论著·智能医学与预测模型

MoACG:一种基于自注意力机制和门控融合的多组学与临床数据整合模型用于多癌种预后预测

秦茂洋,沈俊杰,王龙昊,郭泓麟,伍亚舟 陆军军医大学(第三军医大学)军事预防医学系军队卫生统计学教研室,重庆

［**摘要**］ **目的** 旨在对癌症多组学数据和临床数据进行深度融合和分析,从而提高癌症预后的预测能力。**方法** 获取TCGA数据库卵巢癌、肝癌及结直肠癌3个癌种的组学数据(mRNA、lncRNA和miRNA数据)和临床数据;构建一种新的癌症预后预测模型MoACG(multi-omics attention clinical gating Model),其基于自注意力机制(self-attention mechanism)探讨不同组学之间的潜在关联,利用门控融合(gated fusion)自适应地融合多组学信息和临床信息(年龄、性别、治疗方式);通过与多种机器学习方法对比验证模型的有效性,最后利用可解释性算法DeepLIFT量化基因对模型贡献,筛选核心预后基因。**结果** 在卵巢癌、肝癌及结直肠癌数据集中,五折交叉验证的AUC值分别为(0.793±0.042)、(0.791±0.065)和(0.789±0.086);AUPR值分别为(0.915±0.020)、(0.855±0.058)和(0.917±0.039);综合性能优于其他9种机器学习模型。消融实验证明,3种组学数据整合模型在各癌种中均展现出最优预测性能。DeepLIFT算法筛选出与肝癌相关的*MED8*、*DLGAP4*、*NABP2*等基因,与已有研究结论高度吻合,且能够通过表达水平有效区分患者的生存风险(P<0.005)。**结论** 相较于既往研究,利用组学数据和临床数据构建的MoACG可有效提高癌症预后预测能力,为癌症诊疗、预后研究提供新方案。

［**关键词**］ 自注意力机制;癌症多组学;门控融合;癌症预后;可解释性
［**中图法分类号**］ R730.21;TP391;R319 ［**文献标志码**］ A

MoACG: A self-attention and gated fusion-based multi-omics and clinical data integration model for pan-cancer prognosis prediction

QIN Maoyang, SHEN Junjie, WANG Longhao, GUO Honglin, WU Yazhou Department of Military Health Statistics, College of Military Preventive Medicine, Army Medical University (Third Military Medical University), Chongqing, China

［**Abstract**］ **Objective** To perform in-depth integration and analysis of cancer multi-omics data and clinical data to enhance the predictive capability for cancer prognosis. **Methods** Multi-omics data (mRNA, lncRNA, and miRNA profiles) and clinical data for 3 cancer types, namely ovarian cancer, liver cancer, and colorectal cancer, were retrieved from The Cancer Genome Atlas (TCGA) database. A novel cancer prognosis prediction model, MoACG (Multi-omics Attention Clinical Gating Model), was constructed based on the self-attention mechanism to explore potential associations among different omics layers, and gated fusion was employed to adaptively integrate multi-omics information with clinical information(age, sex, and treatment modality). The effectiveness of the model was validated through comparison with multiple machine learning

［**基金项目**］ 国家自然科学基金项目(82173621,82574207)
［**通信作者**］ 伍亚舟, E-mail: asiawu@tmmu.edu.cn

2097-0927/Copyright © 2026, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议 (https://creativecommons.org/licenses/by/4.0/)
引用格式:秦茂洋,沈俊杰,王龙昊,等. MoACG:一种基于自注意力机制和门控融合的多组学与临床数据整合模型用于多癌种预后预测[J]. 陆军军医大学学报,2026,48(06):809-821. DOI:10.16016/j.2097-0927.202512061.
Citation:QIN M Y,SHEN J J,WANG L H,*et al*. MoACG: A self-attention and gated fusion-based multi-omics and clinical data integration model for pan-cancer prognosis prediction[J]. J Army Med Univ,2026,48(06):809-821. DOI:10.16016/j.2097-0927.202512061.

methods, and the interpretability algorithm DeepLIFT was utilized to quantify gene contributions to the model and identify core prognostic genes. **Results** In the ovarian cancer, liver cancer, and colorectal cancer datasets, five-fold cross-validation yielded the area under curve (AUC) values of receiver operating characteristic (ROC) curve of (0.793±0.042), (0.791±0.065), and (0.789±0.086), respectively, and the AUC values of precision-recall curve (AUPR) were (0.915±0.020), (0.855±0.058), and (0.917±0.039), respectively. The comprehensive performance surpassed that of 9 other machine learning models. Ablation experiments demonstrated that the 3-omics data integration model exhibited optimal predictive performance across all cancer types. The DeepLIFT algorithm identified *MED8*, *DLGAP4*, and *NABP2* as genes associated with liver cancer, showing high concordance with existing research findings and effectively stratifying patient survival risk based on expression levels ($P<0.005$). **Conclusion** Compared with previous studies, the MoACG model, constructed by integrating multi-omics and clinical data, effectively enhances the predictive performance for cancer prognosis, thereby providing a novel approach for cancer diagnosis, treatment, and prognostic research.

[**Key words**] self-attention mechanism; cancer multi-omics; gated fusion; cancer prognosis; interpretability

Supported by the National Natural Science Foundation of China (82173621, 82574207). Corresponding author: WU Yazhou, E-mail: asiawu@tmmu.edu.cn

全球范围内,癌症是仅次于心血管疾病的第二大死因,2022年全球近2 000万新发癌症病例,970万人死于癌症[1]。其中,结直肠癌(colon adenocarcinoma,COAD)、肝癌(liver hepatocellular carcinoma,LIHC)和卵巢癌(ovarian cancer,OV)死亡率排名第2、第3和第14[2],严重威胁人类生命健康。COAD、LIHC作为高发病高死亡癌种,前者术后复发率高[3]、后者易多中心复发[4];OV虽死亡率相对较低,但大多数病例确诊时已达晚期[5];若无法提前预判预后风险,易错失干预窗口。精准预后预测对于开发有效的癌症治疗方法和潜在的个性化药物至关重要[6]。然而,传统预后评估仅依赖病理分期、临床指标等单一维度信息,难以应对肿瘤分子异质性带来的挑战[7]。即使同一病理类型、相同分期的癌症患者,其基因表达、突变模式等分子特征差异仍会导致预后分化显著[8-9]。高通量测序技术的发展为突破这一瓶颈提供了支撑,基因组学、转录组学等多组学数据,能够从分子层面揭示癌症发生、发展及转移的调控机制[10]。

尽管多组学数据在癌症预后预测中展现出显著价值,但现有研究仍存在局限。一方面,传统多组学数据整合方法难以有效挖掘不同组学间的潜在相关。而随着深度学习技术中的发展,特别是自注意力机制可通过动态计算注意力权重,精准捕捉不同组学间的潜在关联,从而提升预测性能。如BEAUDE[11]提出了一种新的基于交叉注意机制的深度学习架构crosssattomics可以通过捕获患者不同组学图谱之间的相互作用来准确预测癌症的类型;SUN[12]提出一个基于自注意力机制的深度学习框架SADLN不仅可以整合多组学数据,还可以自适应地建模样本之间的关系,从而准确地学习潜在的低维表示。另一方面,大多数模型仅依赖多组学数据构建,未充分融合临床数据[13]。忽视临床数据会导致模型脱离实际诊疗场景,难以指导实际诊疗决策。门控融合通过引入门控单元来控制不同数据源或特征融合的过程[14],能依据数据贡献度动态调节组学与临床数据的融合权重[15-16]。如GONG[17]等提出一种多级门控模态融合模型,通过门控机制充分保留各模态特有的解剖学特征与空间剂量分布信息,最终实现治疗剂量的精准预测。

深度学习模型虽在预测中展现出优异性能,但其"黑箱"特性导致的可解释性差,严重制约了模型筛选生物标志物的临床转化价值[18]。而DeepLIFT(deep learning important feaTures)算法可通过计算输入特征对模型预测结果的贡献度,量化解析关键特征,有效弥补了模型可解释性不足的缺陷[19]。如TASAKI等[20]借助DeepLIFT算法计算基因结合位点对预测结果的累计贡献分数,以此量化不同调控因子的重要性。

因此,本研究构建了基于自注意力机制和门控融合的癌症多组学与临床数据整合模型MoACG(multi-omics attention clinical gating model),应用于OV、LIHC及COAD的mRNA、miRNA、lnRNA组学数据和临床数据的整合研究。利用注意力机制探讨不同组学之间的潜在关联,然后利用门控融合自适应地融合多组学信息和临床信息,旨在解决不同组学间关联性挖掘不足、组学信息与临床信息割裂、模型可解释性不足等问题,进而提升癌症预后预测的精准性与临床适用性,为个体化诊疗决策提供理论支撑。

1　材料与方法

1.1　数据收集与预处理

为了验证MoACG模型的有效性,本研究利用TCGA数据库中的OV、LIHC、COAD的多组学数据。具体来说,从TCGA数据库中下载3种癌症的mRNA、lncRNA和miRNA数据。本文对基因进行的处理是针对无意义特征,即如果某一特征超过20%的样本中信息值为0,将其从数据中剔除。匹配出同时具有三种数据的样本,并进行归一化处理。基于数据完整性与临床意义双重考量,本研究纳入的临床信息包括性别、年龄和治疗方式,具体如表1所示。然后根据癌症样本的临床数据,将存活超过24个月的患者(无论生存状态如何)视为长期生存(long-term survivals,LTS),将死亡且存活时间少于24个月的患者视为短期生存(short-term survivals,Non-LTS),生存时间少于24个月的在世患者被排除在本研究之外[21],如表2所示。

1.2　模型构建

为解决组学数据间关联性挖掘不足、组学与临床信息未充分融合、模型可解释性欠缺等问题。本研究构建的模型主要包括四个模块:降维模块、组学注意力模块、门控融合与分类模块和可解释性模块,模型结构图如图1所示。

1.2.1　降维模块　　组学数据通常具有高维、高噪声的特性,易引发“维度诅咒”,而通过降维可在保留关键生物信息的基础上精简数据维度,不仅能减轻计算负担,还能提升预测任务的准确性与高效性[22]。随机森林可通过构建T棵决策树集成学习模型,基于特征重要性Gini系数增益筛选特征。Gini系数增益的计算如公式如下:

$$I(g)=\frac{1}{T}\sum_{t=1}^{T}[G(t)-\frac{N_1}{N}G(t_1)-\frac{N_2}{N}G(t_2)] \tag{1}$$

其中,$I(g)$为特征g的重要性得分,T为决策树数量,t_1、t_2为节点t分裂后的子节点,N为节点t的样本数,N_1、N_2为子节点t_1、t_2的样本数。$G(t)$为决策树节点t的Gini不纯度,计算公式如下:

$$G(t)=1-\sum_{c=1}^{C}p(c|t)^2 \tag{2}$$

其中,C为类别数,$p(c|t)$为节点t中属于类别c的样本比例。

本研究根据随机森林特征重要性排序筛选出重要性排名前200的mRNA数据和lncRNA数据以及排名前100的miRNA数据,3种癌症数据组学特征如表2所示。

1.2.2　组学自注意力模块　　通过计算不同组学特征间的注意力权重来探讨不同组学间的潜在关联[23]。具体来说,先将三个组学特征映射至低维空间

$$h=Linear_k(\sigma(Linear_m(X))) \tag{3}$$

其中σ为非线性激活函数,m为中间维度,k为目标维度,此处k值为128维,X为组学特征,$X\in\mathbb{R}^{n\times d}$($n$为样本数,$d$为特征数)。

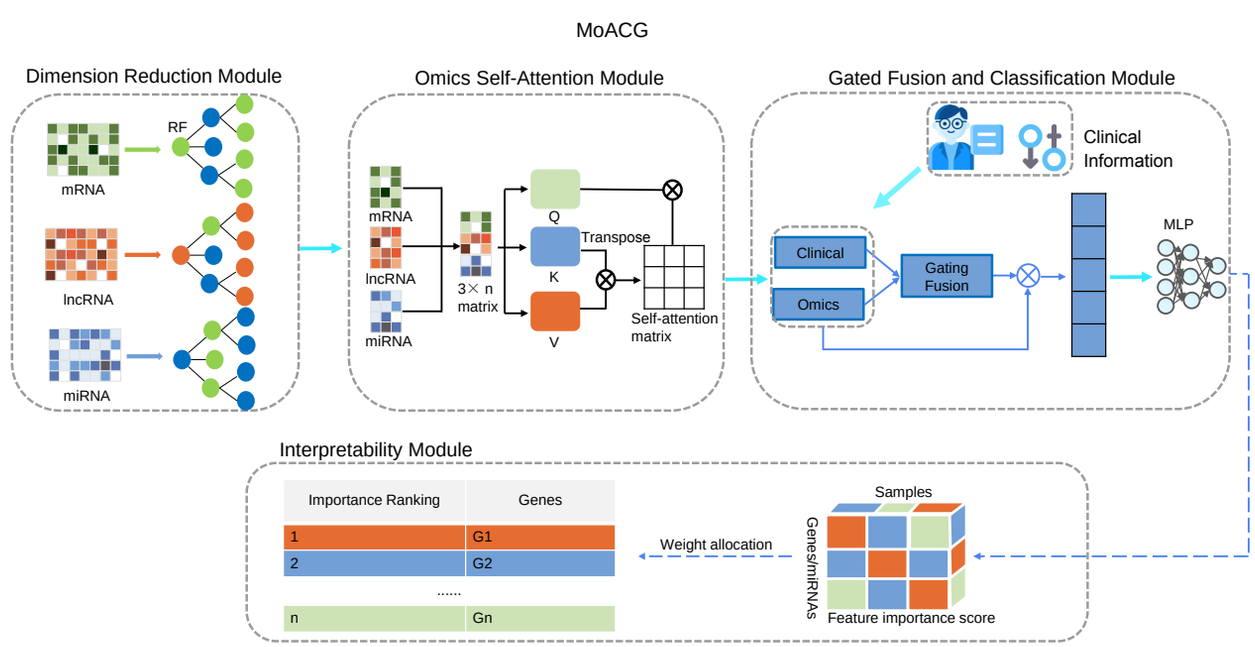


图1　MoACG模型结构图

Figure 1　Schematic diagram of the MoACG model

表1 3种不同类型癌症数据集的临床特征

Table 1 Clinical features of 3 different types of cancer datasets

Datasets	Clinical features	Non-LTS	LTS
OV	Age/years		
	<60	29(17.68%)	135(82.32%)
	≥60	49(30.63%)	111(69.37%)
	Treatment		
	Pharmaceutical	33(22.30%)	115(77.70%)
	Radiation	45(25.57%)	131(74.43%)
	Median survival time/d	486	1 662
LIHC	Age/years		
	<60	40(39.22%)	62(60.78%)
	≥60	44(36.67%)	76(63.33%)
	Gender		
	Female	28(35.44%)	51(64.56%)
	Male	56(39.16%)	87(60.84%)
	Treatment		
	Pharmaceutical	44(36.67%)	76(63.33%)
	Radiation	40(39.22%)	62(60.78%)
	Median survival time/d	302	2 532
COAD	Age/years		
	<60	13(22.41%)	45(77.59%)
	≥60	34(19.43%)	141(80.57%)
	Gender		
	Female	21(19.81%)	85(80.19%)
	Male	26(20.47%)	101(79.53%)
	Treatment		
	Pharmaceutical	22(19.64%)	90(80.36%)
	Radiation	25(20.66%)	96(79.34%)
	Median survival time/d	256	3 042

LTS: Long-term survivals; Non-LTS: Short-term survivals; OV: Ovarian cancer; LIHC: Liver hepatocellular carcinoma; COAD: Colon adenocarcinoma colorectal cancer

得到低维度的组学特征之后,将三个组学通过线性投影生成查询(Q)、键(K)、值(V)向量

$$Q = hW_Q + b_1 \tag{4}$$

$$K = hW_K + b_2 \tag{5}$$

$$V = hW_V + b_3 \tag{6}$$

其中h为组学特征,W为权重矩阵,b为偏置项。

通过自注意力机制公式计算出不同组学之间的相关性,再将得到的三个矩阵通过*concat*函数进行拼接,通过线性层压缩到128维,形成统一的组学融合特征,为后续与临床特征结合做准备。

$$AttentionX = soft\max\left(\frac{Q\times K^T}{\sqrt{d_K}}\right)V \tag{7}$$

$$h_{omics} = Linear_{128}(concat(AttentionX_{mRNA}, AttentionX_{miRNA}, AttentionX_{\ln cRNA})) \tag{8}$$

1.2.3 门控融合与分类模块 组学特征与临床特征存在维度异构,直接拼接易导致临床特征维度尺度失衡。本研究通过门控机制生成动态权重,对拼接后的特征逐维度分配权重平衡两组数据的贡献[24]。先将组学特征和临床特征*concat*函数进行拼接,得到拼接后的特征h_{comb}

$$h_{comb} = concat(h_{omics} + h_{clinical}) \tag{9}$$

通过一个全连接层进行线性转换,应用Sigmoid激活函数得到门控值g

$$g = Sigmoid(Linear_{131}(h_{comb})) \tag{10}$$

使用门控权重筛选有效特征

$$h_{fused} = g \odot h_{comb} \tag{11}$$

其中,h_{fused}表示融合后的特征,⊙表示逐元素相乘

将融合后的特征输入到多层感知机(MLP, Multilayer Perceptron),用于预测癌症预后

$$y_i = MLP(h_{fused}) \tag{12}$$

损失函数采用交叉熵损失函数

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)] \tag{13}$$

1.2.4 可解释模块 可解释性模块采用DeepLIFT算法,DeepLIFT的核心思想是计算输入特征相对于一个参考输入的变化,以及这种变化如何导致模型输出相对于参考输出的变化[25]。在MLP中,神经元的输出y与输入$x_1, x_2, \cdots\cdots, x_n$的函数$y=f(x_1, x_2, \cdots\cdots, x_n)$。参

表2 3种不同类型癌症数据集的组学特征

Table 2 Omics features of 3 different types of cancer datasets

Datasets	Sample	LTS	Non-LTS	Before dimensionality reduction (mRNA, lncRNA, miRNA)	After dimensionality reduction (mRNA, lncRNA, miRNA)
OV	324	246	78	16 685,6 531,1 879	200,200,100
LIHC	222	138	84	15 358,3 996,1 879	200,200,100
COAD	233	186	47	15 775,4 105,1 879	200,200,100

LTS: Long-term survivals; Non-LTS: Short-term survivals; OV: Ovarian cancer; LIHC: Liver hepatocellular carcinoma; COAD: Colon adenocarcinoma colorectal cancer

考输入记为 $x_1^0,x_2^0,...,x_n^0$，参考输出记为 $y^0=f(x_1^0,x_2^0,...,x_n^0)$。输出变化量$\Delta y=y-y^0$可以通过输入的变化量$\Delta x_i=x_i-x_i^0$来解释。DeepLIFT定义了一个贡献分数$\delta y_i$,表示输入$x_i$的变化对输出变化$\Delta y$的贡献,并且满足加性规则:

$$\Delta y=\sum_{i=1}^{n}\delta y_i \quad (14)$$

根据贡献度排序筛选出排名前30的mRNA作为核心基因,然后进行GO功能注释分析和KEGG信号通路分析。并通过核心基因,采用中位数分层策略,将样本按基因表达水平划分为低危组(表达值低于中位数)与高危组(表达值高于中位数),通过Kaplan-Meier生存曲线与对数秩检验,评估基因表达与患者生存结局的关联性。

1.3 模型评价指标

本文采用准确率(accuracy, ACC)、召回率(recall)、F1分数(F1-Score)、受试者工作特征(receiver operating characteristic, ROC)曲线下面积(area under curve, AUC)和精确率-召回率曲线(precision-recallcurve, PR)下面积(AUPR)来评价模型对预后的预测性能[26]。

1.4 模型比较与消融实验

为全面评估MoACG模型的有效性,本研究选取9种机器学习分类方法,开展系统性性能对比实验,具体包括逻辑回归(logistic regression,LR)、随机森林(random forest,RF)、支持向量机(support vector machine,SVM)、极端梯度提升(extreme gradient boosting,XGBoost)、决策树(decision tree,DT)、K近邻(K-nearest neighbors,KNN)、朴素贝叶斯(naive Bayes,NB)、多层感知器(multi-layer perceptron,MLP)及自适应提升(adaptive boosting,AdaBoost),所有机器学习方法均基于直接拼接组学信息和临床信息作为输入。且都采用5折交叉验证策略,确保实验结果的可靠性与泛化性。然后开展消融实验,评估不同组学数据组合和不同模块组合下模型在癌症预后预测任务中的性能表现。

1.5 实现工具及参数设置

本研究基于Python 3.12平台和PyTorch 2.2.2框架构建和训练模型,采用R 4.2.0完成生信分析和生存分析。本研究模型训练过程中,核心超参数设置如下:学习率(learning rate)设置为0.000 1;权重衰减系数(weight decay)为0.000 1,用于缓解模型过拟合;批次大小(batch size)设为64;训练迭代轮数(epoch)为150轮,直至模型收敛且性能稳定。

2 结果

2.1 MoACG模型预后预测性能优于其他模型

实验结果表明,相比于其他9种模型,MoACG在各个癌症数据集的主要指标上都具有较好的性能。在OV、LIHC、COAD数据中,五折交叉的AUC平均值分别为(0.793±0.042)、(0.791±0.065)和(0.789±0.086);AUPR值分别为(0.915±0.020)、(0.855±0.058)和(0.917±0.039)。MoACG模型的综合性能均优于其他模型,MoACG在3个数据集的具体性能表现如下表3、表4和表5。

2.2 miRNA+lncRNA+mRNA三组学结合的MoACG模型性能最优

不同组学数据对模型预测性能的贡献可能存在差异。本研究以OV、LIHC数据为例,通过比较单一组学、两组学构建的模型,比较不同组学组合的模型性能。三组学整合模型在OV、LIHC数据中预测效能最佳,AUC值分别达0.793、0.791。不同组学数据组合

表3 MoACG模型在OV数据的性能比较

Table 3 Comparison of the MoACG model's performance on OV data

Models	AUC	ACC	Recall	F1-score	AUPR
DT	0.505±0.013	0.735±0.017	0.947±0.025	0.844±0.013	0.761±0.007
KNN	0.589±0.060	0.765±0.011	1.000±0.000	0.866±0.006	0.795±0.025
AdaBoost	0.635±0.053	0.744±0.036	0.878±0.043	0.838±0.027	0.826±0.036
NB	0.669±0.027	0.732±0.026	0.784±0.053	0.815±0.025	0.838±0.012
XGboost	0.702±0.075	0.735±0.026	0.923±0.015	0.841±0.014	0.881±0.036
RF	0.745±0.096	0.756±0.024	0.988±0.024	0.860±0.016	0.886±0.056
SVM	0.752±0.072	0.790±0.048	0.903±0.036	0.867±0.030	0.900±0.022
LR	0.776±0.074	0.827±0.031	0.976±0.024	0.896±0.018	0.906±0.026
MLP	0.781±0.057	0.806±0.040	0.943±0.015	0.881±0.023	0.908±0.014
MoACG	0.793±0.042	0.800±0.060	0.907±0.035	0.873±0.037	0.915±0.020

OV: Ovarian cancer; AUC: Area under receiver operating characteristic curve; ACC: Accuracy; AUPR: Area under precision-recall curve

表4 MoACG模型在LIHC数据的性能比较

Table 4 Comparison of the MoACG model's performance on LIHC data

Models	AUC	ACC	Recall	F1-Score	AUPR
DT	0.597±0.021	0.626±0.027	0.727±0.127	0.703±0.046	0.673±0.019
KNN	0.639±0.047	0.640±0.022	0.906±0.048	0.758±0.013	0.713±0.040
NB	0.700±0.032	0.712±0.020	0.805±0.073	0.775±0.021	0.747±0.029
SVM	0.760±0.059	0.716±0.078	0.791±0.110	0.773±0.074	0.819±0.060
XGboost	0.760±0.036	0.676±0.032	0.783±0.073	0.749±0.025	0.845±0.038
AdaBoost	0.762±0.012	0.689±0.030	0.791±0.081	0.758±0.036	0.827±0.029
MLP	0.771±0.060	0.748±0.042	0.856±0.059	0.808±0.032	0.826±0.069
LR	0.784±0.054	0.712±0.050	0.827±0.096	0.779±0.046	0.844±0.064
RF	0.789±0.025	0.685±0.036	0.791±0.098	0.754±0.044	0.858±0.031
MoACG	0.791±0.065	0.738±0.043	0.804±0.055	0.791±0.065	0.855±0.058

LIHC: Liver hepatocellular carcinoma; AUC: Area under receiver operating characteristic curve; ACC: Accuracy; AUPR: Area under precision-recall curve

表5 MoACG模型在COAD数据的性能比较

Table 5 Comparison of the MoACG model's performance on COAD data

Models	AUC	ACC	Recall	F1-Score	AUPR
DT	0.512±0.028	0.777±0.024	0.957±0.021	0.873±0.014	0.802±0.018
KNN	0.600±0.052	0.803±0.008	1.000±0.000	0.890±0.005	0.834±0.015
NB	0.682±0.031	0.781±0.046	0.866±0.065	0.862±0.032	0.865±0.013
XGboost	0.693±0.083	0.785±0.012	0.941±0.010	0.875±0.007	0.899±0.041
LR	0.709±0.045	0.833±0.015	0.973±0.034	0.903±0.011	0.871±0.032
MLP	0.716±0.041	0.824±0.017	0.957±0.027	0.897±0.009	0.878±0.033
SVM	0.717±0.056	0.811±0.016	0.925±0.020	0.886±0.011	0.875±0.041
AdaBoost	0.721±0.019	0.803±0.035	0.946±0.034	0.885±0.020	0.920±0.006
RF	0.766±0.069	0.807±0.031	0.973±0.024	0.890±0.016	0.928±0.021
MoACG	0.789±0.086	0.837±0.050	0.952±0.020	0.904±0.028	0.917±0.039

COAD: Colon adenocarcinoma colorectal cancer; AUC: Area under receiver operating characteristic curve; ACC: Accuracy; AUPR: Area under precision-recall curve

中,OV中lncRNA单组学(AUC:0.740)与LncRNA+miRNA双组学(AUC:0.773)和LIHC中mRNA单组学(AUC:0.763)与mRNA+miRNA双组学(AUC:0.774)都分别表现突出,三组学结合的MoACG模型性能最优。不同组学组合性能比较如图2和图3所示。

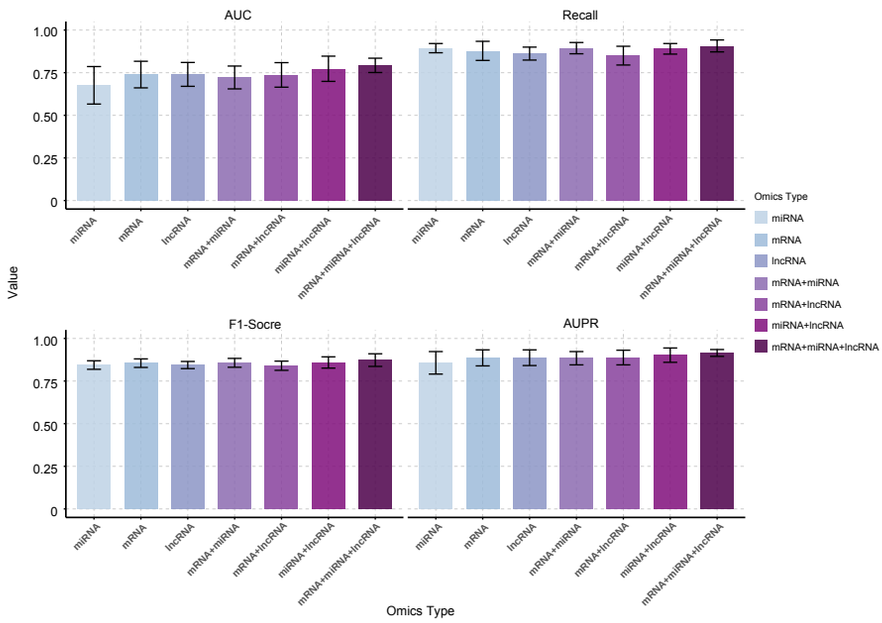
2.3 消融实验显示各个模块都对模型预测性能具有关键作用

为探究组学注意力模块、门控融合模块及临床信息对模型预测性能的影响,本研究开展系统性消融实验,以OV数据为例,通过不同模块组合进行性能评估,如图4。结果显示,组学注意力模块与临床信息结合模型MoAC的AUC为0.764;门控融合模块联合临床信息模型MoCG的AUC值达0.749;而组学注意力模块与门控融合模块组合模型MoAG的AUC值相对较低,为0.747。相较于其他模块组合,临床信息融入模型后,均展现出较高的预测效能,表明其在提升癌症预后预测性能方面具有关键作用。

2.4 模型可解释性分析结果

在本研究中,应用DeepLIFT算法对模型输入的基因特征进行重要性评估,以LIHC数据为例,依据计算所得的贡献度得分对基因进行排序。

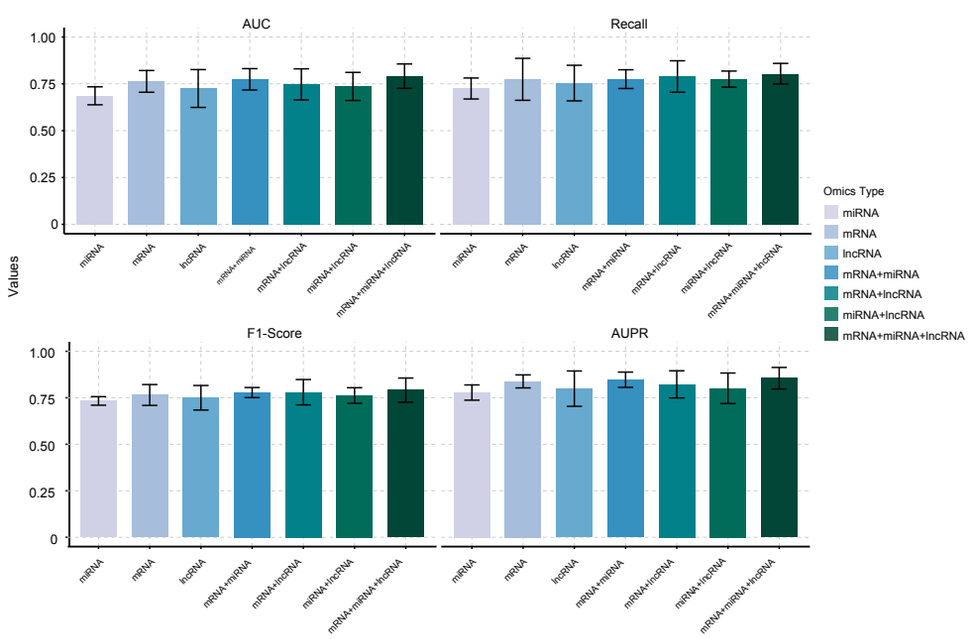
通过对各个基因的贡献得分分布进行分析,如图5所示,我们发现其符合生物网络中典型的幂律分布特征:少数基因表现出显著的重要性(长尾分布),而多数基因贡献度相对较低。进一步分析显示(表6),贡献度最高的前12个基因中包含多个已被既往研究验证的与肝癌预后和进展有关的分子。LIHC核心基因GO功能注释分析和KEGG信号通路分析如图6、图7所示。



OA: Ovarian cancer

图2　OV不同组学性能比较

Figure 2　Comparison of different omics' performance in OV



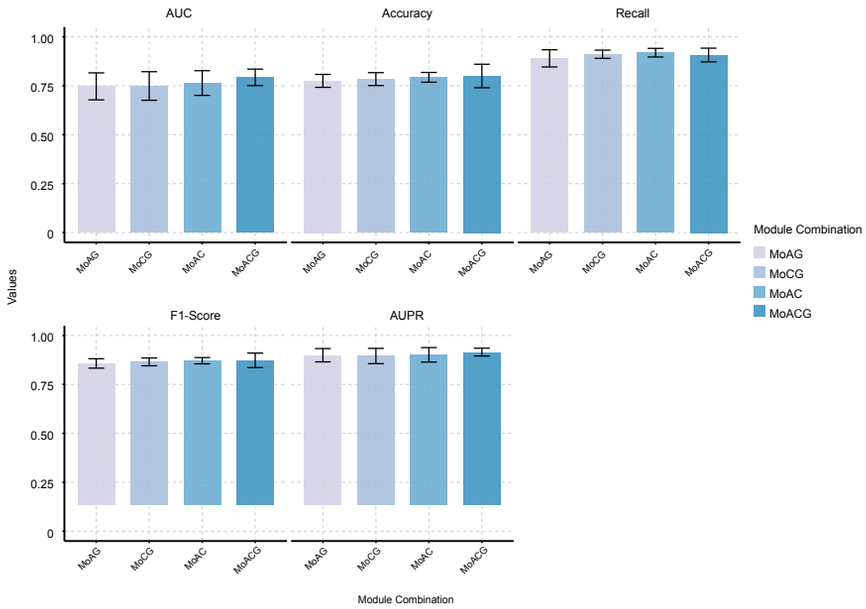
LIHC: Liver hepatocellular carcinoma

图3　LIHC不同组学性能比较

Figure 3　Comparison of different omics' performance in LIHC

2.5　MoACG模型筛选出的高排名基因在高低危组间生存差异显著

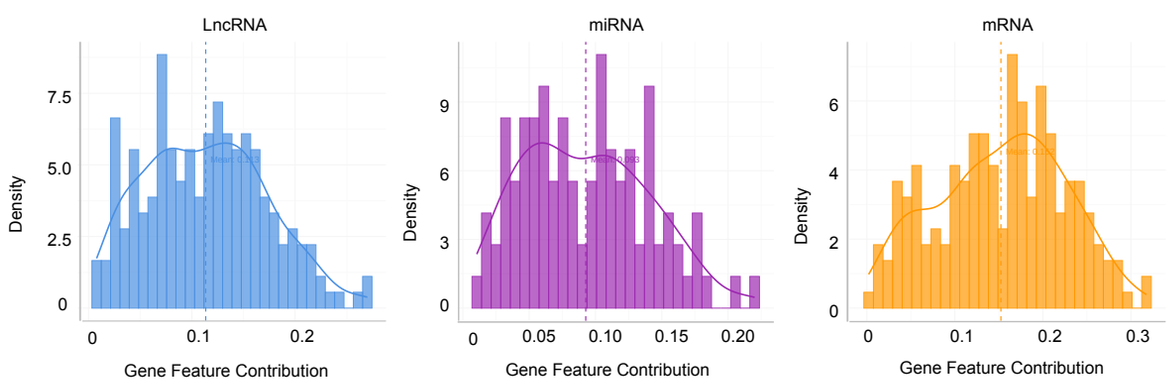
为验证MoACG模型所鉴定生物标志物对患者生存的影响,本研究基于LIHC的mRNA数据表达开展生存分析。Kaplan-Meier生存曲线与对数秩检验分析结果显示,MoACG模型筛选出的部分高排名基因(如*ZCRB1*、*MED8*、*DLGAP4*、*NABP2*)在高低危组间呈现显著的生存差异(P<0.005),Kaplan-Meier生存曲线如图8所示。



OV: Ovarian cancer

图4 消融实验(以OV数据为例)

Figure 4 Ablation experiments (using OV data as an example)



LIHC: Liver hepatocellular carcinoma

图5 LIHC基因贡献度得分分布

Figure 5 Distribution of gene contribution scores in LIHC

3 讨论

癌症预后是遗传变异、分子异常及临床干预共同作用的复杂过程[27],精准的预后评估对于制定个体化治疗方案、优化医疗资源至关重要。相较于主要依赖临床指标数据的探索研究,当前基于基因组、转录组、蛋白组与代谢组等多维组学数据的研究能够更系统揭示疾病特异性细胞群体的功能,有利于识别癌症诊断标志物和发现新的癌症治疗靶点[28-29]。基于组学数据的预后预测对临床诊疗决策具有重要指导意义,已成为当前研究的热点问题[30]。在OV、LIHC、COAD等多种癌症的研究中,虽已有部分预后预测探索,但对多组学数据与临床信息的融合分析仍不够深入。本研究通过整合上述癌种的多组学与临床数据,并引入自注意力机制和门控融合构建融合预测模型,旨在提升癌症预后预测的准确性,为个体化诊疗提供新的分析框架。

3.1 自注意力机制能捕捉不同组学之间的潜在关联

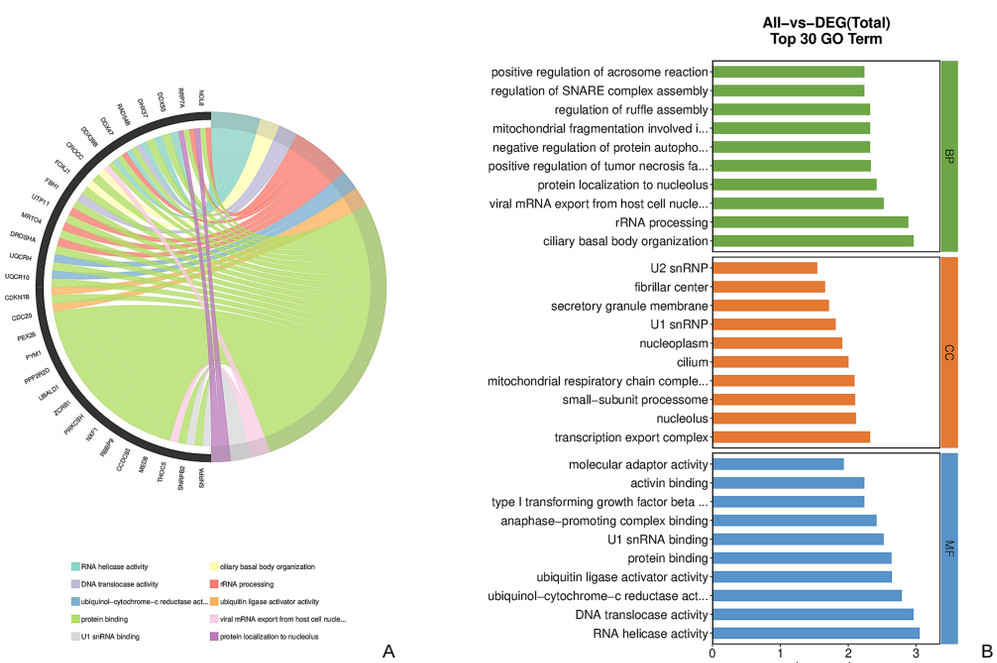
在既往研究中,随着高通量测序技术和人工智能技术的快速发展,以深度学习架构为代表的组学分析方法被广泛用于癌症的预后预测[31]。MAO等[32]通过构建患者相似图、嵌入蛋白质编码基因特征并结合多层图卷积建立癌症预后预测模型Cox-Sage;JOSE等[33]提出了DeepGraphMut方法,该方法利用图自编码器整

表6　LIHC数据的mRNA重要性排序

Table 6　mRNA importance ranking on LIHC data

Importance ranking	Gene	Importance score	Literature evidence
1	*PEX26*	0.317737350944182	DAHABIEH, *et al*[41](2022)
2	*PYM1*	0.3143433331717848	
3	*PPP2R2D*	0.29392422249565137	ZHOU, *et al*[42](2014)
4	*UBALD1*	0.28999300628514113	
5	*ZCRB1*	0.28891978621220493	SONG, *et al*[43](2022)
6	*PRKCSH*	0.27972822561875693	WILLS, *et al*[40](2016)
7	*NXF1*	0.2754720085324621	CHI, et al[36](2022)
8	*RBBP9*	0.27088858931924636	HAMADA, *et al*[44](2025)
9	*CCDC93*	0.2687791777913094	
10	*MED8*	0.2633822612576154	NI, *et al*[37](2024)
11	*DLGAP4*	0.2622264222264224	DONG, *et al*[38](2022)
12	*NABP2*	0.2606472463714001	DAI, *et al*[39](2025)

LIHC: Liver hepatocellular carcinoma; *PEX26*: Peroxisomal biogenesis factor 26; *PYM1*: PYM homolog 1 exon junction complex associated factor; *PPP2R2D*: Protein phosphatase 2 regulatory subunit bdelta; *UBALD1*: UBA-like domain-containing 1; *ZCRB1*: Zinc finger CCHC-type and RNA binding motif containing 1; *PRKCSH*: Protein kinase C substrate 80K-H; *NXF1*: Nuclear RNA export factor 1; *RBBP9*: Retinoblastoma binding protein 9; *CCDC93*: Coiled-coil domain containing 93; *MED8*: Mediator complex subunit 8; *DLGAP4*: DLG associated protein 4; *NABP2*: Nucleic acid binding protein 2



A: Chord diagram of GO enrichment analysis; B: Bar plot of GO enrichment analysis. GO: Gene ontology; LIHC: Liver hepatocellular carcinoma

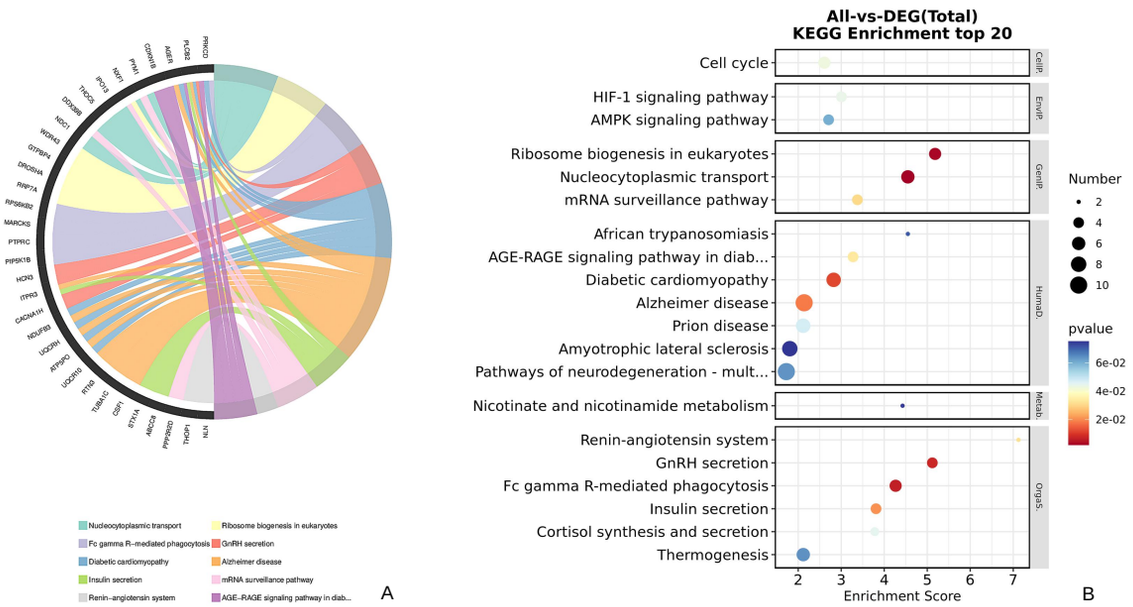
图6　LIHC核心基因GO功能注释分析

Figure 6　GO functional annotation analysis of core genes in LIHC

合体细胞突变数据与蛋白质-蛋白质相互作用(PPI)网络。虽然这些方法在癌症预后预测方面表现出良好的性能,但是他们忽略了不同组学之间的潜在关联,可能会丢失部分有意义的信息。MoACG模型的注意力模块通过“查询-键-值”动态权重分配机制,能够自适应学习不同癌症类型组学数据之间的潜在关联,优化特征提取并剔除冗余噪声,进而提升模型的预测性能、泛化能力,为多组学整合分析提供支撑。

3.2　门控融合模块的优势

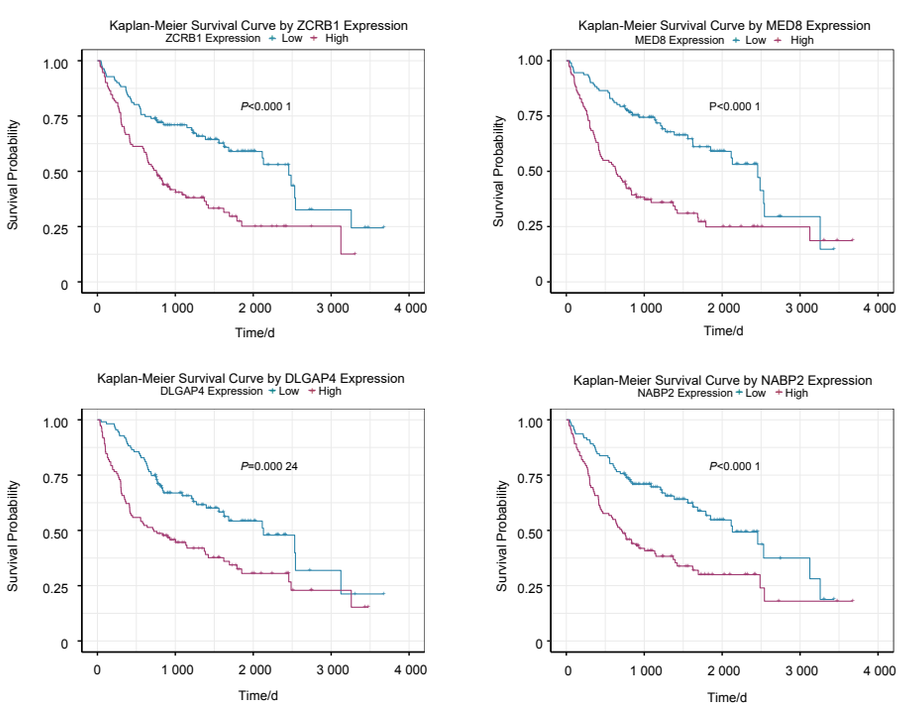
本文预测结果揭示,MoACG模型可有效预测癌症



A: Chord diagram of KEGG enrichment analysis; B: Bubble plot of KEGG enrichment analysis. KEGG: Kyoto encyclopedia of genes and genomes; LIHC: Liver hepatocellular carcinoma

图7 LIHC核心基因KEGG信号通路分析

Figure 7 KEGG signaling pathway analysis of core genes in LIHC



LIHC: Liver hepatocellular carcinoma; *ZCRB1*: Zinc finger CCHC-type and RNA binding motif containing 1; *MED8*: Mediator complex subunit 8; *DLGAP4*: DLG associated protein 4; *NABP2*: Nucleic acid binding protein 2

图8 LIHC核心基因表达的Kaplan-Meier生存曲线

Figure 8 Kaplan-meier survival curves of core gene expression in LIHC

预后,在OV、LIHC、COAD 3个独立数据集的预后预测中均展现最优性能,进一步证实MoACG模型在癌症研究中的实用性和重要性。在去除门控融合模块后,模型效果大幅下降,这也证明了门控融合模块的重要性。门控融合模块的优势可总结如下[34-35]:①平衡高维组学数据与低维临床数据异构特征的贡献度,避免临床信息被组学维度掩盖。②动态分配特征权重,自适应挖掘多源数据的互补价值。③替代数据简单拼接,提升融合特征质量,为模型精准预测奠定基础。

3.3 可解释性模块能筛选出癌症预后相关的基因特征

本研究引入DeepLIFT可解释性算法,量化解析MoACG模型在癌症数据集预后预测中的关键决策依据,筛选出的关键特征,其结果不仅与已有研究结论高度吻合,更揭示了癌症预后的特异性分子机制。以肝癌的mRNA为例,贡献度最高的前12个基因中包含多个已被既往研究验证的与肝癌预后和进展有关的分子。如根据CHI等[36]的研究显示,*RAE1*是一种肝细胞癌患者预后生物标志物,而蛋白-蛋白互作分析显示,*RAE1*与*NXF1*密切相关。NI等[37]证实*MED8*不仅与肝细胞癌的病理进程密切相关,其表达水平对肝细胞癌的预后预测也展现出强大诊断价值。DONG[38]等证实*DLGAP4*作为肝细胞癌的有效预后预测因子,与肿瘤进展密切相关。DAI[39]等证实*NABP2*可能既是肝细胞癌的生物标志物,也是潜在的治疗靶点。常染色体显性遗传性多囊肝病由*PRKCSH*[40]等基因变异引起。另有部分基因与肝癌的关联性虽尚未经实验验证明确证实,但其在其他恶性肿瘤中的异常表达模式、功能调控作用及临床预后价值已被现有研究证实。如*PEX26*[41]的低表达水平表达与癌症患者生存期延长显著相关,*PPP2R2D*[42]的低表达可抑制T细胞凋亡、增强T细胞的增殖能力及细胞因子分泌水平,*ZCRB1*[43]的高表达能有效抑制胶质母细胞瘤异种移植瘤的生长。*RBBP9*[44]的缺失会增加结直肠癌的发生易感性。

在筛选与肝癌预后相关的核心基因后,本文进行了两类富集分析。GO分析基因功能集中在核酸转运与加工、细胞器组织及酶活性调控等,而KEGG分析提示基因功能集中在细胞基础生命活动、神经内分泌调节及疾病相关病理过程等。这些生信分析结果揭示了LIHC细胞在核糖体生物合成、核质转运及RNA代谢等细胞活动中的异常性,并提示了与癌症进展相关的通路。

当然,本研究构建的MoACG模型仍存在一定局限性:其一,模型开发与验证数据均来源于公共数据库,样本存在地域集中性与检测平台异质性,且缺乏多中心、前瞻性临床队列的外部验证,可能限制模型的临床泛化能力;其二,临床信息纳入维度相对有限,当前仅涵盖性别、年龄、治疗方法等基础指标,未纳入肿瘤分级分期、淋巴结转移状态、术后并发症及动态随访数据等关键临床变量。

综上,MoACG模型通过"组学关联挖掘-异构特征平衡-可解释性验证"的三层次框架,不仅提升了癌症预后预测性能,更构建了"算法预测-生物标志物解析-临床意义阐释"的完整研究路径,为实现癌症精准诊疗提供了具有可解释性与临床转化潜力的方法学支持。

利益冲突 本文所有作者声明不存在利益冲突

参考文献:

[1] FORCE L M, KOCARNIK J M, MAY M L, *et al*. The global, regional, and national burden of cancer, 1990-2023, with forecasts to 2050: a systematic analysis for the Global Burden of Disease Study 2023[J]. Lancet, 2010, 406(10512): 1565-1586. DOI: 10.1016/S0140-6736(25)01635-6.

[2] BRAY F, LAVERSANNE M, SUNG H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2024, 74(3): 229-263. DOI: 10.3322/caac.21834.

[3] TAKAGI M, KIM S, SUZUKI M, *et al*. Factors associated with postoperative recurrence in perforated colorectal cancer: unraveling the high recurrence rate of perforated colorectal cancer[J]. World J Surg Oncol, 2025, 23(1): 139. DOI: 10.1186/s12957-025-03783-5.

[4] GIANNONE F, CASSESE G, CUBISINO A, *et al*. Patterns of site and timing of recurrence after curative resection in single and multiple large hepatocellular carcinoma: a multicenter international comprehensive analysis[J]. Ann Surg Oncol, 2025, 32(9): 6312-6323. DOI: 10.1245/s10434-025-17673-3.

[5] SALAS BOLÍVAR P, GONZALEZ-BENITEZ C, CARBONELL LÓPEZ M, *et al*. Prognostic factors after the first recurrence of ovarian cancer[J]. J Clin Med, 2025, 14(2): 470. DOI: 10.3390/jcm14020470.

[6] 李鹏鹏, 陈锡程, 黄金羽, 等. 基于转录组学数据和堆叠监督自编码器的肺腺癌预后预测研究[J]. 陆军军医大学学报, 2023, 45(6): 579-585. DOI: 10.16016/j.2097-0927.202212025.
LI P P, CHEN X C, HUANG J Y, *et al*. Prognostic prediction of lung adenocarcinoma based on transcriptomic data and stacked supervised autoencoder[J]. J Army Med

Univ, 2023, 45(6): 579-585. DOI: 10.16016/j.2097-0927.202212025.

[7] SEOANE J, DE MATTOS-ARRUDA L. The challenge of intratumour heterogeneity in precision medicine[J]. J Intern Med, 2014, 276(1): 41-51. DOI: 10.1111/joim.12240.

[8] 李佳芮, 钱力, 沈俊杰, 等. 基于转录组学和变分自编码器的癌症分期诊断研究[J]. 陆军军医大学学报, 2025, 47(6): 613-622. DOI: 10.16016/j.2097-0927.202412079.
LI J R, QIAN L, SHEN J J, *et al*. Cancer staging diagnosis based on transcriptomics and variational autoencoder[J]. J Army Med Univ, 2025, 47(6): 613-622. DOI: 10.16016/j.2097-0927.202412079.

[9] QIAO T, YIN B N, LIANG J, *et al*. Implications of T cell-mediated tumor killing genes for molecular heterogeneity and clinical stratification in lung adenocarcinoma[J]. Genes Dis, 2024, 11(4): 101162. DOI: 10.1016/j.gendis.2023.101162.

[10] TEREKHANOVA N V, KARPOVA A, LIANG W W, *et al*. Epigenetic regulation during cancer transitions across 11 tumour types[J]. Nature, 2023, 623(7986): 432-441. DOI: 10.1038/s41586-023-06682-5.

[11] BEAUDE A, AUGÉ F, ZEHRAOUI F, *et al*. CrossAttOmics: multiomics data integration with cross-attention[J]. Bioinformatics, 2025, 41(6): btaf302. DOI: 10.1093/bioinformatics/btaf302.

[12] SUN Q W, CHENG L, MENG A, *et al*. SADLN: Self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition[J]. Front Genet, 2023, 13: 1032768. DOI: 10.3389/fgene.2022.1032768.

[13] YE X C, SHI T Y, HUANG D, *et al*. Multi-Omics clustering by integrating clinical features from large language model[J]. Methods, 2025, 239: 64-71. DOI: 10.1016/j.ymeth.2025.03.017.

[14] PENG Q Q, HE Z H, LIU S C, *et al*. Integrating multi-omics data into a gated graph convolutional networks for identifying cancer driver genes and function modules[C]// 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). December 5-8, 2023. Istanbul, Turkiye. IEEE, 2023. DOI: 10.1109/bibm58861. 2023.10385796.. DOI: 10.1109/bibm58861.2023.10385796.

[15] YU Z G, LI Z X, LI P. MultiRepPI: a cross-modal feature fusion-based multiple characterization framework for plant peptide-protein interaction prediction[J]. BMC Plant Biol, 2025, 25(1): 933. DOI: 10.1186/s12870-025-06878-z.

[16] XU Y, ZHANG L, SHEN X. Multi-modal adaptive gated mechanism for visual question answering[J]. PLoS One, 2023, 18(6): e0287557. DOI: 10.1371/journal. pone.0287557.

[17] GONG C F, HUANG Y L, JIAN J M, *et al*. Measurement-guided therapeutic-dose prediction using multi-level gated modality-fusion model for volumetric-modulated arc radiotherapy[J]. Front Oncol, 2025, 15: 1468232. DOI: 10.3389/fonc.2025.1468232.

[18] WANG F G, ZHUANG Z F, GAO F, *et al*. TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology[J]. Genome Biol, 2024, 25(1): 149. DOI: 10.1186/s13059-024-03293-9.

[19] SHRIKUMAR A, GREENSIDE P, KUNDAJE A B. Learning important features through propagating activation differences [C]//International Conference on Machine Learning., 2017

[20] TASAKI S, GAITERI C, MOSTAFAVI S, *et al*. Deep learning *Decodes* the principles of differential gene expression[J]. Nat Mach Intell, 2020, 2(7): 376-386. DOI: 10.1038/s42256-020-0201-6.

[21] HAO J, KIM Y, KIM T K, *et al*. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data [J]. BMC Bioinformatics, 2018, 19(1): 510. DOI: 10.1186/s12859-018-2500-z.

[22] YAN H X, WENG D W, LI D G, *et al*. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation *via* multi-omics data integration [J]. Brief Bioinform, 2024, 25(3): bbae184. DOI: 10.1093/bib/bbae184.

[23] ZHANG G, PENG Z, YAN C K, *et al*. MultiGATAE: a novel cancer subtype identification method based on multi-omics and attention mechanism[J]. Front Genet, 2022, 13: 855629. DOI: 10.3389/fgene.2022.855629.

[24] ZHAO C M, GAO F, SHEN Z H. Multi-motion sensor behavior based continuous authentication on smartphones using gated two-tower transformer fusion networks [J]. Comput Secur, 2024, 139: 103698. DOI: 10.1016/j.cose.2023.103698.

[25] ELMARAKEBY H A, HWANG J, ARAFEH R, *et al*. Biologically informed deep neural network for prostate cancer discovery[J]. Nature, 2021, 598(7880): 348-352. DOI: 10.1038/s41586-021-03922-4.

[26] LAN W, LIAO H B, CHEN Q F, *et al*. DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery[J]. Brief Bioinform, 2024, 25(3): bbae185. DOI: 10.1093/bib/bbae185.

[27] SUN R, JUN A, YU H L, *et al*. Proteomic landscape profiling of primary prostate cancer reveals a 16-protein panel for prognosis prediction[J]. Cell Rep Med, 2024, 5(8): 101679. DOI: 10.1016/j.xcrm.2024.101679.

[28] ZHENG Y Y, WANG G H, JIANG A. Explore the expression of mitochondria-related genes to construct prognostic risk model for ovarian cancer and validate it, so as to provide optimized treatment for ovarian cancer[J]. Front Immunol, 2024, 15: 1458264. DOI: 10.3389/fimmu.2024.1458264.

[29] WU Y, XU Y J, CHEN Y T, *et al*. Development of a prognostic prediction model based on damage-associated

molecular pattern for colorectal cancer applying bulk RNA-seq analysis[J]. Sci Rep, 2025, 15(1): 25792. DOI: 10.1038/s41598-025-10592-z.

[30] CUI H T, WANG C, MAAN H, *et al*. scGPT: toward building a foundation model for single-cell multi-omics using generative AI[J]. Nat Methods, 2024, 21(8): 1470-1480. DOI:10.1038/s41592-024-02201-0.

[31] DONG B T, ZHANG H, DUAN Y Y, *et al*. Development of a machine learning-based model to predict prognosis of alpha-fetoprotein-positive hepatocellular carcinoma [J]. J Transl Med, 2024, 22(1): 455. DOI:10.1186/s12967-024-05203-w.

[32] MAO R J, WAN L, ZHOU M H, *et al*. Cox-Sage: enhancing Cox proportional hazards model with interpretable graph neural networks for cancer prognosis [J]. Brief Bioinform, 2025, 26(2): bbaf108. DOI: 10.1093/bib/bbaf108.

[33] JOSE A, SRIVASTAVA A, ANSARI A, *et al*. DeepGraphMut: a graph-based deep learning method for cancer prognosis using somatic mutation profile [J]. Brief Bioinform, 2025, 26(4): bbaf409. DOI: 10.1093/bib/bbaf409.

[34] YING Y, WU N, HUO J H, *et al*. scUCAF: an uncertainty-aware cross-omics alignment and fusion network for single-cell multi-omics data clustering [J]. Comput Biol Chem, 2026, 120: 108631. DOI: 10.1016/j.compbiolchem.2025.108631.

[35] WANG J H, LIAO N Q, DU X F, *et al*. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks[J]. BMC Genomics, 2024, 25(1): 86. DOI:10.1186/s12864-024-09985-7.

[36] CHI G, PEI J H, LI X Q. RAE1 is a prognostic biomarker and is correlated with clinicopathological characteristics of patients with hepatocellular carcinoma [J]. BMC Bioinformatics, 2022, 23(1): 252. DOI: 10.1186/s12859-022-04806-8.

[37] NI Y, LU M Q, LI M, *et al*. Unraveling the underlying pathogenic factors driving nonalcoholic steatohepatitis and hepatocellular carcinoma: an in-depth analysis of prognostically relevant gene signatures in hepatocellular carcinoma [J]. J Transl Med, 2024, 22(1): 72. DOI: 10.1186/s12967-024-04885-6.

[38] DONG C R, HUANG S L, SUN L, *et al*. DLGAP4 acts as an effective prognostic predictor for hepatocellular carcinoma and is closely related to tumour progression [J]. Sci Rep, 2022, 12(1): 19775. DOI:10.1038/s41598-022-23837-y.

[39] DAI C C, TONG Y, BAI N, *et al*. Decoding the role of nucleic acid binding protein 2 in lipid dysregulation and hepatocellular carcinoma progression through LKB1-mediated mitochondrial dysfunction[J]. Cell Signal, 2025, 132: 111820. DOI:10.1016/j.cellsig.2025.111820.

[40] WILLS E S, CNOSSEN W R, VELTMAN J A, *et al*. Chromosomal abnormalities in hepatic cysts point to novel polycystic liver disease genes[J]. Eur J Hum Genet, 2016, 24(12): 1707-1714. DOI:10.1038/ejhg.2016.97.

[41] DAHABIEH M S, HUANG F, GONCALVES C, *et al*. Silencing PEX26 as an unconventional mode to kill drug-resistant cancer cells and forestall drug resistance [J]. Autophagy, 2022, 18(3): 540-558. DOI: 10.1080/15548627.2021.1936932.

[42] ZHOU P H, SHAFFER D R, ALVAREZ ARIAS D A, *et al*. *In vivo* discovery of immunotherapy targets in the tumour microenvironment [J]. Nature, 2014, 506(7486): 52-57. DOI:10.1038/nature12988.

[43] SONG J, ZHENG J, LIU X B, *et al*. A novel protein encoded by ZCRB1-induced circHEATR5B suppresses aerobic glycolysis of GBM through phosphorylation of JMJD5 [J]. J Exp Clin Cancer Res, 2022, 41(1): 171. DOI: 10.1186/s13046-022-02374-6.

[44] HAMADA K, NAKANISHI Y, YU M T, *et al*. Retinoblastoma-binding protein 9 suppresses intestinal inflammation and inflammation-induced tumorigenesis in mice[J]. Cell Mol Gastroenterol Hepatol, 2025, 19(3): 101435. DOI:10.1016/j.jcmgh.2024.101435.

(收稿:2025-12-11;修回:2026-01-26)

(编辑 邓强庭)