https://crad.ict.ac.cn
63(11): 2976−2985, 2026

基于文本-图模态的癌症驱动基因识别方法

韩有瑞[1] 萨娅·赛立克[1] 艾克热木·乃比江[1] 孟维华[1] 胡悠鹏[1] 李子源[1] 尚学群[1,2] 陈伯林[1,2]

[1](西北工业大学计算机学院 西安 710072)

[2](大数据存储与管理工信部重点实验室(西北工业大学) 西安 710072)

(yrhan@mail.nwpu.edu.cn)

Text-Image Modal-Based Method for Identifying Cancer Driver Genes

Han Yourui[1], Saya Sailike[1], Aikeremu Naibijiang[1], Meng Weihua[1], Hu Youpeng[1], Li Ziyuan[1], Shang Xuequn[1,2], and Chen Bolin[1,2]

[1] (*School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an* 710072)

[2] (*Key Laboratory of Big Data Storage and Management* (*Northwestern Polytechnical University*), *Ministry of Industry and Information Technology*, *Xi'an* 710072)

Abstract The identification of cancer driver genes is an important research area in oncology research, holding significant value for a deeper understanding of the mechanisms of cancer development and achieving precision medicine. Currently, most research methods primarily rely on graph data to capture the characteristics of driver genes. However, the evolution process of cancer development involves many complex interactions among various biomolecules, making it difficult to fully reveal its essence using information from only single modality. Therefore, how to efficiently integrate multi-modal biological data to comprehensively identify driver genes has become a critical issue that needs to be addressed. To this end, we propose a cancer driver gene identification method, titled as TGMNN (text-graph modal neural network), based on a combination of text and graph modality. This method builds upon the graph modality by further incorporating text semantic information related to genes and employs an interactive attention mechanism to achieve collaborative fusion of multi-modal data, thereby reducing the differences between modalities. Compared with existing methods, TGMNN can capture the information interactions more effectively between modalities, significantly improving the accuracy of driver gene predictions. Experimental results show that AUC value of this method reaches 0.897, outperforming existing methods and high-lighting the model's effectiveness.

Key words driver gene identification; graph neural networks; multimodal fusion; interactive attention; text-graph modal

摘 要 癌症驱动基因的识别是肿瘤学研究中的一个重要方向,对于深入解析癌症的发生机制及实现精准医疗具有重要价值。目前,大多数研究方法主要依赖图数据来捕捉驱动基因的特征,但是癌症的发生过程涉及多种生物分子之间复杂的相互作用,仅凭单一模态信息难以全面揭示其本质。因此,如何高效整合多模态生物数据以全面刻画驱动基因,成为亟待解决的关键问题。为此,提出了一种基于文本-图模态的癌症驱动基因识别方法 TGMNN(text-graph modal neural network)。该方法在图模态的基础上,进一

收稿日期: 2025-10-21; **修回日期**: 2026-07-09

基金项目: 国家重点研发计划项目(2021YFA1000402); 国家自然科学基金项目(62433016); 中央高校基本科研业务费专项资金(D5000250279, G25KY0612610)

This work was supported by the National Key Research and Development Program of China (2021YFA1000402), the National Natural Science Foundation of China (62433016), and the Fundamental Research Funds for the Central Universities (D5000250279, G25KY0612610).

通信作者: 陈伯林(blchen@nwpu.edu.cn)

步引入与基因相关的文本语义信息，并采用交互注意力机制实现多模态的协同融合，从而降低模态之间的差异。与现有方法相比，TGMNN 能够更有效地捕捉多模态之间的信息交互，显著提升驱动基因预测的准确性。实验结果表明，该方法的 AUC 值达到 0.897，优于现有方法，突显了模型的有效性。

关键词 驱动基因识别；图神经网络；多模态融合；交互注意力；文本-图模态

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550749 **CSTR:** 32373.14.issn1000-1239.202550749

癌症驱动基因的识别是肿瘤学研究的核心科学问题之一，对阐明癌症发生机制、推动精准医疗发展具有重要意义。早期驱动基因识别方法主要聚焦于序列特征与突变频率分析，例如 MutSigCV[1] 通过检测显著突变基因实现驱动基因筛选，但此类方法未能充分考虑基因间的相互作用关系。随着网络生物学的发展，基于网络的癌症驱动基因识别方法逐渐成为主流。该类方法将基因嵌入生物网络中，通过分析基因的拓扑结构特征与功能关联实现驱动基因筛选，典型代表如 DriverNet[2] 等。然而，现有多数基于网络的方法依赖人工设计特征，难以有效捕捉生物网络中复杂的高维信息，限制了识别性能的进一步提升。

为了解决传统基于网络方法的局限性，深度学习方法被引入到驱动基因的识别中。这类方法通过自动化学习特征提取过程，能够处理高维和非线性的数据，为突破传统方法瓶颈提供了新途径。在深度学习方法中，图神经网络（graph neural network, GNN）作为结合网络与深度学习的工具，在驱动基因识别中展现出强大的潜力[3]。

Song 等人[4] 提出图采集与聚合方法（GraphSage）采用半监督学习，首先通过图注意力（GAT）网络为每个节点生成新的表征，随后利用 2 层图采样聚合获得最终的节点表征[5]。为 GNN 在该领域的应用提供了重要参考。此后，多款基于 GNN 的驱动基因识别方法相继提出，例如 EMOGI[6] 通过图卷积网络整合多组学数据和蛋白质相互作用网络，有效提升了驱动基因预测精度；HGDC[7] 则结合层次化图结构和注意力机制，强化了关键基因的特征提取。然而，这些方法通常依赖于单阶矩信息传递，难以捕捉生物网络中基因之间的复杂关系，容易忽视患者驱动基因与正常基因丰度之间的差异。因此，Chen 等人[8] 基于混合矩图神经网络，通过多阶矩嵌入的信息传递与注意力机制，提出了 DMGNN 方法，以捕获更丰富的邻域分布信息。

然而，现有驱动基因识别方法仍局限于单一模态数据源，例如基因组变异数据、表观遗传信息或基因表达数据。尽管这些方法在特定场景下取得了一定成效，但无法全面捕捉驱动基因所处生物系统的复杂性。癌症发生机制具有多因素、多层次的特点，涉及多种生物分子间的相互作用及动态调控过程，仅依赖单一模态数据无法充分揭示其内在关联。因此，如何有效整合不同模态生物数据的优势，全面刻画驱动基因的潜在特性，已成为当前癌症研究领域的重要挑战之一。

为此，提出了一种基于文本-图模态的癌症驱动基因识别方法 TGMNN（text-graph modal neural network）。该方法利用基因功能描述的文本数据，采用 BERT 模型[9] 提取基因的语义嵌入；同时，以生物网络和多组学数据为基础，构建基因图数据，并采用 DMGNN 学习基因的网络表示。最终，通过对 2 种模态数据的分层融合，结合 2 分类任务的优化目标，实现对癌症驱动基因的有效识别。TGMNN 在公开的癌症驱动基因数据集上验证了其有效性，与单模态方法相比，该模型在 AUC 和 AUPRC 指标上达到了当前最优水平。

1 基础模型介绍

1.1 BERT 模块

BERT 模型是基于 Transformer 编码器的预训练语言模型[9]，通过双向上下文建模能够有效捕捉文本语义信息。其核心优势在于预训练过程中采用的掩码语言模型和下一句预测任务，使其能够学习到词语间的依赖关系和句子级语义关联。

基于预训练 BERT 模型，采用全参数微调，来适配癌症驱动基因的识别任务。全参数微调在训练中未冻结任何参数（包括在嵌入层与编码器各层），在保留 BERT 模型上下文表示能力的基础上，通过梯度反向传播动态调整参数实现端到端优化，进而有效识别文本数据中与基因相关的功能注释、信号通路及相互作用关系。此外，BERT 模型的分层结构支持多粒度语义建模：底层解析基因符号的局部语法；中

层整合短语与句子级语义,揭示功能间生物学逻辑;高层通过全局上下文建模,抽象功能与表型的跨层级关联。如此协同作用能够有效提取基因从术语到概念的全谱系信息,从而提升驱动基因的识别能力。

1.2 DMGNN 模块

DMGNN 模型是针对驱动基因识别设计的图神经网络模型[8],其核心在于结合多阶矩嵌入的信息传递模块与注意力机制,能够有效捕捉基因网络中的复杂拓扑结构和节点间的依赖关系,如图 1 所示。

DMGNN 的网络结构分为多个层次,第 1 层主要捕获节点的局部信息,通过聚合节点自身及其直接邻居的特征更新节点表示;随着网络层数加深,模型逐渐整合来自更远节点的特征,从而捕获更广泛的图结构信息。该模型通过多阶矩嵌入能够刻画节点邻域的分布特征,提升了对基因网络中复杂关系的建模能力,为图模态特征提取提供了可靠的基础架构。

为了与 BERT 模型各层的语义信息进行有效融合,因而移除了 DMGNN 模型中的隐层聚合操作,直接输出所需的隐层表征,以适配与 BERT 模型的特征融合。

2 数据处理

2.1 图模态数据

基于与 EMOGI[6] 方法相同的多组学数据源,整合了来自 TCGA[10] 项目的 16 种癌症类型共 30 000 例样本的基因突变频率、DNA 甲基化及基因表达数据。在生物网络构建方面,研究不仅采用了 STRING[11] 数据库,还整合了 CPDB[12], InBioMap[13], IntAct[14-15], iRefIndex[16] 这 4 个数据库进行训练。具体而言,针对 CPDB 数据库的边进行筛选,仅保留置信度评分高于 0.85 的相互作用关系,并剔除未在组学数据中出现的基因节点。

数据预处理后,最终构建的蛋白质相互作用 PPI 网络[17] 规模如表 1 所示。

2.2 文本模态数据

为构建驱动基因识别任务的文本数据集,从

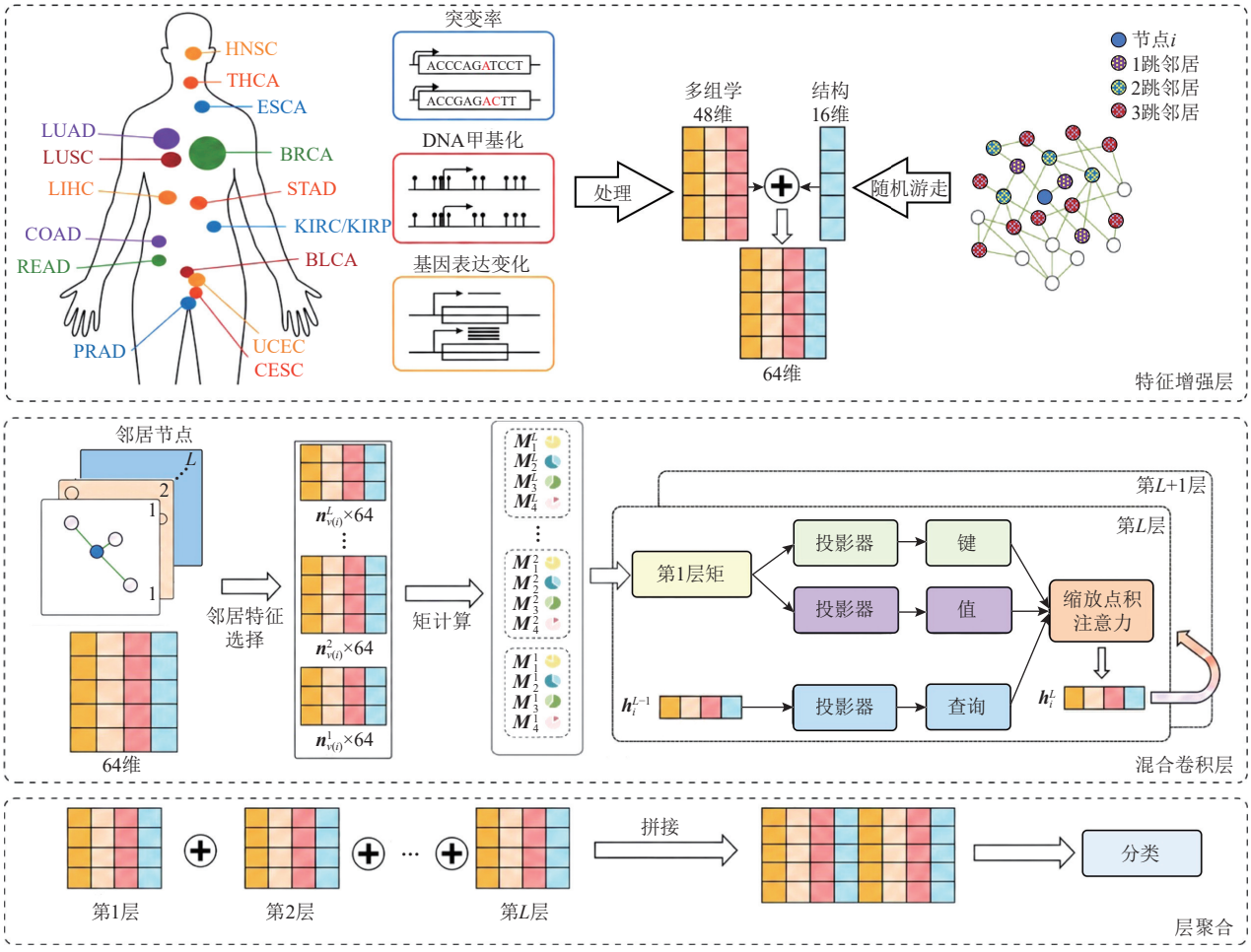


Fig. 1 DMGNN framework

图 1 DMGNN 框架

Table 1 Size of PPI Network Used by TGMNN

表 1 TGMNN 使用的 PPI 网络规模

网络名称	节点数量	数量
STRING	11 395	181 877
CPDB	12 590	313 707
IntAct	10 987	94 606
InBioMap	13 043	501 122
iRefIndex	11 784	115 431

GeneCards 数据库[18-19]爬取了基因功能描述(例如 CSMD1 的功能描述包含参与学习和记忆、乳腺分支发育及膜结构组成等内容)，并将其作为模型输入。然而，部分基因(如 TP53)的描述直接提及“肿瘤抑制蛋白”“癌症相关突变”等疾病信息，存在标签泄露风险。

因此，通过文本清洗流程，删除包含“癌”“肿瘤”等关键词的句子，仅保留蛋白应激响应机制、剪接变异等中性分子机制描述，以消除显性标签干扰，提升模型泛化能力。

3 TGMNN 模型

基于文本-图模态的癌症驱动基因识别方法 TGMNN，如图 2 所示，其核心架构包含特征提取模块、交叉注意力融合模块、层次融合模块及分类模块。

3.1 特征提取模块

TGMNN 模型集成了用于提取文本和图这 2 种模态特征的 BERT 模型和 DMGNN 模型，为后续融合提供基础。

针对文本模态特征提取，TGMNN 模型采用预训练 BERT 模型进行全参数微调以适配驱动基因识别任务。根据 BERT 模型的分层特性，分别提取第 1 层(首层)、第 6 层(中层)和第 12 层(末层)的特征输出，其中首层特征主要反映基因功能描述中的局部词汇信息，中层特征体现短语级语义关联，末层特征则刻画全局上下文语义。

针对图模态特征提取，TGMNN 模型采用 DMGNN 模型处理基因相互作用网络的拓扑结构特征。DMGNN 模型包含 3 层，首层主要聚合节点自身及直接邻居的局部特征，中层进一步整合 2 阶邻居节点信息，末层则捕获整个网络中的全局拓扑特征，形成不同粒度的图模态特征表征。

3.2 交叉注意力融合模块

TGMNN 模型引入跨模态注意力机制实现特征协同，通过动态建模文本与图模态间的交互关系，精准定位 2 类特征中具有语义关联的关键区域，并基于注意力权重对齐特征空间。以 BERT 模型第 l 层文本模态特征及 DMGNN 模型第 l 层图模态特征为例，交叉注意力融合模块以图模态特征作为查询 $\boldsymbol{Q}_l$，文本模态特征表征作为键 $\boldsymbol{K}_l$ 和值 $\boldsymbol{V}_l$，计算交叉注意力权重矩阵：

$$\boldsymbol{A}_l = softmax\left(\frac{\boldsymbol{Q}_l\boldsymbol{K}_l^{\mathrm{T}}}{\sqrt{d}}\right)。\tag{1}$$

进一步地，基于交叉注意力权重矩阵 $\boldsymbol{A}_l$ 计算跨

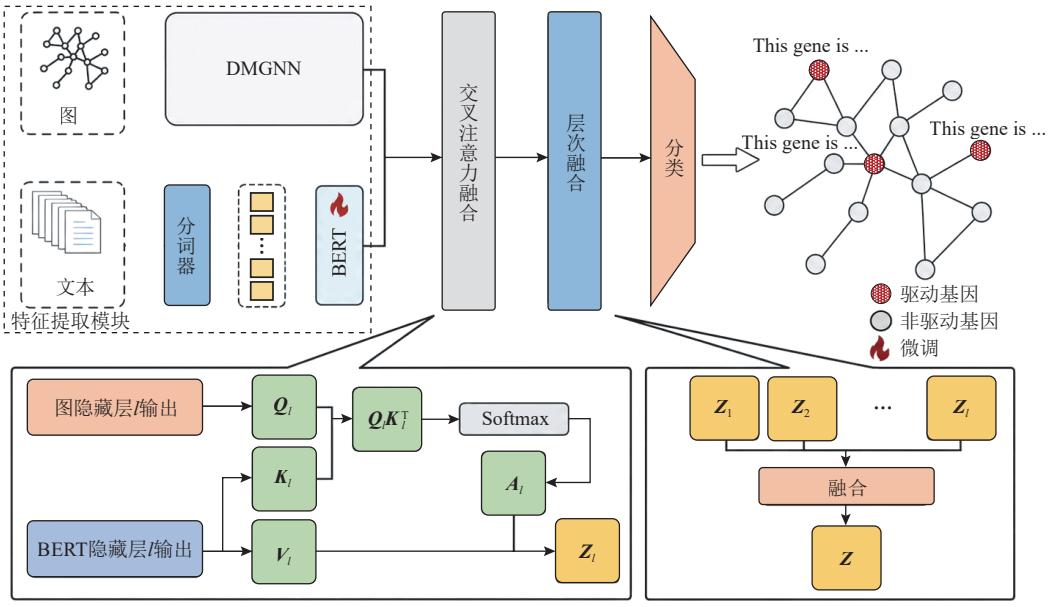


Fig. 2 TGMNN model architecture

图 2 TGMNN 模型架构

模态融合特征:

$$\boldsymbol{Z}_l = \boldsymbol{A}_l \boldsymbol{V}_l 。 \tag{2}$$

3.3 层次融合模块

为了有效整合不同层次隐层跨模态融合特征,本研究对不同隐层交叉融合后的跨模态融合特征进行聚合操作。常见的聚合机制如下:

1)最大融合机制

最大融合机制是一种广泛应用的池化方法,其原理在于从每个区域中提取最大值。在多模态融合后的多层次不同语义融合任务中,最大融合能够有效地从各层表示中筛选出最为重要的特征,从而辅助模型聚焦于最显著的信息。

$$MaxPool(\boldsymbol{X}) = \max(\boldsymbol{X}) , \tag{3}$$

其中$\boldsymbol{X}$是当前层的输出。

2)平均融合机制

与最大融合机制相似,平均池化通过计算每个池化区域的均值来实现数据降维。该方法能够平滑不同层次的信息,尤其在特征分布较为均匀的情况下,这种融合方式有助于更有效地保留信息。

$$AvgPool(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i , \tag{4}$$

其中$\boldsymbol{X}$表示融合后的输出,$\boldsymbol{X}_i$表示第i层输入的特征,n为需要融合的层数。

3)加权融合机制

加权融合机制在图神经网络中被广泛应用,旨在根据节点的重要性对各个元素进行加权处理。此机制能够使某些特征在融合过程中占据更为重要的地位,从而增强模型对关键特征的敏感性。权重的获取通常依赖于其他学习机制,例如注意力机制。

$$WeightedPool(\boldsymbol{X}) = \sum_{i=1}^{n}\alpha_i \boldsymbol{X}_i , \tag{5}$$

其中α_i为每个特征分配的权重,$\boldsymbol{X}_i$为节点的特征。

在层次融合过程中,采用了加权融合机制,其中每个权重由模型自主学习生成。通过对不同隐层交叉融合后的跨模态融合特征进行加权聚合,充分挖掘各层级特征的互补优势,提升模型表征能力。

3.4 分类模块及模型训练

分类模块使用全连接神经网络,输入为层次融合后的跨模态融合特征。TGMNN 模型所使用的损失函数与 DMGNN 一致,定义为带有 *Logit*的 2 元交叉熵,其损失函数为:

$$O = -L\log(\sigma(Logit)) - (1-L)\log(1-\sigma(Logit)) , \tag{6}$$

其中

$$\sigma(Logit) = \frac{1}{1+\mathrm{e}^{-Logit}} 。 \tag{7}$$

在文本模态训练中,由于 BERT 预训练大模型对单个样本的计算需求较高,直接输入全部数据容易导致内存溢出,因此,采用分批训练策略和梯度平均降低噪声影响,在训练效果与内存消耗之间取得平衡,以适应有限的硬件资源。在图模态训练中,每个批次均输入整张图及所有节点特征,利用文本模态的对应标签训练 DMGNN 模型,以确保 DMGNN 模型在较少的迭代轮次下获得合理效果。

4 结果分析

4.1 模型效果评估

本研究采用 10 次 5 折交叉验证的平均结果作为评估依据,选用 AUC(area under curve)和 AUPRC(area under the precision-recall curve)这 2 项核心指标量化模型性能。为系统验证 TGMNN 在驱动基因识别任务中的有效性及网络鲁棒性,将其与 EMOGI, MTGCN, HGDC 这 3 种主流基线模型进行对比实验,所有模型均基于 STRING, CPDB 等 5 个不同的 PPI 网络数据集完成训练。实验结果如表 2 所示。

TGMNN 在所有 PPI 数据集上均表现最佳,平均 AUC 较基线模型提升至少 3 个百分点,显示其显著优势。

为进一步验证 TGMNN 在不同癌症类型中的泛化能力,针对癌症异质性影响,选取乳腺癌、肺腺癌、膀胱癌、肝细胞癌这 4 种常见癌症进行单癌种测试。仅保留单癌组学数据后,TGMNN 在 AUC 和 AUPRC 指标上仍优于基线模型。测试结果如表 3 所示。

TGMNN 在这 4 种癌症类型中均保持最优性能,AUC 值均超过 0.87,其中在肝细胞癌中表现最佳,AUC 达到 0.892,较基线模型中表现最好的 HGDC 提升约 5.4 个百分点,显示出对该癌症类型驱动基因的强识别能力。

进一步分析不同癌症类型的预测差异发现,TGMNN 在肝细胞癌和膀胱癌中的性能提升更为显著,这可能与这 2 种癌症的驱动基因在文本功能描述和网络拓扑结构上的特征更为显著有关。而在肺腺癌中,驱动基因的异质性更强,单一模态难以全面刻画,而 TGMNN 通过多模态融合和层次融合机制,仍能有效整合互补信息,保持较高的预测准确性。总体而言,TGMNN 在不同癌症类型中的稳定表现,证明其具有较强的泛化能力,能够适应不同癌症的

Table 2 Comparative Analysis of TGMNN Results

表 2 TGMNN 结果对比分析

PPI 数据库	方法	AUC	AUPRC
STRING	TMGNN（本文）	**0.893±0.016**	**0.778±0.004**
	EMOGI	0.827±0.016	0.733±0.003
	MTGCN	0.810±0.013	0.709±0.003
	HGDC	0.846±0.013	0.772±0.004
CPDB	TMGNN（本文）	**0.881±0.006**	**0.778±0.010**
	EMOGI	0.816±0.007	0.721±0.001
	MTGCN	0.835±0.011	0.740±0.009
	HGDC	0.847±0.014	0.744±0.003
IntAct	TMGNN（本文）	**0.876±0.007**	**0.804±0.010**
	EMOGI	0.812±0.012	0.708±0.010
	MTGCN	0.831±0.014	0.744±0.009
	HGDC	0.840±0.003	0.772±0.007
InBioMap	TMGNN（本文）	**0.879±0.008**	**0.784±0.012**
	EMOGI	0.820±0.015	0.731±0.002
	MTGCN	0.839±0.009	0.740±0.008
	HGDC	0.835±0.005	0.747±0.002
iRefIndex	TMGNN（本文）	**0.880±0.012**	**0.798±0.009**
	EMOGI	0.824±0.010	0.731±0.006
	MTGCN	0.845±0.006	0.755±0.008
	HGDC	0.851±0.007	0.759±0.004

注：黑体数值表示最优值。“±”前表示平均值，“±”后表示标准差。

驱动基因识别需求。

4.2 消融实验

TGMNN 模型集成了用于提取文本和图这 2 种模态特征的 BERT 模型和 DMGNN 模型。为了进一步验证 TGMNN 模型中不同模态特征的必要性，这里采用消融实验对驱动基因的识别效果进行分析。具体的消融实验设计如下：

1）分析不同模态对模型性能的影响

为评估多模态模型中各模态对整体性能的贡献，本实验将通过逐步拆解不同模态的数据输入进行分析。具体而言，消融实验将逐步去除不同模态的数据输入，并观察模型性能的变化，以量化各模态在最终任务中的重要性。该消融实验基于 STRING 数据库，分别对图模态和文本模态进行分析。单模态测试效果如表 4 所示。

不同模态的特征在癌症驱动基因识别任务上各具优势，具体而言，文本模态特征在 AUC 指标上表现更佳，而图模态特征则在 AUPRC 指标上表现更为出色。进一步地，TGMNN 模型通过聚合 2 种模态不同层次的跨模态融合特征，能够有效提升癌症驱动

Table 3 Comparative Analysis of TGMNN Single Cancer Prediction Results

表 3 TGMNN 单癌症预测结果对比分析

癌症类型	方法	AUC	AUPRC
乳腺癌	TMGNN（本文）	**0.883±0.003**	**0.787±0.011**
	EMOGI	0.818±0.018	0.719±0.017
	MTGCN	0.826±0.007	0.723±0.004
	HGDC	0.838±0.003	0.730±0.001
肺腺癌	TMGNN（本文）	**0.879±0.004**	**0.776±0.009**
	EMOGI	0.821±0.012	0.734±0.003
	MTGCN	0.836±0.015	0.745±0.011
	HGDC	0.840±0.005	0.757±0.006
膀胱癌	TMGNN（本文）	**0.884±0.011**	**0.774±0.004**
	EMOGI	0.820±0.013	0.727±0.007
	MTGCN	0.830±0.018	0.741±0.005
	HGDC	0.828±0.002	0.748±0.009
肝细胞癌	TMGNN（本文）	**0.892±0.009**	**0.783±0.004**
	EMOGI	0.821±0.013	0.713±0.005
	MTGCN	0.833±0.003	0.732±0.003
	HGDC	0.838±0.007	0.755±0.012

注：黑体数值表示最优值。“±”前表示平均值，“±”后表示标准差。

Table 4 The Effectiveness of Single Modal

表 4 单模态效果

特征	AUC	AUPRC
图模态	0.856	**0.810**
文本模态	0.882	0.772
文本-图模态	**0.897**	0.791

注：黑体数值表示最优值。

基因的识别效果。

2）分析多模态融合位置对性能的影响

本实验设计了 3 种不同的文本与图像模态融合策略，以探讨文本与图结构信息融合位置的差异对模型性能的影响。

第 1 种策略采用特征预拼接的方法，将 BERT 模型生成的文本嵌入与图中每个节点的特征进行拼接，随后将拼接后的特征输入图神经网络进行学习，最终获得分类结果。第 2 种策略则是进行端到端的联合训练，将文本和图特征同时输入模型，通过统一的损失函数进行反向传播，最终得到分类结果。第 3 种策略则是将文本和图像分别输入各自的模型进行学习和预测，获取各自模型输出的表征后，再将这 2 个表征进行合并以进行分类，最终得到分类结果。这 3 种策略的消融实验分析如表 5 所示。

实验表明，端到端联合训练效果最优。其核心

Table 5 Effectiveness Analysis of Fusion at Different Positions

表 5 不同位置融合效果分析

融合位置策略	AUC	AUPRC
策略 1	0.885	0.776
策略 2	0.897	0.791
策略 3	0.883	0.780

优势在于统一优化文本与图特征表示：通过共享损失函数和交叉注意力机制，自动捕捉跨模态交互关系，实现深层特征融合。这种端到端方式基于整体任务损失优化，有效避免了分阶段训练导致的语义空间不匹配问题。此外，该方法显著提升模型对多模态信息的综合理解能力，精准建模复杂关联。相比特征拼接或分阶段训练，端到端联合训练支持动态调整特征表示，充分挖掘模态互补性，从而提升学习效率和预测性能。

3)分析不同层次信息的作用

TGMNN 在训练过程中利用了 BERT 模型和 DMGNN 模型不同隐层输出的特征，本实验进一步探讨了不同隐层输出特征及其组合方式对模型性能的影响。具体而言，在完整的 TGMNN 框架中，DMGNN 模型为 3 层，而 BERT 模型则为 12 层。因此，进行层次融合的图模态特征分别为 DMGNN 模型首层(第 1 层)、中层(第 2 层)和末层(第 3 层)输出的特征，文本模态特征则对应为 BERT 模型首层(第 1 层)、中层(第 6 层)以及末层(第 12 层)输出的特征。实验结果如表 6 所示。

实验结果表明，末层深层特征在多模态模型中表现尤为突出，是性能提升的核心因素。这类特征能够捕获全局语义信息，反映模态间的高阶语义和复杂关联。例如，末层图与末层 BERT 的组合在 AUC

Table 6 Effectiveness Analysis of Different Levels

表 6 不同层次效果分析

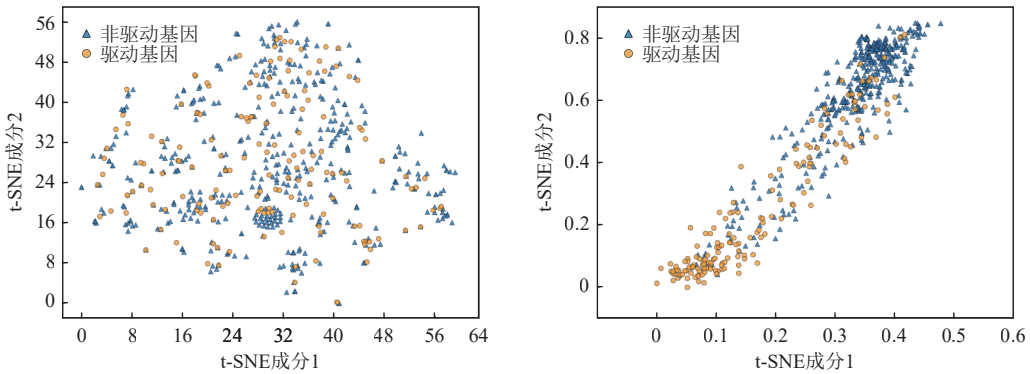
模型使用层次	AUC	AUPRC
首层图 + 首层 BERT	0.848	0.716
中层图 + 中层 BERT	0.859	0.730
末层图 + 末层 BERT	0.868	0.738
首层图 + 中层 BERT	0.866	0.717
首层图 + 末层 BERT	0.870	0.730
末层图 + 首层 BERT	0.853	0.712
末层图 + 中层 BERT	0.861	0.738

和 AUPRC 上分别达到 0.868 和 0.738，显著优于首层图+末层 BERT 的组合。这说明文本和图均采用深层特征时，模型效果最佳。深层特征的优势在于其对全局结构和上下文的强大建模能力。图的深层特征可提取复杂的全局信息，而 BERT 末层则擅长生成高质量的上下文语义表示。通过结合深层特征，模态间的信息融合更加高效，语义一致性增强，从而有效缩小了语义差距。

4.3 模型有效性分析

为验证 TGMNN 模型的有效性，本研究对分类层前最后一层隐层输出特征进行 t-SNE 降维处理，并绘制散点图进行可视化分析，结果如图 3 所示。

图 3(a)为模型训练前的特征分布，非驱动基因与驱动基因的数据点未形成明显聚类结构，特征呈现混杂状态，且这 2 种类型的数据点上的取值范围较广，分布均匀，类别区分度极低。图 3(b)为模型训练后的特征分布，非驱动基因的数据点集中区域发生迁移，驱动基因的数据点在左下角形成显著的紧密聚类，整体特征分布更具有序性，这 2 类数据的类别区分效果明显；同时，数据点呈现明确的分布方向性，尤其在驱动基因上表现出从左到右的线性排列



(a) 模型训练前样本特征可视化　(b) 模型训练后样本特征可视化

Fig. 3 t-SNE dimensionality reduction visualization of sample features before and after TGMNN model training

图 3 TGMNN 模型训练前后样本特征 t-SNE 降维可视化

趋势，这一现象表明模型在训练过程中充分学习了特征间的线性关系或梯度变化规律。

通过对比训练前后的特征散点图可以发现，模型训练前数据点的聚类效果较差，而训练后形成了清晰的聚类结构。结果表明，TGMNN 模型在训练过程中成功学习到了驱动基因与非驱动基因特征之间的内在关联，能够有效聚集相似特征，并在不同类别之间建立明确的决策边界，进一步验证了模型特征提取与分类能力的有效性。

4.4 新驱动基因预测

本研究探讨 TGMNN 在新型癌症驱动基因预测中的能力。实验使用所有已标记基因作为训练集，对未标记基因进行预测并分配驱动概率分数。通过排序提取前 30 个基因进行文献验证，其中排名前 5 的如表 7 所示，这些具有明确发病位置的基因在癌症过程中的作用得到验证。

Table 7 Top Five New Driver Genes Predicted by TGMNN

表 7 TGMNN 预测的前 5 个新驱动基因

基因名称	基因功能	对应发病位置
MCL1	抗细胞凋亡	骨髓瘤、白血病[20]
KITLG	造血干细胞发育	黑色素瘤[21]
CCNG1	细胞周期调控	多种癌症的生物标志物[22]
DDOST	蛋白质 N-糖基化	宫颈癌[23]
GULP1	细胞凋亡的清除	卵巢癌[24]

MCL1 基因属于抗凋亡 Bcl-2 家族成员，在血液系统恶性肿瘤（如多发性骨髓瘤、白血病）中功能异常，通过抑制促凋亡蛋白 BAX/BAK 阻止凋亡，在缺氧或化疗压力下维持癌细胞存活。其过表达与增殖和耐药相关，靶向 MCL1（如 BH3 域抑制剂）成为抗癌研究热点。

KITLG 基因属于干细胞因子，通过结合 KIT 受体激活 PI3K/AKT 和 MAPK 通路，调控细胞增殖与存活。在黑色素瘤和胃肠道间质瘤（GIST）中，其异常高表达促进侵袭和转移。靶向治疗（如 KIT 抑制剂伊马替尼）在 GIST 中疗效显著，凸显 KITLG/KIT 信号的重要性。

尽管部分未标记基因（如 CCNG1）未被纳入标准数据库，但近年研究证实其可作为多种癌症的生物标志物，显著影响癌症发生发展。CCNG1 作为周期蛋白，其抑制剂可治疗相关癌症，且 miRNA-27b 通过下调 CCNG1 能增强肿瘤药物敏感性。

4.5 KEGG 通路和基因本体论分析

对排名前 30 的基因进行 KEGG 通路富集分析[25]。TGMNN 预测的基因集合显著富集于多个与癌症发生和信号转导密切相关的 KEGG 通路，包括癌症通路（pathways in cancer）、MAPK 信号通路和 p53 信号通路，这些通路在肿瘤细胞增殖、凋亡及转移过程中发挥关键调控作用。此外，白细胞穿内皮迁移（leukocyte transendothelial migration）和细胞黏附分子（cell adhesion molecules）等通路的富集提示了免疫细胞迁移和细胞黏附在肿瘤微环境中的重要生物学功能，可能与炎症反应和癌症转移密切相关。该富集结果表明 TGMNN 所预测的基因集合在癌症驱动机制中具有重要的生物学意义。

5 结　　论

本研究提出一种基于文本-图双模态的癌症驱动基因识别网络框架 TGMNN。该框架以 DMGNN 为图模态基础模型，以 BERT 预训练模型为文本模态基础模型，充分利用 PPI 网络数据及基因节点的功能描述信息，通过 2 个基础模型分别学习图模态与文本模态的特征表征；采用交叉注意力机制实现 2 种模态特征的高效融合，并通过层次融合对不同隐层交叉融合后的跨模态融合特征进行聚合操作，最终生成具有强判别力的基因全局表征。TGMNN 模型通过引入文本模态作为额外信息源，与图数据提供的生物网络信息形成高度互补，既能够通过图模态捕捉基因间的结构关联，又能借助文本模态提取基因功能相关的语义特征，为癌症驱动基因识别提供了更为全面、多维的特征支撑。

作者贡献声明：韩有瑞、萨娅·赛立克、艾克热木·乃比江、孟维华、胡悠鹏、李子源负责完成实验并撰写论文；尚学群提出指导意见并修改论文；陈伯林提出了方法思路和实验方案。所有作者都审阅并同意最终版本。

参　考　文　献

[1] Lawrence M, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer associated genes[J]. Nature, 2013, 499(7457): 214−218

[2] Bashashati A, Haffari G, Ding Jiarui, et al. DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer[J]. Genome Biology, 2012, 13(12): R124

[3] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1):

61−80

[4] Song Hongzhi, Yin Chaoyi, Li Zhuopeng, et al. Identification of cancer driver genes by integrating multiomics data with graph neural networks[J]. Metabolites, 2023, 13(3): 339

[5] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint, arXiv: 1710.10903, 2018

[6] Schulte-Sasse R, Budach S, Hnisz D, et al. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms[J]. Nature Machine Intelligence, 2021, 3(6): 513−526

[7] Zhang Tong, Zhang Shaowu, Xie Mingyu, et al. A novel heterophilic graph diffusion convolutional network for identifying cancer driver genes[J]. Briefings in Bioinformatics, 2023, 24(3): bbad137

[8] Chen Bolin, Li Ziyuan, Li Haodong, et al. DMGNN: Cancer driver gene identification based on mix-moment graph neural network[C]// Proc of the 19th IEEE Int Conf on Bioinformatics and Biomedicine. Piscataway, NJ: IEEE, 2024: 6240−6247

[9] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[J]. arXiv prepring, arXiv: 1810.04805, 2019

[10] Tomczak K, Czerwinska P, Wiznerowicz M. The cancer genome atlas (TCGA): An immeasurable source of knowledge[J]. Contemporary Oncology, 2015, 19(1A): A68−A77

[11] Szklarczyk D, Morris J H, Cook H, et al. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible[J]. Nucleic Acids Research, 2017, 45(D1): D362−D368

[12] Herwig R, Hardt C, Lienhard M, et al. Analyzing and interpreting genome data at the network level with ConsensusPathDB[J]. Nature Protocols, 2016, 11(10): 1889−1907

[13] Li Taibo, Wernersson R, Hansen R B, et al. A scored human protein-protein interaction network to catalyze genomic interpretation[J]. Nature Methods, 2017, 14(1): 61−64

[14] Hermjakob H, Montecchi-Palazzi L, Lewington C, et al. IntAct: An open source molecular interaction database[J]. Nucleic Acids Research, 2004, 32(Suppl_1): D452−D455

[15] Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases [J]. Nucleic Acids Research, 2014, 42(D1): D358−D363

[16] Razick S, Magklaras G, Donaldson I M. iRefIndex: A consolidated protein interaction database with provenance[J]. BMC Bioinformatics, 2008, 9: 405

[17] Koh G C K W, Porras P, Aranda B, et al. Analyzing protein-protein interaction networks[J]. Journal of Proteome Research, 2012, 11(4): 2014−2031

[18] Safran M, Dalah I, Alexander J, et al. Genecards version 3: The human gene integrator[J]. Database, 2010, 2010: baq020

[19] Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From gene data mining to disease genome sequence analyses[J]. Current Protocols in Bioinformatics, 2016, 54: 1.30.1−1.30.33

[20] Akgul C. Mcl-1 is a potential therapeutic target in multiple types of cancer[J]. Cellular and Molecular Life Sciences, 2009, 66(8): 1326−1336

[21] Woodman S E, Davies M A. Targeting KIT in melanoma: A paradigm of molecular medicine and targeted therapeutics[J]. Biochemical Pharmacology, 2010, 80(5): 568−574

[22] Al-Shihabi A, Chawla S P, Hall F L, et al. Exploiting oncogenic drivers along the CCNG1 pathway for cancer therapy and gene therapy[J]. Molecular Therapy Oncolytics, 2018, 11: 122−126

[23] Mei Jie, Pan Liuliu, Huang Min, et al. DDOST is associated with tumor immunosuppressive microenvironment in cervical cancer[J]. Discover Oncology, 2024, 15(1): 69

[24] Maldonado L, Brait M, Izumchenko E, et al. Integrated transcriptomic and epigenomic analysis of ovarian cancer reveals epigenetically silenced GULP1[J]. Cancer Letters, 2018, 433: 242−251

[25] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Research, 2000, 28(1): 27−30



Han Yourui, born in 1997. PhD. His main research interests include machine learning and bioinformatics.

韩有瑞, 1997 年生。博士。主要研究方向为机器学习、生物信息学。



Saya Sailike, born in 2004. Undergraduate. Her main research interests include data mining and artificial intelligence.

萨娅·赛立克, 2004 年生。本科生。主要研究方向为数据挖掘、人工智能。



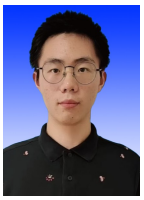
Aikeremu Naibijiang, born in 2000. Master candidate. His main research interests include machine learning and bioinformatics.

艾克热木·乃比江, 2000 年生。硕士研究生。主要研究方向为机器学习、生物信息学。



Meng Weihua, born in 2004. Undergraduate. Her main research interests include artificial intelligence and data mining.

孟维华, 2004 年生。本科生。主要研究方向为人工智能、数据挖掘。



Hu Youpeng, born in 2003. Undergraduate. His main research interests include data mining and bioinformatics.

胡悠鹏, 2003 年生。本科生。主要研究方向为数据挖掘、生物信息学。



Li Ziyuan, born in 2000. Master. His main research interests include machine learning and bioinformatics.

李子源, 2000 年生。硕士。主要研究方向为机器学习、生物信息学。



Shang Xuequn, born in 1973. PhD, professor. Her main research interests include data mining, machine learning, and bioinformatics.

尚学群, 1973 年生。博士，教授。主要研究方向为数据挖掘、机器学习、生物信息学。



Chen Bolin, born in 1984. PhD, associate professor. Senior member of CCF. His main research interests include data mining, machine learning, and bioinformatics.

陈伯林, 1984 年生。博士，副教授。CCF 高级会员。主要研究方向为数据挖掘、机器学习、生物信息学。