DOI:10.11784/tdxbz202508011

基于图对比学习的单细胞多组学癌症亚型识别方法

王永旺，梁　潇，刘　威，骆嘉伟
(湖南大学计算机学院，长沙 410012)

摘　要：随着单细胞测序技术的快速发展，单细胞多组学数据在癌症异质性研究中展现出巨大潜力. 癌症亚型识别作为癌症异质性研究中的关键任务，对于揭示肿瘤的发展具有重要意义. 传统的癌症亚型识别方法多依赖于 bulk 测序数据，虽取得一定成效，但由于其仅反映细胞群体的平均表达水平，难以揭示肿瘤微环境中细胞间的多样性与异质性. 此外，现有多组学融合方法往往未能充分兼顾组学内部的特征关联与组学间的协同整合，导致融合后的特征未能充分捕获组学间的复杂关系. 针对上述问题，提出了一种基于图对比学习的单细胞多组学癌症亚型识别方法(scMOGCL). 该方法为不同组学分别构建了模态特定的图注意力网络，自适应地捕捉细胞之间的相互作用关系并提取潜在表示. 随后，通过引入自监督对比学习策略，在每种组学中构造原始样本与扰动样本之间的对比机制，提升模型在无监督情境下的表征能力与鲁棒性. 同时，引入跨模态一致性约束机制，实现各组学之间的协同训练，从而获取更具一致性的多组学综合表示. 该方法在 3 个真实的单细胞多组学癌症数据集上进行了系统实验，结果表明，相比于其他 6 种基准测试方法，scMOGCL 展现出具有竞争力的性能，验证了其在癌症亚型识别与单细胞多组学整合任务中的有效性与应用前景.

关键词：癌症亚型识别；多组学整合；图注意力网络；图对比学习；一致性约束
中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：0493-2137(2026)11-1208-11

Graph Contrastive Learning-Based Approach for Cancer Subtype Identification Using Single-Cell Multi-Omics Data

Wang Yongwang，Liang Xiao，Liu Wei，Luo Jiawei
(College of Computer Science and Electronic Engineering，Hunan University，Changsha 410012，China)

Abstract：With the rapid technological development of single-cell sequencing，such multi-omics data have shown great potential in cancer heterogeneity research and tumor pathogenesis. A key challenge in such research is subtype identification，the traditional methods for which mostly rely on bulk sequencing data. Although these methods have achieved a certain degree of success，they reflect only average expression intensities at the cell population level but not the diversity or heterogeneity of individual cells within the tumor microenvironment. In addition，integrating existing multi-omics often fails to fully consider the feature correlations within each omics and the collaborative integration across different omics，resulting in a combination of features that are ineffective in capturing complex relationships across omics. Considering these issues，we propose scMOGCL，a graph contrastive learning-based method for cancer subtype identification using single-cell multi-omics data. This method constructs modality-specific graph attention networks for various omics，which can adaptively capture cell-cell interactions and extract latent representations. A self-supervised，contrastive learning strategy is subsequently introduced to construct contrastive pairs of original and perturbed samples within each omics，thereby enhancing the representative ability and robustness of the model under unsupervised conditions. Meanwhile，a mechanism constraining cross-modal consistency is also incorporated

收稿日期：2025-08-13；**修回日期**：2025-09-29.
作者简介：王永旺(1999—　　)，男，硕士研究生，ywwang@hnu.edu.cn.
通信作者：骆嘉伟，luojiawei@hnu.edu.cn.
基金项目：国家自然科学基金资助项目(62372165)；国家自然科学基金重点资助项目(62032007).
Supported by the National Natural Science Foundation of China(No. 62372165)，the Key Program of the National Natural Science Foundation of China(No. 62032007).

to coordinate training across multi-omics and thus represent them more consistently and comprehensively. The proposed method is systematically evaluated for applicability by employing three real-world datasets. Compared with six benchmark methods, scMOGCL exhibited competitive performance, validating its effectiveness and application potential in cancer subtype identification utilizing single-cell multi-omics data.

Keywords: cancer subtype identification; multi-omics integration; graph attention network; graph contrastive learning; consistency constraint

癌症亚型是指在同一种癌症类型中，根据恶性肿瘤所表现出的分子特征、组织学特点、基因表达以及临床表现等因素，将肿瘤进一步细分为具有不同生物学特性和临床意义的亚型. 每个亚型的肿瘤在分子层面具有相似的特征并形成具有共同特征的肿瘤簇，但不同亚型之间又存在显著差异[1]. 不同癌症亚型的预后和治疗方案差异很大，因此确定癌症亚型对于癌症患者的预后和治疗至关重要.

近年来，单细胞测序技术得到了快速发展，推动癌症研究进入了高分辨率时代. 与传统的组织样本“bulk 测序”方法相比，单细胞测序能够对成千上万个细胞的基因表达谱进行逐个分析，从而揭示肿瘤组织内部的高度异质性[2]，而 bulk 测序所得的分子特征仅反映细胞群体的平均表达水平，难以解析肿瘤微环境中的多样性和异质性. 借助深度学习算法，单细胞技术能够对大规模、噪声复杂的数据进行有效建模，为准确识别癌症亚型和探索肿瘤微环境的复杂组成提供了有效手段.

然而，单细胞测序数据常常伴随着高维度、高稀疏性等特点[3]，给单细胞数据的分析带来诸多挑战. SC3[4]、scDeepcluster[5] 以及 scGNN[6]等单细胞聚类方法的出现，代表着单细胞测序技术在深度学习领域的潜力逐渐被挖掘出来. 研究表明，单细胞测序技术能够以单细胞分辨率解析肿瘤组织的异质性，识别不同细胞类型及其转录特征，揭示肿瘤发生、进展及耐药机制背后的细胞间相互作用. 在此基础上，单细胞多组学测序技术的发展，为获取同一细胞不同组学层次的信息提供了机会. 单细胞多组学整合能从多个角度描绘单个细胞的功能状态，有助于深入理解组织微环境中的复杂生物机制. 为有效融合单细胞多模态数据，多种单细胞多组学整合方法被研究者们相继提出. 例如，scMDC[7]、GCN-sc[8]以及 scButterfly[9]等方法，这些方法引入了图神经网络、自编码器、对比学习等深度学习技术，能够对 CITE-seq、scRNA-seq 和 scATAC-seq 等多种单细胞测序技术得到的数据进行融合. 这些方法针对单细胞多组学数据提出了具有创新性的特征提取策略，并在数据整合以及下游分析任务中展现出显著成效. 然而，尽管已有众多基于 bulk 多组学数据的癌症亚型识别方法以及单细胞多组学整合方法，但专门针对 bulk 数据与单细胞数据的联合分析，或完全基于单细胞数据进行癌症亚型识别的研究仍处于探索阶段，这也导致了不同组学的单细胞数据没能得到充分的利用.

为应对上述挑战，近年涌现出了多种应用于单细胞数据的癌症亚型识别方法，如 OmicsGAT[10]、AAFL[11]以及 scCluster[12]等方法. OmicsGAT[10]是一种基于图注意力网络的模型，融合了图学习和注意力机制来深入分析癌症样本的 bulk 与单细胞数据. 该模型通过为每个样本的邻居节点动态分配不同的注意力系数，精准捕捉样本间的复杂关系. 该方法在乳腺癌和膀胱癌等 bulk 数据集上展现了卓越的性能，并通过实验证明了其在单细胞数据分析中的适用性，成功应用于 H3K27M 突变型胶质瘤单细胞数据集和 SARS-CoV-2 抗体单细胞数据集的分析中. AAFL[11]是一种基因选择方法. AAFL 在基因选择前先获取差异表达基因，过滤冗余基因，同时，自动识别不同细胞亚型（癌症亚型）的不同基因特征，该方法将残差网络与低秩网络相结合，有效筛选与特定癌症亚型最相关的基因来提升识别精度，并应用于真实的单细胞癌症数据集，以识别癌症亚型并检测识别出的癌症亚型的基因特征. scCluster[12]则是一种单细胞多组学癌症亚型识别方法. 该方法采用基于 ZINB 分布的双模态自编码器结构进行预训练，并在微调阶段引入深度软 K 均值聚类与跨实例对比聚类损失，从而提升其对潜在表示的判别能力.

研究人员尝试将癌症亚型识别提升至单细胞分辨率，并在临床转化中展现出初步潜力. 例如，研究人员使用单细胞调控网络、聚类分析以及单细胞基因富集分析，探究癌症内部成纤维细胞每个亚型的功能特征，并揭示不同亚型与患者临床结局之间的关联性. 上述方法在单细胞癌症亚型识别领域提出了新的见解，但仍然存在以下局限性：①提取潜在信息时未能自适应地学习细胞间的相互作用关系；②现有多组学融合方法未能充分兼顾组学内部与组学之间两个层面的整合.

针对当前研究中存在的局限性，本文提出一种基

于图对比学习的单细胞多组学方法(contrastive learning-based approach using single-cell multi-omics data, scMOGCL). 该方法采用动态权重的方式提取各组学数据中的潜在表示,充分挖掘不同组学数据中的深层信息. 此外,借鉴对比学习的思想,scMOGCL 在每一组学内构建原始样本与扰动后的负样本之间的对比机制,以增强模型的特征学习能力. 最后,利用一致性约束与平均池化操作融合不同组学数据的潜在信息,从而保证亚型识别的准确性与鲁棒性. 本文在 3 个单细胞多组学癌症数据集上对 scMOGCL 方法进行了全面评估,所使用的组学类型包括基因表达(RNA)与表达相关的单核苷酸多态性(eSNP). 实验内容涵盖模型的性能评估、参数分析、消融实验以及临床相关性分析. scMOGCL 与 6 种基准方法进行了对比,结果显示,scMOGCL 在各指标上均表现出色,验证了其在单细胞多组学癌症亚型识别任务中的有效性与可靠性.

1 scMOGCL 方法框架与模型构建

1.1 方法总体框架

scMOGCL 的整体框架如图 1 所示,该方法主要包含 6 个模块:特征扰动模块、图注意力网络(graph attention networks, GAT)模块、图对比学习模块、一致性约束模块、解码模块以及潜在表示融合模块. 首先,针对每个经过预处理的组学数据 $\boldsymbol{X}^l$,模型会构建其对应的邻接图 $\boldsymbol{A}^l$,同时,对特征矩阵 $\boldsymbol{X}^l$ 进行样本间扰动,得到损坏后的特征矩阵 $\boldsymbol{X}^l_{\text{corrupted}}$,以此得到图注意力网络模块的输入,即 $\boldsymbol{G}^l_{\text{original}}=\{\boldsymbol{X}^l,\boldsymbol{A}^l\}$,$\boldsymbol{G}^l_{\text{corrupted}}=\{\boldsymbol{X}^l_{\text{corrupted}},\boldsymbol{A}^l\}$. 其次,在图注意力网络模块中,本文为每个组学都配备一个独立的 GAT. 对于第 l 个组学来说,将 $\boldsymbol{G}^l_{\text{original}}$ 与 $\boldsymbol{G}^l_{\text{corrupted}}$ 一同输入到该组学对应的 GAT 中,分别得到未扰动特征矩阵的潜在表示 $\boldsymbol{Z}^l_{\text{original}}$ 与扰动后特征矩阵的潜在表示 $\boldsymbol{Z}^l_{\text{corrupted}}$. 随后,在图对比学习模块中,$\boldsymbol{Z}^l_{\text{original}}$ 中的样本经过 Readout 操作后生成正样本对,$\boldsymbol{Z}^l_{\text{corrupted}}$ 中的样本经过 Readout 操作后生成负样本对,并以此得到优化后的组学 l 的潜在表示 $\boldsymbol{Z}^l_{\text{refined}}$. $\boldsymbol{Z}^l_{\text{refined}}$ 将用于解码模块的重构训练,不同组学间的 $\boldsymbol{Z}^l_{\text{refined}}$ 将通过平均池化操作进行融合,最终生成用于下游任务的融合潜在表示 $\boldsymbol{Z}$.

1.2 数据预处理

在单细胞多组学数据分析中,不同组学数据所携带的生物学信息有着高度异质性,例如 RNA 表达数据通常反映转录水平的动态变化,而 eSNP 数据则反映了与基因表达相关的遗传调控变异. 因此,为了充分保留各组学内部的关键特征与组学间的差异性表达,本文对不同组学数据采取了差异化的预处理策略.

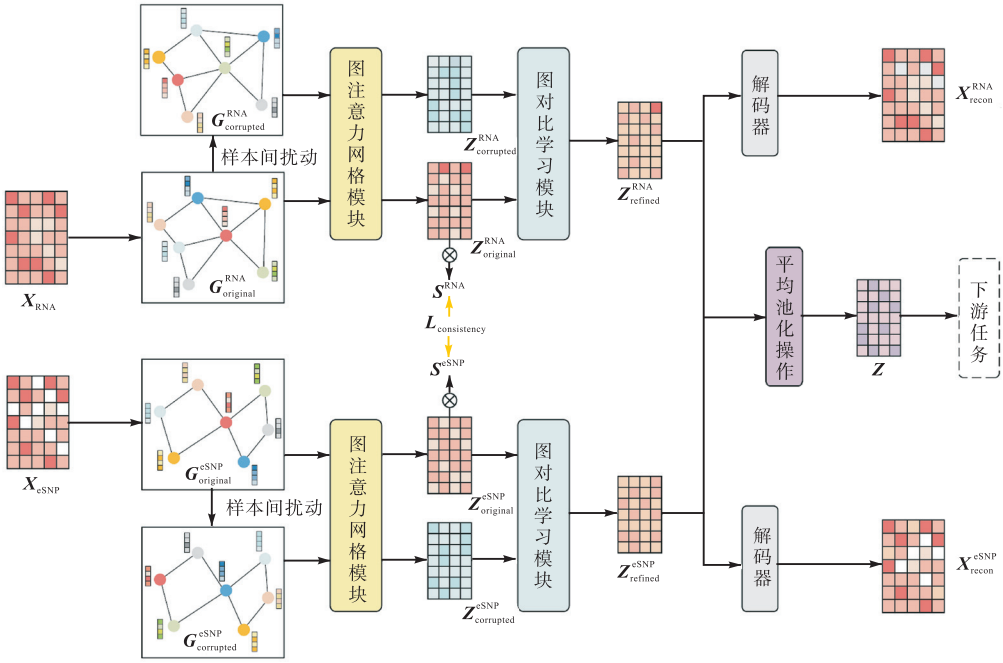


图 1 scMOGCL 架构

Fig.1 Architecture of scMOGCL

本文使用 Wang 等[12]初步处理后的数据集进行实验. 为了消除测序深度和技术噪声带来的影响,本文进一步对这些单细胞癌症数据进行标准化处理.

对于单细胞 RNA 测序数据,首先过滤低表达的

基因，仅保留至少在 10 个细胞中出现非零表达的基因，并对其进行自然对数变换，这种基因过滤的预处理方式在单细胞分析中已被广泛采用，例如 scSemiGAN 也采用了相同的策略[13]. 随后，对于 Chung 等基因表达矩阵维度大于 20 000 的数据集使用 SCANPY[14]中的高可变基因筛选方法，选取变异性最高的前 3 000 个基因作为特征，对其他如 Patel、Gerber 等维度不高的数据集则保留所有基因表达值. 最后，对得到的基因表达矩阵执行标准化处理，以消除基因表达量尺度的差异.

对单细胞 eSNP 数据，则采取了稳健的处理策略. 首先，移除在样本中出现频率极低的 SNP 位点，仅保留至少在 5% 细胞中出现的突变位点数据，以减少稀疏性干扰. 随后，应用分位数归一化方法将每个位点的分布统一至近似均匀分布，有效缓解原始数据中可能存在的极端值或非高斯分布问题，提升多模态数据融合时的稳定性与表达一致性.

1.3　scMOGCL 模型结构

1.3.1　特征扰动模块

以第 l 个组学数据为例. 首先，对于输入的特征矩阵 $\boldsymbol{X}^l$，通过余弦相似性度量样本之间的相似度，且仅保留每个样本的前 k 个最相似的邻居节点，参数 k 控制图的稀疏性与信息聚合范围，当 k 较小时，模型更关注局部结构，能避免噪声干扰；当 k 较大时，模型可整合更多全局信息，但可能引入冗余连接. 第 2.4 节实验中对参数 k 的范围进行了参数分析，基于此构建得到邻接图 $\boldsymbol{A}^l$. $\boldsymbol{A}^l$ 将作为图注意力网络模块的输入，用于捕捉细胞之间的关系，自适应地提取特征矩阵中的潜在信息，即

$$S_{ij}^l = \frac{\boldsymbol{X}_i^l \cdot \boldsymbol{X}_j^l}{\left\|\boldsymbol{X}_i^l\right\|\left\|\boldsymbol{X}_j^l\right\|} \tag{1}$$

式中 $\boldsymbol{X}_i^l \cdot \boldsymbol{X}_j^l$ 表示第 i 个样本与第 j 个样本特征向量的内积，用于衡量它们在特征空间中的相似程度，分母用于归一化，确保相似度 S_{ij}^l 的取值范围在[−1，1]之间.

$$A_{ij}^l = \begin{cases} 1 & j \in N_i \\ 0 & \text{其他} \end{cases} \tag{2}$$

式中：A_{ij}^l 的取值依赖于 k，即每个节点仅与其相似度最高的前 k 个邻居建立连接；N_i 为样本 i 的相似度最高的前 k 个邻居.

与此同时，为了生成对比学习中的负样本，本研究对 $\boldsymbol{X}^l$ 进行样本间扰动. 具体而言，绑定 $\boldsymbol{X}^l$ 中每 5 个样本为一个块，并对这些块执行随机 Shuffle 操作，目的是在生成损坏的特征矩阵 $\boldsymbol{X}_{\text{corrupted}}^l$ 的同时，保留局部的样本间关系. 通过该扰动方式，可在对比学习中提供区分性的负样本，从而增强图对比学习模块进一步优化潜在表示的能力.

1.3.2　图注意力网络模块

在图注意力网络模块中，每个组学的输入数据包括组学特征矩阵 $\boldsymbol{X}^l$、损坏的特征矩阵 $\boldsymbol{X}_{\text{corrupted}}^l$ 与邻居图 $\boldsymbol{A}^l$，即 $\boldsymbol{G}_{\text{original}}^l = \left\{\boldsymbol{X}^l, \boldsymbol{A}^l\right\}$，$\boldsymbol{G}_{\text{corrupted}}^l = \left\{\boldsymbol{X}_{\text{corrupted}}^l, \boldsymbol{A}^l\right\}$. 随后，将 $\boldsymbol{G}_{\text{original}}^l$ 与 $\boldsymbol{G}_{\text{corrupted}}^l$ 输入到同一个 GAT 中，以提取对应的潜在表示. 具体公式为

$$\begin{cases} \boldsymbol{Z}_{\text{original}}^l = \text{GAT}^l\left(\boldsymbol{G}_{\text{original}}^l\right) \\ \boldsymbol{Z}_{\text{corrupted}}^l = \text{GAT}^l\left(\boldsymbol{G}_{\text{corrupted}}^l\right) \end{cases} \tag{3}$$

式中：$\boldsymbol{Z}_{\text{original}}^l$ 为原始特征矩阵经过 GAT 后得到潜在表示；$\boldsymbol{Z}_{\text{corrupted}}^l$ 为损坏的特征矩阵经过 GAT 后得到潜在表示；$\boldsymbol{Z}_{\text{original}}^l, \boldsymbol{Z}_{\text{corrupted}}^l \in \mathbf{R}^{N_{\text{cell}} \times d}$，$N_{\text{cell}}$ 为细胞数量，d 为 GAT 的潜在表示维度.

1.3.3　图对比学习模块

受深度图互信息 DGI[15]的启发，本模块引入了自监督对比学习（SCL）的思想，捕捉局部样本的上下文关系，使潜在表示 $\boldsymbol{Z}_{\text{original}}^l$ 包含更具区分性的信息，其整体流程如图 2 所示.

具体而言，使用 Readout 函数将 $\boldsymbol{Z}_{\text{original}}^l$ 中细胞 i 的邻居表示聚合为细胞 i 的局部上下文 g_i，作为对比学习中的锚点，表示细胞的邻域微环境. 随后，将 $\boldsymbol{Z}_{\text{original}}^l$ 中细胞 i 的潜在表示 z_i 与 g_i 构建为正样本对，而损坏的图对应表示 $\boldsymbol{Z}_{\text{corrupted}}^l$ 中提取细胞 i 的潜在表示 z_i' 与 g_i 构建为负样本对. 最后，最大化正样本对之间的互信息，最小化负样本对之间的互信息来进一步优化细胞 i 潜在表示.

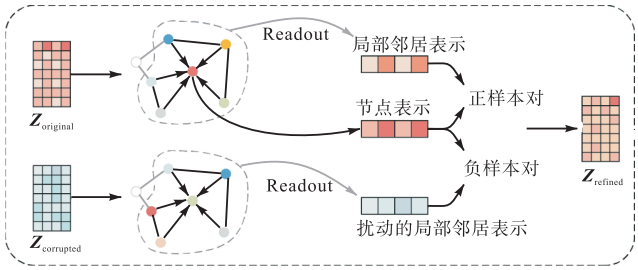


图 2　图对比学习模块

Fig.2　Graph contrastive learning module

通过使用对比学习策略，相邻的细胞中将具有更相似的表达，而非相邻细胞将具有更差异化的表达. 接下来，引入二元交叉熵建模图对比学习模块，以得到每个组学经过图对比学习模块优化后的潜在表示 $\boldsymbol{Z}_{\text{refined}}^l$. 该设计核心在于通过互信息最大化机制

提升潜在表征的判别性与稳定性,使模型在无监督条件下自适应地优化潜在空间的结构分布,即

$$\mathcal{L}_{\mathrm{SCL}}=-\frac{1}{2N_{\mathrm{cell}}}\left(\sum_{i=1}^{N_{\mathrm{cell}}}\left(E_{(\boldsymbol{x}^l,\boldsymbol{G}^l)}\left[\lg\boldsymbol{\Phi}(z_i,g_i)\right]+E_{(\boldsymbol{x}^l_{\mathrm{corrupted}},\boldsymbol{G}^l)}\left[\lg\left(1-\boldsymbol{\Phi}(z_i',g_i)\right)\right]\right)\right) \tag{4}$$

式中 $\boldsymbol{\Phi}(\cdot)$ 是一个用于区分正样本对和负样本对的二元判别神经网络,对于给定的细胞潜在表示 z_i 及其邻居的平均表达 g_i,$\boldsymbol{\Phi}(z_i,g_i)$ 计算得到分配给正样本对的概率分数,$1-\boldsymbol{\Phi}(z_i',g_i)$ 则是 z_i' 和 g_i 不相似(即负样本对)的概率分数. 通过最小化该二元交叉熵损失,模型能够最大化原始样本与其局部上下文之间的一致性,同时抑制扰动样本与真实邻域的相关性,这样设计的目的在于使模型学习到区分性更强、结构更稳定的潜在空间,从而提高模型的判别能力.

1.3.4 一致性约束模块

为了在多组学数据中实现不同模态间的协同学习与潜在空间的一致表达,本文设计了一致性约束模块,用于对齐各组学的潜在表示分布. 此处以第 l 与第 $l+1$ 个组学为例,这两个组学经过图对比学习模块以后分别得到对应的潜在表示 $\boldsymbol{Z}^l_{\mathrm{original}}$ 与 $\boldsymbol{Z}^{l+1}_{\mathrm{original}}$,对这两个嵌入分别执行与自身转置相乘操作,得到各自的相似性矩阵,即

$$\begin{cases}\boldsymbol{S}^l=\sigma\left(\mathrm{Norm}\left(\boldsymbol{Z}^l_{\mathrm{original}}\right)\cdot\mathrm{Norm}\left(\boldsymbol{Z}^l_{\mathrm{original}}\right)^{\mathrm{T}}\right)\\ \boldsymbol{S}^{l+1}=\sigma\left(\mathrm{Norm}\left(\boldsymbol{Z}^{l+1}_{\mathrm{original}}\right)\cdot\mathrm{Norm}\left(\boldsymbol{Z}^{l+1}_{\mathrm{original}}\right)^{\mathrm{T}}\right)\end{cases} \tag{5}$$

式中:$\mathrm{Norm}(\boldsymbol{Z})=\dfrac{\boldsymbol{Z}-\bar{\boldsymbol{Z}}}{\|\boldsymbol{Z}-\bar{\boldsymbol{Z}}\|_2}$,表示中心化后的 L_2 归一化;$\sigma(\cdot)$ 是 Sigmoid 函数. 同样,使用二元交叉熵损失强制约束 $\boldsymbol{S}^l$ 与 $\boldsymbol{S}^{l+1}$ 的分布一致,使单细胞组学获取潜在表示时考虑组学间结构,使得各组学之间进行协同训练. 具体公式为

$$\mathcal{L}^l_{\mathrm{consistency}}=-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\boldsymbol{S}^l_{ij}\lg\boldsymbol{S}^{l+1}_{ij}+\left(1-\boldsymbol{S}^l_{ij}\right)\lg\left(1-\boldsymbol{S}^{l+1}_{ij}\right)\right] \tag{6}$$

1.3.5 解码模块

经过图对比学习模块,可获得各组学优化后的潜在表示,记作第 l 个组学的潜在表示为 $\boldsymbol{Z}^l_{\mathrm{refined}}$. 在解码模块中,scMOGCL 为每个组学配备了含两层隐藏层的多层感知机(MLP)对各组学的 $\boldsymbol{Z}^l_{\mathrm{refined}}$ 进行解码重构,以生成原始输入矩阵 $\boldsymbol{X}^l$ 相似的重构矩阵,这一过程的计算公式为

$$\boldsymbol{X}^l_{\mathrm{recon}}=\mathrm{MLP}_1(\boldsymbol{Z}^l)\qquad l\in1,2,\cdots,V \tag{7}$$

本模块使用 MSE 损失作为重构损失函数,即

$$\mathcal{L}^l_{\mathrm{recon}}=\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{X}^l[i]-\boldsymbol{X}^l_{\mathrm{recon}}[i]^2\right) \tag{8}$$

式中 $\boldsymbol{X}^l[i]$ 和 $\boldsymbol{X}^l_{\mathrm{recon}}[i]$ 分别是原始输入矩阵和重构矩阵中第 i 个细胞的特征向量.

1.3.6 潜在表示融合模块

在经过上述模块后,可获得各组学优化后的潜在表示,记第 l 个组学的潜在表示为 $\boldsymbol{Z}^l_{\mathrm{refined}}$. 为了整合不同组学的潜在表示,scMOGCL 引入了平均池化层对不同组学间潜在表示进行融合,生成最终用于下游任务的共享潜在表示 $\boldsymbol{Z}$,该过程为

$$\boldsymbol{Z}=\frac{1}{V}\sum_{l=1}^{V}\boldsymbol{Z}^l_{\mathrm{refined}} \tag{9}$$

1.3.7 模型训练策略

scMOGCL 模型的损失主要由 3 种损失构成,即重构损失 $\mathcal{L}_{\mathrm{recon}}$、自监督对比损失 $\mathcal{L}_{\mathrm{SCL}}$ 以及一致性约束损失 $\mathcal{L}_{\mathrm{consistency}}$. 综合上述 3 种类型的损失,scMOGCL 的总体损失函数可定义为 $\mathcal{L}_{\mathrm{total}}$,即

$$\mathcal{L}_{\mathrm{total}}=\sum_{l=1}^{V}\left(\lambda^l_{\mathrm{recon}}\mathcal{L}^l_{\mathrm{recon}}+\lambda_l\mathcal{L}^l_{\mathrm{SCL}}\right)+\sum_{k=1}^{(V-1)!}\lambda_k\mathcal{L}^k_{\mathrm{consistency}} \tag{10}$$

式中:V 为组学数量;$\lambda^l_{\mathrm{recon}}$ 为第 l 个组学中重构损失的权重;λ_l 代表第 l 个组学中图对比学习模块的损失权重;λ_k 代表每两种组学之间的一致性约束损失所占的权重. 本文使用两种组学数据,因此涉及的损失权重包括 $\lambda_{\mathrm{recon,\ RNA}}$、$\lambda_{\mathrm{recon,\ eSNP}}$、$\lambda_{\mathrm{RNA}}$、$\lambda_{\mathrm{eSNP}}$ 和 λ_k $(k=1)$,它们的最优值是通过网格搜索确定的. 在 Patel 数据集中,它们的最优值分别为 5、5、0.1、0.1 和 1,学习率为 0.001. 在 Gerber 数据集中,它们的最优值分别为 10、10、1.0、0.1 和 5,学习率为 0.000 1. 在 Chung 数据集中,它们的最优值分别为 1、1、0.1、5.0 和 10,学习率为 0.000 1. scMOGCL 将采用上述最优权重与学习率的设置,并使用权重衰减为默认值的 AdamW 优化器进行迭代优化.

2 实验设计与结果分析

2.1 数据集简介

本文使用了不同类型癌症的 3 个真实的单细胞多组学癌症数据集,分别来自胶质母细胞瘤患者、黑色素瘤患者以及乳腺癌患者. 本文使用的数据集已由 Wang 等[12]进行处理,这 3 个癌症数据集组学数据类型有两种,分别是 RNA 数据与 eSNP 数据,数据集及其特征的详细信息如表 1 所示.

数据集从 NCBI 基因表达综合数据库门户获

取. 具体来说，Patel 数据集可通过 GEO 登录号 GSE57872 获取，其来源于 5 名胶质母细胞瘤患者，共 7 种不同的细胞类型. Gerber 数据集可通过 GEO 登录号 GSE81383 获取，其来源于 3 名黑色素瘤患者，共 4 种不同的细胞类型. Chung 数据集可通过 GEO 登录号 GSE75688 获取，其来自 11 名被诊断为乳腺癌的患者，共包含 13 种细胞类型.

表 1　单细胞癌症数据集详细信息

Tab.1　Detailed information of scRNA datasets

数据集	癌症类型	细胞数量	基因数量	eSNP	细胞类型
Patel	Glioblastoma tumor	543	5 948	5 226	7
Gerber	Melanoma	307	19 019	30 081	4
Chung	Breast cancer	549	57 915	34 800	13

2.2　实验设置

（1）参数设置：在模型训练阶段，scMOGCL 的潜在表示维度被设定为 128，模型优化采用 AdamW 优化器，最大训练轮数为 600 轮.

（2）基线方法：本文选用了 6 种具有代表性的方法作为 scMOGCL 的对比方法，分别是 *K*-Means[16]、SNF[17] 、MOFA+[18] 、scMDC[7] 、scMIC[19] 以及 scCluster[12]. 对比方法的参数设置遵循作者所推荐的参数或默认参数.

（3）评价指标：使用 I_{ARI}、I_{NMI} 和 I_{AMI} 分别表示调整兰德指数[20]（adjusted rand index，ARI）、归一化互信息[21]（normalized mutual information，NMI）以及调整互信息[22]（adjusted mutual information，AMI）的取值. 具体来说，I_{ARI} 用于度量识别结果与真实标签之间的相似度，I_{NMI} 用于度量识别结果与真实标签之间的归一化互信息，I_{AMI} 在互信息的基础上对随机一致性进行了校正，用于评估识别结果与真实标签之间的一致程度，从而更客观地反映识别结果的准确性.

本文实验均基于 Python 3.8 语言进行编写，并在 Pytorch 2.1.0 深度学习框架中进行开发与运行. 所有实验在 NVIDIA GeForce RTX 4090 GPU 上完成，其显存容量为 24 GB.

2.3　癌症亚型识别性能评估

为了全面评估 scMOGCL 方法的识别性能，将其与基准方法在胶质母细胞瘤（Patel）、黑色素瘤（Gerber）以及乳腺癌（Chung）数据集上进行对比实验，并使用 I_{ARI}、I_{NMI} 和 I_{AMI} 对各方法的性能进行评价. 评价指标值越高，表明方法在癌症数据集上的识别性能越好. scMOGCL 与其他方法的性能对比如图 3 所示.

从图 3 中可以看出，在 Patel 与 Gerber 数据集上，scMOGCL 在 I_{ARI}、I_{NMI} 与 I_{AMI} 3 个指标上均取得了最佳性能. 在 Patel 数据集中，scMOGCL 的 I_{ARI}、I_{NMI}、I_{AMI} 分别达到了 0.982 1、0.980 3、0.979 9. 在 Gerber 数据集中，scMOGCL 的 I_{ARI}、I_{NMI} 和 I_{AMI} 分别达到了 0.980 6、0.977 2 和 0.977 0. 在 Chung 数据集中，scMOGCL 的 I_{ARI}、I_{NMI} 和 I_{AMI} 分别达到了 0.620 3、0.723 2 和 0.705 9. 在 Patel 数据集中，相较于传统 *K*-Means 方法，scMOGCL 在 I_{ARI}、I_{NMI} 和 I_{AMI} 指标上分别提升约 34%、28% 和 29%；与次优方法 scMDC 相比，scMOGCL 仍保持约 2.0% ~ 2.5% 的性能优势.

值得注意的是，scMDC、scMIC 与 scCluster 均采用自编码器或图自编码器的结构，而 scMOGCL 采用图注意力网络进行特征提取，这表明 GAT 能够有效地建模细胞间的相互作用关系，自适应地学习细胞间的权重系数，从而为识别癌症亚型提供更符合生物学性质的信息.

此外，scMOGCL 还引入了图对比学习和一致性约束策略. 图对比学习能够增强模型对细胞特征的判别能力，一致性约束确保了在多组学数据融合过程中不同组学的识别结果能够保持一致，这两个模块在其他基准测试方法中均未涉及，体现出了所设计的模块能够生成更可靠的识别结果.

2.4　参数分析

为了进一步评估 scMOGCL 方法在不同的参数设置情况下，是否能够保持其模型性能的稳定，本节在 Patel 数据集上进行参数分析. 因 scMOGCL 方法应用了图注意力网络模块，本节对其中图结构构建时的邻居数 k 与 GAT 中的注意力头数进行实验.

首先对参数 k 进行分析. 在 scMOGCL 的图注意力网络模块中，本节使用余弦相似性计算细胞间的相似性矩阵，并选择每个细胞的前 k 个相似性最高的细胞作为其邻居，在相应的邻接图中，设置邻居的边值为 1，其余设为 0. 不同的 k 值将直接影响所构建图结构的稀疏程度，从而影响 GAT 捕获细胞间关系的能力，同时也会影响模型的计算效率. 因此，本节在 $k\in[3,15]$ 的范围内进行实验.

根据图 4 所示的实验结果，在 [3,15] 这个范围内，scMOGCL 方法的 I_{ARI}、I_{NMI}、I_{AMI} 有较小的波动，且 I_{ARI} 始终能保持 0.940 0 以上，I_{NMI} 与 I_{AMI} 能保持 0.920 0 以上. 这表明 scMOGCL 使用不同的 k 值进行构图时，仍然能够保持优异的性能，也说明 scMOGCL 方法对参数 k 值并不敏感. 值得注意的是，当 $k=10$ 时，scMOGCL 的 3 个性能指标在 $k\in[3,15]$ 的范围内最优，因此，在本文所有实验中，

将选择 $k=10$ 作为 scMOGCL 方法的默认 k 值.

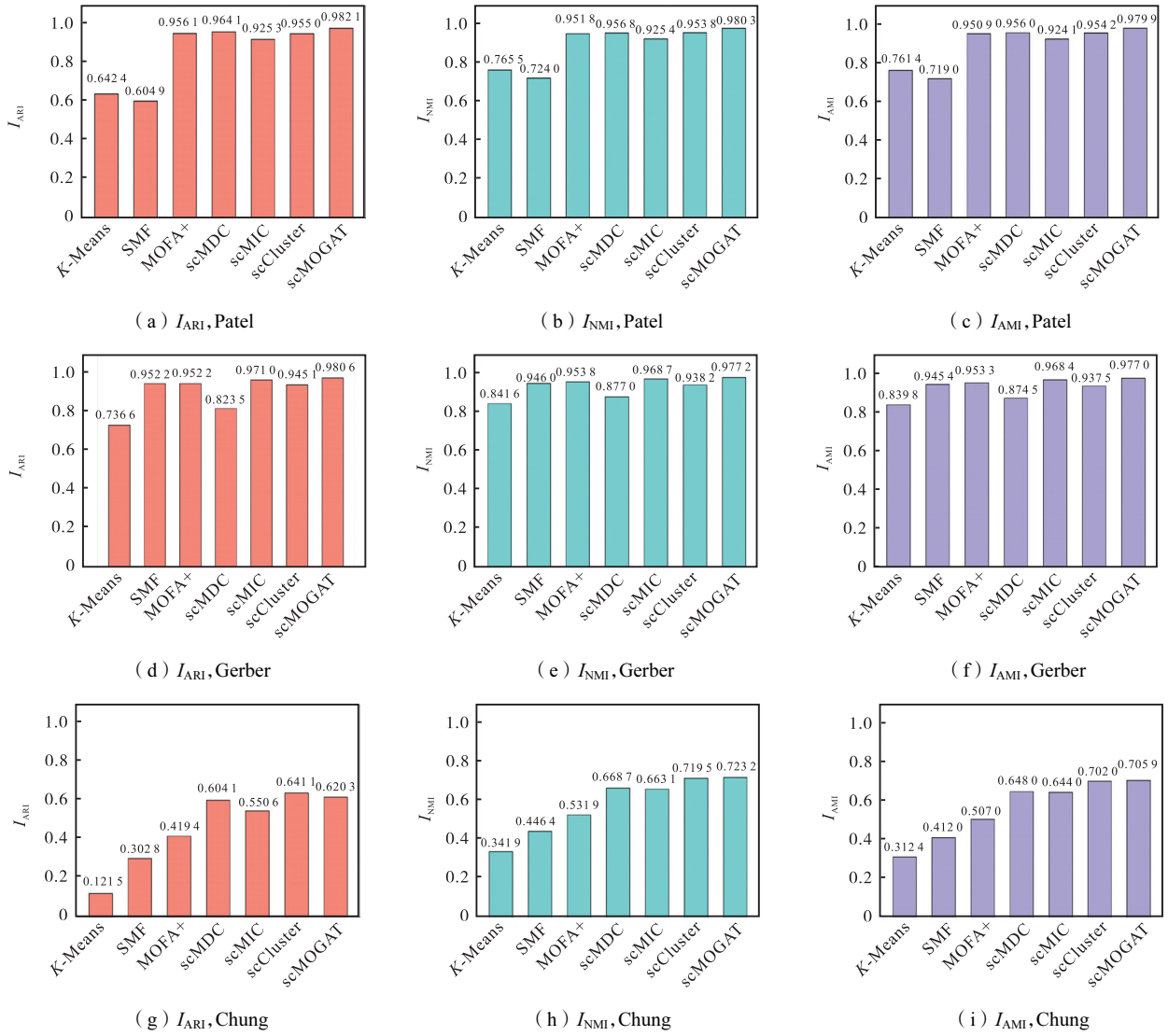


图 3 不同方法在 3 个单细胞癌症数据集上的性能对比

Fig.3 Comparison of the performance metrics of different methods using three single-cell cancer datasets

其次,本节对图注意力网络中的注意力头数进行参数分析. 实验选取了注意力头数范围在 [1, 8] 范围内进行分析. 根据图 5 所示的分析结果,随着注意力头数量的增加,scMOGCL 的表现并未出现显著变化. 仅当注意力头的数量从 7 增加到 8 时,性能有轻微下降,但是 3 个指标依然能够保持在 0.900 0 左右,这表明了 scMOGCL 方法对 GAT 中的注意力头数的数量不敏感.

值得注意的是,当注意力头数量为 1 时,模型已经展现出了出色的性能表现,3 个评估指标均超过了 0.970 0,这表明单一注意力头已能有效聚合多组学特征信息. 因此,从模型的性能与计算效率考虑,本文中所有实验均采用注意力头数为 1 的设置.

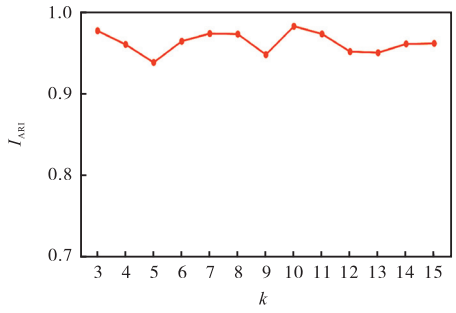
综上所述,scMOGCL 模型在图结构的细胞邻居数 k 和注意力头数等关键参数上表现出了显著的不敏感性,这表明了模型在不同的参数设置下能保持一定的可靠性和稳定性.

2.5 消融实验

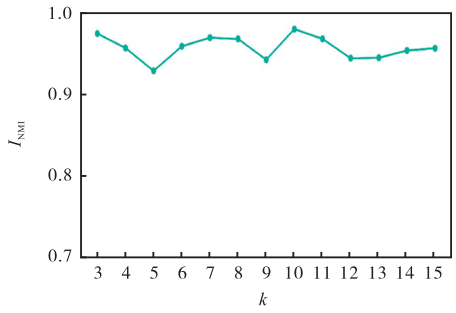
本文通过消融实验验证 scMOGCL 中各个模块的有效性,该实验在 Patel 数据集上进行. 具体而言,本文对 scMOGCL 方法构建了 5 个变体,包括图对比学习模块、一致性约束模块和潜在表示融合模块的消融.

scMOGCL-w/o-$\mathcal{L}_{SCL}$:移除每个组学数据中的图对比学习模块. 这意味着,在处理组学数据时,不再生成特征扰动后的矩阵作为负样本输入至图对比学习模块中. 因此,缺乏了图对比学习模块对每个组学潜在表示进行优化的过程.

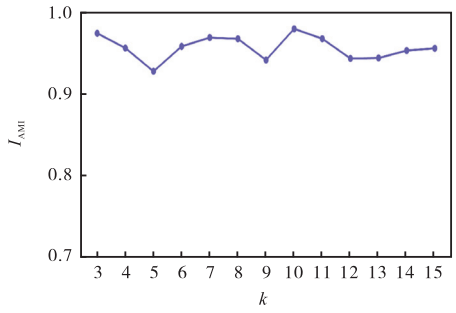
scMOGCL-w/o-RNA-$\mathcal{L}_{SCL}$:仅移除 RNA 数据通道中的图对比学习模块,保留 eSNP 组学通道中的图



（a）I_{ARI}



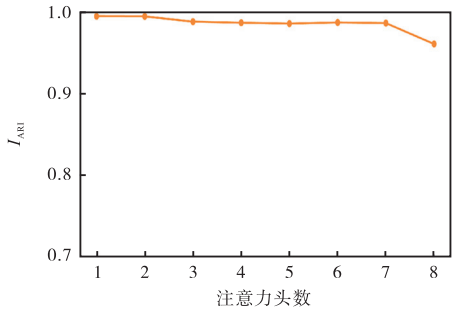
（b）I_{NMI}



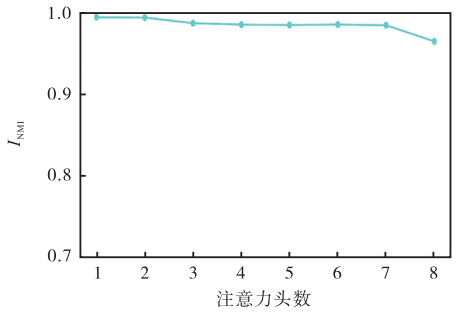
（c）I_{AMI}

图 4　不同 *k* 值下模型性能

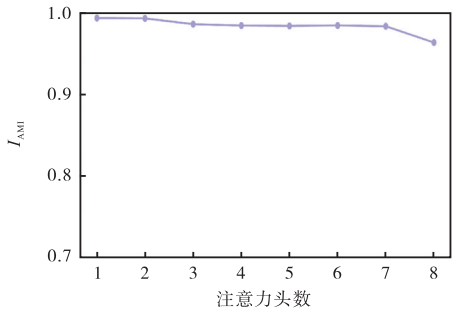
Fig.4　Model performance at different *k* values



（a）I_{ARI}



（b）I_{NMI}



（c）I_{AMI}

图 5　图注意力网络中不同注意力头数量下的模型性能

Fig.5　Model performance in relation to the number of attention heads within the graph attention network

对比学习模块.

scMOGCL - w / o-eSNP - $\mathcal{L}_{SCL}$：仅移除 eSNP 数据通道中的图对比学习模块，保留 RNA 组学通道中的图对比学习模块.

scMOGCL - w/o - $\mathcal{L}_{consistency}$：该变体中移除了一致性约束模块. 该模块能够促进不同模态之间的协同训练，其计算复杂度为 $O(n^2)$. 移除该模块意味着各组学数据的潜在表示在融合之前是独立生成的，缺乏组学间的交互作用.

scMOGCL - w/o - Average- Pool：移除了用于融合多个组学潜在表示的平均池化操作，改为采用直接拼接的操作方式.

scMOGCL 变体在 Patel 数据集上运行的结果如表 2 所示. 从表 2 中可以看出各个模块的重要性程度. 图对比学习模块的消融结果显示，该模块的缺失会使得 I_{ARI} 下降约 2% ~ 3%，这表明图对比学习模块在模型良好性能的情况下，主要起进一步的优化潜在表示的作用. 潜在表示平均池化机制的缺失会导致 I_{ARI} 下降约 4%，这表明其在进行各组学潜在表示融合上发挥了重要作用.

表 2　scMOGCL 及其变体在 Patel 数据集上的性能比较

Tab.2　Comparison of the performance metrics of scMOGCL and its variants on the Patel dataset

方法	I_{ARI}	I_{NMI}	I_{AMI}
scMOGCL - w/o-$\mathcal{L}_{SCL}$	0.951 4	0.946 8	0.945 8
scMOGCL - w/o-RNA- $\mathcal{L}_{SCL}$	0.960 1	0.955 9	0.955 1
scMOGCL - w/o - eSNP- $\mathcal{L}_{SCL}$	0.959 6	0.957 1	0.956 4
scMOGCL - w/o - $\mathcal{L}_{consistency}$	0.916 6	0.913 4	0.911 8
scMOGCL - w/o - Average - Pool	0.948 1	0.947 0	0.946 0
scMOGCL	0.982 1	0.980 3	0.979 9

值得注意的是，移除一致性约束模块对 scMOGCL 方法的性能产生了较为显著的影响（I_{ARI} 下降了约 7%），这很大程度上是因为一致性约束能

够使不同组学之间进行互补式训练. 这些结果证明了跨组学协同训练策略的有效性以及该方法设计上的合理性.

2.6 临床相关性分析

本文在 Gerber 数据集上进行了临床相关性分析. Gerber 数据集是从 3 名黑色素瘤患者收集的 scRNA-seq 数据集,其中共包含 307 个细胞,根据癌症亚型可分为 BRAF/NRAS 双野生型、BRAF V600E 突变型以及 NRAS G13R 突变型,而根据细胞类型可分为 Mel、Mel2、Mel3 和 Mel4[12]. 基于 scMOGCL 的识别结果,本文将所有细胞分为不同的 4 个簇,并进行差异表达基因分析.

本文利用 Seurat[23]方法对每个细胞类型中差异表达显著的标记基因进行筛选,并设置相应的筛选阈值,具体为:每个簇相对于所有其他簇的基因表达 $|\mathrm{lbFoldChange}|>1$,且调整后的 p 值 <0.05. 其中,lbFoldChange 表示两组细胞之间基因表达变化的对数倍数,$|\mathrm{lbFoldChange}|>1$ 表示基因表达差异超过两倍;调整后的 p 值用于校正多重假设检验带来的假阳性结果,从而更准确地评估基因表达差异的统计显著性.

在此基础上,本节对每个簇中差异表达最显著的前 10 个基因的表达模式进行展示,如图 6 中差异表达基因分析所示,该图清晰地显示了 scMOGCL 在 Gerber 数据集上识别结果的差异表达模式,这表明 scMOGCL 识别得到的每个簇都有着独特的高表达标记基因,这一发现进一步证明了 scMOGCL 的识别结果的可靠性.

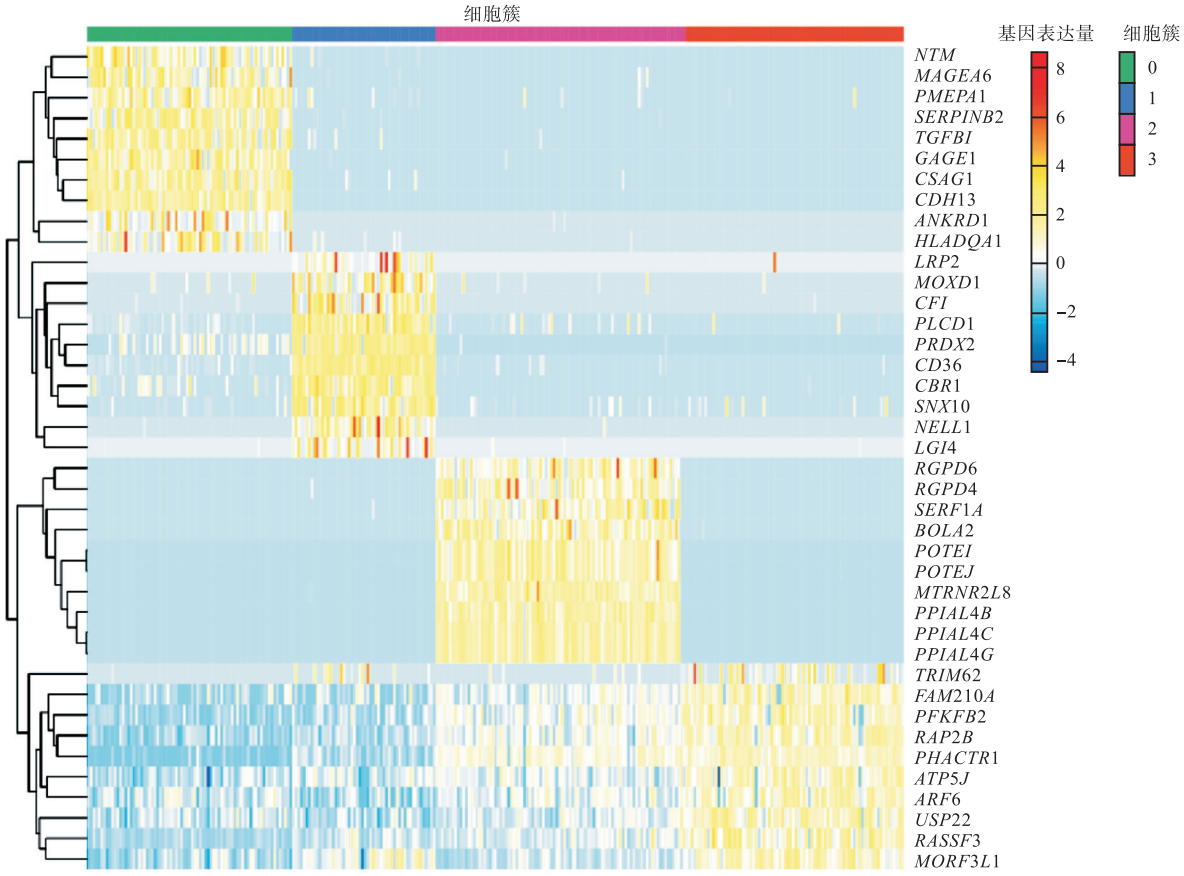


图 6 各细胞簇差异表达基因热图

Fig.6 Heatmap of differentially expressed genes across cell clusters

基于上述差异表达分析中筛选得到的标志基因,进一步进行了 Gene Ontology(GO)富集分析,旨在了解不同细胞类型相关的生物学功能,并通过 GO 富集分析选出每个细胞类型中最相关的前 5 个生物学功能条目进行展示,具体结果见图 7 中的 GO 富集分析所示. 其中,“focal adhesion”条目在“cluster 0”中显著富集,“focal adhesion”的富集结果与黑色素瘤的侵袭性高度相关,这在文献[24]中得以证明,文中表明,黑色素瘤最相关的整合素组合之一 $\alpha_v\beta_3$,黑色素瘤细胞中过表达的 $\alpha_v\beta_3$ 整合素将会显著增强黏着斑激酶(focal adhesion kinase)和 Src 的活化水平,这表明了这些激酶在整合素介导的侵袭中发挥了作用.

此外,“cell-cell adhesion via plasma-membrane adhesion molecules”在“cluster 1”中的结果能够说明

细胞黏附分子赋予黑色素瘤细胞转移的潜能，这在文献[25]中得以证明. 文中表明，黑色素瘤细胞通过动态地调控质膜黏附分子的表达（如 Cadherins、Mel-CAM/MUC18、Integrins），来改变细胞间相互作用，从而影响黑色素瘤细胞的转移.

综上所述，scMOGCL 方法在黑色素瘤单细胞多组学数据上能够实现有效的无监督识别，结果显示其 GO 富集条目与已知的黑色素瘤分子机制具有一致性，该实验结果表明了 scMOGCL 方法的有效性与可靠性.

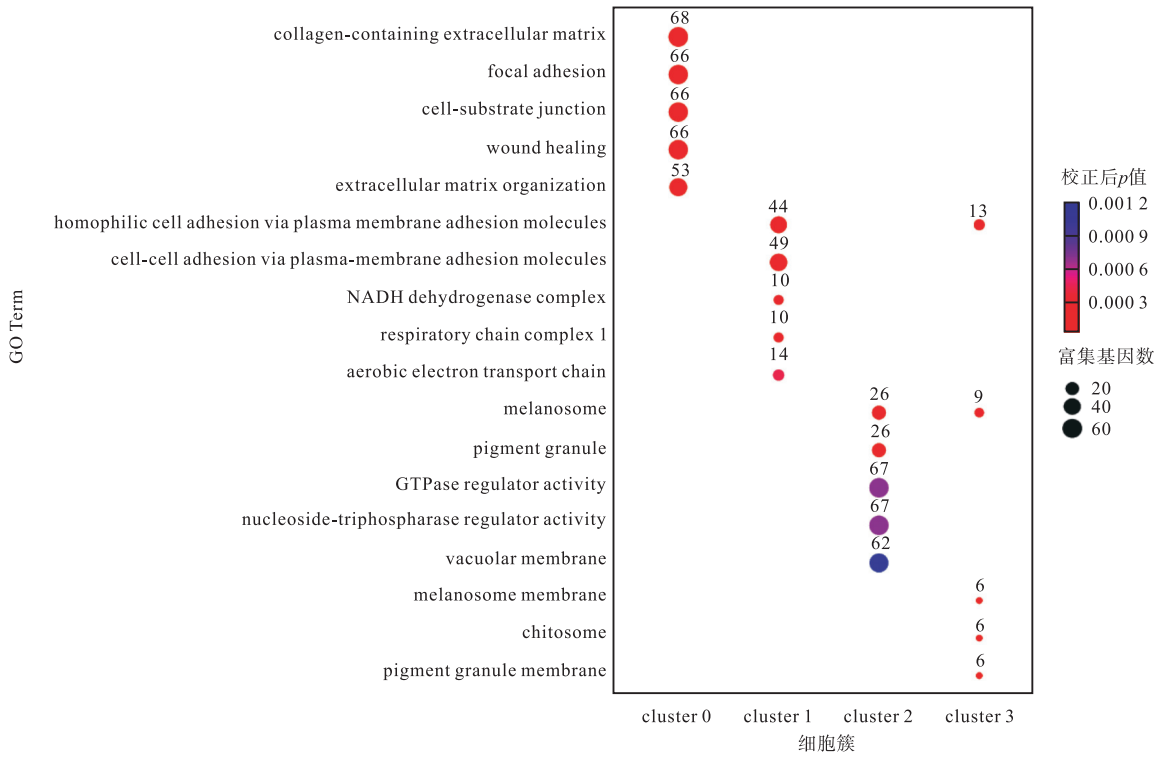


图 7　不同细胞簇的 GO 富集分析

Fig.7　GO enrichment analysis of different cell clusters

3　结　语

本文提出了 scMOGCL 方法，通过图注意力网络模块，使得模型能够从原始表达矩阵中自适应地提取高质量潜在表示. 同时，本文引入一致性约束模块，实现了各组学之间的协同训练，保持组学间的相对一致性. 最后，通过图对比学习优化各组学潜在表示，并使用平均池化操作对各组学的潜在表示进行融合，充分整合了各组学数据中的信息. 在 3 个真实的单细胞癌症数据集上的实验结果充分证明了 scMOGCL 方法的出色性能以及模型设计的有效性. 该方法在单细胞 RNA 数据与 eSNP 数据上的融合取得了出色的性能表现，但尚未在如 scATAC-seq 或 ADT 等其他单细胞组学进行整合并验证效果，未来将进一步探索该方法的整合范围.

参考文献：

[1] Hoadley K A，Yau C，Wolf D M，et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin[J]. Cell，2014，158(4)：929-944.

[2] Zhang Y J，Wang D，Peng M，et al. Single-cell RNA sequencing in cancer research[J]. Journal of Experimental & Clinical Cancer Research，2021，40(1)：81.

[3] Vallejos C A，Risso D，Scialdone A，et al. Normalizing single-cell RNA sequencing data：Challenges and opportunities[J]. Nature Methods，2017，14(6)：565-571.

[4] Kiselev V Y，Kirschner K，Schaub M T，et al. SC3：Consensus clustering of single-cell RNA-seq data[J]. Nature Methods，2017，14(5)：483-486.

[5] Tian T，Wan J，Song Q，et al. Clustering single-cell RNA-seq data with a model-based deep learning approach[J]. Nature Machine Intelligence，2019，1(4)：191-198.

[6] Wang J X，Ma A J，Chang Y Z，et al. scGNN is a novel graph neural network framework for single-cell

RNA-seq analyses[J]. Nature Communications, 2021, 12(1): 1882.

[7] Lin X, Tian T, Wei Z, et al. Clustering of single-cell multi-omics data with a multimodal deep learning method[J]. Nature Communications, 2022, 13: 7705.

[8] Gao H L, Zhang B, Liu L, et al. A universal framework for single-cell multi-omics data integration with graph convolutional networks[J]. Briefings in Bioinformatics, 2023, 24 (3): bbad081.

[9] Cao Y C, Zhao X M, Tang S M, et al. scButterfly: A versatile single-cell cross-modality translation method via dual-aligned variational autoencoders[J]. Nature Communications, 2024, 15: 2973.

[10] Baul S, Ahmed K T, Filipek J, et al. omicsGAT: Graph attention network for cancer subtype analyses[J]. International Journal of Molecular Sciences, 2022, 23 (18): 10220.

[11] Huang M, Long C Z, Ma J T. AAFL: Automatic association feature learning for gene signature identification of cancer subtypes in single-cell RNA-seq data[J]. Briefings in Functional Genomics, 2023, 22(5): 420-427.

[12] Wang T, Mai D D, Shu H, et al. Enhancing cell subpopulation discovery in cancer by integrating single-cell transcriptome and expressed variants[J]. Fundamental Research, 2026, 6(1): 88-98.

[13] Xu Z Y, Luo J W, Xiong Z H. scSemiGAN: A single-cell semi-supervised annotation and dimensionality reduction framework based on generative adversarial network[J]. Bioinformatics, 2022, 38(22): 5042-5048.

[14] Wolf F A, Angerer P, Theis F J. SCANPY: Large-scale single-cell gene expression data analysis[J]. Genome Biology, 2018, 19(1): 15.

[15] Veličković P, Fedus W, Hamilton W L, et al. Deep graph infomax[C]//International Conference on Learning Representations(ICLR). New Orleans, USA, 2019.

[16] Hartigan J A, Wong M A. Algorithm AS 136: A K-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100.

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

[18] Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data[J]. Genome Biology, 2020, 21(1): 1-17.

[19] Zhan Y L, Liu J H, Ou-Yang L. scMIC: A deep multi-level information fusion framework for clustering single-cell multi-omics data[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(12): 6121-6132.

[20] Hubert L, Arabie P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

[21] Strehl A, Ghosh J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3: 583-617.

[22] Vinh N X, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance[J]. Journal of Machine Learning Research, 2010, 11: 2837-2854.

[23] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data[J]. Cell, 2019, 177 (7): 1888-1902.

[24] Haass N K, Smalley K S M, Li L, et al. Adhesion, migration and communication in melanocytes and melanoma[J]. Pigment Cell Research, 2005, 18(3): 150-159.

[25] McGary E C, Lev D C, Bar-Eli M. Cellular adhesion pathways and metastatic potential of human melanoma [J]. Cancer Biology & Therapy, 2002, 1(5): 459-465.

(责任编辑：王晓燕)