生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.230902

Jul. 25, 2024, 40(7): 2235-2245

·医药生物技术·

基于基因组学特征分布对齐和药物结构信息的癌症药物敏感性预测方法

廉令航，杨旭华*

浙江工业大学 计算机科学与技术学院，浙江 杭州 310023

廉令航, 杨旭华. 基于基因组学特征分布对齐和药物结构信息的癌症药物敏感性预测方法[J]. 生物工程学报, 2024, 40(7): 2235-2245.
LIAN Linghang, YANG Xuhua. Prediction of cancer drug sensitivity based on genomic feature distribution alignment and drug structure information[J]. Chinese Journal of Biotechnology, 2024, 40(7): 2235-2245.

摘　要：近年来精准医学在癌症治疗中得到了广泛的应用，其重点在于如何准确地预测不同的患者对药物治疗的反应。本研究设计了一种基于基因组学特征分布对齐和药物结构信息的癌症药物敏感性预测方法，该方法首先对齐来自细胞系的基因组学特征与来自患者的基因组学特征的分布并去除基因表达数据中的噪声，之后融合药物结构信息，使用多任务学习的方式进行患者药物敏感性预测。结果表明，在癌症相关药物敏感性基因组学数据集(genomics of drug sensitivity in cancer, GDSC)上，此方法的预测结果中均方误差降至 0.905 2，相关系数提升至 0.875 4，准确率提升至 0.836 0，显著优于最近发表的方法，在癌症基因组图谱数据集(the cancer genome atlas, TCGA)上，此方法预测药物敏感性的平均召回率提升至 0.571 4，F1-分数提升至 0.658 0，表现出优秀的泛化性能。这展现了本方法未来用于辅助选择临床治疗方案的潜力。

关键词：精准医学；基因组学特征分布对齐；药物结构信息；噪声去除

Prediction of cancer drug sensitivity based on genomic feature distribution alignment and drug structure information

LIAN Linghang, YANG Xuhua*

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, Zhejiang, China

Abstract: In recent years, precision medicine has demonstrated wide applications in cancer

资助项目：国家自然科学基金(62176236)
This work was supported by the National Natural Science Foundation of China (62176236).
*Corresponding author. E-mail: xhyang@zjut.edu.cn
Received: 2023-12-29; Accepted: 2024-04-25; Published online: 2024-04-28

therapy, and the focus of precision medicine lies in accurately predicting the responses of different patients to drug treatment. We propose a model for predicting cancer drug sensitivity based on genomic feature distribution alignment and drug structure information. This model initially aligns the genomic features from cell lines with those from patients and removes noise from gene expression data. Subsequently, it integrates drug structure features and employs multi-task learning to predict the drug sensitivity of patients. The experimental results on the genomics of drug sensitivity in cancer (GDSC) dataset indicates that this method achieved a reduced mean square error of 0.905 2, an increased correlation coefficient of 0.875 4, and an enhanced accuracy rate of 0.836 0 which significantly outperformed the recently published methods. On the cancer genome atlas (TCGA) dataset, this method demonstrates an improved average recall rate of 0.571 4 and an increased F1-score of 0.658 0 in predicting drug sensitivity, exhibiting excellent generalization performance. The result demonstrates the potential of this method to assist in the selection of clinical treatment plans in the future.

Keywords: precision medicine; genomic feature distribution alignment; drug structure characteristics; noise removal

癌症是当今世界重要的全球性健康问题[1]。尽管癌症的诊断与治疗在过去的几十年里取得了重要的进展，但是由于癌症的复杂性和异质性，罹患同种癌症的不同患者对于相同药物的反应可能不同，这导致临床治疗用药时很难预测药物的实际疗效，实验性的用药往往治疗效果不佳，甚至还会对病人造成健康上的损害，以至于加重病情。在过去的 10 年中，随着分子医学和基因测序技术的进步，研究人员已对癌症的分子特征进行深入研究[2-3]，探索患者肿瘤组织的生物分子特征与药物反应之间的联系已成为可能，个性化的癌症治疗正逐步应用。与传统的癌症治疗方法相比，个性化的癌症治疗具有更好的治疗效果和更小的副作用。然而，个性化的治疗策略也对癌症药物敏感性的预测提出了前所未有的高要求。高通量测序技术的最新进展和大规模数据集的公布[4-6]使得对于模型未知的药物和细胞系的双盲药物反应预测问题[7]取得了巨大的进展。目前已经有大量文献报道了基于机器学习的药物反应预测方法，如差异甲基化分析[8]、支持向量机[9]、随机森林[10]、XGBoost 算法[11]、贝叶斯推理方法[12]、矩阵分解方法[13]、集成学习方法[14]等。然而，在预测性能和模型泛化能力方面，这些基于经典机器学习的方法仍显薄弱，文献[15]指出深度学习方法在药物反应预测领域具有更高的准确性。DeepDSC 模型[16]使用堆叠自编码器和深度神经网络(deep neural network, DNN)架构，对输入网络的 mRNA 表达数据和药物分子特征谱进行降维来预测药物敏感性。陆家兴等[17]使用堆叠自编码器(stacked autoencoders, SAE)学习关键基因特征，并将随机游走算法(random walk)引入 XGBoost 学习算法中来预测癌症药物敏感性。MOLI 模型[18]设计了 3 个编码子网络融合 3 种不同组学类型的数据预测药物敏感性。CODE-AE 模型[19]通过自监督预训练特征编码模块对齐体外癌细胞系和患者组学特征的分布。scDEAL 模型[20]通过域自适应神经网络对齐单细胞嵌入和批量细胞嵌入(bulk)的分布，增强了模型在单细胞数据集上预测癌症药物敏感性的预测性能。

尽管基于不同模型结构和输入数据模态的药物反应预测方法不断涌现，然而，面向精准医学实际应用需求，现有的药物反应预测方法仍存在显著的局限性。首先，目前大多数模型忽视了药物分子空间结构信息，即图级结构特征。这些信息往往蕴含着药物分子的化学性质，对提高模型预测性能起着至关重要的作用，例如同分异构体乙醇和乙醚，化学表达式相同但空间结构不同，导致化学性质迥异。采用低复杂度模型直接处理高维药物图数据会导致模型表示学习效率低下[21]。其次，肿瘤细胞往往经历了大量的增殖分裂过程，加之恶性肿瘤细胞易产生基因上的突变，其产生的子代细胞呈现出分子生物学或基因方面的改变，影响肿瘤的生长速度、侵袭能力、对药物的敏感性及预后等特性[22]。肿瘤的异质性导致模型学习得到的来自体内和体外癌组织的基因组学特征分布产生差异，反映在嵌入空间中即为基因嵌入分布的异质性。这会产生分布外泛化(out of domain, OOD)问题，且患者反应数据量不足以支持训练高性能的预测模型[19]。因此，癌症药物反应预测模型不仅需要对细胞系数据有良好的预测效果，并且要对来自患者的数据也有很好的泛化能力。最后，由于样品污染、测序错误和不恰当的读段定位策略等系统因素以及G+C含量[23]、RNA二级结构[24]等导致读段分布不均匀，还有数据批次效应等其他因素仍然会使得基因测序的结果与真实基因表达产生偏差，进而影响药物反应预测的准确性。

为了解决上述局限性，本文提出了基于基因组学特征分布对齐结合药物结构信息的癌症药物敏感性预测深度学习方法(prediction of cancer drug sensitivity based on genomic feature distribution alignment and drug structure information, DADS)，该方法首先将来自体外组织和患者体内组织的基因表达嵌入进行分布对齐，之后降维分布对齐后的基因组学特征并从抗癌药化合物的 r-半径子图中学习药物的图级结构特征，最后将两者的结果拼接后输入预测器分别预测药物的 IC_{50} 值(回归任务)和敏感性标签(分类任务)。

1 方法

本研究提出的方法融合了药物图级结构特征且对基因组学数据进行噪声去除，提高了模型在药物敏感性预测任务上的性能，并通过数据分布对齐的方法提高模型在患者数据集上的泛化能力。

1.1 问题定义

药物敏感性预测问题可以被看作为标签预测问题。在细胞系/患者基因特征和药物结构信息都已知的情况下，对于给定的已知药物反应矩阵 R，从 R 中学习得到药物对特定基因特征的细胞系/患者的反应函数，并通过此反应函数实现准确预测未知细胞系/患者对药物的反应标签。

1.2 模型框架

模型框架如图1所示，主要由数据预处理模块和预测模块组成。

1) 数据预处理模块：如图2所示，该模块设计了一个去噪自编码器(denoise auto encoder, DAE)，用来降低基因组学特征中由于人为因素和系统因素所产生的基因组学特征中表达量过低或过高、序列不完整等噪声对模型预测性能所产生的影响，对齐来自细胞系和来自患者的基因组学特征在嵌入空间中的分布。

2) 预测模块：使用全连接神经网络对预处理后的基因特征进行降维，使用图卷积神经网络(graph convolutional network, GCN)学习药物的原子特性以及分子间的空间结构特征，拼接这2个特征并输入到预测器预测该药物的半抑制浓度值以及敏感性标签。

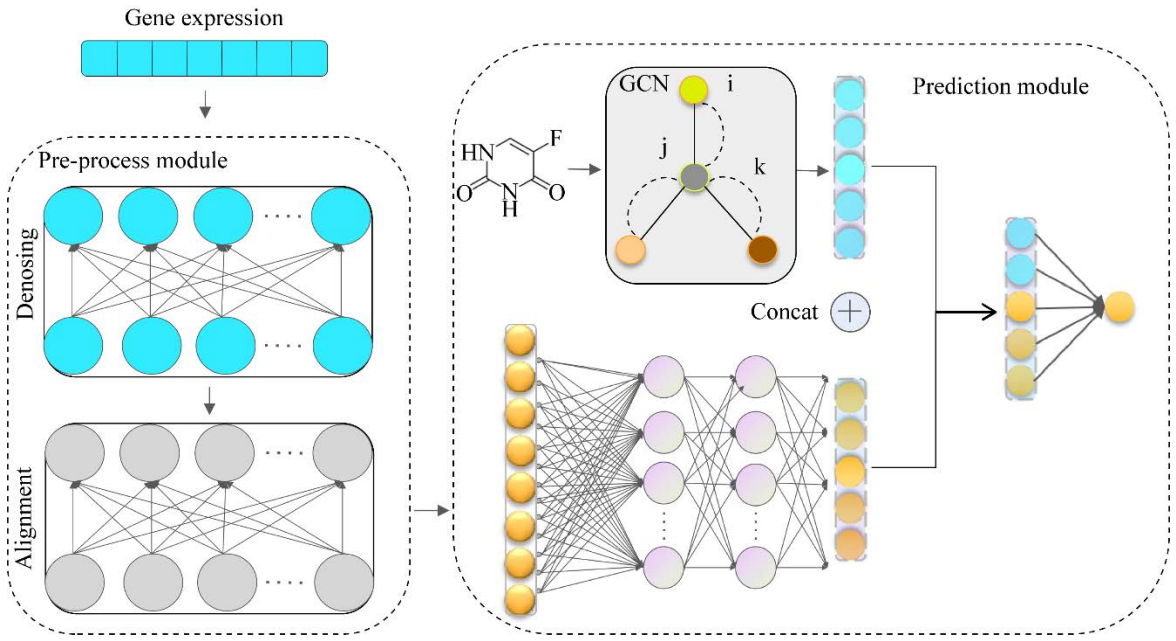

图 1　DADS 模型框架图

Figure 1　Model framework diagram of DADS.

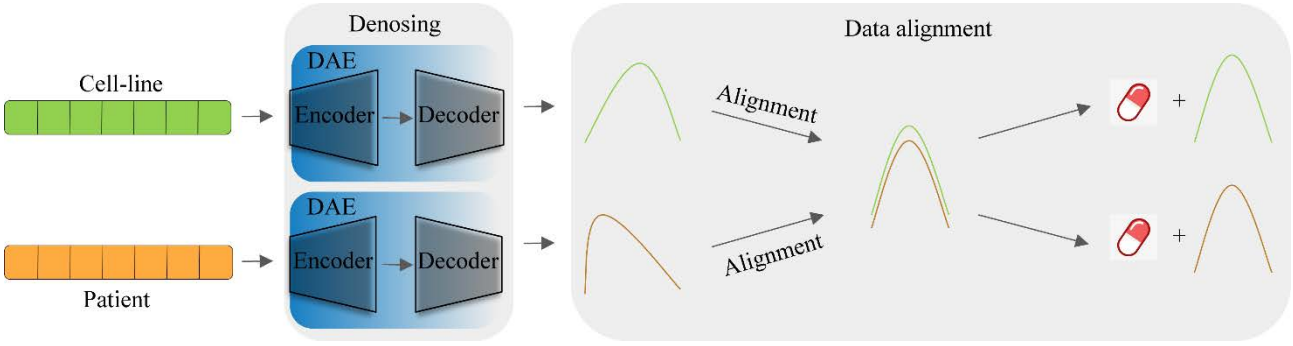

图 2　数据预处理模块

Figure 2　Data pre-process module.

1.3　数据预处理模块

数据预处理模块由 2 个子模块构成，分别是去噪子模块和基因组学特征分布对齐子模块。首先，本文提出了用于去除基因表达数据中噪声的去噪自编码器。该去噪自编码器由 1 个编码器(encoder)和 1 个解码器(decoder)构成，可以表示为式(1)：

$$G' = Decoder(Encoder(G + N_r)) \quad (1)$$

其中，N_r 表示随机生成的符合高斯分布的噪声，G'表示去噪之后的结果。编码器由卷积神经网络构成，由于数据中噪声占比较低而激活函数的作用是为网络添加非线性特性使得网络可以学习并处理更复杂的模式，故此处无需使用激活函数。编码器和解码器可以具体表示为式(2)和式(3)：

$$E_{out} = E^{(l)} \cdot W^{(l)} \quad (2)$$

$$D_{out} = D^{(l)} \cdot W^{(l)} \tag{3}$$

其中，$E_{out} \subseteq R^{N \times c}$，$D_{out} \subseteq R^{N \times c}$，$c$ 表示输出层向量的维度，$W^{(l)}$是第 l 层的参数。由于我们期望该模块的实际输出与预期输出完全一致，因此，我们使用均方误差(mean square error, MSE)评估模型的损失，具体的损失函数可表示为式(4)：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 \tag{4}$$

其中，y_i是标签数据，$f(x_i)$为模型的预测值，$f(\cdot)$为输入值到预测值的映射函数。

其次，本文使用基因特征分布对齐模块对齐来自不同数据集的基因组学特征，该模块由 1 个全连接神经网络构成，具体可以表示为式(5)：

$$E^{(l+1)} = \sigma(E^{(l)} \cdot W^{(l)}) \tag{5}$$

其中，$E^{(l)}$为编码器第 l 层的输出，$W^{(l)}$层是第 l 层的参数矩阵，σ 为该层网络的激活函数。最大均值化差异(maximum mean discrepancy, MMD)可以有效地度量 2 个概率分布之间差异。该网络的目标是使来自不同数据域的输入数据的嵌入分布一致，故此处使用 MMD 作为该模块的训练损失，计算方式如式(6)：

$$loss_{mmd} = \left|\frac{1}{n}\sum_{i=1}^{n}\varphi(G_c^i) - \frac{1}{m}\sum_{j=1}^{m}\varphi(G_p^j)\right| \tag{6}$$

其中，$G_c = \{G_c^i\}_{i=1,2,...,n}$ 和 $G_p = \{G_p^i\}_{i=1,2,...,n}$ 分别代表来自细胞系和患者的基因表达数据，$\varphi(\cdot)$表示从基因表达数据到嵌入空间的映射关系，使用最大均值化差异作为训练损失，可以更好地对齐源域数据(体外癌组织)和目标域数据(体外癌组织)在嵌入空间中的特征分布。

1.4 预测模块

预测模块由 1 个全连接神经网络和图卷积神经网络组成，全连接神经网络用于学习预处理后的基因组学特征，并将结果作为癌组织特征用作进行下游的预测任务，具体过程可表示为式(7)：

$$G^{(l+1)} = RELU(G^{(l)} \cdot W^{(l)}) \tag{7}$$

其中，$G^{(l)}$是 l 的输出，$W^{(l)}$是第 l 层的参数矩阵。在得到网络的输出后，为了加快模型的收敛速度和加强模型的稳定性，我们通过 BatchNorm 层将输出向量归一化，并且使用 ReLU 激活函数为网络添加非线性特性使之能够学习更加复杂的信息，之后通过 Dropout 层屏蔽部分神经元，防止模型过拟合。本文使用了一个 2 层的图卷积神经网络提取药物的图级结构特征，具体过程可表示为式(8)和式(9)：

$$D^{(1)} = RELU(W^{(1)} \cdot D) \tag{8}$$

$$D^{(2)} = RELU(W^{(2)} \cdot D^{(1)}) \tag{9}$$

其中 D、$D^{(1)}$、$D^{(2)}$分别为图卷积神经网络的输入、第 1 层输出和最终输出。$W^{(1)}$、$W^{(2)}$分别是第 1 层和第 2 层网络的参数矩阵，该网络同样通过 ReLU 函数激活输出。在此之后，我们使用相同的方法学习所有的药物结构特征并将所有药物的特征通过求平均值的方法将其压缩至一维。最后，将学习到的药物图级结构特征拼接在药物特征矩阵中，作为药物特征输入预测器，进行下游的预测任务。回归任务使用 MSE 作为训练损失，具体计算方法如式(4)所示，分类任务使用交叉熵损失(cross entropy loss)作为训练损失，具体计算方法见式(10)：

$$crossEntropy(x_i, y_i) = -\sum_{i=1}^{c} x_i \log(y_i) \tag{10}$$

其中，x_i表示第 i 个真实标签值，y_i表示第 i 个模型预测值，c 表示标签向量的维度。

2 结果与分析

实验的主要目标有以下 3 个。首先，对于细胞系数据集，我们评估了 DADS 模型预测药物半抑制浓度(IC_{50})值的预测性能，并与类似的研究方法进行了比较。其次，评估了本文提出的模

型在癌症基因组图谱数据集(the cancer genome atlas, TCGA)[4]上的性能，验证我们的模型是否能够缓解由于来自体外细胞系癌组织癌症相关药物敏感性基因组学数据集(genomics of drug sensitivity in cancer, GDSC)的基因组学特征和来自患者体内癌组织(TCGA)的基因组学特征分布差异导致的分布外泛化问题。最后，我们对本文模型的数据预处理模块中的 2 个子模块分别进行了消融实验，判断每一部分对结果的影响。

2.1 数据集和预处理

本研究中涉及的药物反应标签数据均来自 GDSC 数据集以及 TCGA 数据集。去除标签值为空值的细胞系/患者-药物对后，本研究参考已有工作[25-26]在 GDSC 数据集中的泛癌数据集上选取了 1 018 个癌细胞系以及 221 种药物组成的 177 988 条细胞系-药物反应数据用于模型的训练和评估，在 TCGA 数据集中的泛癌数据集上选取了 1 255 条反应数据用于评估模型的泛化能力。

2.1.1 药物结构数据集

本文使用的药物结构数据来自简化分子线性输入系统数据集(simplify the molecular linear input system dataset, SMILES)，我们借鉴 SWnet[26]的处理方法对来自 SMILES 的药物结构数据进行预处理，利用 RDKit 将 SMILES 转换为 Morgan 指纹，通过 Tanimoto 距离计算化合物之间的相似度。

2.1.2 细胞系数据集

GDSC 数据集[5]是癌症细胞药物敏感性和药物反应分子标志物信息的最大公共资源。GDSC 目前包含近 75 000 个实验的药物敏感性数据，描述了近 700 种癌细胞系中 138 种抗癌药物的反应。GDSC 共包括 3 个部分的数据，分别是细胞系药物敏感性数据(cell line drug sensitivity data)、细胞系的基因组数据集(genomic datasets for cell lines)和药物敏感性基因特征分析(analysis of genomic features of drug sensitivity)。

1) 细胞系药物敏感性数据集是由英国桑格研究所(Wellcome Trust Sanger Institute, WTSI)的癌症基因组计划和马萨诸塞州总医院分子治疗中心进行的高通量筛选，使用大于 1 000 个细胞系的集合产生的。

2) 目前可用于每个细胞系的基因组数据集包括关于 75 个癌症基因中的体细胞突变的信息、基因组广泛的基因拷贝数扩增和缺失、转录数据等信息。

3) 药物敏感性基因特征分析目前使用两种补充分析方法[27]。使用多变量方差分析(multivariate analysis of variance, MANOVA)将药物敏感性(半抑制浓度值和剂量-反应曲线的斜率)与癌症中的基因组改变相关联，应用弹性网络回归法确定影响每种药物反应的多个相互作用的基因组特征。根据 Broad L1000 项目[28]，我们选择了 1 458 个基因位点的表达作为 1 018 个细胞系的基因特征，这组基因已被证明能够有效地预测药物治疗后的转录组变化[26]。我们使用该基因集来降低输入数据的维度，减少模型过拟合的可能性。

2.1.3 患者数据集

TCGA 数据集是由美国国家癌症研究所(National Cancer Institute, NCI)和美国国家人类基因组研究所(National Human Genome Research Institute, NHGRI)于 2006 年联合启动的项目，收录了各种人类癌症(包括亚型在内的肿瘤)的临床数据、基因组变异、mRNA 表达、miRNA 表达、甲基化等数据，是癌症研究者很重要的数据来源。我们从 TCGA 数据集的泛癌数据中选取了反应信息不为空值的 1 255 条反应信息并采用与 GDSC 相同的基因集合来降低患者基因特征维度，使用 mRNA 表达作为患者基因特征，以

此确保基因组学特征分布对齐的效果。TCGA 数据集中的药物反应数据以“临床进展性疾病(clinical progressive disease)” “稳定的疾病(stable disease)” “部分响应(partial response)”和“完整的反应(complete response)”4 种方式表示。本文将“临床进展性疾病(clinical progressive disease)”和“稳定的疾病(stable disease)”这 2 种标签视作该患者对该药物不敏感，将“部分响应(partial response)”和“完整的反应(complete response)”这 2 种标签视作该患者对该药物敏感。

2.2 回归任务验证

本文选取了一些传统机器学习方法作为基线模型，例如 KBMTL[29]、SRMF[30]及 WGRMF[31]。此外，还选取了输入信息与本文模型类似且基于相同数据集的工作作为基线模型进行对比，例如 Precily[32]、GraphDRP[33]及 SWnet[26]等。这些模型都是以基因组学特征以及药物的化学结构作为输入来预测药物的半抑制浓度值。使用均方误差(MSE)和相关系数(R^2)来评估我们的方法与基线模型在 GDSC 数据集上的预测性能。均方误差用于衡量模型预测值和真实标签值之间的偏差，相关系数用于衡量预测值和标签值之间的线性相关程度。均方误差的计算方法如式(4)所示，相关系数的具体的计算方式如式(11)和式(12)所示：

$$R^2 = 1 - \frac{\sum_i^N (y_i - \widehat{y_i})^2}{\sum_i^N (y_i - \overline{y_i})^2} \tag{11}$$

$$\overline{y_i} = \frac{1}{N}\sum_1^N y_i \tag{12}$$

其中，y_i为模型预测的标签值，$\overline{y_i}$为真实标签值(ground truth)，N为标签向量的维度。本文将样本数据 10%和 90%的比例划分为测试集和训练集，采用五折交叉验证的方法训练模型，结果如表 1 所示。本文提出的模型预测药物半抑制浓度的均方误差低至 0.905 2，相较于基线方法分别提升了 28.40%、8.32%、8.05%、28.08%和 3.54%，说明 DADS 的预测结果与真实值之间的偏差较小，在相关系数指标上达到 0.875 4，相较于其他基线方法提升了 6.43%、1.63%、1.58%、6.38%和 0.82%，说明我们的模型在预测药物半抑制浓度时预测结果与标签值的线性相关程度较高，更加贴近药物反应的真实标签值。可见，DADS 模型在预测药物半抑制浓度方面的表现相较于基线方法有显著提高。

表 1 不同模型预测结果的均方误差值和相关系数值

Table 1 Mean square error values and correlation values of the prediction results of different models

Methods	MSE	R^2
KBMTL	1.264 2	0.822 5
SRMF	0.987 4	0.861 4
WGRMF	0.984 4	0.861 8
GraphDRP	1.258 6	0.822 9
SWnet	0.938 4	0.868 3
DADS	**0.905 2**	**0.875 4**

2.3 分类任务验证

本实验同样选取了传统机器学习方法：Deep Forest 以及具有类似输入的深度学习方法，例如 DeepCDR[34]、HNMDRP[35]以及 MOFGCN[36]。同时，采用曲线下面积(area under curve, AUC)值以及准确率(precision)作为评估指标，在二分类模型的预测结果中，最后的输出是预测结果标签为 1 的概率值，对于相同的概率值，不同的阈值会导致不同的划分结果，将阈值从 0 移动到 1 会产生许多误报率-召回率(false positive rate-true positive rate, FPR-TPR)值对，根据这些值制作坐标图，形成受试者工作特征(receiver operateing characteristic, ROC)曲线，获得曲线下面积的值，代表随机抽取一对正负样本，模型预测出的正样本概率值大于负样本概率值的概率。TPR、FPR 的计算方式如式(13)和式(14)所示：

$$TPR = \frac{FP}{FP + TN} \tag{13}$$

$$FPR = \frac{TP}{TP + FN} \tag{14}$$

准确率代表了正确预测的预测数占总样本的比例，能够很好地衡量模型的分类性能，具体的计算方式如式(15)所示：

$$presicion = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

其中，TP、TN、FP、FN 所代表的含义如表 2 所示。

本文按照 10%和 90%的比例划分测试集和训练集，采用五折交叉验证的方法训练我们的模型。结果如表 3 所示，DADS 预测药物敏感性的 AUC 值达到 0.860 1，显著大于传统机器学习方法以及 HNMDRP 等现有方法，分别提升了 23.14%、8.13%、19.64%，略低于 MOFGCN；在准确率指标上达到 0.836 0，相较于基线方法分别提升了 34.64%、16.83%、33.40%及 3.47%。可见，我们的方法在预测癌症药物敏感性标签时的预测性能相较于基线方法有所提高。

2.4 泛化能力验证

本文使用 TCGA 数据集来评估 DADS 在患者数据集上执行预测任务的性能。依据细胞系数据集的筛选标准，我们以同样的标准[28]在 TCGA 数据集中的泛癌数据集中选取了 1 458 个基因位点的表达作为基因组学特征，药物结构信息与基于细胞系数据集上的实验一致。我们观察到 TCGA 数据集中的患者-药物反应数据存在数据标签不平衡的问题。文献[37]表明，不平衡的数据标签下使用平均召回率(mean recall, MR)作为评价指标更加科学。因此，本文采用平均召回率作为评价指标来评估模型在患者数据集上的泛化性能，平均召回率的具体计算方式如式(16)所示：

$$Mean\ recall = \frac{\sum TP}{\sum TP + \sum FN} \tag{16}$$

本实验选取了传统机器学习方法随机森林(random forest, RF)[9]和弹性网络(elastic-net, E-net)[10]作为基线方法，同时也比较了近期发表的方法 MOLI[18]以及 CODE-AE[19]，本实验按照 10%和 90%的比例划分测试集以及训练集，采用五折交叉验证的方法训练模型。在 3 种常用的抗癌药物吉西他滨(gemcitabine)、替莫唑胺(temozolomide)以及顺铂(cisplatin)上与上述基线模型对比预测性能。结果如图 3 所示，在吉西他滨和替莫唑胺 2 种药物的反应预测中，我们的模型均优于基线方法，在顺铂的反应预测中，我们的模型优于随机森林、弹性网络以及 MOLI 这 3 个模型，但略逊于 CODE-AE。DADS 在患者数据集上表现优于基线方法的原因是其相较于基线方法能够有效地降低基因组学特征中噪声对模型预测性能的影响，并且能够更好地学习药物分子内部的复杂关系，例如药物化合物原子间的化学键类型和化合物分子的空间结构等影响药物生物活性以及作用机制的信息。通过去噪和融合药物图级结构特征，模型可以为患者提供更加精准的药物推荐。

表 2 FP、FN、TP 及 TN 的含义

Table 2 FP, FN, TP and TN

	Positive sample	Negative sample
Positive sample	TP	FN
Negative sample	FP	TN

表 3 不同模型预测结果的 AUC 值和准确率

Table 3 The AUC values and precisions of the prediction results of different models

Methods	AUC	Precision
DF	0.698 5	0.620 9
DeepCDR	0.795 4	0.715 6
HNMDRP	0.718 9	0.626 7
MOFGCN	0.891 8	0.808 0
DADS	0.860 1	0.836 0

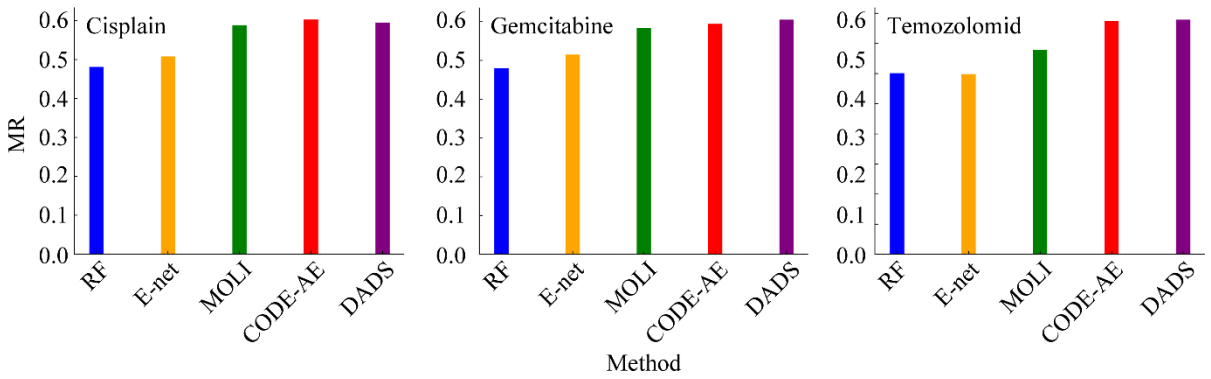

图 3　预测单种化疗药物临床反应的性能

Figure 3　Performances for predicting clinical responses to single chemotherapy.

这显示出 DADS 模型在患者数据集上的优秀预测性能，以及未来用于临床患者药物反应预测的可能性。

2.5　消融实验

为了验证 DADS 中数据预处理模块对模型预测性能的贡献，设计了 2 个消融实验来分别验证数据预处理模块中 2 个子模块的作用，结果如表 4、表 5 所示。

首先，我们去除去噪子模块，并将细胞系数据集(GDSC)按 10%和 90%的比例划分为测试集和训练集，使用五折交叉验证的方法训练不带去噪模块的模型，将测试结果与之前的实验结果对比(表 4)。结果表明，不带去噪模块的模型的预测性能在均方误差(0.938 6>0.905 2)和相关系数(0.868 2<0.875 4)这 2 个指标上表现均有下降，表明了去噪模块对模型性能的贡献。之后，我们去除基因组学特征分布对齐子模块，并在细胞系数据集(GDSC)上训练，在患者数据集(TCGA)上进行验证(表 5)。结果表明，去掉基因组学特征分布对齐模块后，模型在平均召回率(0.484 9< 0.571 4)和 F1_score (0.500 1<0.658 0)上均出现大幅下降，表明了该模块对提高模型泛化能力的贡献。

表 4　去噪子模块消融实验

Table 4　Denosing sub-module ablation experiment

Denosing	MSE	R^2
Without	0.938 6	0.868 2
With	0.905 2	0.875 4

表 5　基因特征分布对齐子模块消融实验

Table 5　Align sub-module ablation experiment

Alignment	MR	F1_score
Without	0.484 9	0.500 1
With	0.571 4	0.658 0

3　讨论与结论

本研究提出了新的癌症药物敏感性预测方法(DADS)。首先，对齐来自细胞系和患者的基因组学特征在嵌入空间中的分布，使模型能够以同患者基因数据具有相似嵌入的细胞系基因数据进行训练，提高模型在预测临床药物敏感性时的泛化性能，缓解了因患者数据不足导致无法支撑模型训练的问题。其次，设计了一种去噪自编码器用于缓解基因测序结果中潜在的误差和噪声对药物敏感性预测结果的负面影响，进一步提升模型预测药物反应的准确性。最后，该模型融合了药物的图级结构特征，采用了 2 层的图卷积神经网络来提取药物特征，并采用 r-半径子图[21]来解决可能由于模型复杂度低导致的表示学习低效的问题。结果表明，我们的模型在癌症药物敏感性预测方面取得了较好的结果。在未来的工作中，我们希望结合转录组学、表观组学等多组

学信息来预测癌症药物敏感性，并将细胞系/患者的相似性纳入模型，从而提高模型的预测性能，同时进一步拓展模型预测临床治疗方案疗效的应用场景，使其能够预测肿瘤对于化疗药物、靶向药物联合放疗的治疗方案的敏感性。

REFERENCES

[1] CHEN WQ, XIA CF, ZHENG RS, ZHOU MG, LIN CQ, ZENG HM, ZHANG SW, WANG LJ, YANG ZX, SUN KX, LI H, BROWN MD, ISLAMI F, BRAY F, JEMAL A, HE J. Disparities by province, age, and sex in site-specific cancer burden attributable to 23 potentially modifiable risk factors in China: a comparative risk assessment[J]. The Lancet Global Health, 2019, 7(2): e257-e269.

[2] GAGAN J, van ALLEN EM. Next-generation sequencing to guide cancer therapy[J]. Genome medicine, 2015, 7: 1-10.

[3] TATE JG, BAMFORD S, JUBB HC, SONDKA Z, BEARE DM, BINDAL N, BOUTSELAKIS H, COLE CG, CREATORE C, DAWSON E, FISH P, HARSHA B, HATHAWAY C, JUPE SC, KOK CY, NOBLE K, PONTING L, RAMSHAW CC, RYE CE, SPEEDY HE, et al. COSMIC: the catalogue of somatic mutations in cancer[J]. Nucleic Acids Research, 2019, 47(D1): D941-D947.

[4] WEINSTEIN JN, COLLISSON EA, MILLS GB, SHAW KRM, OZENBERGER BA, ELLROTT K, SHMULEVICH I, SANDER C, STUART JM. The cancer genome atlas pan-cancer analysis project[J]. Nature genetics, 2013, 45(10): 1113-1120.

[5] YANG WJ, SOARES J, GRENINGER P, EDELMAN EJ, LIGHTFOOT H, FORBES S, BINDAL N, BEARE D, SMITH JA, THOMPSON IR, RAMASWAMY S, FUTREAL PA, HABER DA, STRATTON MR, BENES C, MCDERMOTT U, GARNETT MJ. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells[J]. Nucleic Acids Research, 2013, 41(database issue): D955-D961.

[6] GHANDI M, HUANG FW, JANÉ-VALBUENA J, KRYUKOV GV, LO CC, McDONALD ER 3rd, BARRETINA J, GELFAND ET, BIELSKI CM, LI HX, HU K, ANDREEV-DRAKHLIN AY, KIM J, HESS JM, HAAS BJ, AGUET F, WEIR BA, ROTHBERG MV, PAOLELLA BR, LAWRENCE MS, et al. Next-generation characterization of the cancer cell line encyclopedia[J]. Nature, 2019, 569(7757): 503-508.

[7] CHEN YR, ZHANG LX. How much can deep learning improve prediction of the responses to drugs in cancer cell lines?[J]. Briefings in Bioinformatics, 2022, 23(1): 1-8.

[8] 张乃千. 差异甲基化分析和抗癌药物敏感性预测中的计算模型[D]. 上海: 上海师范大学博士学位论文, 2016.
ZHANG NQ. Computational models in differential methylation analysis and sensitivity prediction of anticancer drugs[D]. Shanghai: Doctoral Dissertation of Shanghai Normal University, 2016 (in Chinese).

[9] SUTHAHARAN S. Machine learning models and algorithms for big data classification[J]. Integrated Series in Information Systems, 2016, 36: 1-12.

[10] LIAW A, WIENER M. Classification and regression by randomForest[J]. R News, 2002, 2(3): 18-22.

[11] 李苗苗. 基于 XG-BOOST 和多数据源的药物重定位预测[J]. 软件导刊, 2020, 19(2): 110-113.
LI MM. Drug reposition prediction based on XG-BOOST and multi-source data[J]. Software Guide, 2020, 19(2): 110-113 (in Chinese).

[12] AMMAD-UD-DIN M, KHAN SA, WENNERBERG K, AITTOKALLIO T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression[J]. Bioinformatics, 2017, 33(14): i359-i368.

[13] ZHANG NQ, WANG HY, FANG Y, WANG J, ZHENG XQ, LIU XS. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model[J]. PLoS Computational Biology, 2015, 11(9): e1004498.

[14] 杨晨雨, 刘振浩, 代培斌, 张钰, 黄鹏杰, 林勇, 谢鹭. 基于多组学数据的肿瘤药物敏感性预测[J]. 生物工程学报, 2022, 38(6): 2201-2212.
YANG CY, LIU ZH, DAI PB, ZHANG Y, HUANG PJ, LIN Y, XIE L. Predicting tumor drug sensitivity with multi-omics data[J]. Chinese Journal of Biotechnology, 2022, 38(6): 2201-2212 (in Chinese).

[15] CHEN JY, ZHANG LX. A survey and systematic assessment of computational methods for drug response prediction[J]. Briefings in Bioinformatics, 2021, 22(1): 232-246.

[16] LI M, WANG YK, ZHENG RQ, SHI XH, LI YH, WU FX, WANG JX. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(2): 575-582.

[17] 陆家兴, 陈明, 秦玉芳, 于晓庆. 基于 LINCS-L1000 扰动信号通过 SAE-XGBoost 算法预测药物诱导下的细胞活性[J]. 生物工程学报, 2021, 37(4): 1346-1359.
LU JX, CHEN M, QIN YF, YU XQ. Prediction of drug-induced cell viability by SAE-XGBoost algorithm based on LINCS-L1000 perturbation signal[J]. Chinese Journal of Biotechnology, 2021, 37(4): 1346-1359 (in Chinese).

[18] SHARIFI-NOGHABI H, ZOLOTAREVA O, COLLINS CC, ESTER M. MOLI: multi-omics late integration with deep neural networks for drug response prediction[J]. Bioinformatics, 2019, 35(14): i501-i509.
[19] HE D, LIU Q, WU Y, XIE L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening[J]. Nature Machine Intelligence, 2022, 4: 879-892.
[20] CHEN JY, WANG XY, MA AJ, WANG QE, LIU BQ, LI L, XU D, MA Q. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-Seq data[J]. Nature Communications, 2022, 13: 6494-6506.
[21] COSTA F, de GRAVE K. Fast neighborhood subgraph pairwise distance Kernel[C]//Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, Israel. ACM, 2010: 255-262.
[22] YANG C, ZHANG SQ, CHENG ZA, LIU ZC, ZHANG LM, JIANG K, GENG HG, QIAN RL, WANG J, HUANG XW, CHEN M, LI Z, QIN WX, XIA Q, KANG XN, WANG C, HANG HL. Multi-region sequencing with spatial information enables accurate heterogeneity estimation and risk stratification in liver cancer[J]. Genome Medicine, 2022, 14(1): 142-160.
[23] DOHM JC, LOTTAZ C, BORODINA T, HIMMELBAUER H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing[J]. Nucleic Acids Research, 2008, 36(16): e105-e114.
[24] MORTAZAVI A, WILLIAMS BA, McCUE K, SCHAEFFER L, WOLD B. Mapping and quantifying mammalian transcriptomes by RNA-Seq[J]. Nature Methods, 2008, 5: 621-628.
[25] TAN M. Prediction of anti-cancer drug response by kernelized multi-task learning[J]. Artificial Intelligence in Medicine, 2016, 73: 70-77.
[26] ZUO ZR, WANG PL, CHEN XW, TIAN L, GE H, QIAN DH. SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures[J]. BMC Bioinformatics, 2021, 22: 1-16.
[27] GARNETT MJ, EDELMAN EJ, HEIDORN SJ, GREENMAN CD, DASTUR A, LAU KW, GRENINGER P, THOMPSON IR, LUO X, SOARES J, LIU QS, IORIO F, SURDEZ D, CHEN L, MILANO RJ, BIGNELL GR, TAM AT, DAVIES H, STEVENSON JA, BARTHORPE S, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells[J]. Nature, 2012, 483: 570-575.
[28] SUBRAMANIAN A, NARAYAN R, CORSELLO SM, PECK DD, NATOLI TE, LU XD, GOULD J, DAVIS JF, TUBELLI AA, ASIEDU JK, LAHR DL, HIRSCHMAN JE, LIU ZH, DONAHUE M, JULIAN BN, KHAN M, WADDEN D, SMITH IC, LAM D, LIBERZON A, et al. A next generation connectivity map: L1000 platform and the first 1 000 000 profiles[J]. Cell, 2017, 171(6): 1437-1452.e17.
[29] GÖNEN M, MARGOLIN AA. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning[J]. Bioinformatics, 2014, 30(17): i556-i563.
[30] WANG L, LI XZ, ZHANG LX, GAO Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization[J]. BMC Cancer, 2017, 17: 1-12.
[31] GUAN NN, ZHAO Y, WANG CC, LI JQ, CHEN X, PIAO X. Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization[J]. Molecular Therapy Nucleic Acids, 2019, 17: 164-174.
[32] CHAWLA S, ROCKSTROH A, LEHMAN M, RATTHER E, JAIN A, ANAND A, GUPTA A, BHATTACHARYA N, POONIA S, RAI P, DAS N, MAJUMDAR A, JAYADEVA, AHUJA G, HOLLIER BG, NELSON CC, SENGUPTA D. Gene expression based inference of cancer drug sensitivity[J]. Nature Communications, 2022, 13: 5680-5694.
[33] NGUYEN T, NGUYEN GTT, NGUYEN T, LE DH. Graph convolutional networks for drug response prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(1): 146-154.
[34] LIU Q, HU ZQ, JIANG R, ZHOU M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response[J]. Bioinformatics, 2020, 36(supplement_2): i911-i918.
[35] ZHANG F, WANG MH, XI JN, YANG JH, LI A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines[J]. Scientific Reports, 2018, 8: 3355-3363.
[36] PENG W, CHEN TL, DAI W. Predicting drug response based on multi-omics fusion and graph convolution[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(3): 1384-1393.
[37] TANG KH, NIU YL, HUANG JQ, SHI JX, ZHANG HW. Unbiased scene graph generation from biased training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 3716-3725.

(本文责编 郝丽芳)