文章编号：1671-7554（2025）08-0051-10 DOI：10.6040/j.issn.1671-7554.0.2025.0510

基于多模态解耦对比学习的癌症亚型聚类方法

张润泽[1]，薛付忠[1,2,3]，杨帆[1,2,3]
（1.山东大学齐鲁医学院公共卫生学院医学数据学系，山东 济南 250012；
2.国家健康医疗大数据研究院，山东 济南 250003；3.山东大学齐鲁医院，山东 济南 250012）

摘要：目的 基于癌症基因组图谱（the cancer genome atlas，TCGA）中 5 种癌症的多组学数据，提出一种融合图卷积网络、自注意力机制与解耦对比学习的癌症亚型聚类模型。**方法** 模型以 TCGA 数据库中 5 种癌症的 4 种组学数据为输入，分别构建每类组学中样本之间的关系网络，利用图卷积网络提取组学内部的结构信息，更好地保留样本之间的特征差异。将不同组学下的特征进行拼接，并通过注意力机制进行加权融合，自动学习各组学的重要程度与互补关系。最后采用解耦对比学习方法，利用样本增强后的不同视角进行无监督训练，引导模型在没有真实标签的情况下识别出潜在的癌症亚型。**结果** 模型在 5 种癌症数据中均表现出良好的聚类效果，能够将样本有效划分为不同的亚型。在生存分析中，各亚型之间的生存曲线呈现显著分离，说明模型识别的亚型预后存在差异。部分亚型在临床特征上也表现出较强的区分能力。与多种现有方法相比，本研究模型在多项评价指标上均取得良好结果，聚类结果具有更高的稳定性，同时展现出更强的生物学解释能力。**结论** 本研究提出的癌症亚型聚类模型通过图卷积网络、自注意力机制与对比学习的协同作用，有效整合多组学数据，显著提升了癌症亚型聚类的准确性和临床解释力，该模型为癌症异质性研究提供了新思路，有助于精准医疗的个性化治疗策略制定。

关键词：癌症亚型聚类；多组学；图卷积网络；自注意力机制；解耦对比学习
中图分类号：R730.43 **文献标志码：**A

Cancer subtype clustering via multimodal decoupled contrastive learning

ZHANG Runze[1], XUE Fuzhong[1,2,3], YANG Fan[1,2,3]
（1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China;
2. National Institute of Health and Medical Big Data, Jinan 250003, Shandong, China;
3. Qilu Hospital of Shandong University, Jinan 250012, Shandong, China）

Abstract: Objective To propose a cancer subtype clustering model that integrates graph convolutional networks, self-attention mechanisms, and decoupled contrastive learning, based on multi-omics data from five cancer types in the cancer genome atlas (TCGA). **Methods** The model took four types of omics data from five cancer types in the TCGA database as input. For each omics type, it constructed a sample-wise relational graph and employed a graph convolutional network (GCN) to extract intra-omics structural information, thereby better preserving inter-sample feature differences. The features from different omics were concatenated and further fused through an attention mechanism, which automatically learned the relative importance and complementary relationships among omics modalities. Finally, a decoupled contrastive learning strategy was applied, and different augmented views of the same sample were used for unsupervised training, guiding the model to identify potential cancer subtypes in the absence of ground-truth labels. **Results** The model demonstrated good clustering performance across five cancer datasets, effectively dividing samples into distinct subtypes. In survival analysis, the survival curves of different subtypes showed significant separation, indicating that the identified subtypes were associated with different prognoses. Some subtypes also exhibited strong differentiation in clinical characteristics. Compared with several existing methods, the proposed

收稿日期：2025-04-30
基金项目：国家自然科学基金（82273736，62272278）
通信作者：杨帆。E-mail：fanyang@ sdu.edu.cn
薛付忠。E-mail：xuefzh@ sdu.edu.cn

model achieved favorable results on multiple evaluation metrics, yielding more stable clustering outcomes and demonstrating stronger biological interpretability. **Conclusion** This study proposes a cancer subtype clustering model that effectively integrates multi-omics data through the synergistic use of GCN, self-attention mechanisms, and contrastive learning. The model significantly improves the accuracy and clinical interpretability of cancer subtype clustering, offering a new perspective for cancer heterogeneity research and contributing to the development of personalized treatment strategies in precision medicine.

Key words: Cancer subtype clustering; Multi-omics; Graph convolutional network; Self-attention mechanism; Decoupled contrastive learning

癌症已成为全球范围内最为严重的公共卫生问题之一[1]。根据世界卫生组织的统计数据，癌症每年造成全球六分之一的死亡，造成了重大的健康和经济负担。癌症是一类高度异质性的复杂疾病，其在形态学、分子特征等方面呈现出差异，这种生物学和临床表型上的异质性不仅体现在不同类型的癌症之间，更存在于同一种癌症的不同亚型中[2]。因此，识别癌症亚型对于精确诊断、预后评估和个体化治疗方案的制定至关重要。

传统的癌症亚型分类方法主要依赖于组织病理学特征、大小、分级以及癌症分期等信息。这些方法在临床中发挥了一定作用，有助于降低癌症死亡率。然而，由于它们无法揭示癌症的深层分子机制，对于辅助治疗决策、判断预后等方面存在明显局限[3]。随着高通量测序技术的飞速发展，大规模癌症组学数据逐渐积累，有助于更加全面地了解各种癌症的复杂机制[4]。

近年来，多模态组学数据融合成为癌症亚型研究的关键方向之一。通过整合不同组学维度中的互补信息，能够更全面地理解癌症的发生发展机制。研究者已在多个癌种中证明，基于全转录组、表观组和其他多组学信息的亚型识别方法比传统单一维度的分类更具生物学解释力和预测价值[5]。

多模态组学数据的复杂异质性和高维性使得它们的有效整合具有挑战性。研究者提出了多种基于统计方法的多模态组学聚类方法，如非负矩阵分解（non-negative matrix factorization, NMF）[6]、多组学典型相关分析（multiple canonical correlation analysis, MCCA）[7]以及基于潜变量模型的 iCluster[8]等。这些方法在一定程度上实现了对多模态组学数据的联合建模，推动了亚型发现的进展。但是这类传统方法通常依赖线性建模假设，难以充分捕捉高维、多模态组学数据中复杂的非线性结构。为弥补传统线性模型在建模非线性多组学关系中的不足，可使用深度学习算法进行多模态组学聚类，如自动编码器（autoencoder, AE）[9]、变分自动编码器（variationalautoencoder, VAE）[10]、生成对抗网络（generative adversarial network, GAN）[11]，但这些方法往往需要较大的数据量以保证模型的稳定性和泛化能力。在癌症研究中，由于真实样本获取成本高、组学数据缺失严重，小样本问题普遍存在，因此这些方法的性能与稳定性均面临显著挑战。

本研究基于癌症基因组图谱中的多组学数据，提出融合图卷积网络、自注意力机制与解耦对比学习的癌症亚型聚类模型，以提升无监督亚型识别性能。该方法有助于深入揭示癌症的分子异质性，为个性化治疗策略的制定提供技术支持。

1 资料与方法

1.1 资料

1.1.1 数据来源

癌症基因组图谱（the cancer genome atlas, TCGA）[12]作为全球权威的肿瘤多组学数据集，涵盖了 33 种癌症的基因组、转录组、表观遗传等多维度数据，为肿瘤研究提供了丰富资源和重要参考标准。为了评估 GSCC 的聚类性能，本研究选取了 5 个已有研究中已明确亚型数量的典型癌症数据集，包括乳腺浸润性癌（breast invasive carcinoma, BRCA）、肺腺癌（lung adenocarcinoma, LUAD）、胰腺癌（pancreatic adenocarcinoma, PAAD）、肾透明细胞癌（kidney renal clear cellcarcinoma, KIRC）、胃腺癌（stomach adenocarcinoma, STAD），每个癌症数据集包含来自拷贝数变异（copy number variation, CNV）、mRNA、miRNA 和 DNA 甲基化 4 个分子平台的多组学数据。

1.1.2 数据预处理

由于癌症的组学数据具有极高的特征维度，因此本研究针对不同组学数据分别进行特征筛选，并将特征选择过程应用于所有数据集：针对 CNV 数据，采用 GISTIC2.0[13]工具对原始 CNV 矩阵进行分析，识别在群体水平上具有统计学意义的扩增与缺失区域。GISTIC2.0 可剔除背景扰动并识别潜在驱动变异，从而有效过滤无生物学意义的随机变异位点，保留与肿瘤发生发展相关的重要拷贝数特征。针对 DNA 甲基化数据，为避免正常组织背景信号

对分析结果的干扰,利用 t 检验对正常组织和癌症组进行差异甲基化 CpG 位点分析,为控制多重检验的影响,对结果进行 FDR 校正,筛选标准为差异均值阈值>0.15,P<0.05。针对 mRNA 表达数据,剔除在<10%样本中表达量≥1 的低表达基因,计算基因表达变异系数,选取前 15%的高变异基因用于后续分析;针对 miRNA 表达数据,首先剔除丰度极低的 miRNA,RPM<10 且在<10%样本中检测到,以减少测序误差的干扰。随后,通过差异表达分析方法,筛选在癌组织与对应正常组织之间差异有统计学意义的 miRNA,阈值设定为 FDR<0.05 且$|\log_2 FC|>1$[14]。

对样本量或特征数缺失超过 20%的数据进行删除[15],当缺失值过多时会导致特征无法真实反映样本特性,大量无效特征增加计算复杂度,而且在后续缺失值插补过程中会掩盖真实的数据分布。对缺失值采用 K 近邻算法(K-nearest neighbors, KNN)[16]进行填补,KNN 算法利用邻近样本的完整数据进行填充,保留高维空间的局部结构,不依赖整体分布,更加适合复杂的多组学数据。采用 Z-score 方法对各组学数据进行标准化,以消除不同特征间的量纲差异。

对每个癌症数据集均采用上述方法进行处理,处理后样本量及各组学特征数量见表 1。

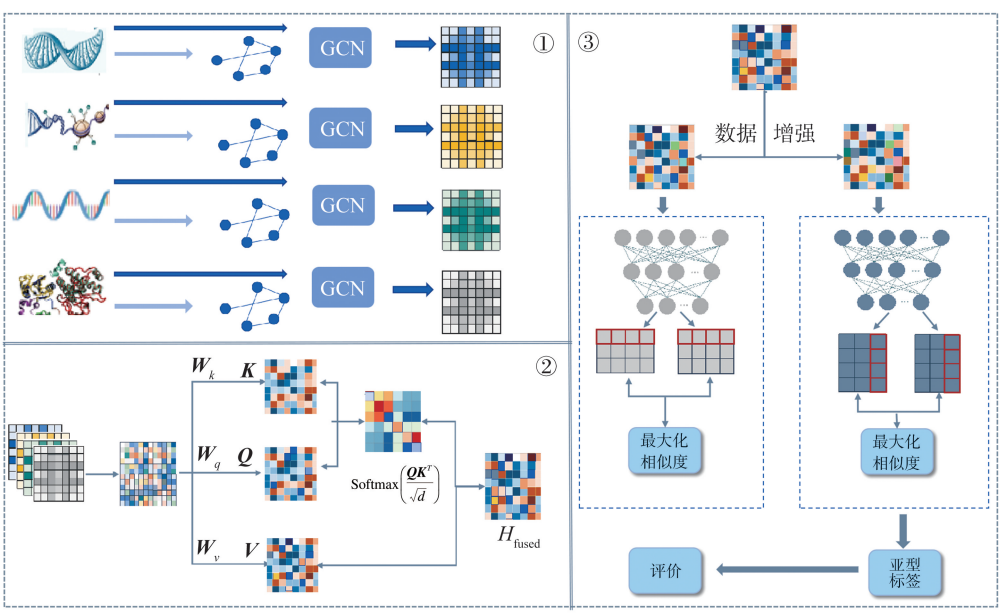
表 1 癌症样本数及组学特征数

Table 1 Number of cancer samples and omics feature

癌症类型	样本量	多组学特征数			
		CNV	DNA 甲基化	mRNA	miRNA
BRCA	920	3 150	3 980	3 989	357
LUAD	440	3 010	3 895	3 096	320
PAAD	150	3 240	3 760	2 231	297
KIRC	430	3 120	2 606	3 882	339
STAD	360	3 350	3 825	3 417	315

1.2 方法

本研究组合了图卷积神经网络(graph convolutional network, GCN)、自注意力机制和解耦对比学习(disentangled contrastive learning, DCL)三部分,提出了 GCN-自注意力对比聚类(GCN-self-attention mechanism contrastive clustering, GSCC)模型,模型框架图如图 1 所示。GCN 通过学习基于余弦相似度构建的样本关系图,整合邻近样本的信息,提取组学数据的结构特征,从而更精准地提取关键特征;引入自注意力机制,其根据数据的实际内容,自适应地调整不同组学间的交互权重与融合方式,捕捉组学间互补信息,实现多模态组学融合。DCL 能够在没有明确标签、样本数量有限的前提下,通过比较样本之间的相似性,引导模型聚出结构清晰的亚型,从而有效缓解小样本数据难以训练的问题。



①特征提取。利用 GCN 对各组学数据构建图结构并提取结构感知的特征表示。②多组学融合。对不同组学的特征进行拼接并通过自注意力机制加权融合。③解耦对比学习模块。通过数据增强生成不同视图,采用样本级与聚类级对比损失进行无监督训练,最终输出亚型标签。

图 1 本研究整体框架

Figure 1 Overall framework of the study

1.2.1 组学特征提取

由于组学内部样本之间存在潜在联系[17]，但是传统的特征提取方法，如主成分分析[18]、线性判别分析[19]等，通常假设样本间相互独立，无法有效捕获组学数据内部样本间的关联。因此，本研究采用GCN方法，如图1-①，通过输入原始特征矩阵$\boldsymbol{X}_m \in \mathbf{R}^{N\times d_m}$和基于样本相似度构建的邻接矩阵$\boldsymbol{A}_m \in \mathbf{R}^{N\times N}$，对每一组学进行独立建模和特征提取，从而使模型能够在图结构中学习组学内部样本之间的潜在关联。

GCN通过聚合每个节点及其邻居节点的特征信息，实现对图结构数据中节点表征的学习。原始特征矩阵$\boldsymbol{X}_m \in \mathbf{R}^{N\times d_m}$提供节点的特征信息，邻接矩阵$\boldsymbol{A}_m \in \mathbf{R}^{N\times N}$刻画节点间的结构依赖关系，图卷积操作通过图结构引导的特征聚合机制，使每个样本的表示融合邻居样本的信息，从而学习出结构保持的嵌入特征。其中，m表示第m个组学，N为样本数，d_m为第m个组学的特征维度。

图结构通常基于样本间的相似性关系构建，本研究采用样本特征之间的余弦相似度作为构图依据，以刻画其在同一组学下的关系强度，余弦相似度的计算公式为：

$$S_{ij}^{m}=\frac{\boldsymbol{X}_i^m \cdot \boldsymbol{X}_j^m}{|\boldsymbol{X}_i^m|\,|\boldsymbol{X}_j^m|}, \tag{1}$$

其中，$\boldsymbol{X}_i^m$表示第i个样本在组学m中的特征向量。为抑制冗余边和噪声连接，仅保留相似度高于阈值τ的边，并排除自连接，构建出的图结构保留了局部显著的相似性结构，有助于提升GCN提取特征时的判别性与鲁棒性。参考相关研究[20-21]通常采用0.6~0.8范围内的相似度阈值，以确保在去除噪声边的同时保持节点间必要的连通性，因此本研究选择$\tau=0.7$作为阈值。图结构$\boldsymbol{A}_m$的定义如下：

$$\boldsymbol{A}_m(i,j)=\begin{cases} S_{ij}^{m}, & \text{if } s_{ij}^{m}\geqslant\tau \text{ and } i \text{ neq } j, \\ 0, & \text{otherwise}。\end{cases} \tag{2}$$

然后将原始特征矩阵和图结构共同输入到GCN模块中，通过图卷积操作提取每个样本在图结构引导下的深层表示。其第l层输出为：

$$\boldsymbol{H}_m^{l+1}=\sigma(\boldsymbol{A}_m\boldsymbol{H}_m^{l}\boldsymbol{W}_m^{l}), \tag{3}$$

其中，$\boldsymbol{H}_m^{l+1}$为第l层特征，$\boldsymbol{W}_m^{l}$为可学习权重矩阵，σ为ReLU激活函数。经过两层GCN的堆叠和传播，模型能够逐步捕获组学内部的高阶语义结构信息，最终得到各组学的高阶特征表示$\boldsymbol{H}_m$。

1.2.2 多组学融合

对于同一样本，不同组学从多个分子层面对样本进行刻画，提供了互补的生物信息，这些特征经过GCN提取后形成了各自独立的语义表示，为整合组学之间的互补信息，将4种组学经过GCN提取的特征按列拼接成一个整体特征向量，为每个样本构建出包含多组学信息的联合表示，有利于模型在统一空间内捕捉不同组学之间的互补性与一致性，多组学联合特征矩阵定义为：

$$\boldsymbol{H}=[\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_m], \tag{4}$$

其中，$\boldsymbol{H}_m$表示第m个组学的高阶表示特征。

为进一步增强不同组学之间的交互建模能力，并捕捉跨组学的语义结构，引入了自注意力机制，自动学习各组学的权重分布，并根据其重要性对组学特征进行加权聚合，最终得到统一的融合表示，如图1-②所示。

首先，将联合特征矩阵$\boldsymbol{H}$通过线性映射为查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）、值（$\boldsymbol{V}$）三个矩阵：

$$\boldsymbol{Q}=\boldsymbol{H}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{H}\boldsymbol{W}_K,\ \boldsymbol{V}=\boldsymbol{H}\boldsymbol{W}_V, \tag{5}$$

其中，$\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$为可学习的线性投影矩阵。

计算组学之间的语义相似性，形成注意力权重矩阵$\boldsymbol{A}$，该矩阵表示每个组学在对其表示进行更新时，对其他组学的关注程度，是注意力机制中的核心中间变量。然后利用该注意力权重矩阵对值矩阵$\boldsymbol{V}$进行加权求和，得到最终融合后的表征输出$\boldsymbol{H}_{\text{fused}}$。

$$\boldsymbol{A}=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\in\mathbf{R}^{N\times N}, \tag{6}$$

$$\boldsymbol{H}_{\text{fused}}=\boldsymbol{A}\cdot\boldsymbol{V}。 \tag{7}$$

1.2.3 DCL

对比学习的核心在于通过构建正负样本对，促进模型区分相似样本与不相似样本，在表征空间中使语义相近的样本更加接近，语义差异较大的样本相对远离[22]。在传统对比学习中，所有表示都通过同一个空间进行对比，这可能导致聚类信息被掩盖。为解决这一问题，GSCC引入了解耦机制[23]，将对比目标拆解为两个子任务：样本级对比学习与聚类级对比学习。样本级对比学习帮助模型更好地区分相似但不完全一样的样本，聚类级对比学习引导同一类的样本聚在一起，共同提升模型的区分性与聚类结构的稳定性，如图1-③所示。

1.2.3.1 数据增强

由于缺乏真实标签，模型难以明确区分不同样本之间的语义关系，GSCC通过数据增强的方式为每个样本构建正负样本对，采用3种样本增强策略—噪声注入（Noise）、掩码遮蔽（Mask）和随机丢弃（Dropout），对每个融合后的样本$\boldsymbol{x}_i$随机施加两种不同的数据增强策略。其中，Noise通过向原始

特征添加均值为0、标准差为0.1的高斯噪声生成噪声样本,Mask以概率0.2随机屏蔽部分特征维度生成遮蔽样本,Dropout以概率0.3随机丢弃部分特征得到随机缺失样本。经过数据增强后得到两个增强样本 $\boldsymbol{X}_i^a$ 和 $\boldsymbol{X}_i^b$。为统一对比空间并兼顾样本相似性与聚类结构,在样本级和聚类级分别构建了两个独立的多层感知机(multilayer perceptron, MLP)投影网络,两个增强样本被分别输入到共享的样本级和聚类级投影头中进行特征映射:

$$q_i^a = f_I(\boldsymbol{x}_i^a),\ q_i^b = f_I(\boldsymbol{x}_i^b), \tag{8}$$

$$y_i^a = \mathrm{softmax}(f_C(\boldsymbol{x}_i^a)),\ y_i^b = \mathrm{softmax}(f_C(\boldsymbol{x}_i^b)), \tag{9}$$

1.2.3.2 样本级对比损失

在癌症亚型聚类任务中,由于缺乏先验标签,正负样本对由数据增强生成的伪标签构建[24],即从同一样本中增强的样本形成正样本对,而其他样本形成负样本对。对比损失是一种通过拉近语义相似样本对、推远语义不相似样本对之间的距离,引导模型在无监督条件下学习判别性特征表示的损失函数。样本级对比损失关注单个样本间的局部相似性,通过最大化正样本对之间的相似度,同时最小化与其他样本之间的相似度,从而学习区分性的特征表示。对应的损失函数定义如下:

$$\mathscr{L}_{\mathrm{ins}} = -\frac{1}{N}\sum_{i=1}^{n}\log\frac{\exp(\mathrm{sim}(q_i^a, q_i^b)/\tau)}{\sum_{j=1, j\neq i}^{N}\left[\exp(\mathrm{sim}(q_i^a, q_i^b)/\tau) + \exp(\mathrm{sim}(q_i^a, q_i^b)/\tau)\right]}, \tag{10}$$

其中,$\mathrm{sim}(\cdot,\cdot)$表示余弦相似度,$\tau$ 为温度系数。

1.2.3.3 聚类级对比损失

聚类级对比损失的目标是通过优化样本在聚类空间中的分布一致性,促使同类样本在嵌入表示空间中聚集于潜在簇中心周围,同时拉远不同类别之间的分离度。聚类级对比损失函数定义如下:

$$\mathscr{L}_{\mathrm{ins}} = -\frac{1}{N}\sum_{i=1}^{n}\log\frac{\exp(\mathrm{sim}(y_i^a, y_i^b)/\tau_c)}{\sum_{j=1}^{N}\left[\exp(\mathrm{sim}(y_i^a, y_i^b)/\tau_c) + \exp(\mathrm{sim}(y_i^a, y_i^b)/\tau_c)\right]} + \lambda H(Y), \tag{11}$$

其中,τ_c 为温度系数,y_i^a、y_i^b 分别表示样本在两个视角下的聚类级表示。$H(Y)$为信息熵正则项,用于防止多数类主导聚类分布,定义如下:

$$H(Y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log y_{ik}, \tag{12}$$

其中,y_{ik}为样本 i 在第 k 个簇上的聚类概率。信息熵正则项能够抑制模型过度偏向某些簇的现象,促使模型生成更加平衡的聚类分布,从而提升聚类结果的可区分性与稳健性。

1.2.3.4 对比损失函数

GSCC的总损失函数整合了上述两个对比路径的优化目标,形成最终联合训练目标:

$$\mathscr{L}_{\mathrm{total}} = \mathscr{L}_{\mathrm{ins}} + \alpha \cdot \mathscr{L}_{\mathrm{clu}}, \tag{13}$$

其中,α 为控制样本级与聚类级损失的权重。该模块可实现表示空间与聚类结构的协同自监督优化,提升模型对未知亚型结构的感知能力。

在本研究中,为确定每个癌症数据集中最优的聚类数量,采用内部评价指标进行聚类效果评估。基于轮廓系数和Davies-Bouldin指数对GSCC聚类结果进行综合评价。其中,轮廓系数越大表示聚类效果越好,Davies-Bouldin指数越小表示类间分离度越高。通过在多个候选聚类数下计算这两个指标,并结合关键指标(轮廓系数)的表现,最终确定在评价指标上表现最优的聚类数量作为每个数据集的最优亚型数[21]。

1.2.4 实验参数设置

为了实现模型的最佳性能,本研究重点调整了6个关键超参数:实例级温度、聚类级温度、批量大小、特征维度、学习率和训练轮数。这些参数对性能影响显著。按上述顺序依次优化各参数,每次仅调整1个,其余5个保持不变。最终确定了用于模型训练的超参数设置,其中实例级温度设置为0.5,聚类级温度设置为1,批量大小设置为64,特征维度设置为128,学习率设置为3.0×10^{-4},训练轮次设置为500。

1.2.5 实验指标

为了全面衡量模型所识别癌症亚型的临床相关性与解释能力,本研究采用两类互补的评价指标:$-\lg P$值和显著的临床参数数量。$-\lg P$ 值用于衡量模型聚类结果与关键临床特征之间的统计关联性。通过log-rank检验计算聚类标签与生存数据之间的 P 值,并将其取负对数,该指标数值越大,表示模型划分出的亚型在生存时间的上差异越显著,能够更好地区分患者的实际生存与病理特征。为进一步反映模型的临床解释力,还统计了与每个聚类结果显著相关的显著临床参数数量。在本研究中,纳入评估的临床参数包括性别、年龄、病理总分期、病理T分期(肿瘤原发部位大小)、病理N分期(区域淋巴结转移)以及病理M分期(远处转移)。$P<0.05$ 作为显著性标准,计算每种方法在每个癌症数据集下

被识别为显著的临床参数数量。

1.2.6 对比实验

为了验证 GSCC 在癌症亚型聚类方面的性能,将 GSCC 与当前 6 种应用较多的亚型聚类方法进行比较:相似性网络融合(similarity network fusion, SNF)[25]、多典型相关分析(multiple canonical correlation analysis, MCCA)[7]、*K* 均值聚类算法(*K*-means clustering, *K*-means)[26]、Subtype-GAN[27]、RISynG[28]。

在对比实验中,SNF、MCCA、*K*-means、iCluster 使用 R 语言中对应的软件包来实现相应的算法;Subtype-GAN 和 RISynG 没有提供可安装的 R 包,因此从 https://github.com/haiyang1986/Subtype-GAN 和 https://github.com/xfurna/RISynG 下载两种方法的源代码并进行实验。为保证实验的公平性,所有方法的参数设置都遵循算法开发人员的默认推荐值。

1.2.7 消融实验

为了评价不同组学在癌症亚型聚类中的作用,并验证多组学融合策略的有效性,本研究在 5 个癌症数据集上设计并开展了消融实验。具体做法为:在原始输入包含 mRNA、miRNA、DNA 甲基化和 CNV 4 类组学的基础上,分别构造 4 组对照实验,每次去除其中 1 类组学数据,其余 3 类保持不变,重新训练并评估模型性能。通过观察聚类评价指标的变化,分析每一组学在模型中的作用强弱以及对最终聚类效果的影响,从而进一步验证多组学融合对于模型性能提升的重要性。

1.3 统计学处理

采用 R 4.3.1 进行统计分析,两独立样本 *t* 检验用于正常组织和癌症组进行差异甲基化 CpG 位点分析,检验水准 $\alpha=0.05$。

2 结 果

2.1 聚类结果

基于内部聚类评估指标确定的最优聚类数见表 2。不同癌症数据集在内部评价指标下对应的最优聚类数有所差异。其中 BRCA、PAAD 和 KIRC 的最优聚类数分别为 4、2、3,对应的轮廓系数为 0.88、0.92、0.72,DB 指数为 0.12、0.15 和 0.24,表明其聚类结构最紧凑、分离度最佳,聚类效果最优。STAD 的最优聚类数为 4,在两项指标中也表现良好。LUAD的最优聚类数为 3,其轮廓系数为 0.83,DB 指数为 0.19,显示出一定的聚类有效性。

表 2 聚类内部评价指标

Table 2 Internal validation metrics

数据集	聚类数	轮廓系数	DB 指数
BRCA	3	0.65	0.15
	4	0.88	0.12
	5	0.41	0.23
PAAD	2	0.92	0.15
	3	0.86	0.17
	4	0.79	0.35
KIRC	3	0.72	0.24
	4	0.67	0.27
	5	0.53	0.36
LUAD	3	0.83	0.19
	4	0.75	0.20
	5	0.56	0.25
STAD	3	0.45	0.33
	4	0.80	0.24
	5	0.72	0.19

本研究最终将 BRCA 分为 4 类、LUAD 分为 3 类、PAAD 分为 2 类、STAD 分为 4 类、KIRC 分为 3 类。图 2 为 5 种癌症数据集亚型聚类的 t-SNE 可视化结果,可以清晰地观察到,模型在各数据集上均能够将样本有效区分为多个结构清晰的簇,体现出较强的聚类效果。在 BRCA、LUAD、PAAD、STAD 和 KIRC 这 5 个数据集的不同类别之间界限明显,簇间分离度较高;展现出良好的样本区分能力。

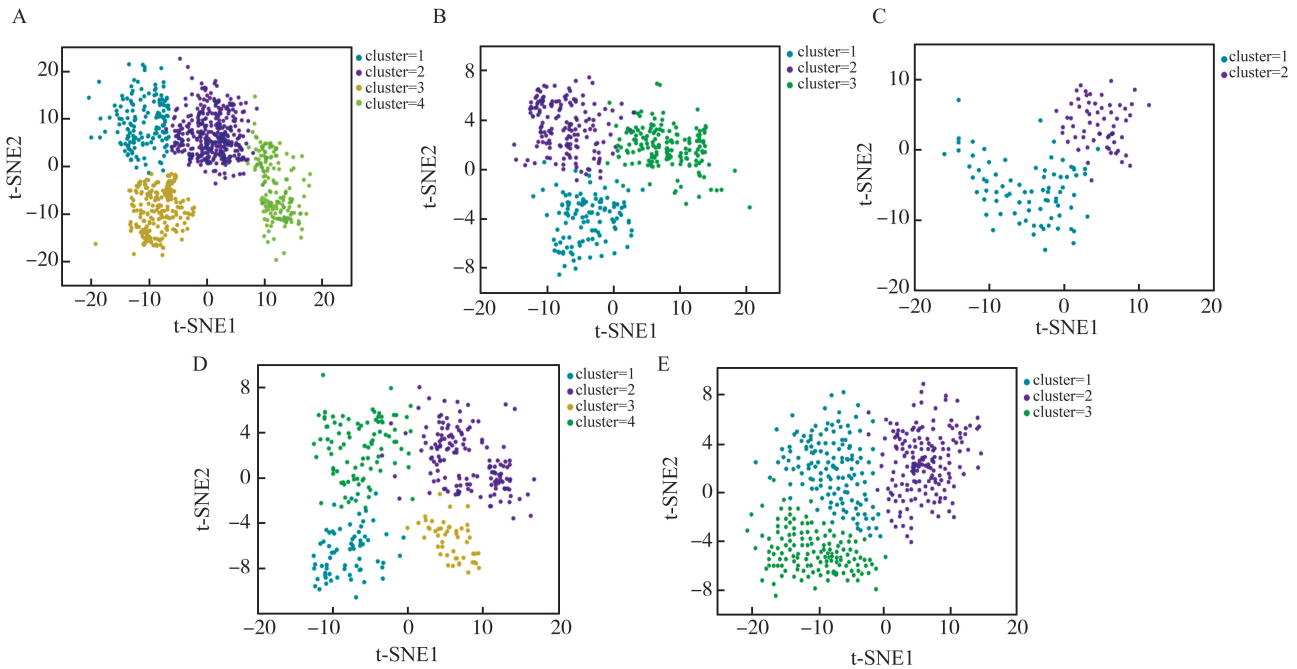


图 2 5 种癌症数据集亚型聚类 t-SNE 可视化结果

A: BRCA; B: LUAD; C: PAAD; D: STAD; E: KIRC。

Figure 2 t-SNE visualization of subtype clustering results for five cancer datasets

A: BRCA; B: LUAD; C: PAAD; D: STAD; E: KIRC.

2.2 对比实验

见表 3。在 $-\lg P$ 指标中,GSCC 在 BRCA、LUAD、PAAD 3 种癌症数据集中取得了最佳结果,分别是 2.04、4.10、4.25;在显著临床参数的数量这个指标中,GSCC 在 BRCA、PAAD、STAD 3 种癌症数据集中取得了最佳结果,分别为 5、4、4。并且,GSCC 在 BRCA、PAAD 聚类的两个评价指标中都取得了最佳结果,显示出 GSCC 在不同癌症类型中都具有稳定且优异的聚类能力,表明其不仅能够有效发现亚型与临床变量间的潜在关联,具备较强的生存区分能力、生物学解释力与泛化能力。其他的对比方法,SNF 仅在 KIRC 和 STAD 的显著临床参数数量上取得最佳结果。MCCA 在 BRCA 数据集的显著临床参数数量和 STAD 数据集的 $-\lg P$ 值上取得最佳结果,RISynG 在 LUAD 数据集显著临床参数数量上取得最佳结果。iCluster 和 *K*-means 方法未在任何数据集上取得最佳结果。Subtype-GAN 和 GSCC 均为深度模型,Subtype-GAN 在多个数据集上的性能仅次于 GSCC,具备一定的建模能力但仍存在差距。

表 3 GSCC 与其他方法的性能比较

Table 3 Performance comparison of GSCCwith other methods

方法/数据集	$-\lg P$(显著临床参数)				
	BRCA	LUAD	PAAD	KIRC	STAD
GSCC	2.04/5	4.10/3	4.25/4	1.91/2	1.08/4
SNF	1.63/3	2.23/2	1.10/2	1.79/3	1.02/4
MCCA	1.52/5	3.31/1	0.79/1	1.52/2	1.28/3
iCluster	0.91/3	0.23/3	0.33/2	1.07/1	0.76/1
K-means	1.22/2	1.01/2	0.98/0	1.19/0	0.01/2
Subtype-GAN	1.93/5	3.56/4	2.37/3	1.98/3	0.91/3
RISynG	0.97/4	2.24/4	0.96/1	1.43/2	0.87/2

GSCC 所识别的亚型在 5 个数据集中均呈现出显著的生存差异,表明该模型能够有效地将患者划分为具有不同预后风险的亚组。在 LUAD 和 PAAD 数据集中,不同亚型之间的生存曲线分离最为明显,随着时间的推移,生存概率的差异逐渐扩大,显示出 GSCC 在这两种癌症中的亚型分层具有较强的预后预测能力。见图 3。

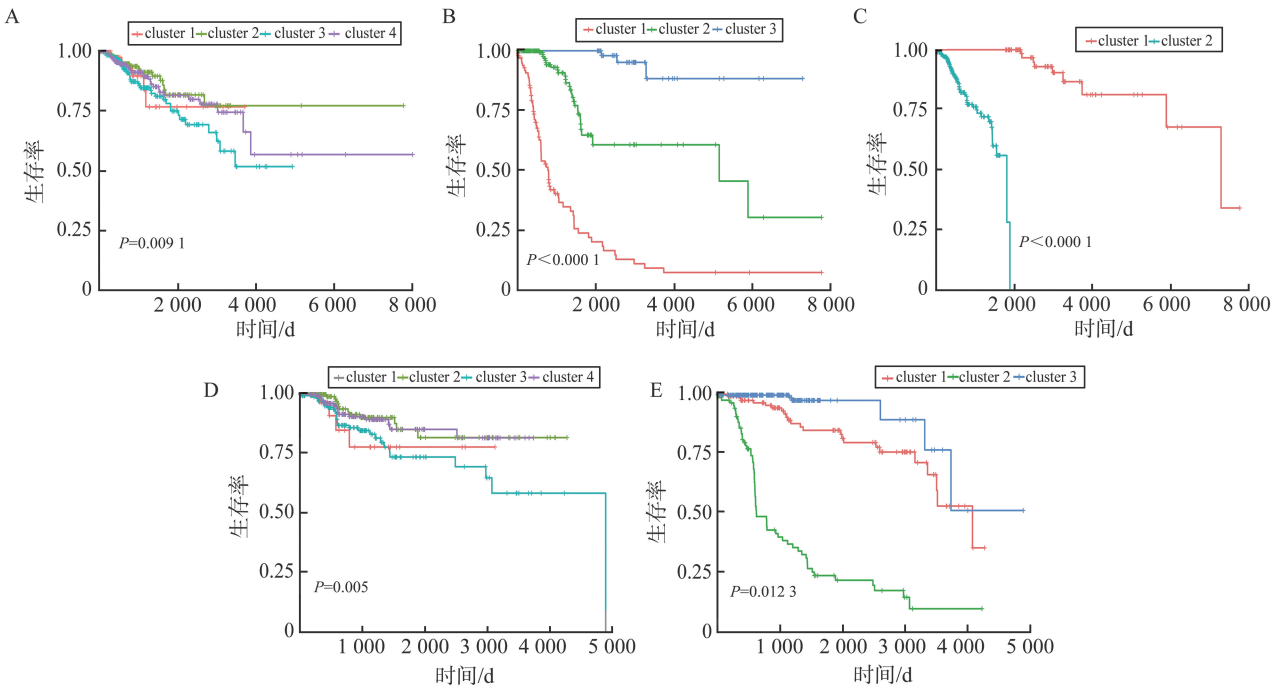


图 3 生存分析结果

A: BRCA; B: LUAD; C: PAAD; D: STAD; E: KIRC。

Figure 3 Results of survival analyses

A: BRCA; B: LUAD; C: PAAD; D: STAD; E: KIRC.

2.3 消融实验

消融实验结果(表 4)表明,去除任意一个组学后,模型在两个评价指标上均表现出不同程度的性能下降。其中,去除 mRNA 和 DNA 甲基化所造成的性能减低尤为显著,说明这两类数据对揭示潜在亚型结构具有较强的表征能力。CNV 和 miRNA 虽在某些癌种中单独作用相对较弱,但在融合建模中仍发挥了重要的互补作用。

表 4 消融实验结果

Table 4 Results of ablation study

消融实验设置	BRCA	LUAD	PAAD	KIRC	STAD
去除 CNV	1.84/4	3.72/2	4.01/3	1.02/3	0.92/2
去除 DNA 甲基化	1.23/2	2.25/1	2.43/1	0.98/2	0.57/1
去除 mRNA	1.31/1	2.18/0	2.39/2	0.74/2	0.60/1
去除 miRNA	1.96/3	3.95/2	3.87/3	1.31/2	1.03/3

3 讨 论

本研究提出一种基于多组学数据的癌症亚型聚类模型—GSCC,其以 GCN、自注意力机制和 DCL 为核心,实现了对多组学数据的结构感知建模、跨组学信息融合和亚型结构的自监督识别。通过在 TCGA 5 种典型癌症数据集上的实验验证,GSCC 在聚类效果、生存区分能力和临床变量关联性等方面相较于其他多组学聚类方法表现出更优性能。

从模型设计层面上,GSCC 通过 GCN 对每种组学数据构建样本相似性图并进行独立建模,有效挖掘了组学内的样本内部潜在联系,提升了特征表达的结构感知能力。相比传统线性特征提取方法,GCN 能更好地适应高维异构组学数据中存在的复杂非线性结构。而通过引入注意力机制,模型能够在统一空间中自适应学习不同组学间的依赖强度,实现对互补信息的充分利用,增强了跨组学对齐能力。GSCC 在缺乏标签监督的前提下引入 DCL 策略,通过样本级与聚类级的联合优化,有效识别出潜在的癌症亚型。在样本量相对有限的真实癌症数据场景中,GSCC 展现出较强的鲁棒性和稳定性,解决了以往深度聚类模型对大规模样本依赖性强的问题。

对比实验结果表明,GSCC 不仅在多个癌症数据集中取得了最优的$-\lg P$ 值和显著临床参数数量,还能够有效区分不同亚型患者的预后差异,具备较强的生物学解释性。对比分析进一步表明,传统的

亚型聚类方法在处理组学异质性及信息融合等方面仍存在局限,而 GSCC 通过结合图结构建模与多组学特征融合,有效提升了模型对组学异质性的适应能力,缓解了传统方法在信息整合方面的局限。消融实验的结果也进一步证明了多组学在亚型聚类中的信息价值,多组学融合建模在提高聚类稳定性、增强生物学解释力方面具有关键意义。单一组学往往难以全面刻画癌症异质性,而多组学联合学习能够更充分地挖掘疾病内部的分层结构,为精准亚型分类提供有力支撑。

尽管 GSCC 在多组学癌症亚型聚类方面表现出一定的优势,但仍存在一些局限性:①本研究的对比范围仍有限,尚未涵盖近年来所有表现优秀的聚类模型,未来可在更大范围内开展系统性评估以进一步验证 GSCC 的优势。②本研究仅使用来自 TCGA 的癌症多组学数据,一定程度上限制了模型的泛化能力,未来可进一步在 ICGC[29]、CPTAC[30] 其他公开组学数据库上进行迁移验证与微调适配,以评估 GSCC 在多来源数据下的鲁棒性。③GSCC 仅使用 CNV、DNA 甲基化、mRNA、miRNA 四个组学数据作为模型的输入,没有引入如蛋白质组学、代谢组学等其他组学的信息,在组学覆盖维度上的局限性可能导致对某些生物学机制的识别能力不足,进而影响对亚型结构的全面揭示;蛋白质组数据作为癌症分子分型的重要信息来源,能够直接反映基因表达后的翻译水平和修饰状态,对疾病机制和潜在治疗靶点具有独特的补充价值。但是现有公开蛋白质组学数据的存在显著稀疏性问题,一方面体现在样本量相对有限,另一方面体现在蛋白质组学数据缺失严重且噪声水平较高。这些特性使得当前模型在扩展到整合蛋白质组学时面临一定挑战,包括对缺失数据的敏感性增加、模型复杂度上升、以及对小样本场景下稳定性的冲击。未来工作可以重点探索基于稀疏数据友好的多模态融合策略,以更全面地挖掘多模态组学之间的协同关系。④尽管 GSCC 在多数癌症数据分析中表现优异,但在数据量极少的罕见癌症类型(如某些肉瘤或儿童肿瘤)中,模型性能可能出现波动,未来可结合小样本学习策略[31] 进一步对模型进行优化。

综上所述,GSCC 为多组学组学数据驱动的癌症亚型识别提供了一种新颖的结构感知与融合对比建模框架,在提升聚类稳定性、生物学解释性以及模型鲁棒性方面表现出显著优势,展现出良好的应用潜力与拓展空间。针对不同癌症亚型采取差异化治疗策略,有望为精准医疗提供参考。尽管本研究聚焦于癌症亚型的无监督识别,GSCC 的建模思想与模块化设计也具备良好的通用性,未来有望拓展应用于其他类型疾病的亚型挖掘,进一步提升对复杂疾病异质性的理解和干预能力。

参考文献:

[1] Cao W, Qin K, Li F, et al. Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (GLOBOCAN)[J]. J Natl Cancer Cent, 2024, 4(2): 128-134.

[2] Duan R, Gao L, Gao Y, et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping[J]. PLoS Comput Biol, 2021, 17(8): e1009224.doi: 10.1371/journal.pcbi.1009224

[3] Ellrott K, Wong CK, Yau C, et al. Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets[J]. Cancer Cell, 2025, 43(2): 195-212.

[4] 司呈坤. 面向组学数据的癌症亚型分类及特征选择技术研究[D]. 济南: 齐鲁工业大学, 2024.

[5] Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology[J]. Cancer Cell, 2022, 40(10): 1095-1110.

[6] Wang YX, Zhang YJ. Nonnegative matrix factorization: a comprehensive review[J]. IEEE Trans Knowl Data Eng, 2012, 25(6): 1336-1353.

[7] Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review[J]. Front Genet, 2022, 13: 854752. doi: 10. 3389/fgene. 2022. 854752

[8] 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[J]. Bioinformatics, 2009, 25(22): 2906-2912.

[9] Lim KL, Jiang X, Yi C. Deep clustering with variational autoencoder[J]. IEEE Signal Process Lett, 2020, 27: 231-235. doi: 10.1109/LSP.2020.2965328

[10] Rong Z, Liu Z, Song J, et al. MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data[J]. ComputBiol Med, 2022, 150: 106085. doi: 10. 1016/j. compbiomed.2022.106085

[11] Zhou T, Li Q, Lu H, et al. GAN review: models and medical image fusion applications[J]. Inf Fusion, 2023, 91: 134-148. doi:10.1016/j.inffus.2022.10.017

[12] Ganini C, Amelio I, Bertolo R, et al. Global mapping of cancers: The Cancer Genome Atlas and beyond[J]. Mol Onco, 2021, 15(11): 2823-2840.

[13] Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers[J]. Genome biol, 2011, 12: 1-14. doi: 10.1186/gb-2011-12-4-r41

[14] 李阳. 基于自注意力机制和多组学数据整合的癌症亚型识别与分类研究[D]. 重庆: 中国人民解放军陆军军医大学, 2024.

[15] 宁斌. 基于深度学习的多组学癌症亚型识别方法研究[D]. 长沙: 湖南大学, 2023.

[16] Veena EV, Pushpalatha KP. Enhanced KNN imputation for missing data[C]//International Conference on Information Technology and Applications. Singapore: Springer Nature Singapore, 2024: 583-592.

[17] Ponzi E, Thoresen M, Haugdahl Nøst T, et al. Integrative, multi-omics, analysis of blood samples improves model predictions: applications to cancer[J]. BMC bioinformatics, 2021, 22: 1-17. doi: 10.1186/s12859-021-04296-0

[18] Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction[J]. Journal of Soft Computing and Data Mining, 2021, 2(1): 20-30.

[19] Zhao S, Zhang B, Yang J, et al. Linear discriminant analysis[J]. Nature Reviews Methods Primers, 2024, 4(1): 70. doi: 10.1038/s43586-024-00346-y

[20] Steck H, Ekanadham C, Kallus N. Is cosine-similarity of embeddings really about similarity? [EB/OL]. (2024-03-08)[2025-04-26]. http://arxiv.org/abs/2403.05440

[21] Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification[J]. Nat commun, 2021, 12(1): 3445. doi: 10.1038/s41467-021-23774-w

[22] Wang X, Qi GJ. Contrastive learning with stronger augmentations[J]. IEEE Trans Anal Mach Intell, 2022, 45(5): 5549-5560.

[23] Zhao J, Zhao B, Song X, et al. Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data[J]. Brief Bioinform, 2023, 24(2): bbad025. doi: 10.1093/bib/bbad025

[24] Li Y, Hu P, Liu Z, et al. Contrastive clustering[EB/OL]. (2020-09-21)[2025-04-26]. http://arxiv.org/abs/2009.09687

[25] Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nat Methods, 2014, 11(3): 333-337.

[26] Ikotun AM, Ezugwu AE, Abualigah L, et al. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data[J]. Inf Sci, 2023, 622: 178-210. doi: 10.1016/j.ins.2022.11.139

[27] Yang H, Chen R, Li D, et al. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data[J]. Bioinformatics, 2021, 37(16): 2231-2237.

[28] Madhumita, Dwivedi A, Paul S. Recursive integration of synergised graph representations of multi-omics data for cancer subtypes identification[J]. Sci Rep, 2022, 12(1): 15629. doi: 10.1038/s41598-022-17585-2

[29] International Cancer Genome Consortium. International network of cancer genome projects[J]. Nature, 2010, 464(7291): 993-998.

[30] Li Y, Dou Y, Leprevost FDV, et al. Proteogenomic data and resources for pan-cancer analysis[J]. Cancer Cell, 2023, 41(8): 1397-1406.

[31] Li A, Huang W, Lan X, et al. Boosting few-shot learning with adaptive margin loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE. 2020: 12576-12584. doi: 10.1109/CVPR42600.2020.01259

(编辑:相峰)