DOI:10.3969/j.issn.1005-202X.2026.04.020

医学人工智能

对比学习驱动的多组学卵巢癌分子分型

韩晓鑫[1],刘庆晨[1],胡雨辰[1],邝竞凡[1],王建林[2]
1.甘肃中医药大学医学信息工程学院,甘肃 兰州 730000;2.兰州大学第一医院信息中心,甘肃 兰州 730013

【摘要】卵巢癌分子分型对个体化治疗与预后判断至关重要,但高度异质性与高维少样本问题限制了传统方法的准确性。针对上述挑战,本研究提出一种融合对比学习与深度聚类的端到端多组学模型CDCM。CDCM整合TCGA-OV的RNA-seq、CNV与DNA甲基化数据输入至自动编码器,捕捉数据中复杂的非线性相互作用。接着通过4种数据增强策略构建正负样本对,利用对比学习机制有效提升表征的鲁棒性与判别性,从而缓解高维小样本下的过拟合问题。最后基于Student's *t*分布的聚类损失,与对比损失联合优化,直接驱动样本向聚类中心聚集并获得分离度更高的清晰亚型。增加消融实验以XGBoost量化评估组学贡献,证明多组学整合能够显著提升分型效果。为增强模型的生物学可解释性,联合XGBoost与WGCNA识别出与卵巢癌相关的12个候选生物标志物,其中有10个已在现有的文献中得到验证。CDCM在轮廓系数0.579、CH指数344.85及生存差异显著性-lg *P*=1.771上均优于K-Means等经典模型,为卵巢癌的精准诊疗提供新的方法学思路。

【关键词】卵巢癌;分子分型;多组学数据;对比学习;深度聚类
【中图分类号】R318;R73 【文献标志码】A 【文章编号】1005-202X(2026)04-0553-08

Contrastive learning-driven multi-omics molecular subtyping of ovarian cancer

HAN Xiaoxin[1], LIU Qingchen[1], HU Yuchen[1], KUANG Jingfan[1], WANG Jianlin[2]
1. School of Medical Information Engineering, Gansu University of Chinese Medicine, Lanzhou 730000, China; 2. Information Center, Lanzhou University First Hospital, Lanzhou 730013, China

Abstract: Molecular subtyping of ovarian cancer is essential for personalized treatment and prognostic assessment, but high tumor heterogeneity and the high-dimensional, low-sample size problem compromise the accuracy of traditional methods. An end-to-end multi-omics model called the contrastive deep clustering model (CDCM) which integrates contrastive learning and deep clustering is proposed to address the aforementioned challenges. CDCM fuses RNA-seq, CNV and DNA methylation data from TCGA-OV, and inputs them into an autoencoder to capture complex nonlinear interactions in the data. Subsequently, 4 data-augmentation strategies are used to construct positive and negative sample pairs, and a contrastive learning mechanism is employed to effectively enhance representation robustness and discriminability, thereby alleviating overfitting under the high-dimensional, low-sample size condition. Finally, a clustering loss based on the Student's *t*-distribution is jointly optimized with the contrastive loss to directly drive samples toward cluster centers and obtain more separable, well-defined subtypes. Ablation experiments using XGBoost quantify the contributions of omics modality demonstrate that multi-omics integration can substantially improve subtyping performance. To enhance the model's biological interpretability, XGBoost and WGCNA are combined to identify 12 candidate biomarkers associated with ovarian cancer, 10 of which have been validated in existing literature. CDCM outperforms classical models such as K-Means, and achieves a silhouette score of 0.579, a Calinski-Harabasz index of 344.85, and a survival-difference significance of -lg *P*=1.771, providing a new methodological avenue for precision diagnosis and treatment of ovarian cancer.

Keywords: ovarian cancer; molecular subtyping; multi-omics data; contrastive learning; deep clustering

前 言

卵巢癌是妇科恶性肿瘤中致死率最高的癌种之一,其发病率位居女性癌症的第八位,而死亡率高居第五[1]。卵巢癌的高度异质性使得患者在临床表现、病理特征、治疗反应及预后方面存在显著差异。因

【收稿日期】2025-12-22
【基金项目】甘肃省重点研发计划(20YF8FA080)
【作者简介】韩晓鑫,硕士研究生,研究方向:医学数据挖掘,E-mail: han110999@163.com
【通信作者】王建林,硕士生导师,正高级工程师,研究方向:医学信息,E-mail: 375763325@qq.com

此，精准的分子分型对于制定个体化治疗策略和改善患者生存至关重要[2]。近年来，基因组学、转录组学、蛋白质组学和代谢组学等多组学技术发展迅速。这些技术为深入探索卵巢癌的分子机制以及实现更精准的诊断和治疗提供了新的策略和靶点[3]。整合多组学数据可以更全面地捕捉肿瘤的生物过程。此外，整合分析还能揭示不同分子层面的交互作用，从而为卵巢癌亚型识别提供更可靠的依据。然而，多组学数据通常维度极高且稀疏，特征数远大于样本数。这会导致直接进行聚类或下游分析时出现维度灾难和严重的拟合风险。并且，肿瘤数据的精准人工标记高度依赖病理医生的专业判读与多组学数据的整合注释。标注过程繁琐且难以大规模获取，导致有标签数据极为稀缺，限制了监督学习方法的应用规模。因此，如何利用大量易获得的未标注数据来鲁棒地识别癌症亚型，是目前多组学分型研究中的核心问题。传统的方法像是直接聚类[4]、线性降维后聚类[5]、基于统计或整合模型[6]以及层次聚类与共识聚类[7]等。这些方法在可解释性和简便性上各有优势，但面对高维多组学数据中复杂的非线性关系和大量噪声时，线性降维和浅层方法往往难以捕捉关键信号。

随着深度学习的飞速发展，许多基于深度学习的方法被开发出来用于癌症分型[8-10]。神经网络具有出色的非线性特征提取能力，能够获得更有效的数据表示。例如，Guo等[11]提出一种带有局部结构保存模块的深度聚类网络IDEC来聚类高维复杂数据。Chaudhary等[12]提出AE-KM模型，应用自编码器从多组学数据中提取患者信息用于肝癌亚型识别。研究表明，深度神经网络在高维特征和非线性特征中提取信息方面优于传统方法。为了更好地表征不同分布的患者数据，Yang等[10]提出利用共识聚类和高斯混合模型的Subtype-GAN。尽管基于深度学习的方法在性能上通常优于传统方法，但有限的样本量依然可能导致肿瘤亚型不清晰并出现重叠。

对比学习的方法不依赖于数据同质性假设，因此在数据受限的情况下更具优势。其通过比较不同样本之间的差异来学习数据的特征，从而避免了建立深度学习模型时常见的数据限制问题[13-15]。该方法能够处理不同来源、不同特征分布的数据集。对比学习通过最大化相似样本之间的相似性和最小化不相似样本之间的相似性来学习特征。这一机制有助于捕捉到数据中的细微差异，从而生成更具区分性的特征表示，提升模型性能。因此，本研究提出了一种融合对比学习与深度聚类的多组学分型模型（Contrastive Deep Clustering Model, CDCM）。CDCM是一种端到端的多组学分型模型，其整体架构可分为3大模块：表征学习模块、对比增强模块和聚类优化模块。表征学习模块使用自动编码器非线性压缩，缓解维度灾难，并保留了复杂的非线性关系，从根本上提升亚型的可分离性，便于下游判别与聚类。对比增强模块大幅提高了在小样本、高噪声与批次效应条件下的泛化鲁棒性。聚类优化模块采用软分配与目标分布锐化策略，形成更清晰的簇结构，提升各簇的分子特征可区分性，从而获得更稳健的候选亚型。

1 材料与方法

1.1 实验数据与数据预处理

本研究基于UCSC Xena浏览器平台（https://xenabrowser. net/datapages/）提供的TCGA（The Cancer Genome Atlas）公共数据（https://portal. gdc. cancer.gov/），获取卵巢癌的RNA-seq基因表达量、基因拷贝数和DNA甲基化3种组学数据进行实验，3种组学数据的具体维度如表1所示。这3种组学能够互补性地捕捉癌症发生发展的多层次关键生物学机制，从而提供更全面、更鲁棒的分型信息[16]。笔者采用三步法对下载的数据进行预处理，预处理后的数据详情如表2所示。首先，对RNA-seq、CNV和DNA甲基化3种组学数据进行缺失值处理，剔除缺失值比例超过20%的样本和特征，并使用KNN均值填充法填补数据集中的其余缺失值。接着，使用z-score对数据进行标准化处理。最后，对3种组学数据进行交集处理并进行早期融合，得到292个样本，共计70 271维特征。

表1 卵巢癌3种组学的初始维度

Table 1 Initial dimensionality of the 3 omics datasets for ovarian cancer

组学类型	样本数量	特征数量
RNA-seq	308	20 531
CNV	579	24 777
DNA甲基化	616	27 579

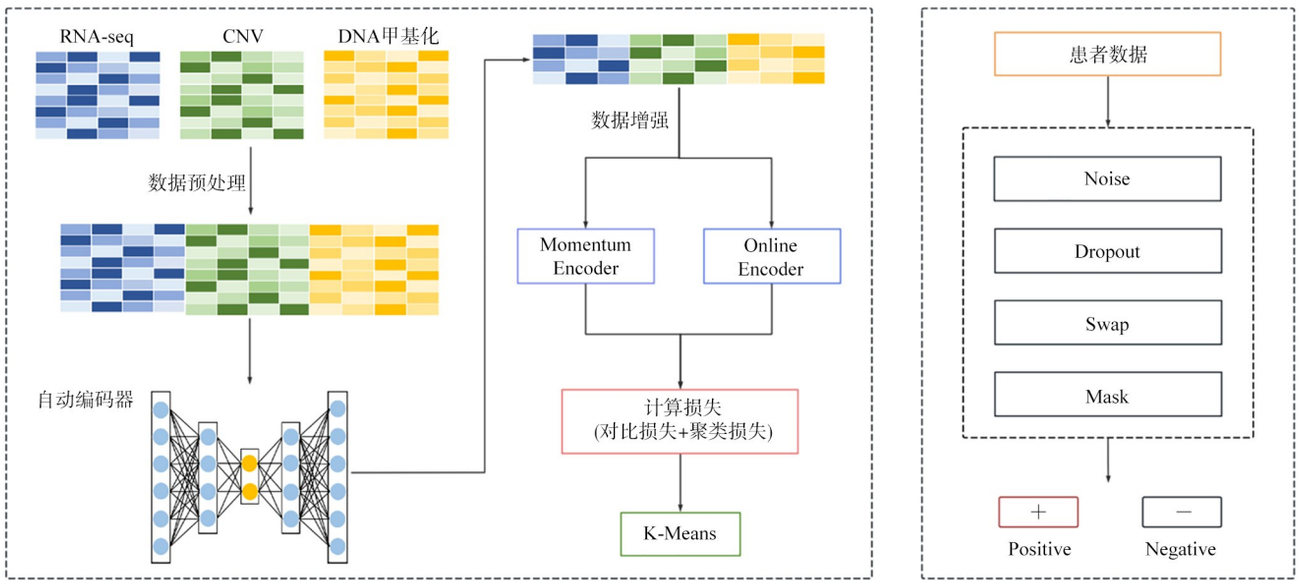
表2 数据预处理后的组学数据维度

Table 2 Dimensionality of the omics datasets after data preprocessing

组学类型	样本数量	特征数量
RNA-seq	305	20 530
CNV	579	24 776
DNA甲基化	596	24 965
Multi-omics	292	70 271

1.2 方法

1.2.1 CDCM框架的体系结构　在本研究中，笔者构建了一个融合对比学习与深度聚类的多组学模型，用于卵巢癌多组学数据的分子分型，以下简称为CDCM模型。其端到端的整体架构如图1所示，主要包含3大核心模块：表征学习模块、对比增强模块和聚类优化模块。首先，如图1a所示，将整合的高维多组学特征通过自动编码器进行非线性压缩。接着，对每个样本采用高斯噪声注入、模拟Dropout、数据交换、数据掩码4种数据增强技术生成多视图，其中数据增强的参数分别为高斯噪声 σ=0.08、模拟Dropout概率 ρ=0.10、特征交换概率swap_prob=0.10、特征掩码概率mask_prob=0.10，如图1b所示。通过动量编码器和在线编码器架构生成嵌入向量，计算对比损失。利用对比损失使得同一原始样本的不同增强视角在嵌入空间中尽可能靠近，而不同样本之间则保持较大距离。最后，聚类优化模块将嵌入的聚类损失纳入深度神经网络，联合优化对比损失和聚类损失，进一步细化对比学习得到的特征，驱动同类样本再嵌入空间聚集。为客观选择簇数，在聚类数为 K=2~8的范围内进行生存学显著性分析，最终确定 K=4为最终分型数。



a：用于癌症患者聚类的CDCM深度神经网络　　b：基于对比学习机制生成正负样本对的方法

图1　用于卵巢癌分子分型的CDCM框架图

Figure 1 CDCM framework for molecular subtyping of ovarian cancer

1.2.2 联合优化框架　对比损失函数：假设 $X=(x_1,x_2,\cdots,x_g)$ 表示一组多组学特征，其中 g 表示特征数量，n 表示样本量。在对比学习的数据增强阶段，样本数量 n 将扩展至 $2n$，即得到 x_i 经两次随机增强得到视图 x_i^A 和 x_i^B。在每个训练批次中，将源自同一样本生成的两个相似数据点定义为正样本对，而将其他组合视为负样本对。在对比学习模块中，采用如下损失函数来区分正样本对与负样本对之间的差异：

$$L_{\mathrm{cl}}=-\frac{1}{2n}\sum_{i=1}^{n}\left[\log\frac{\exp\left(z_i^A\cdot z_i^{B'}/\tau\right)}{\sum_{k=1}^{2N}\exp\left(z_i^A\cdot z_k/\tau\right)}+\log\frac{\exp\left(z_i^B\cdot z_i^{A'}/\tau\right)}{\sum_{k=1}^{2N}\exp\left(z_i^B\cdot z_k/\tau\right)}\right] \tag{1}$$

其中，z_i^A 和 z_i^B 分别为 x_i^A 和 x_i^B 经在线编码器得到的嵌入向量，$z_i^{A'}$ 和 $z_i^{B'}$ 是视图在动量编码器下的嵌入表示，z_k 表示嵌入队列的第 k 个样本，利用温度系数 τ 来调节模型对负样本的区分度。

聚类损失函数：为了区分患者的疾病亚型，完成相似性模块中的聚类任务，以最小化以下目标：

$$L_{\mathrm{cluser}}=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_{ij}} \tag{2}$$

其中，p_{ij} 是目标分布，通过平方标准化增强置信度：

$$p_{ij}=\frac{q_{ij}^2\Big/\sum_i q_{ij}}{\sum_j\left(q_{ij}^2\Big/\sum_i q_{ik}\right)} \tag{3}$$

其中，q_{ij} 是嵌入空间的软分配概率用于描述聚类中心 μ_j 与嵌入点之间的相似性，α 为Student's t 分布自由度

超参数，写为：

$$q_{ij}=\frac{\left(1+\left\|z_i-\mu_j\right\|^2/a\right)^{-1}}{\Sigma_j\left(1+\left\|z_i-m_k\right\|^2/a\right)^{-1}} \quad (4)$$

联合优化目标：将对比损失与聚类损失加权合并，得到CDCM的最终优化目标：

$$L=L_{\mathrm{cl}}+\gamma L_{\mathrm{cluser}} \quad (5)$$

在本研究中，为了平衡对比学习损失L_{cl}和聚类损失L_{cluser}的重要性，γ值被设置为1×10^{-5}。此外，该参数可通过五折交叉验证进行动态校准。

1.3 实验设计与统计学分析

为系统评估CDCM的性能以及多组学数据整合优势，笔者设计了以下几项实验，并采用严格的统计学分析方法对结果进行验证。

聚类性能评估：为了验证所识别的分子亚型与患者生存预后的相关性，使用Kaplan-Meier生存曲线可视化不同亚型的生存率随时间的变化。随后，采用log-rank检验比较各亚型之间生存曲线的差异。将log-rank检验给出检验统计量（χ^2）及对应的P值转换为负对数指标，即-lg P以便于比较。较高的-lg P值表明各亚型之间存在更显著的生存差异，说明模型在分型方面具有更好的区分能力。轮廓系数定义为所有样本轮廓系数的均值。轮廓系数越接近1，表明簇内更为紧密且簇间分离良好，从而表明样本聚类结果更合理。需要指出的是，多组学肿瘤数据常呈现高维、噪声大、批次效应显著且生物学上常存在连续性或过渡样本。因此在该情境下获得很高的二维或低维聚类指标本身就更加困难。在此场景下，经验上通常认为轮廓系数大于0.5表示较好的聚类分离[17-19]。CH指数是簇间离散程度与簇内离散程度的比，CH指数取值越大，聚类分离性越好。聚类质量指标的轮廓系数和CH指数与生存学检验的Kaplan-Meier曲线和log-rank检验的联合使用，既能衡量簇间分离度又能评估分型的临床相关性。通过对比不同模型的3个指标，可以直观了解模型性能的优劣。

消融实验：为验证多组学数据整合是否能提升分型效果，笔者分别使用单组学数据RNA-seq、CNV和DNA甲基化独立进行分型。然后将每种单组学的分型结果与整合多组学数据的分型结果进行比较。对于每一种组学数据，均保持相同的数据预处理和模型配置，以确保结果的可比性。记录各单一组学数据所得到的-lg P值，并统计多组学整合后的结果，观察是否在生存分析中表现出更高的显著性。若多组学整合后的-lg P值高于单一组学数据，表明整合策略有效地利用了不同数据类型间的互补信息，有效提升了患者亚型的区分度和生物学意义。

组学重要性评估：为了进一步解释多组学数据对卵巢癌亚型分类的贡献，以聚类标签为响应变量，训练XGBoost分类模型，计算各个组学的重要性评分。重要性分数能够反映模型在现有数据和预处理下对不同组学依赖程度，为后续的独立队列验证与生物学功能实验提供指导，从而推动结果向临床转化方向发展。

生物标志物识别：为增强分型的生物学可解释性，笔者采用“机器学习筛选+共表达网络验证”的整合方案来识别候选生物标志物。该流程旨在结合模型驱动的特征筛选与网络级别的模块分析以提高可信度。首先，通过XGBoost计算每个特征的重要性并取均值，然后选取排名前100的特征作为候选列表。接着，对RNA-seq表达数据使用WGCNA挖掘基因模块，获取相关的枢纽基因。最后，将XGBoost筛选得到的候选特征与WGCNA识别的枢纽基因进行交叉验证，取两者的交集作为最终候选标志物列表。

为保证结论稳健性，所有分析均通过交叉验证、bootstrap重抽样及必要的置换检验下重复验证，并对多重比较采取适当校正。

2 结 果

2.1 卵巢癌分子分型的可视化结果

应用CDCM对整合的多组学数据进行端到端训练后，样本被划分为4个分子亚型。为了直观展示模型分析，卵巢癌患者被划分为局域显著生存差异的分子亚型。如图2所示，通过主成分分析（Principal Component Analysis, PCA）将CDCM学习的高维表征投影至二维空间，不同亚型样本用颜色区分呈现颜色集中密集、清晰分离，表明模型成功捕获了数据的非线性结构。

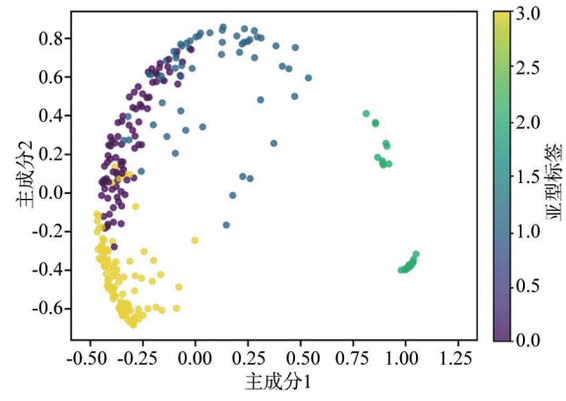


图2 卵巢癌分子分型聚类可视化结果

Figure 2 Visualization of clustering results for molecular subtypes of ovarian cancer

生存分析如图3所示，不同亚型的Kaplan-Meier生存曲线存在明显差异。采用log-rank检验计算的总体显著性-lg P为1.771，即P=0.017。表明4类亚型在总体生存上具有统计学差异，从而能够证明所识别的分子亚型具有临床预后相关性。为直观展示CDCM联合优化效果，如图4的收敛曲线所示，联合损失L在训练过程中总体呈持续下降趋势，表明模型在训练过程中表现良好，能够有效地学习数据。

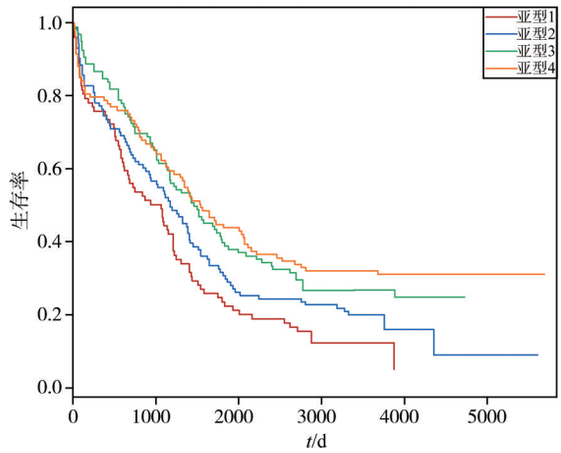


图3 CDCM绘制的Kaplan−Meier生存曲线

Figure 3 Kaplan−Meier survival curves generated by CDCM

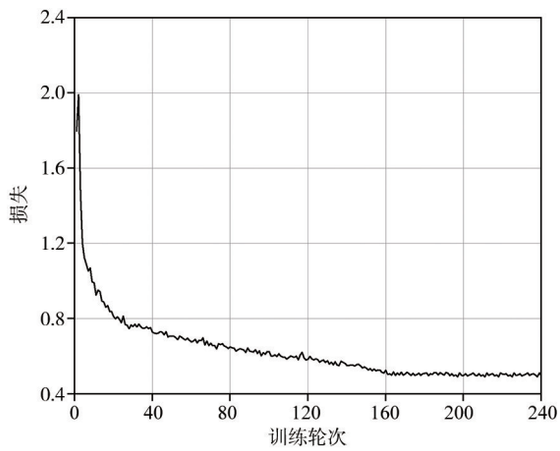


图4 联合损失收敛曲线

Figure 4 Convergence curve of the joint loss function

2.2 聚类性能比较

为定量评估CDCM在分型任务中的表现，将CDCM与多种代表性方法进行对比，如表3所示。评价指标包括轮廓系数、CH指数与生存显著-lg P。在总体表现上，CDCM的轮廓系数与CH指数分别为0.579和344.85。这两个值均在所有对比方法中最高，表明CDCM学得的嵌入既能实现较高的簇内紧密性，又能实现良好的簇间分离。在生存分析方面，CDCM的-lg P的值为1.771，显著优于其他方法。该结果表明CDCM识别的簇与生存结局之间存在明显的统计学相关性。将聚类结构质量和临床预后区分能力与常见的传统方法和其他深度学习相关方法进行比较。比较结果表明，将对比学习与聚类目标联合优化并结合多视图增强，在多组学卵巢癌分型问题上具有明显优势。

表3 不同聚类方法的性能比较

Table 3 Performance comparison of different clustering methods

方法	轮廓系数	CH指数	-lg P
K-means[20]	0.237	175.14	0.754
AE-KM[12]	0.334	200.37	0.945
PCA-KM[5]	0.288	199.47	0.455
iCluster[21]	0.313	233.66	0.739
SNF[22]	0.443	316.97	1.237
MOFA[23]	0.299	217.63	0.744
HDBSCAN[24]	0.476	299.55	1.268
IDEC[11]	0.518	335.46	1.520
S-GAN[10]	0.439	292.77	1.073
CDCM	0.579	344.85	1.771

2.3 消融实验分析

为验证多组学整合是否提升分型效果，通过使用单一组学RNA-seq、CNV和DNA甲基化数据重复CDCM流程，并与整合多组学的结果进行比较，结果如图5所示。多组学整合后得到的-lg P值明显高于任一单模态结果，说明通过融合不同组学的互补信息，CDCM能更好地刻画影响预后的分子特征组合，从而提高亚型的判别力与临床相关性。

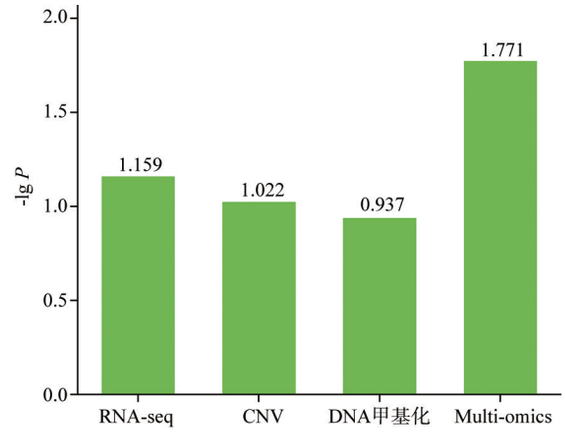


图5 CDCM消融实验中不同类型组学数据得到的−lg P值

Figure 5 The −lg P values obtained by COLCS using different type of omics data in ablation experiments

2.4 组学重要性评估

为定量评估不同数据模态在卵巢癌分子分型中的贡献，利用XGBoost对聚类标签进行分类建模，并计算各类组学特征的重要性评分，如图6所示。结果显示，RNA-seq在所有组学类型中贡献度最高，其平均重要性评分为0.582，与消融实验中RNA-seq单独建模时表现最佳的结果一致，提示转录组信号在卵巢癌亚型区分中具有核心作用。相比之下，DNA甲基化和CNV的重要性评分相对较低，但仍在整体模型中提供了互补信息。这一结果进一步说明，多组学整合不仅提升了分型的判别力，同时不同组学在模型中承担着差异化的角色，其中RNA-seq提供主要判别信号，而DNA甲基化与CNV数据在亚型细化和增强鲁棒性方面发挥辅助作用。

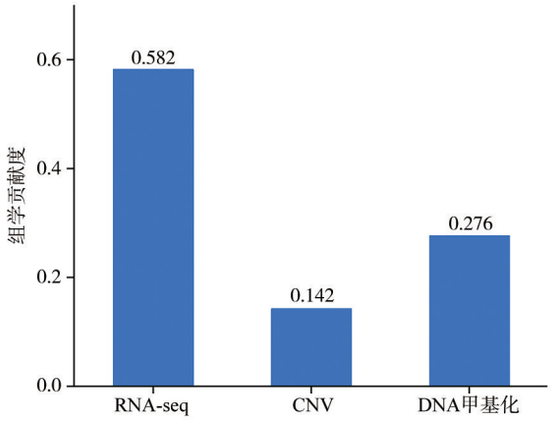


图6 基于XGBoost评估的各类组学数据对卵巢癌亚型识别的贡献

Figure 6 Contribution of each omics data to ovarian cancer subtyping identification evaluated by XGBoost

2.5 生物标志物识别

为进一步提升分型的生物学可解释性并识别潜在的候选生物标志物，笔者采用WGCNA共表达网络分析与基于XGBoost的特征重要性筛选的互补策略。WGCNA采用平均连锁聚类将表达模式相似的基因分组为模块，其结果如图7所示。基于各自的特征向量（Module Eigengene, ME），笔者计算每个卵巢癌分型与对应模块之间的相关性，如图8所示，确定10个与卵巢癌分型生物学上显著相关的模块，同时展示了不同模块的相关性，如图9所示。接着，在这10个模块中，分别计算了每个基因与卵巢癌分型的相关度（Gene Significance, GS）以及与所属模块特征向量的相关度（Module Membership, MM）。GS反映了基因表达与癌症亚型的相关强度，GS值越高，说明该基因与目标亚型的关系越重要；GS等于0表示无关联。MM表示基因表达与其所属模块ME之间的相关系数，MM值越高，说明该基因在模块中越核心。在筛选时，笔者将阈值设定为|GS|>0.5且|MM|>0.8，优先保留那些与分型显著相关且在各自模块内居中的候选基因。最终，将筛选后的基因与XGBoost选择的重要特征进行交集分析，识别出12个重叠基因作为潜在的卵巢癌生物标志物，如表4所示。其中，有3个基因RBX1、PLEKHM3和CKS2在近期文献中已被报道与卵巢癌相关，另有7个基因在其他癌种中具有明确作用。综上所述，这些结果不仅验证了使用CDCM研究卵巢癌分子分型具有生物学意义，也凸显了该分析框架在识别癌症相关基因方面的优势。

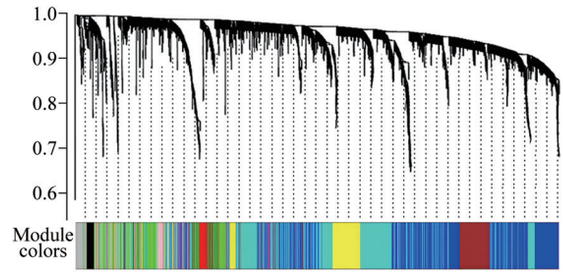


图7 基因树突图和识别模块

Figure 7 Gene dendrogram and identified modules

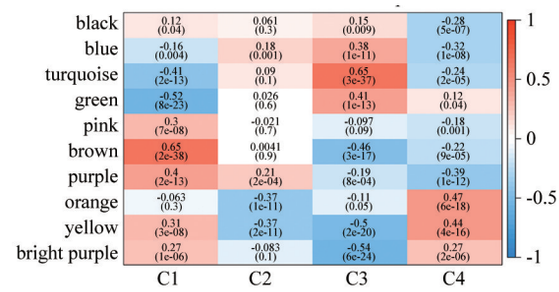


图8 聚类模块与分子亚型之间的模块–特征关系

Figure 8 Module–trait relationships between the clustered modules and molecular subtypes

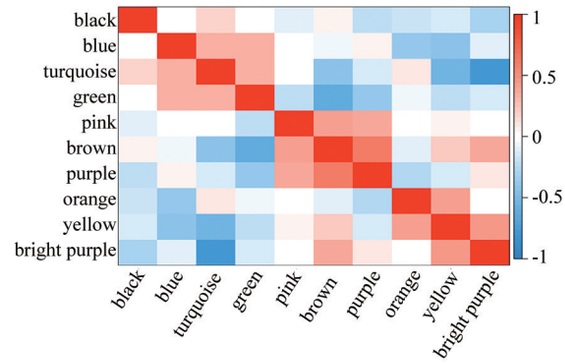


图9 不同模块之间的相关性

Figure 9 Correlations between different modules

表4 与卵巢癌生存相关的生物标志物

Table 4 Biological markers related to the survival of ovarian cancer

序号	基因	相关度	特征重要性权重	序号	基因	相关度	特征重要性权重
1	RBX1[25]	0.587	4.4×10^{-5}	7	SSSCA1	0.501	4.4×10^{-4}
2	UQCR11[26]	0.569	6.0×10^{-5}	8	CKS2[27]	0.566	2.5×10^{-4}
3	UQCRQ[28]	0.560	6.2×10^{-5}	9	ZBTB40[29]	0.549	1.5×10^{-4}
4	PLEKHM3[30]	0.590	6.8×10^{-6}	10	GON4L[31]	0.559	1.5×10^{-4}
5	MRPL36[32]	0.543	8.7×10^{-7}	11	CHD6[33]	0.600	1.9×10^{-4}
6	SERF2	0.523	1.4×10^{-5}	12	USP34[34]	0.540	2.5×10^{-6}

3 讨 论

本研究构建的CDCM模型创新性地将对比学习与深度聚类相结合，实现了端到端的多组学表征学习与亚型划分。在表征学习阶段，CDCM通过自动编码器对高维多组学特征进行非线性压缩，以挖掘复杂的跨组学关联。在此基础上，引入多视角数据增强生成正负样本对，利用动量编码器和在线编码器框架计算对比损失。对比损失驱动同一样本的不同增强视图在嵌入空间尽可能靠近，而不同样本间保持距离，从而显著提升表征的鲁棒性与判别性。在聚类优化阶段，CDCM将嵌入空间中的聚类损失纳入网络训练，与对比损失联合优化，使特征学习同时兼顾判别性与簇内一致性，直接驱动模型捕获清晰分离的亚型结构。CDCM通过融合自编码器的非线性映射能力和对比学习的自监督机制，在高维小样本情形下显著缓解过拟合，能够有效挖掘多组学异质性信息。

多组学整合可以从生物学和算法性能两方面带来显著优势。从生物学角度看，不同组学层面的数据具有互补性：RNA-seq反映基因转录活性，CNV反映基因剂量效应，DNA甲基化反映表观遗传调控，这些信息共同描绘了肿瘤发生发展的多层次机制。在卵巢癌研究中，综合考虑基因组、表观组和转录组数据的多组学分析可发现传统单组学难以识别的亚型特征，从而提供对疾病机理的新认识。从算法表现来看，整合多组学往往提升聚类质量指标和生存分析的显著性。CDCM的实验结果显示，与单组学相比，整合后的亚型在log-rank统计量上显著提高，这表明多组学融合有效利用了不同数据类型间的互补信息，提高了亚型划分的鲁棒性和临床相关性。类似地，Wang等[35]的研究中使用SNF方法融合多组学样本相似性网络，实验证明融合后在多个癌症数据集上的亚型识别和生存预测效果明显优于单组学分析。此外，Chen等[36]提出的DMCL模型在10个多组学癌症数据集上的比较中也显示出优于其他方法的性能，进一步验证了对比学习和聚类损失联合优化在多组学子型识别中的有效性。多组学的整合使得CDCM能够结合多源信息，在保持模型判别能力的同时提高了临床预后的显著性与聚类质量。

CDCM模型具备良好的移植性和表达能力，因此可以应用于其他癌种的分子分型。但不同癌种在分子驱动机制上存在显著差异，需要针对性地选择和优化待整合的组学模态，并根据不同的组学模态选择不同的预处理方式。使用该模型应用到新的癌种时，建议先开展小规模的模态贡献评估，用于量化每种组学对分型性能和临床相关性的增益，并据此确定优先整合的模态组合。

虽然CDCM已经被证明能够提供可靠的卵巢癌亚型标签，但在聚类性能上仍然存在一定的改进空间。本工作主要利用分子组学数据进行分析，尚未整合放射组学或病理图像等多模态信息，因此无法利用影像学所提供的肿瘤形态学和组织学特征，可能遗漏这部分对临床诊断具有价值的信息。对于肿瘤内和肿瘤间的高度异质性，本模型虽然通过对比学习增强了鲁棒性，但对极端小样本类别或离群样本的识别能力仍有提升空间。在未来研究中，笔者将通过改进组学数据的整合策略，将癌症影像信息与组学数据相融合，持续提升方法性能。

【参考文献】

[1] Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023[J]. CA Cancer J Clin, 2023, 73(1): 17-48.

[2] 廉鑫, 范典, 郑博豪, 等. 卵巢癌的分子分型及其在临床应用中的研究进展[J]. 现代妇产科进展, 2022, 31(7): 542-544.
Lian X, Fan D, Zheng BH, et al. Advances in molecular subtyping of ovarian cancer and its clinical applications[J]. Progress in Obstetrics and Gynecology, 2022, 31(7): 542-544.

[3] Correa-Aguila R, Alonso-Pupo N, Hernández-Rodríguez EW. Multi-omics data integration approaches for precision oncology[J]. Mol Omics, 2022, 18(6): 469-479.

[4] Huo ZG, Ding Y, Liu S, et al. Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies [J]. J Am Stat Assoc, 2016, 111(513): 27-42.

[5] Qarmiche N, El Kinany K, Otmani N, et al. Cluster analysis of dietary patterns associated with colorectal cancer derived from a Moroccan case-control study[J]. BMJ Health Care Inform, 2023, 30(1): e100710.
[6] Shen RL, 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.
[7] Monti S, Tamayo P, Mesirov J, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data[J]. Mach Learn, 2003, 52(1): 91-118.
[8] International Business Machines Corporation, Rensselaer Polytechnic Institute. Semi-supervised vertical federated learning: US 2023/0342655 A1[P]. 2023-10-26.
[9] Tian T, Zhang J, Lin X, et al. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data[J]. Nat Commun, 2021, 12(1): 1873.
[10] Yang H, Chen R, Li DD, et al. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data [J]. Bioinformatics, 2021, 37(16): 2231-2237.
[11] Guo LY, Wu AH, Wang YX, et al. Deep learning-based ovarian cancer subtypes identification using multi-omics data[J]. Bio Data Min, 2020, 13: 10.
[12] Chaudhary K, Poirion OB, Lu LQ, et al. Deep learning-based multi-omics integration robustly predicts survival in liver cancer[J]. Clin Cancer Res, 2018, 24(6): 1248-1259.
[13] Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 14313-14323.
[14] Gong RL, Wang LL, Wang J, et al. Self-distilled supervised contrastive learning for diagnosis of breast cancers with histopathological images [J]. Comput Biol Med, 2022, 146: 105641.
[15] Han WK, Cheng YQ, Chen JY, et al. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis[J]. Brief Bioinform, 2022, 23(5): bbac377.
[16] Zheng MJ, Hu YX, Gou R, et al. Integrated multi-omics analysis of genomics, epigenomics, and transcriptomics in ovarian carcinoma[J]. Aging (Albany NY), 2019, 11(12): 4198-4215.
[17] Dalmaijer ES, Nord CL, Astle DE. Statistical power for cluster analysis [J]. BMC Bioinformatics, 2022, 23(1): 205.
[18] Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis[J]. Bioinformatics, 2005, 21(15): 3201-3212.
[19] Lengyel A, Botta-Dukát Z. Silhouette width using generalized mean-a flexible method for assessing clustering efficiency[J]. Ecol Evol, 2019, 9(23): 13231-13243.
[20] Kakushadze Z, Yu W. K-means and cluster models for cancer signatures [J]. Biomol Detect Quantif, 2017, 13: 7-31.
[21] Zhang XY, Zhou ZW, Xu HF, et al. Integrative clustering methods for multi-omics data[J]. Wiley Interdiscip Rev Comput Stat, 2022, 14(3): e1553.
[22] Chierici M, Bussola N, Marcolini A, et al. Integrative network fusion: a multi-omics approach in molecular profiling[J]. Front Oncol, 2020, 10: 1065.
[23] Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets[J]. Mol Syst Biol, 2018, 14(6): e8124.
[24] Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates [C]//Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 160-172.
[25] Tang SY, Guo T, Song CC, et al. MGA loss-of-function variants cause premature ovarian insufficiency[J]. J Clin Invest, 2024, 134(22): e183758.
[26] Zhao Y, Zhang DN, Meng B, et al. Integrated proteomic and glycoproteomic analysis reveals heterogeneity and molecular signatures of brain metastases from lung adenocarcinomas[J]. Cancer Lett, 2024, 605: 217262.
[27] Xu JH, Wang Y, Xu D. CKS2 promotes tumor progression and metastasis and is an independent predictor of poor prognosis in epithelial ovarian cancer[J]. Eur Rev Med Pharmacol Sci, 2019, 23 (8): 3225-3234.
[28] Su F, Zhou FF, Zhang T, et al. Quantitative proteomics identified 3 oxidative phosphorylation genes with clinical prognostic significance in gastric cancer[J]. J Cell Mol Med, 2020, 24(18): 10842-10854.
[29] Cai XP, Zhou JM, Deng JW, et al. Prognostic biomarker SMARCC1 and its association with immune infiltrates in hepatocellular carcinoma [J]. Cancer Cell Int, 2021, 21(1): 701.
[30] Zhang L, Zhou Q, Qiu QZ, et al. CircPLEKHM3 acts as a tumor suppressor through regulation of the miR-9/BRCA1/DNAJB6/KLF4/AKT1 axis in ovarian cancer[J]. Mol Cancer, 2019, 18(1): 144.
[31] Agarwal N, Dancik GM, Goodspeed A, et al. GON4L drives cancer growth through a YY1-androgen receptor-CD24 axis[J]. Cancer Res, 2016, 76(17): 5175-5185.
[32] Luo WX, Han YS, Li X, et al. Breast cancer prognosis prediction and immune pathway molecular analysis based on mitochondria-related genes[J]. Genet Res (Camb), 2022, 2022: 2249909.
[33] Zhang BY, Liu QX, Wen WJ, et al. The chromatin remodeler CHD6 promotes colorectal cancer development by regulating TMEM65-mediated mitochondrial dynamics via EGF and Wnt signaling[J]. Cell Discov, 2022, 8(1): 130.
[34] Liao D, Cui YM, Shi LJ, et al. USP34 regulates PIN1-cGAS-STING axis-dependent ferroptosis in cervical cancer via SUMOylation[J]. Int Immunopharmacol, 2025, 147: 113968.
[35] 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.
[36] Chen WL, Wang H, Liang C. Deep multi-view contrastive learning for cancer subtype identification [J]. Brief Bioinform, 2023, 24(5): bbad282.

（编辑：薛泽玲）