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.

Multi-omics cancer subtyping via attention-guided variational representation learning and deep clustering

Xinning Liu
Dalian Neusoft University of Information
Li Han
Dalian Neusoft University of Information
Ling Kang
Dalian Neusoft University of Information
Lei Zhao
Dalian Neusoft University of Information
Quan Guo
guoquan@neusoft.edu.cn
Dalian Neusoft University of Information
Huadong Miao
Independent Researcher

Article

Keywords: multi-omics integration, cancer subtyping, variational autoencoder, attention-based fusion, deep embedded clustering, survival stratification

Posted Date: April 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9259481/v1

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

Additional Declarations: No competing interests reported.

Title Page

Title

Multi-omics cancer subtyping via attention-guided variational representation learning and deep clustering

Authors

Xinning Liu[1,*],Li Han[1],Ling Kang[2],Lei Zhao[2],Quan Guo[2,*],Huadong Miao[3]

Affiliations

1.School of Software, Dalian Neusoft University of Information,Dalian, Liaoning, 116023, PR China

2.Neusoft Research Institute, Dalian Neusoft University of Information, Dalian, Liaoning,116023, PR China

3.Independent Researcher, Dalian, Liaoning,116023, PR China

Corresponding Author

Quan Guo

Email: guoquan@neusoft.edu.cn

Xinning Liu

Email: liuxinning@neusoft.edu.cn

Multi-omics cancer subtyping via attention-guided variational representation learning and deep clustering

Xinning Liu[1,*],Li Han[1],Ling Kang[2],Lei Zhao[2],Quan Guo[2,*],Huadong Miao[3]

[1]School of Software, Dalian Neusoft University of Information,Dalian, Liaoning, 116023, PR China

[2]Neusoft Research Institute, Dalian Neusoft University of Information, Dalian, Liaoning,116023, PR China

[3]Independent Researcher, Dalian, Liaoning,116023, PR China

[*]Corresponding authors:Quan Guo. Email: guoquan@neusoft.edu.cn

[*]Corresponding authors:Xinning Liu. Email: liuxinning@neusoft.edu.cn

Abstract

Tumor heterogeneity is a major challenge for accurate molecular subtyping and precision oncology, as complex variations across and within tumors can lead to distinct biological behaviors, treatment responses, and clinical outcomes. To address this issue, we propose MAVDC, a multi-omics integration framework that combines unsupervised feature selection, variational representation learning, attention-based fusion, contrastive learning, and adaptive deep embedded clustering. MAVDC first ranks features within each omics modality to suppress noise and retain informative signals, and then determines the optimal number of clusters using a composite scoring strategy. Compact variational autoencoders are employed to learn modality-specific latent representations, which are further integrated through a multi-head attention mechanism. A progressive training strategy jointly optimizes reconstruction, variational regularization, clustering refinement, and cross-omics consistency. Experiments on TCGA BRCA and LUSC cohorts integrating RNA expression, DNA methylation, and copy number variation data identified four molecular subtypes in each cancer. These subtypes exhibited clear geometric separation, distinct functional and immune characteristics, and significant survival differences, outperforming representative single-omics and multi-omics baseline methods. MAVDC provides an effective tool for discovering clinically meaningful cancer subtypes.

Keywords: multi-omics integration; cancer subtyping; variational autoencoder; attention-based fusion; deep embedded clustering; survival stratification

Introduction

Tumor heterogeneity remains one of the fundamental challenges in cancer research and clinical management. Substantial molecular and phenotypic differences can arise not only across patients with the same cancer type (inter-tumoral heterogeneity) but also among distinct cell populations within a single tumor (intra-tumoral heterogeneity) [1]. Such complexity undermines accurate tumor classification, complicates prognosis assessment, and contributes to variable therapeutic responses and drug resistance, particularly in targeted and precision therapies [2].

Conventional cancer classification systems are largely based on histopathological characteristics. Although clinically informative, tumors with similar morphological features often exhibit markedly different molecular mechanisms and clinical outcomes, highlighting the limitations of morphology-driven stratification [3]. Advances in high-throughput sequencing technologies have enabled systematic profiling of tumors across multiple molecular layers, including genomics, transcriptomics, epigenomics and proteomics, providing unprecedented opportunities to characterize tumor heterogeneity at a systems level [4]. However, single-omics analyses capture only partial aspects of tumor biology and are insufficient to resolve the complex regulatory interactions underlying cancer progression.

Early computational frameworks for multi-omics cancer subtyping were primarily based on statistical modeling and similarity-based integration. Representative methods include iCluster, which jointly models multiple omics data through latent variable modeling [5], and Similarity Network Fusion (SNF), which integrates patient similarity networks derived from individual omics layers and has shown strong performance in breast cancer and glioblastoma subtyping [6]. Multi-Omics Factor Analysis (MOFA) further extended this paradigm by identifying shared latent factors

across omics layers and has been widely applied in pan-cancer studies and biomarker discovery [7]. Perturbation-based approaches such as PINSPlus emphasized clustering stability and robustness by incorporating resampling strategies and automatic cluster number estimation [8]. Despite their success, these methods often struggle with high-dimensional data, nonlinear cross-omics interactions and missing values, limiting their ability to uncover complex latent biological structures.

Recent advances in deep learning have substantially reshaped the landscape of multi-omics integration. Deep neural networks can automatically learn abstract representations and capture nonlinear regulatory relationships across omics layers, making them well suited for modeling complex cancer systems. Chaudhary et al. [9] first demonstrated the potential of deep learning for multi-omics survival prediction in hepatocellular carcinoma, achieving robust performance across multiple independent cohorts. Subsequently, Chai et al. [10] proposed an autoencoder-based framework for pan-cancer survival prediction, integrating mRNA, miRNA and DNA methylation data and outperforming conventional approaches. Disease-specific deep learning models have also been developed for prostate cancer relapse prediction [11], multi-task learning of prognosis and drug response [12], and ensemble-based survival modeling across multiple cancer types [13]. More recently, AutoSurv introduced an interpretable deep learning framework that jointly integrates clinical and multi-omics data [14], while multimodal architectures incorporating histopathology images and omics profiles further improved pan-cancer survival prediction [15].

Although deep learning-based multi-omics models have demonstrated strong predictive performance, their effectiveness often depends critically on upstream feature selection and data preprocessing. Many existing studies rely on simple variance-based or univariate filtering strategies, which may inadequately distinguish informative biological signals from noise and fail to account for heterogeneity across omics layers. To address these limitations, attention mechanisms have been increasingly adopted to enable adaptive feature weighting and long-range dependency modeling. Transformer-based and self-attention architectures have been shown to improve subtype classification and feature robustness by capturing global and local omics dependencies [16,17]. Integrating biological prior knowledge with attention mechanisms, such as pathway-guided self-attention, has further enhanced interpretability and predictive accuracy [18]. Other studies have combined attention with graph-based learning to model heterogeneous interactions across omics layers, achieving improved performance in cancer subtype identification and driver gene prediction [19,20].

In parallel, variational autoencoders (VAEs) have gained increasing attention as powerful generative models for robust low-dimensional representation learning in multi-omics integration. VAE-based frameworks enable probabilistic modeling of noisy, high-dimensional omics data and facilitate the discovery of latent structures shared across modalities. Applications of VAEs in ovarian cancer [21], pan-cancer subtyping [22], and adaptive multi-omics fusion [23] have demonstrated improved clustering stability and survival prediction. Extensions such as modality-aware VAEs and factorized VAEs further preserve modality-specific characteristics while enhancing biological consistency and interpretability [24 – 26]. Beyond VAEs, generative adversarial networks and graph attention networks have also been explored to capture complex inter-omics relationships and improve subtype robustness [27 – 29].

Despite these advances, several key challenges remain in multi-omics integration for cancer subtyping. First, feature selection strategies often overlook the intrinsic heterogeneity and noise characteristics of different omics layers, potentially introducing bias or instability. Second, many fusion strategies apply uniform or global weighting schemes, which may fail to capture sample-specific and cross-modal interactions. Third, clustering performance is frequently evaluated using technical metrics alone, without sufficient integration of clinical outcomes such as survival. In order to overcome the challenges of heterogeneous noise, nonlinear cross-modal relationships, and the disconnection between clustering structure and clinical outcomes in multi-omics data, we developed Multi-omics Attention-based Variational Deep Clustering (MAVDC), a unified framework that integrates feature selection, representation learning, and clustering into a single optimization process. To retain informative signals, we designed a rank-based feature selection strategy that combines dispersion and information measures across modalities. To ensure that latent representations and subtype structures reinforce each other, we jointly optimized variational encoding, contrastive alignment, and deep clustering. Furthermore, we introduced a structural consistency constraint to align clustering geometry across omics layers. To evaluate performance, we applied MAVDC to

BRCA and LUSC datasets, where it consistently identified stable and prognostically relevant subtypes, outperforming existing integration methods.

Method

MAVDC is a unified multi-omics framework for cancer subtyping that integrates feature selection, variational representation learning, attention-based fusion, and deep clustering. The overall architecture and workflow are illustrated in Figure 1.

Phase 1: Enhanced Full-based Pretraining

To reduce dimensionality and suppress noise in high-dimensional multi-omics data, an unsupervised variability-based feature selection strategy is applied independently to each modality. Features are ranked using multiple dispersion and information criteria, including variance, median absolute deviation (MAD), entropy, and interquartile range (IQR) [30], and aggregated through a rank-based scheme. All metrics are normalized within each modality to ensure comparability, with entropy estimated using a fixed discretization scheme. For downstream modeling, the top 5,000 features per modality are retained, and sensitivity analysis confirms stable clustering performance across a broad range of feature counts. Each omics modality (DNA methylation, RNA expression, and copy number variation) is then encoded separately using a compact variational encoder to control model capacity and reduce overfitting. The resulting modality-specific latent representations are integrated via a lightweight attention-based fusion module that adaptively weights different modalities on a per-sample basis to capture cross-modal interactions while preserving modality-specific information. For clarity, the attention-based fusion procedure is summarized in Algorithm 1.

Algorithm 1. Muti-Modal Attention Fusion

Require: Latent representations from multiple modalities: $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ where $\mathbf{z}_i \in \mathbb{R}^{B\times d}$ for modalities $i \in \{DNA\ methylation, RNA-seq, CNV\}$
Require: Number of attention heads: H, Latent dimension: d
Ensure: Fused multi-modal representation: $\mathbf{Z}_{fused} \in \mathbb{R}^{B\times d}$
Stack modalities: $\mathbf{Z}_{stacked} \leftarrow \text{Stack}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$ {Shape: $[B, 3, d]$}
for $h = 1$ to H **do**
  Project to query, key, value: $\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h \leftarrow \text{Linear}_h(\mathbf{Z}_{stacked})$
end for
for $h = 1$ to H **do**
  Compute attention scores: $\mathbf{S}_h \leftarrow \frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d/H}}$
  Apply softmax: $\mathbf{A}_h \leftarrow \text{Softmax}(\mathbf{S}_h)$ {Attention weights: $[B, H, 3, 3]$}
  Weighted aggregation: $\mathbf{C}_h \leftarrow \mathbf{A}_h \mathbf{V}_h$ {Context vectors: $[B, H, 3, d/H]$}
end for
Concatenate heads: $\mathbf{C} \leftarrow \text{Concat}(\mathbf{C}_1, \ldots, \mathbf{C}_H)$ {Shape: $[B, 3, d]$}
Residual connection and normalization: $\mathbf{Z}_{attended} \leftarrow \text{LayerNorm}(\mathbf{C} + \mathbf{Z}_{stacked})$
Global average pooling: $\mathbf{Z}_{fused} \leftarrow \frac{1}{3}\sum_{i=1}^{3} \mathbf{Z}_{attended}[:, i, :]$ {Average over modalities}
Apply dropout: $\mathbf{Z}_{fused} \leftarrow \text{Dropout}(\mathbf{Z}_{fused}, p = 0.1)$
return $\mathbf{Z}_{fused}$

The fused representation is optimized within a unified variational learning framework using a multi-objective loss that jointly enforces reconstruction fidelity, latent regularization, clustering refinement, cross-modal alignment, and representation consistency:

$$L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{KL} + \lambda_3 L_{cluster} + \lambda_4 L_{contrastive} + \lambda_5 L_{consistency} \quad (1)$$

Here, the contrastive term promotes cross-modal alignment [31] and is implemented using an InfoNCE-based objective [32], which aligns representations from different modalities of the same sample while separating those from different samples.

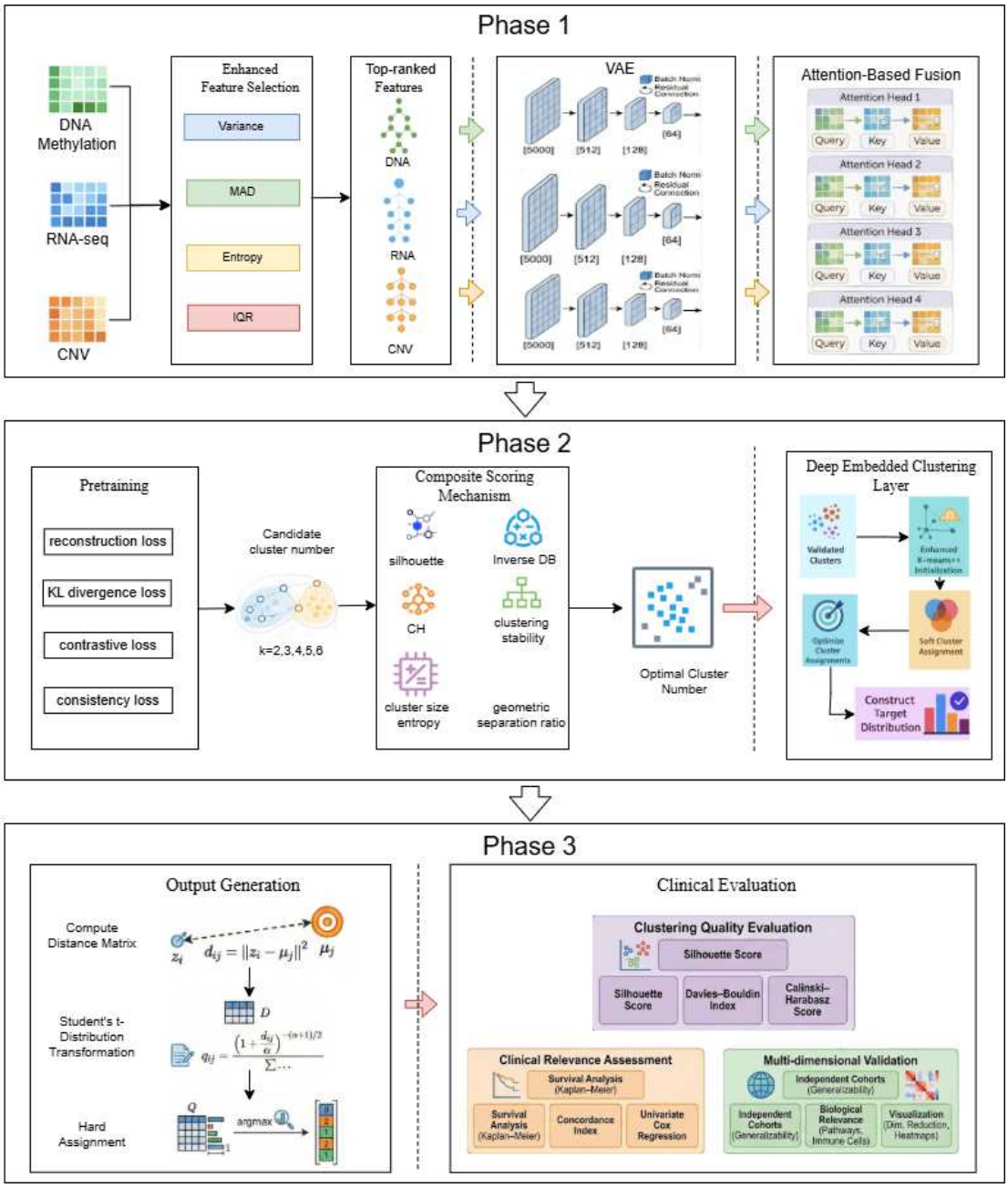


Figure 1. The flowchart of MAVDC

Phase 2: Progressive Training Strategy and Quality Control

1. Pretraining Strategy

Initially, the model is pre-trained without clustering loss to learn robust representations through reconstruction, KL divergence, contrastive, and consistency objectives. This progressive training strategy separates representation learning from clustering optimization to prevent cluster collapse.

2. Clustering Number Determination Based on a Composite Scoring Mechanism

To identify the optimal number of clusters, we employ an adaptive cluster discovery framework that integrates six complementary metrics reflecting geometric structure, stability, size distribution, and geometric separation. These include the Silhouette coefficient, Calinski–Harabasz index, inverse Davies-Bouldin index, clustering stability , cluster size entropy and geometric separation ratio [33-38]. For each candidate $k \in \{2, 3, 4, 5\}$, K-means clustering is repeatedly applied, this range was selected based on prior pan-cancer and disease-specific studies reporting that most solid tumors can be stratified into 2-5 major molecular subtypes. and a composite score is computed as:

$$S(k) = w_1 \cdot \mathrm{sil}(k) + w_2 \cdot \mathrm{ch}(k) + w_3 \cdot \mathrm{db}(k) + w_4 \cdot \mathrm{stability}(k) + w_5 \cdot \mathrm{entropy}(k) + w_6 \cdot \mathrm{geom_sep(k)} \quad (2)$$

where each metric is normalized, and the weights balance complementary geometric and stability-related criteria. In all experiments, the weights w_i were set to equal values to avoid introducing subjective bias. The optimal cluster number k^* is selected by maximizing S_k, ensuring robust and biologically meaningful stratification.

3. Construction and Optimization of the Deep Embedded Clustering Layer

Based on the validated cluster configuration, cluster assignments are refined using a deep embedded clustering (DEC) strategy in the latent representation space, as formalized in Algorithm 2. This stage aims to improve intra-cluster compactness and inter-cluster separability by jointly optimizing cluster centroids and soft assignments.

Once the cluster configuration is determined, we refine cluster assignments using Deep Embedded Clustering (DEC) in the learned latent space. Initial centroids are derived from K-means on the fused representations z_{fused}. Soft assignments are calculated using a Student's t-distribution kernel:

$$q_{ij} = \frac{\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}}{\sum_{j'=1}^{k}\left(1+\left\|z_i-\mu_{j'}\right\|^2\right)^{-1}} \quad (3)$$

where z_i denotes the latent embedding of sample i, μ_j is the centroid of cluster j, and q_{ij} represents the probability that sample i belongs to cluster j. To sharpen cluster assignments and emphasize confident memberships, a target distribution p_{ij} is constructed by squaring and normalizing the soft assignments:

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{n} q_{ij}}{\sum_{j'=1}^{k}\left(q_{ij'}^2 / \sum_{i=1}^{n} q_{ij'}\right)} \quad (4)$$

The clustering objective minimizes the Kullback–Leibler divergence between the target distribution p and the current soft assignments q[39]:

$$L_{cluster} = \mathrm{KL(P||}Q) = \sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (5)$$

This loss is jointly optimized with other objectives during the main training phase. Cluster centroids μ_j are treated as learnable parameters, allowing dynamic adjustment of cluster boundaries, thereby improving both compactness and separability[40].

4. Cross-Omics Consistency Optimization

To promote coherent clustering structure across omics modalities, we introduce a cross-omics consistency loss comprising two components: a modality-wise clustering quality term and a cross-modality consistency term.

Because the standard Silhouette score is non-differentiable[33], we employ a differentiable surrogate Silhouette computed from soft cluster assignments and continuous distance measures, enabling end-to-end optimization. Let Z_m denote the latent embedding of modality m, and let q denote the shared soft cluster assignment inferred from the fused representation. The modality-wise quality loss is defined as:

$$L_{\mathrm{quality}} = -\frac{1}{M}\sum_{m=1}^{M} \widetilde{\mathrm{Sil}}(z_m, q) \quad (6)$$

where $s_m = \widetilde{\mathrm{Sil}}(\cdot)$ denotes the differentiable Silhouette surrogate.

To encourage consistency in clustering quality across modalities, we minimize the variance of the surrogate Silhouette scores:

$$L_{cross} = \frac{1}{M-1}\sum_{m=1}^{M}(s_m - \bar{s})^2 \ , \ s_m = \widetilde{\mathrm{Sil}}(z_m, q) \ , \ \bar{s} = \frac{1}{M}\sum_{m=1}^{M} s_m \quad (7)$$

The overall cross-omics consistency loss is:

$$L_{omics} = \alpha_1 \cdot L_{quality} + \alpha_2 \cdot L_{cross} \quad (8)$$

This term is incorporated into the total objective with a scheduled weight $\lambda_6(t)$ that increases after the initial representation learning phase.

To balance representation learning and modality consistency, an adaptive scheduling strategy is applied to λ_6, gradually increasing its contribution over training epochs:

$$\lambda_6(t) = \lambda_{init} + (\lambda_{final} - \lambda_{init}) \cdot \min\left(1, \frac{t}{T_{warmup}}\right) \quad (9)$$

Algorithm 2. Deep Embedded Clustering with Adaptive Cluster Discovery

Require: Fused representations $Z \in \mathbb{R}^{n \times d}$, candidate cluster numbers $K_{\text{candidates}} = \{2, 3, 4, 5\}$
Ensure: Optimal cluster number k^*, cluster centroids $\{\mu_j\}_{j=1}^{k^*}$, soft assignments Q
Parameters: Number of stability runs R, metric weights $\{w_1, \ldots, w_6\}$

function AdaptiveClusterDiscovery(Z, $K_{\text{candidates}}$)
  ▷ Evaluate each candidate cluster number
  for $k \in K_{\text{candidates}}$ **do**
    ▷ Run K-means multiple times for stability assessment
    for $r = 1$ **to** R **do**
      $\{\mu_j^{(r)}\}, \{c_i^{(r)}\} \leftarrow \text{KMeans}(Z, k, \text{seed} = r)$
    end for
    $c \leftarrow c^{(1)}$ ▷ Use first run as reference labels
    ▷ Compute clustering quality metrics
    $s(k) \leftarrow \text{SilhouetteScore}(Z, c)$
    $d(k) \leftarrow \text{DaviesBouldinIndex}(Z, c)$
    $c(k) \leftarrow \text{CalinskiHarabaszScore}(Z, c)$
    ▷ Compute stability (average ARI across runs)
    $r(k) \leftarrow \frac{1}{\binom{R}{2}} \sum_{i<j} \text{ARI}(c^{(i)}, c^{(j)})$
    ▷ Compute size distribution entropy
    $e(k) \leftarrow -\sum_{j=1}^{k} p_j \log p_j / \log k$ ▷ $p_j = |C_j|/n$
    ▷ Compute geometric separation ratio
    $v(k) \leftarrow \frac{\bar{d}_{\text{inter}}}{\bar{d}_{\text{intra}}}$ ▷ Inter-cluster / intra-cluster distance
    ▷ Normalize metrics to [0, 1]
    $s_n, c_n, d_n, v_n \leftarrow \text{Normalize}(s, c, 1/d, v)$
    ▷ Compute composite score
    $S(k) \leftarrow w_1 s_n + w_2 c_n + w_3 d_n + w_4 r + w_5 e + w_6 v_n$
  end for
  ▷ Select optimal cluster number
  $k^* \leftarrow \arg\max_{k \in K_{\text{candidates}}} S(k)$
  return k^*
end function

function DeepEmbeddedClustering(Z, k^*, $\{\mu_j\}_{j=1}^{k^*}$)
  ▷ Compute soft cluster assignments using Student's t-distribution
  for $i = 1$ **to** n **do**
    for $j = 1$ **to** k^* **do**
      $q_{ij} \leftarrow \frac{(1+||z_i - \mu_j||^2)^{-1}}{\sum_{j'=1}^{k^*}(1+||z_i - \mu_{j'}||^2)^{-1}}$
    end for
  end for
  ▷ Compute target distribution
  for $i = 1$ **to** n **do**
    for $j = 1$ **to** k^* **do**
      $f_j \leftarrow \sum_{i'=1}^{n} q_{i'j}$ ▷ Cluster frequency
      $p_{ij} \leftarrow \frac{q_{ij}^2 / f_j}{\sum_{j'=1}^{k^*}(q_{ij'}^2 / f_{j'})}$
    end for
  end for
  ▷ Compute clustering loss
  $L_{\text{cluster}} \leftarrow \sum_{i=1}^{n} \sum_{j=1}^{k^*} p_{ij} \log \frac{p_{ij}}{q_{ij}}$ ▷ KL(P——Q)
  return Q, P, L_{cluster}
end function

This warm-up mechanism allows the model to prioritize stable embedding learning before focusing on cross-omics constraints[41].

The overall loss function jointly integrates objectives that promote clustering quality, enhance cross-modal alignment, and enforce cross-omics consistency:

$$L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{KL} + \lambda_3 L_{cluster} + \lambda_4 L_{contrastive} + \lambda_5 L_{consistency} + \lambda_6 L_{omics} \quad (10)$$

where L_{recon} is the reconstruction loss ensuring faithful data reconstruction, L_{KL} is the KL divergence regularizing latent distributions toward a standard normal prior, $L_{cluster}$ is the deep embedded clustering loss optimizing cluster assignments, $L_{contrastive}$ is the contrastive learning loss enhancing cross-modal alignment, $L_{consistency}$ aligns modality-specific embeddings with the fused representation, and L_{omics} ensures cross-omics consistency across all modalities.

To ensure stable optimization and generalization, the final training stage also incorporates multiple convergence control mechanisms: (1) Learning Rate Scheduling (2) Early Stopping (3) Gradient Clipping.

Phase 3: Output Generation and Clinical Evaluation

1. Output Generation and Clinical Evaluation

The final model outputs clinically interpretable results, including subtype assignments and fused multi-omics representations. Subtypes are derived by selecting the cluster with the highest soft assignment probability:

$$c_i = arg\max_j q_{ij} \tag{11}$$

The fused embeddings z_{fused} provide an integrated latent space for downstream analyses such as survival prediction and biomarker identification.

Clustering quality is evaluated using the Silhouette score, Davies–Bouldin index, and Calinski–Harabasz score, offering complementary assessments of structural validity. Clinical relevance is assessed through survival analysis, including Kaplan–Meier curves, concordance index, and Cox proportional hazards regression[42-44]. In this study, Cox regression was primarily conducted in a univariate manner, with subtype indicators as the covariates, to evaluate the association between molecular subtypes and patient survival outcomes.

2. Quality Assurance and Validation

A multi-dimensional validation framework is employed to ensure the robustness and clinical relevance of the clustering results. The model is assessed across independent cancer cohorts to verify generalizability, and biological relevance is established through differential analysis of pathway activity and immune cell infiltration, confirming distinct molecular characteristics among subtypes. Visualization techniques, including dimensionality reduction and heatmaps, further support clustering integrity by revealing clear subtype separation and coherent omics patterns. Final outputs are retained only when all structural, prognostic, and biological quality standards are satisfied.

Datasets and data preprocessing

Multi-omics data were retrieved from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) portal. Two independent cohorts were analyzed: breast invasive carcinoma (BRCA) and lung squamous cell carcinoma (LUSC)[45]. Raw TCGA cohorts were initially downloaded and subsequently filtered by retaining only samples with complete DNA methylation, RNA-seq expression, copy number variation (CNV), and survival information. After removing samples with missing values and aligning the three omics modalities, the final datasets used for integrative clustering analysis consisted of 1,036 BRCA patients and 467 LUSC patients.

For each patient, three molecular layers were integrated: DNA methylation (beta values), RNA-seq expression (TPM), and copy number variation (CNV). CNV profiles were derived from ASCAT segmentation and encoded as gene-associated copy number features for integrative modeling[46]. Only samples with complete measurements across all three omics layers were included for integrative analyses.

The BRCA cohort exhibited substantial histological diversity. Infiltrating duct carcinoma was the predominant subtype (n = 748, 71.0%), followed by lobular carcinoma (n = 194, 18.4%), with the remaining cases distributed among multiple rare subtypes, reflecting the well-recognized heterogeneity of breast cancer. In contrast, the LUSC cohort was highly homogeneous: squamous cell carcinoma (all variants) accounted for 96.2% of cases (n = 455), with only minor representation of basaloid and papillary variants. The complementary design of a heterogeneous (BRCA) and a homogeneous (LUSC) cohort enables robust evaluation of the proposed multi-omics clustering framework across diverse biological contexts and facilitates assessment of its ability to capture molecular heterogeneity beyond conventional histopathological classification.

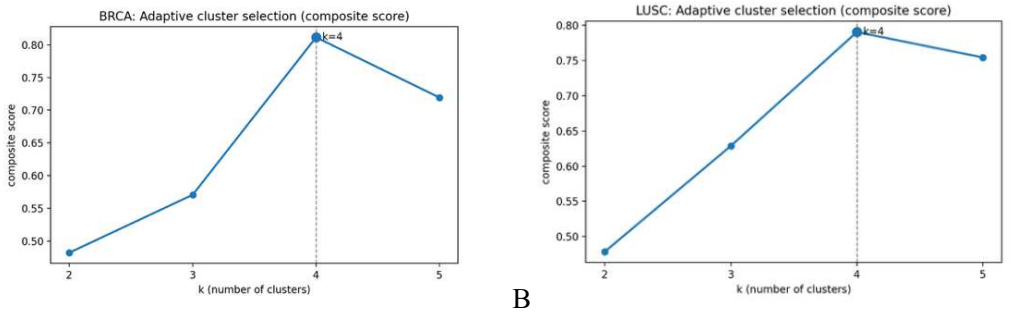
Results

Identification of molecular subtypes

To evaluate whether integrative multi-omics profiling can identify coherent molecular subtypes across distinct cancer types, we performed unsupervised clustering on integrated RNA-seq expression, DNA methylation, and CNV profiles from TCGA-BRCA (n = 1,036) and TCGA-LUSC (n = 467).

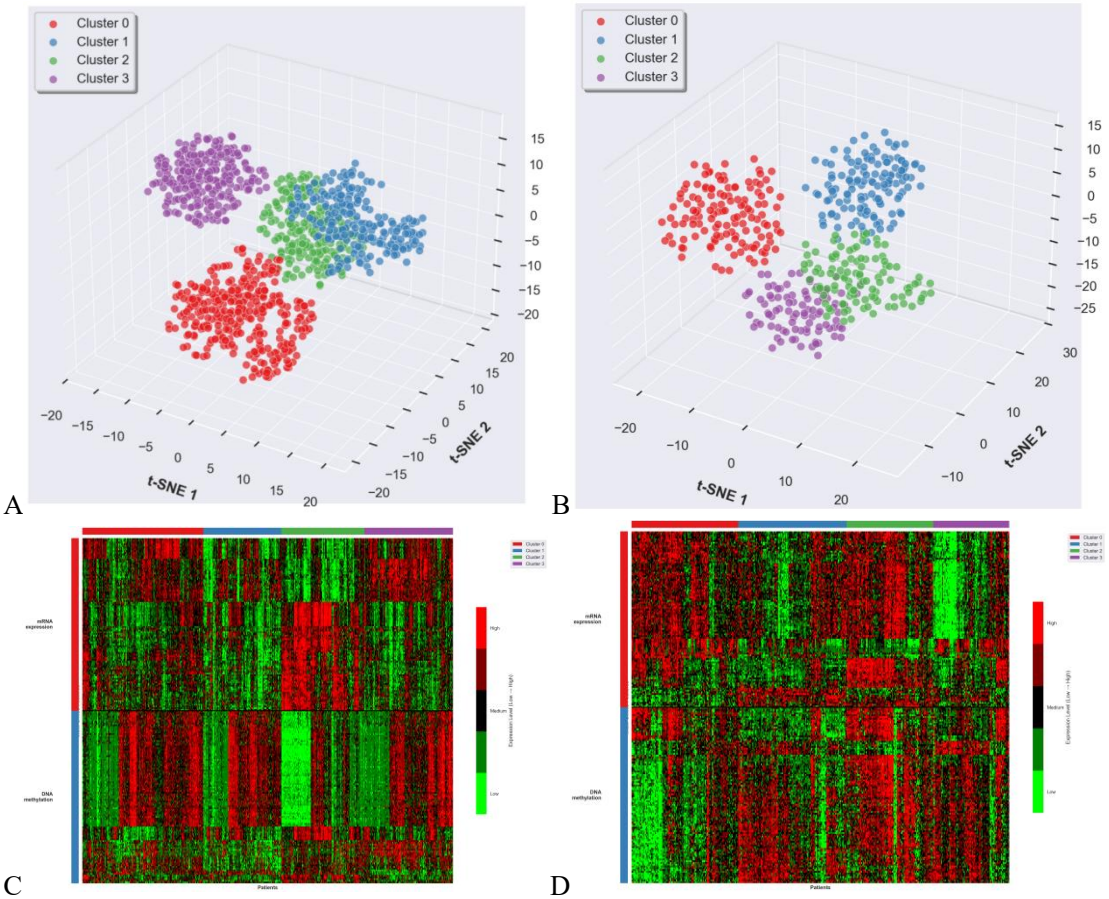
The optimal number of clusters was determined using an unsupervised, stability-guided cluster discovery strategy. Candidate cluster numbers (k = 2, 3, 4, 5) were systematically evaluated using a composite scoring function that integrated complementary geometric and stability-based criteria.

This composite evaluation consistently identified k = 4 as the optimal solution for both cohorts. (Figure 2A-B).



A B

Figure 2. Adaptive selection of the optimal number of clusters. (A) Optimal number of clusters for BRCA. (B) Optimal number of clusters for LUSC



A B

C D

Figure 3. Multi-omics subtyping of BRCA and LUSC. (A) 3D t-SNE of BRCA colored by subtype. (B) 3D t-SNE of LUSC colored by subtype. (C) Integrated multi-omics heatmap of BRCA (top 200 features). (D) Integrated multi-omics heatmap of LUSC (top 200 features).

In both BRCA and LUSC, the resulting four molecular subtypes exhibited relatively balanced sample distributions, indicating that the identified partition was not driven by degenerate or highly imbalanced clustering (Table 1). Quantitative evaluation further demonstrated favorable geometric clustering properties for the four-subtype solution in both cancer types, with clear separation between clusters and compact intra-cluster structure as reflected by standard internal validity indices (Table 1).

Low-dimensional visualization of the fused latent representations provided intuitive support for the inferred subtype structure. In three-dimensional t-SNE space, samples from different subtypes formed four spatially distinct groups in both BRCA and LUSC cohorts, with limited overlap (Figure 3A–B). In addition, integrative heatmap analysis based on the top 200 most discriminative transcriptomic and DNA methylation features revealed coherent block-wise molecular patterns, in which samples belonging to the same subtype shared concordant multi-omics signatures (Figure 3C–D). CNV features were not included in the heatmap visualization because,

although mapped to gene-level values for modeling, their genomic alteration patterns and value distributions differ fundamentally from those of gene expression and DNA methylation, making direct visual comparison potentially misleading.

Collectively, evidence from internal geometric validity metrics (Table 1), low-dimensional embeddings of the fused latent space (Figure 3A–B), and integrative molecular patterning (Figure 3C–D) consistently supports the existence of four structured and coherent multi-omics molecular subtypes in both BRCA and LUSC.

Table 1. Clustering summary and internal validity of MAVDC-derived subtypes

Dataset	Subtype	N	Percentage (%)	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
BRCA	Cluster 0	286	27.6	—	—	—
BRCA	Cluster 1	246	23.7	—	—	—
BRCA	Cluster 2	282	27.2	—	—	—
BRCA	Cluster 3	222	21.4	—	—	—
BRCA	**Overall**	**1036**	**100.0**	**0.3182**	**1.2364**	**336.06**
LUSC	Cluster 0	132	28.3	—	—	—
LUSC	Cluster 1	134	28.7	—	—	—
LUSC	Cluster 2	107	22.9	—	—	—
LUSC	Cluster 3	94	20.1	—	—	—
LUSC	**Overall**	**467**	**100.0**	**0.3410**	**1.1805**	**164.05**

Survival association of multi-omics subtypes

To investigate whether the identified multi-omics subtypes are associated with clinically meaningful differences in patient outcomes, overall survival was evaluated across subtypes in both TCGA-BRCA and TCGA-LUSC cohorts using Kaplan–Meier analysis.

In the BRCA cohort, Kaplan–Meier analysis demonstrated significant survival differences among the four subtypes (Figure 4A). Using Cluster 3 as the reference group, Cox regression identified subtype-specific risk heterogeneity (Table 2), with Cluster 0 showing the strongest protective association relative to Cluster 3 (HR = 0.3999, $p < 0.001$). The resulting model achieved a concordance index (C-index) of 0.5830, indicating moderate prognostic discrimination.

In the LUSC cohort, despite pronounced histopathological homogeneity—where over 96% of samples belong to a single histological subtype—significant survival stratification was still observed among the four molecular subtypes (Figure 4B). Cox regression similarly supported subtype-associated hazard differences (Table 2), with a C-index of 0.5796, highlighting that survival-associated molecular variation can be captured beyond conventional histopathological classification.

Collectively, consistent Kaplan–Meier separation (Figure 4A–B) and subtype-associated hazard patterns (Table 2) across both cohorts support the relevance of the identified multi-omics subtypes to patient survival outcomes, even under the stringent conditions imposed by histologically homogeneous cancer populations.

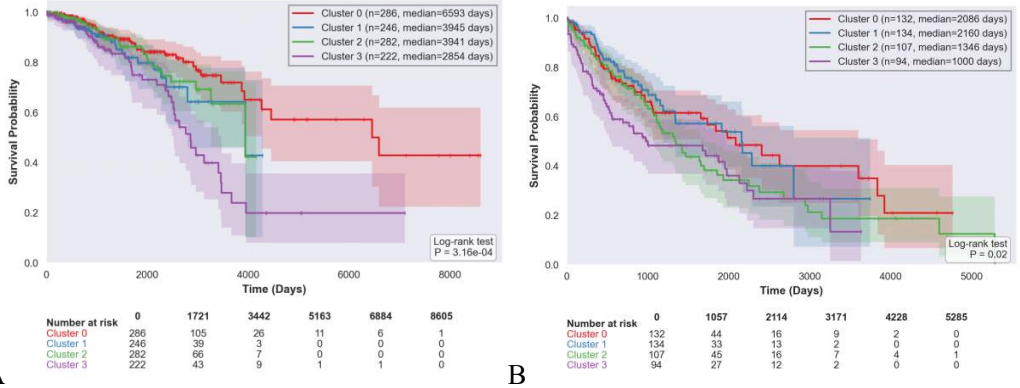


Figure 4. Survival relevance of multi-omics subtypes. (A) Kaplan–Meier curves for BRCA subtypes. (B) Kaplan–Meier curves for LUSC subtypes. (C) Forest plot of subtype hazard ratios (reference: Cluster 3) for BRCA

subtypes. (D) Forest plot of subtype hazard ratios (reference: Cluster 3) for LUSC subtypes.

Table 2. Log-rank test and Cox proportional hazards analysis of molecular subtypes

Dataset	Subtype	Median Survival (days)	HR (vs Cluster 3)	Cox p-value (Wald/Global)	Log-rank p-value	C-index
BRCA	Cluster 0	6,593	0.3999	**3.68×10^{-5}**	—	—
BRCA	Cluster 1	3,945	0.6654	0.092	—	—
BRCA	Cluster 2	3,941	0.5490	0.012	—	—
BRCA	Cluster 3	2,854	Reference	—	—	—
BRCA	**Overall**	**—**	**—**	**6.11×10^{-4}**	**3.16×10^{-4}**	**0.5830**
LUSC	Cluster 0	2,086	0.6067	0.012	—	—
LUSC	Cluster 1	2,160	0.5589	0.006	—	—
LUSC	Cluster 2	1,346	0.8025	0.256	—	—
LUSC	Cluster 3	1,000	Reference	—	—	—
LUSC	**Overall**	**—**	**—**	**0.019**	**0.0166**	**0.5796**

To further evaluate the survival relevance of the identified subtypes, we compared MAVDC with several representative multi-omics clustering methods, including SNF, iClusterPlus, and CIMLR, each under its respective cluster configuration. As summarized in Table 3, different methods resulted in different numbers of clusters across cancer types. Across both TCGA-BRCA and TCGA-LUSC cohorts, MAVDC-derived subtypes consistently exhibited statistically significant survival separation. These results suggest that MAVDC can identify clinically relevant molecular stratifications with stable prognostic associations across cancer types, despite methodological differences in cluster resolution among integrative approaches.

Table 3. Comparison of subtyping results of different integration methods

Cancer types	Item	MAVDC (Ours)	SNF	iClusterPlus	CIMLR
BRCA	Number of clusters	4	5	3	5
	Log-rank P-value	3.16E-04	2.41E-03	1.05E-02	5.20E-04
LUSC	Number of clusters	4	4	3	4
	Log-rank P-value	0.0166	0.0424	0.1152	0.0382

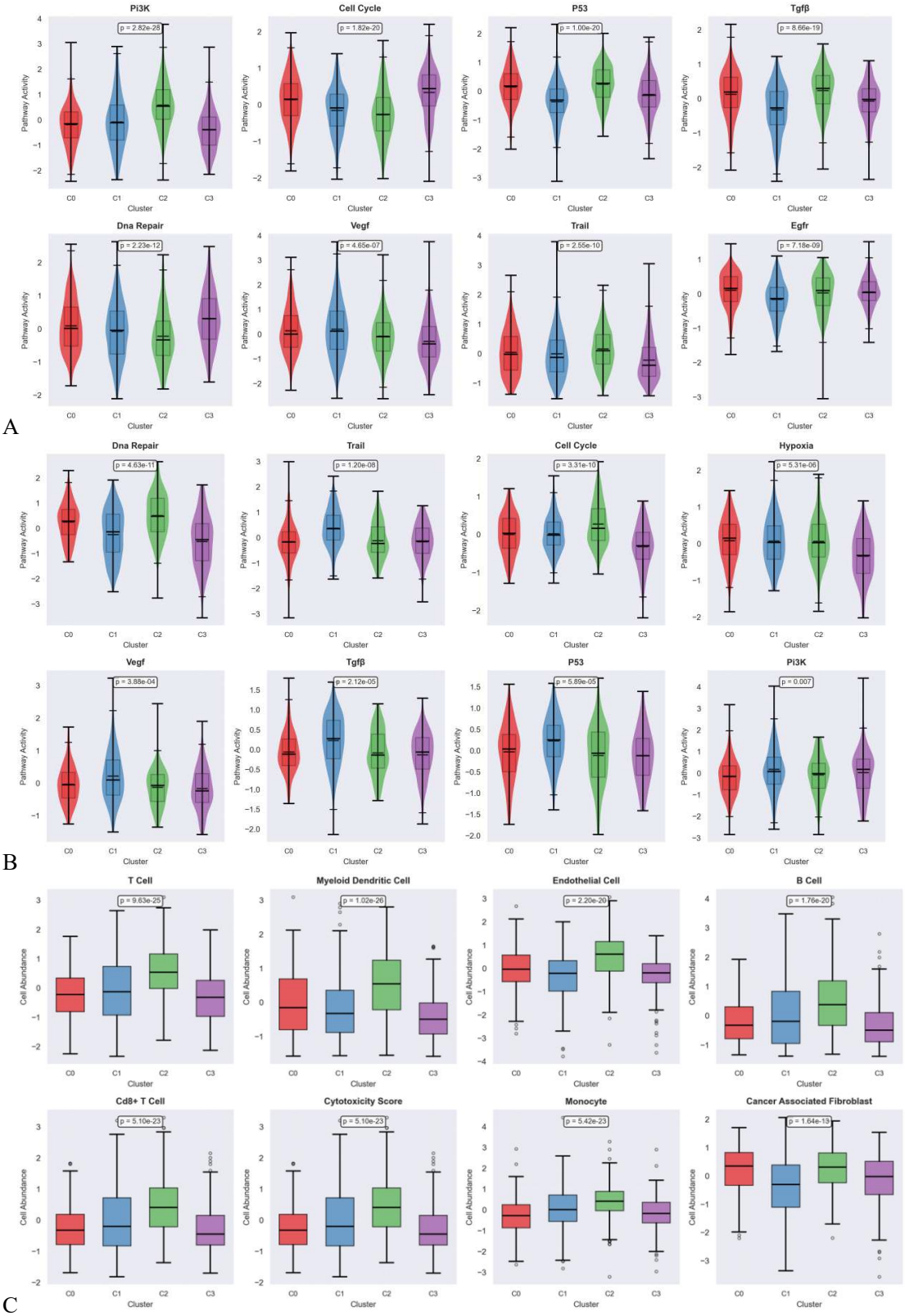
Functional and immune characteristics of subtypes

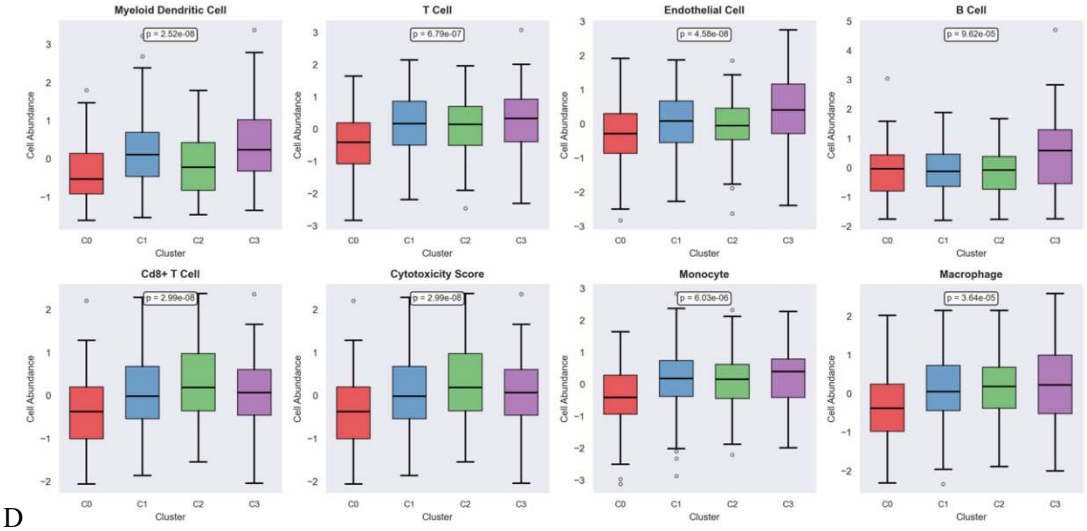
To elucidate the biological basis underlying the survival-associated multi-omics subtypes, we conducted pathway-level functional profiling and immune microenvironment characterization in both BRCA and LUSC cohorts.

Pathway activity was quantified using single-sample gene set enrichment analysis (ssGSEA) across a curated collection of cancer-related signaling pathways. In both cohorts, the majority of evaluated pathways exhibited significant subtype-dependent differences as assessed by one-way analysis of variance followed by Benjamini–Hochberg false discovery rate (FDR) correction (FDR < 0.05), indicating widespread functional heterogeneity among the four molecular subtypes (Figure 5A–B). These findings demonstrate that the integrative subtypes are characterized by coordinated divergence across multiple oncogenic and stress-response programs, rather than by isolated gene-level alterations.

To characterize the tumor immune landscape, immune and stromal components were estimated using marker gene expression–based abundance scoring. Significant subtype-specific differences were observed across both cancer types in multiple immune cell populations, encompassing both lymphoid and myeloid lineages, as well as in composite cytotoxicity scores (Figure 5C-D). These variations in immune contexture suggest that the integrative subtypes capture not only tumor-intrinsic molecular programs but also distinct tumor microenvironmental states.

Collectively, the coordinated divergence of oncogenic signaling pathways (Figure 5A–B) and immune/stromal contexts (Figure 5C-D) across subtypes supports an integrated biological framework in which tumor-intrinsic signaling programs and microenvironmental states jointly contribute to the observed survival stratification.


A
Pi3K
p = 2.82e-28
Cell Cycle
p = 1.82e-20
P53
p = 1.00e-20
Tgfβ
p = 8.66e-19
Dna Repair
p = 2.23e-12
Vegf
p = 4.65e-07
Trail
p = 2.55e-10
Egfr
p = 7.18e-09
B
Dna Repair
p = 4.63e-11
Trail
p = 1.20e-08
Cell Cycle
p = 3.31e-10
Hypoxia
p = 5.31e-06
Vegf
p = 3.88e-04
Tgfβ
p = 2.12e-05
P53
p = 5.89e-05
Pi3K
p = 0.007
C
T Cell
p = 9.63e-25
Myeloid Dendritic Cell
p = 1.02e-26
Endothelial Cell
p = 2.20e-20
B Cell
p = 1.76e-20
Cd8+ T Cell
p = 5.10e-23
Cytotoxicity Score
p = 5.10e-23
Monocyte
p = 5.42e-23
Cancer Associated Fibroblast
p = 1.64e-13
Pathway Activity
Cell Abundance
Cluster
C0
C1
C2
C3



D

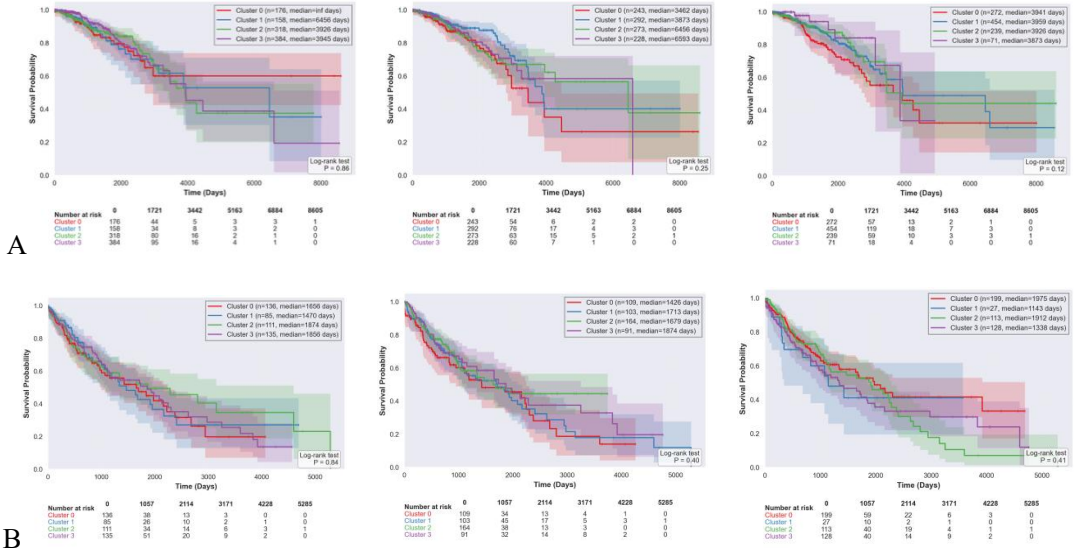
Figure 5. Functional and immune characterization of molecular subtypes. (A-B) Violin plots of ssGSEA pathway activities in BRCA and LUSC. (C-D) Box plots of immune/stromal component scores across subtypes in both BRCA and LUSC.

Comparison with single-omics baselines

To evaluate whether integrative multi-omics fusion provides added value beyond individual molecular layers, we compared the proposed integrative subtyping framework with clustering solutions derived from RNA expression, DNA methylation, and CNV data alone under identical analytical settings.

Across both the BRCA and LUSC cohorts, single-omics–based clustering exhibited consistently poorer geometric structure than the integrative model, as reflected by lower Silhouette scores and higher Davies–Bouldin indices (Figure 6; Table 4). In contrast, the integrative approach achieved the most balanced combination of cluster compactness and inter-cluster separation across both cancer types.

Although CNV-only clustering showed relatively high Calinski–Harabasz values, it failed to demonstrate consistent prognostic stratification, indicating that geometric separation alone is insufficient to capture survival-relevant heterogeneity.



A

B

Figure 6. Comparison of multi-omics integration with single-omics baselines. (A) Kaplan–Meier curves for BRCA based on single-omics clustering (RNA-only, methylation-only and CNV-only; k = 4). (B) Kaplan–Meier curves for LUSC based on single-omics clustering (RNA-only, methylation-only and CNV-only; k = 4).

Importantly, integrative multi-omics subtypes demonstrated more stable and significant survival discrimination than any single-modality baseline in both cohorts. While clustering based on individual omics layers produced weak or inconsistent survival separation, integrative modeling yielded reproducible subtype-associated survival differences across both BRCA and LUSC cohorts

(Table 4).

Table 4. Effects of different omics combinations on BRCA clustering performance and survival outcomes

Dataset	Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score	Log-rank p-value	C-index	Cox p-value
BRCA	Multi-omics	0.3182	1.2364	336.06	3.16×10^{-4}	0.5830	6.11×10^{-4}
BRCA	RNA-only	0.1440	2.0309	160.02	0.864	0.5591	0.410
BRCA	Meth-only	0.1035	2.1741	116.99	0.248	0.5224	0.443
BRCA	CNV-only	0.2547	1.8797	628.48	0.116	0.5588	0.026
LUSC	Multi-omics	0.3410	1.1805	164.05	0.017	0.5796	0.019
LUSC	RNA-only	0.0955	2.4535	51.44	0.835	0.5145	0.547
LUSC	Meth-only	0.1232	2.1432	65.90	0.404	0.5361	0.124
LUSC	CNV-only	0.2617	1.6639	261.64	0.410	0.5282	0.125

Collectively, these findings demonstrate that clinically meaningful tumor subtypes arise from coordinated alterations across multiple molecular layers, highlighting the necessity of integrative modeling to robustly capture survival-associated molecular heterogeneity.

Discussion

Cancer subtyping remains a central challenge in precision oncology, as clinically similar tumours can originate from distinct molecular programs and tumour microenvironmental states. In this study, we established a harmonized multi-omics framework that jointly integrates RNA expression, DNA methylation, and copy-number variation to derive data-driven patient strata across independent cohorts, aiming to attenuate platform-specific noise while preserving complementary biological signals across molecular layers. The resulting subtype structure was highly reproducible across two biologically distinct cancer contexts, supporting the robustness of multi-omics fusion in capturing shared disease axes rather than cohort-specific artefacts.

A central finding of this work is that the inferred four-subtype solution captures clinically meaningful risk heterogeneity. Subtype membership consistently stratified overall survival in both cohorts, demonstrating reproducible prognostic relevance. Importantly, we did not treat subtype labels as purely geometric partitions; instead, we anchored them to biological mechanisms through integrative functional and immune profiling, providing an interpretable bridge from unsupervised stratification to mechanistic hypotheses. Functional characterization using ssGSEA across canonical cancer-related pathways revealed coordinated subtype divergence in growth factor signaling, cell-cycle regulation, and DNA damage and repair programs, suggesting that the identified subtypes correspond to distinct functional tumour states with potentially different aggressiveness and therapeutic vulnerabilities. The consistency of these pathway patterns across two tumour types further implies that a component of the subtype structure reflects transferable biological programs rather than dataset-specific statistical structure.

In parallel, immune profiling revealed systematic subtype-specific differences across lymphoid, myeloid, and stromal compartments, together with composite cytotoxicity measures, indicating that the subtypes capture not only tumour-intrinsic programs but also tumour microenvironmental heterogeneity. This observation supports an integrated interpretation of subtype-associated prognosis, whereby outcome differences may be driven by the joint effects of oncogenic signaling programs and immune–stromal context rather than by either dimension alone. Moreover, these immune-contexture differences generate clinically testable hypotheses: immune-active subtypes may potentially benefit from immunotherapy-oriented strategies, whereas immune-cold subtypes may require rational combinatorial approaches targeting both tumour signaling and immune exclusion. However, these therapeutic implications require prospective validation in independent clinical cohorts.

Comparison with single-omics baselines further underscores the practical value of integration.

Under matched analytical settings, multi-omics fusion consistently yielded more coherent cluster structures and stronger survival stratification than analyses restricted to RNA expression, DNA methylation, or CNV alone, highlighting that clinically relevant tumour heterogeneity is distributed across molecular layers and that integrative modeling can mitigate modality-specific noise, sparsity, and missingness.

Several limitations warrant consideration. First, the current analyses are retrospective and based on public cohorts; broader validation across additional tumour types and independent datasets will be necessary to establish generalizability and clinical transportability. Second, as an unsupervised approach, the current subtype discovery framework is not explicitly optimized for clinical outcomes; semi-supervised or outcome-guided extensions may further enhance prognostic utility while preserving interpretability. Third, functional and immune characterizations are inference-based summaries; orthogonal validation using alternative immune cell quantification methods and experimental assays would strengthen biological credibility and translational relevance.

In summary, our integrative framework delineates reproducible molecular subtypes with coherent functional programs and distinct immune microenvironments, providing an interpretable foundation for risk stratification, biological hypothesis generation, and the development of subtype-aware strategies in precision oncology.

Funding

This work was supported by Dalian Science and Technology Innovation Fund Program [2022 JJ12GX017], the Foundation of Liaoning Education Ministry [LJ232513631003], and Technology Innovation Project of Dalian Neusoft University of Information [TIFP202302].

Data Availability

The datasets used in this study are publicly available from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Marusyk, A., & Polyak, K. (2010). *Tumor heterogeneity: causes and consequences*. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1805(1), 105–117. https://doi.org/10.1016/j.bbcan.2009.11.002.
2. Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology, 15(2), 81 – 94. https://doi.org/10.1038/nrclinonc.2017.166.
3. Hoadley, K. A., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4), 929 – 944. https://doi.org/10.1016/j.cell.2014.06.049.
4. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 83. https://doi.org/10.1186/s13059-017-1215-1.
5. Shen, R., Olshen, A. B., & Ladanyi, M. (2009).Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.Bioinformatics, 25(22), 2906 – 2912.
6. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale.Nature Methods, 11(3), 333 – 337.
7. Argelaguet, R., Velten, B., Arnol, D., et al. (2018). Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6), e8124. https://doi.org/10.15252/msb.20178124.
8. Nguyen, N. D., & Wang, D. (2020). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics, 36(3), 922 – 924.

9. Chaudhary K, Poirion O B, Lu L, et al. Deep learning–based multi-omics integration robustly predicts survival in liver cancer[J]. Clinical cancer research, 2018, 24(6): 1248-1259.
10. Chai H, Zhou X, Zhang Z, et al. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction[J]. Computers in biology and medicine, 2021, 134: 104481.
11. Wei Z, Han D, Zhang C, et al. Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer[J]. Frontiers in oncology, 2022, 12: 893424.
12. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer[J]. BMC genomics, 2021, 22(1): 214.
13. Poirion O B, Jing Z, Chaudhary K, et al. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data[J]. Genome medicine, 2021, 13(1): 112.
14. Jiang L, Xu C, Bai Y, et al. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data[J]. NPJ precision oncology, 2024, 8(1): 4.
15. Hu Y, Li X, Yi Y, et al. Deep learning-driven survival prediction in pan-cancer studies by integrating multimodal histology-genomic data[J]. Briefings in Bioinformatics, 2025, 26(2): bbaf121.
16. Wang, J., Liao, N., Du, X., et al. (2024). A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. BMC Genomics, 25(1), 86.
17. Cheng L, Huang Q, Zhu Z, et al. MoAGL-SA: a multi-omics adaptive integration method with graph learning and self attention for cancer subtype classification[J]. BMC bioinformatics, 2024, 25(1): 364.
18. Lan, W., Liao, H., Chen, Q., et al. (2024). DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. Briefings in Bioinformatics, 25(3), bbae185.
19. Liu, Z., & Park, T. (2024). DMOIT: denoised multi-omics integration approach based on transformer multi-head self-attention mechanism. Frontiers in Genetics, 15, 1488683.
20. Peng, W., Wu, R., Dai, W., et al. (2023). Identifying cancer driver genes based on multi-view heterogeneous graph convolutional network and self-attention mechanism. BMC Bioinformatics, 24(1), 16.
21. Hira, Muta Tah, et al. "Integrated multi-omics analysis of ovarian cancer using variational autoencoders." Scientific reports 11.1 (2021): 6265.
22. Cai, Y., & Wang, S. (2024). Deeply integrating latent consistent representations in high-noise multi-omics data for cancer subtyping. Briefings in Bioinformatics, 25(2), bbae061.
23. Benkirane H, Pradat Y, Michiels S, et al. CustOmics: A versatile deep-learning based strategy for multi-omics integration[J]. PLOS Computational Biology, 2023, 19(3): e1010921.
24. Rahmanian, M., & Mansoori, E. G. (2024). MoVAE: Multi-Omics Variational Auto-Encoder for Cancer Subtype Detection. IEEE Access, 12, 58945–58957.
25. Ballard, J. L., Wang, Z., Li, W., et al. (2024). Deep learning-based approaches for multi-omics data integration and analysis. BioData Mining, 17(1), 38.
26. Wang, L., Zhang, H., Yi, B., et al. (2025). FactVAE: a factorized variational autoencoder for single-cell multi-omics data integration analysis. Briefings in Bioinformatics, 26(2), bbae132.
27. Tanvir, R.B.; Islam, M.M.; Sobhan, M.; Luo, D.; Mondal, A.M. MOGAT: A Multi-Omics Integration Framework Using Graph Attention Networks for Cancer Subtype Prediction. Int. J. Mol. Sci. 2024, 25, 2788.
28. Yang, H., Chen, R., Li, D., et al. (2021). Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics, 37(16), 2231–2237.
29. Omran, M. M., Emam, M., Gamaleldin, M., et al. (2025). Comparative analysis of statistical and deep learning-based multi-omics integration for breast cancer subtype classification. Journal of Translational Medicine, 23(1), 709.
30. Bourgon, R., Gentleman, R., & Huber, W. (2010).Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences of the United States of America, 107(21), 9546–9551.
31. Xie, J., Girshick, R., & Farhadi, A. (2016).Unsupervised deep embedding for clustering analysis.International Conference on Machine Learning (ICML), 2016, 478–487.

32. van den Oord, A., Li, Y., & Vinyals, O. (2018).Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748.
33. Rousseeuw, P. J. (1987).Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.Journal of Computational and Applied Mathematics, 20, 53 – 65.
34. Davies, D. L., & Bouldin, D. W. (1979).A cluster separation measure.IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224 – 227.
35. Calinski, T., & Harabasz, J. (1974).A dendrite method for cluster analysis.Communications in Statistics, 3(1), 1 – 27.
36. Von Luxburg, U. (2010).Clustering stability: An overview.Foundations and Trends in Machine Learning, 2(3), 235 – 274.
37. Gionis, A., Mannila, H., & Tsaparas, P. (2007).Clustering aggregation.ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 4.
38. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001).On clustering validation techniques.Journal of Intelligent Information Systems, 17(2 – 3), 107 – 145.
39. Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). survey of clustering with deep learning: From the perspective of network architecture.IEEE Access, 6, 39501 – 39514.
40. Zhu, L., Liu, Z., & Liu, G. (2025).A deep multiple self-supervised clustering model based on autoencoder networks.Scientific Reports, 15, 18372 (2025).
41. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Liu, C. (2020).Supervised contrastive learning.Advances in Neural Information Processing Systems, 33, 18661 – 18673.
42. Kaplan, E. L., & Meier, P. (1958).Nonparametric estimation from incomplete observations.Journal of the American Statistical Association, 53(282), 457 – 481.
43. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982).Evaluating the yield of medical tests.Journal of the American Medical Association (JAMA), 247(18), 2543 – 2546.
44. 44.Cox, D. R. (1972).Regression models and life-tables.Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187 – 220.
45. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016).Toward a shared vision for cancer genomic data. New England Journal of Medicine, 375(12):1109 – 1112.
46. Van Loo, P., Nordgard, S. H., Lingjærde, O. C., et al. (2010).Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America, 107(39), 16910 – 16915.

Key Points

1. We propose MAVDC, a progressive three-phase training strategy. Separates representation learning from clustering optimization to prevent cluster collapse. Pre-training learns representations without clustering loss, followed by adaptive cluster discovery, then joint optimization.
2. An adaptive cluster discovery strategy is introduced to determine the optimal number of molecular subtypes by jointly considering clustering geometry, stability, and separation.
3. Multi-head attention for cross-modal fusion combined with InfoNCE contrastive learning enhances modality alignment and improves survival differences.
4. Applied to BRCA and LUSC cohorts, MAVDC consistently identifies four molecular subtypes with clear geometric separation, distinct functional and immune characteristics, and significant survival differences.
5. Comparative analyses show that MAVDC outperforms single-omics and representative multi-omics methods in clustering quality and prognostic relevance, highlighting the value of integrative modeling for cancer heterogeneity analysis.

Author Biographies

Xinning Liu is an Associate Professor at the School of Software, Dalian Neusoft University of

Information, with research interests in multi-omics data integration and deep learning for cancer subtyping and precision medicine.

Li Han is a researcher at the School of Software, Dalian Neusoft University of Information, focusing on multi-omics data analysis, machine learning, and bioinformatics modeling.

Ling Kang is a Professor at the Neusoft Research Institute, Dalian Neusoft University of Information, working on bioinformatics, medical data analysis, and multi-omics‑driven studies of disease mechanisms.

Lei Zhao is a researcher at the Neusoft Research Institute, Dalian Neusoft University of Information, with interests in multi-omics integration and deep learning methods for complex disease research.

Quan Guo is a Professor at the Neusoft Research Institute, Dalian Neusoft University of Information, whose research focuses on bioinformatics, statistical learning, and multi-omics data analysis in cancer research.

Huadong Miao is an independent researcher with interests in multi-omics data integration, machine learning, and computational approaches for cancer subtyping.