correct} - mean_{incorrect}$.

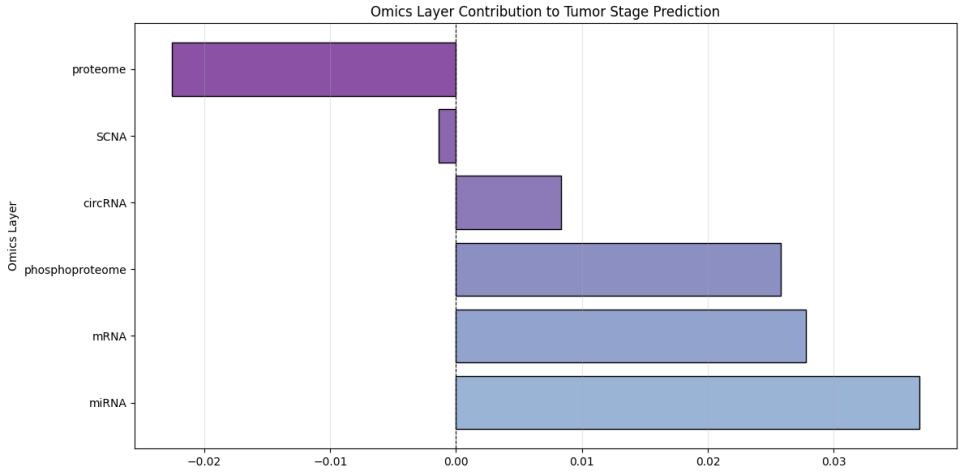

Figure 17: Omics layers ranked by contribution

A positive contribution score indicates that the corresponding omics layer captures similarity structures aligned with tumor stage, meaning that correctly predicted patients tend to be more similar to one another within that layer than patients whose stage was incorrectly assigned. This analysis provides a biologically interpretable perspective on which data modalities most strongly support accurate tumor stage discrimination within the integrated network.

5 Limitations and Future Directions

Several limitations warrant consideration. First, the modest sample size (n=137) limits statistical power for detecting cluster-clinical associations and training robust predictive models, particularly for rare Stage IV cases. Second, the absence of survival outcome

data prevented assessment of prognostic value. Third, validation in independent cohorts is essential to confirm molecular subtype reproducibility. Fourth, functional interpretation of molecular subtypes requires pathway analysis and biological validation.

Future work should focus on: (1) expansion to larger multi-institutional cohorts, (2) integration of treatment response data to identify therapy-relevant subtypes, (3) incorporation of spatial omics and single-cell data to resolve intratumoral heterogeneity, (4) mechanistic investigation of subtype-defining molecular features, and (5) development of clinically applicable subtype classifiers for prospective patient stratification.

References

[1] Charaighn Sesock; *Pancreatic Cancer Diagnoses and Mortality Rates Climb* Jan, 2025. https://pancan.org/press-releases/

[2] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A.(2014); *Similarity network fusion for aggregating data types on a genomic scale.*. Nature Methods, 11(3), 333-337