# 基于多通路神经网络模型预测药物敏感性响应

李 晴[1] 闫效莺[1] 靳艳春[1]
LI Qing YAN Xiaoying JIN Yanchun

**摘 要** 准确预测药物敏感性响应是当前个性化治疗的关键。利用深度学习强大的特征学习能力，提出一种基于多通道神经网络模型的预测方法。首先，采用深度学习算法对细胞系的多组学特征分别处理，采用多个图神经网络模块提取药物分子图的多级子结构特征；然后，引入共同注意力机制评价各通路特征组合对药物 - 细胞系敏感性响应的影响，优化细胞系和药物特征；最后，通过深层神经网络模型预测。通过基于 GDSC 和 CCLE 数据集的测试，并与 RefDNN、DeepCDR 和 GraphCDR 算法进行比较，验证算法性能。

**关键词** 深度学习；药物 - 细胞系敏感性；基因表达；图同构网络；共同注意力

doi：10.3969/j.issn.1672-9528.2024.05.004

## 0 引言

随着基因组测序技术、生物信息与计算机的交叉应用，精准医疗作为新的医疗模式成为现实。精准医疗的核心目标之一是实现个体化治疗，即为每个患者提供最有效的治疗方案，以提高治疗效果和生存率。在这一背景下，研究药物敏感性响应成为医学领域的焦点之一。传统的试错式治疗方法可能导致患者在治疗过程中遭受不必要的痛苦，同时浪费宝贵的时间和资源。因此，发展预测模型来准确预测药物对个体疾病的敏感性响应关系具有重要意义。

抑制浓度 IC50 值[1] 反映了使 50% 的细胞生长受到抑制所需的药物浓度，可用于指导临床选用何种药物和确定所选药物的剂量。采用计算模型预测药物敏感性响应关系，通常包括两种场景，一是基于回归模型预测 IC50 值，另一种是将药物敏感性响应预测看作二分类问题，预测药物与细胞系之间是否存在敏感性和耐药性关系。近年来，相继发布的药物敏感性数据库 CCLE（癌症细胞系百科全书）[2] 和 GDSC（癌症药物敏感性基因组学数据库）[3]，整合收录了大量细胞系与药物之间的敏感性关系数据。基于此，Iorio 等人[4] 提出了基于弹性网络和 LASSO 的回归模型，Yan 等人[5] 提出具有可解释性的三矩阵分解方法，Yang 等人[6] 提出基于随机游走等网络推断模型方法预测药物敏感性响应。

近年来，深度学习算法在特征表示学习方面的优势，使得基于图神经网络的预测模型在社团挖掘、会话推荐，特别是药物敏感性响应预测等领域均取得了较好的应用。如 Zhu 等人[7] 基于化合物分子图网络和基因网络，构建了双图网络模型预测药物敏感性；Choi 等人[8] 使用细胞系的基因表达谱和药物的分子结构相似性，提出基于深层神经网络的敏感性预测方法，记为 RefDNN；Liu 等人[9] 采用一种混合 GCN 模型，融合细胞系的多组特征和药物分子图结构特征进行预测，记为 DeepCDR；基于此，Liu 等人[10] 设计了一种新奇的基于图神经网络的对比学习预测框架，记为 GraphCDR，模型中将敏感性和耐药性关系看作两种对立事件，并分别构建敏感性正网络和耐药性负网络，进而优化节点特征学习，并取得了较好的预测性能。

然而，当前模型大多只是基于药物的分子图特征和细胞系的基因表达特征，或者是直接串联组合多类特征，利用深度学习和图神经网络的强大特征学习能力进行预测。但是，其忽略了细胞系的多组学特征和药物的多级子结构，以及其特征相关性对预测模型的影响。本文将提出一种基于图神经网络的药物子结构与细胞系多组学特征融合策略 MCGNN，算法框架图见图 1。本文主要贡献为：采用基因表达、基因突变和 DNA 甲基化数据三种组学数据作为细胞系特征，并采用三个独立通路分支分别进行处理；采用化合物分子图作为药物特征，基于药物子结构的概念，采用图神经网络模型分别提取三级子结构作为药物的三个通路分支；引入共同注意力模块评价各通路特征组合对药物 - 细胞系敏感性响应的影响，进一步优化细胞系特征和药物特征。

1. 西安石油大学 陕西西安 710065

[ 基金项目 ] 陕西省自然科学基础研究计划“可解释多通道药物响应预测模型研究及其在肺癌中的应用”（2023-JC-YB-591）；西安石油大学研究生创新与实践能力培养计划项目（YCS23213172）

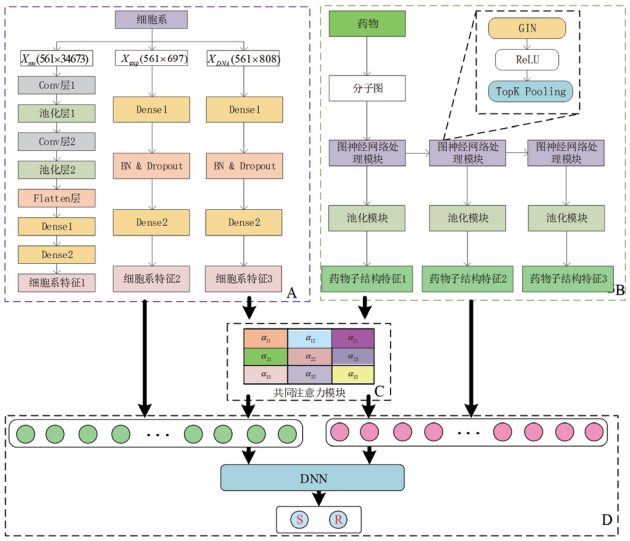



图 1 算法模型框架图

## 1 基于多通路神经网络模型的药物响应预测算法

### 1.1 细胞系特征表征学习

细胞系特征数据中的基因突变数据包含 34 673 个基因的突变位置，所以细胞系 $C_i$ 的突变特征可以表示为 $X_{\text{mu}}=[m_{i1},\cdots,m_{ih},\cdots,m_{ip}]$，$p$=34 673，它是一组二元特征向量，$m_{ih}$ 为“1”表示在该位置发生了突变，“0”代表在这个基因位置没有发生突变。基因表达数据是对基因表达的 TPM 值进行对数归一化获得，表示为 $X_{exp}=[e_{i1},\cdots,e_{ih},\cdots,e_{iq}]$，$q$ = 697, 这里 $e_{ih}$ 表示细胞系 $C_i$ 中基因 1 的表达值。DNA 甲基化数据表示为 $X_{\text{DNA}}=[dna_{i1},\cdots,dna_{ih},\cdots,dna_{ik}]$，$k$ = 808。对这三组不同的特征分别进行不同的处理。对于基因突变数据 $X_{\text{mu}}$，使用卷积神经网络来提取其特征信息，具体操作公式为：

$$N_{\text{mu}}^{(l)}=conv(w_{mu}^{(l)},z_{mu}^{(l-1)})+b_{mu}^{(l)} \tag{1}$$

$$N_{\text{mu}}^{(l)}=\sigma(N_{mu}^{(l)}) \tag{2}$$

基因表达数据 $X_{\text{exp}}$ 以及 DNA 甲基化数据 $X_{\text{DNA}}$ 采用深度神经网络来提取它们的细胞系特征，见公式（3）和（4），得到的特征分别记为 $Z_{\text{exp}}$ 和 $Z_{\text{DNA}}$，具体流程如图 1（A）所示。由此，得到细胞系的三通道特征，并组合为 $Z_c=\{Z_{\text{mu}}, Z_{\text{exp}}, Z_{\text{DNA}}\}$。

$$N_{\text{exp}}^{(l)}=w_{\text{exp}}^{(l)}z_{\text{exp}}^{(l-1)}+b_{\text{exp}}^{(l)} \tag{3}$$

$$Z_{\text{exp}}^{(l)}=\sigma(N_{\text{exp}}^{(l)}),\text{其中,}Z_{\text{exp}}^{0}=x_{\text{exp}} \tag{4}$$

### 1.2 药物特征表示学习

PubChem 数据库提供了 1900 万种经过验证的化合物分子结构信息，本文下载药物的 SMILES 序列，并采用 Rdkit 库 [11] 将其转化为可以进行可视化展示的分子图形式。

图神经网络通过不断聚合邻居节点特征信息，更新当前节点的嵌入表示，在图结构数据学习方面表现优秀。然而当前多数主流的图神经网络模型，如 GCN、GraphSAGE，通常只能利用节点之间的局部邻居信息进行卷积或聚合操作，难以充分捕捉全局图结构的复杂性。而一个强大的模型不仅应该能够区分非同构图，而且应学习如何将不同图结构数据映射到不同的嵌入空间。考虑图同构网络 [12]（graph isomorphism network，GIN）在分子图表示学习中的优秀性能，本文采用 GIN 模型用于分子图中原子特征学习。原子节点聚合更新过程如公式（5）所示，采用多层感知机（MLP）映射函数聚合第 $k$-1 层时节点 $v$，及其邻居节点的特征。读出药物分子特征获取过程如公式（6）所示，读出分子图中包含的所有原子特征，并采用求和、最大池化或平均池化等方式得到药物分子的特征表示 $Z_d$。

$$h_v^{(k)}=MLP[(1+\varepsilon^{(k)})h_v^{(k-1)}+\sum_{u\in N(v)}h_u^{(k-1)}] \tag{5}$$

$$Z_d=CAT(READOUT(\{h_v^{(k)}\mid v\in G_d\})\mid k=0,...,K) \tag{6}$$

式中：$N(v)$ 表示节点 $v$ 的邻居节点集合，MLP 代表多层感知机映射函数，READOUT 和 CAT 分别对应读出和池化操作。

考虑药物对细胞系的敏感性反应通常依赖于药物化合物结构中的部分子结构，本文对上述的 GIN 聚合过程重复三层，初始的药物分子图中节点特征只是原子自身特征，在经过各层 GIN 图神经网络的聚合后，模型会根据中心节点感受野，聚合其周围邻居节点的特征，以更新中心节点的特征表示，因此每个节点的特征均代表一个子结构。而且，子结构的规模随着 GIN 层数的增加而增大，如图 2 所示，本文采用三层 GIN，产生了药物的三个子结构特征，并组合为 $Z_d=\{Z_{d1}, Z_{d2}, Z_{d3}\}$，如图 1（B）所示。

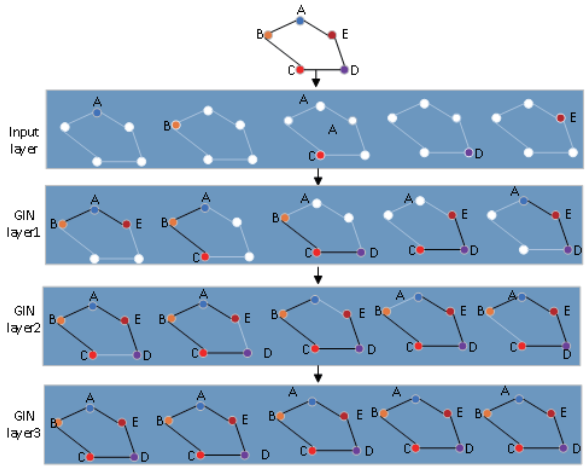



图 2 GIN 聚合得到药物三级子结构

### 1.3 基于共同注意力机制的多通道特征相关性

经上述处理，分别得到了细胞系和药物的三类特征和三级子结构，但是各通路特征对药物 - 细胞系敏感性响应的贡献并不相同。受 SSI-DDI 模型 [13] 启发，本文通过计算各通路特征组合对的相关性，引入共同注意力模块评价各通路特征组合对药物 - 细胞系敏感性响应的影响，具体计算过程为：

$$a_{ij}=b^{T}\tanh(W_{c}z_{c}^{i}+W_{d}z_{d}^{j})Z_{c}^{i},\ i,j=1,2,3 \quad (7)$$

式中：$Z_c$ 和 $Z_d$ 分别为细胞系和药物的三通道特征，$a_{ij}$ 为细胞系和药物的三通道特征两两成对组合对敏感性响应关系的重要性得分，得分越高表示该特征组合对预测贡献更大，反之，则贡献相对较小。

### 1.4 分类预测与损失计算

将上述得到的相关性得分矩阵 $a_{ij}$，通过 $Z_a = a_{ij}Z_c$ 和 $Z_b = a_{ij}Z_d$ 分别融入前面得到的细胞系三通路特征表示和药物的三级子结构特征中，得到优化的细胞系和药物特征，之后将优化的细胞系特征和药物特征串联，并送入深层神经网络模型中预测该细胞系与药物的响应得分值。具体计算公式为：

$$L=-\frac{1}{|T|}\sum_{(a,b)\in T}\left[q_{a,b}\cdot\log\hat{p}_{a,b}+(1-q_{a,b})\cdot\log(1-\hat{p}_{a,b})\right] \quad (8)$$

## 2 实验结果分析

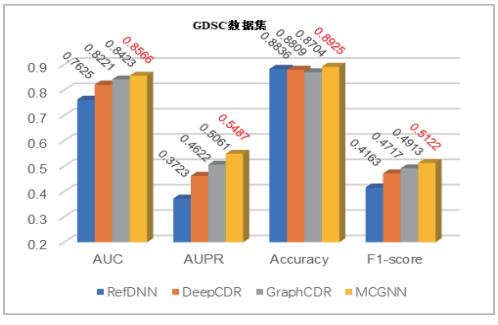
### 2.1 所用数据集

本文采用开源的 GDSC 数据和 CCLE 数据集，其中 GDSC 数据集包含了 265 种药物和 1001 种细胞系之间的抑制浓度 IC50 值。通过 DepMap 接口（https://depmap.org/）下载所需要的三种组学数据；对于药物，则从 Pubchem 下载了药物的 SMILES 序列特征。

### 2.2 实验结果与分析

使用 PyCharm 编写代码，PyTorch 框架对模型性能进行评估。评价指标包括 AUC（area under curve）、AUPR（area under precision-recall curve）、Accuracy（准确率）和 F1_Score（$F_1$ 分数）。其中，学习率设置为 0.000 01，batch = 256，epoch = 500，细胞系和药物各通路特征维数为 128，引入共同注意力模块后深层神经网络中细胞系和药物特征分别为 256 维。

为评价算法性能，将预测结果与 RefDNN、DeepCDR 和 GraphCDR 三种目前性能较优的基于深度学习模型的预测算法进行对比，结果如图 3 所示。MCGNN 相比于 GraphCDR、DeepCDR 和 RefDNN 三个算法，在两个数据集上 AUC、AUPR、Accuracy 和 F1_Scores 值都有明显提升，结果表明 MCGNN 整体预测性能更优。



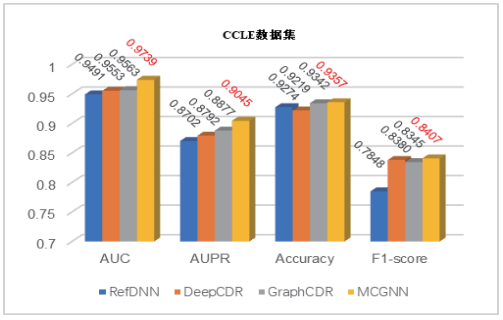




图 3 MBCDR 与其他方法性能对比

在实验过程中发现，在 GDSC 数据集上，模型的 AUPR 值低于其他评估指标的值。查阅资料发现可能是由目前 GDSC 数据集中负样本的数目比较多造成的。为了评估正负样本分布对模型的影响，进行了一系列的实验。通过选取不同倍数的负样本数量进行训练和测试，并将它们与使用全部样本时的结果进行比对，结果表明，随着负样本数量的增多，模型的整体性能明显下降，由此可见样本分布对模型性能有显著影响。实验结果如图 4 所示。图 4 中从左下及右下至上分别是 all AUC、$p$=5、$p$=4、$p$=3、$p$=2 及 $p$=1 的曲线图

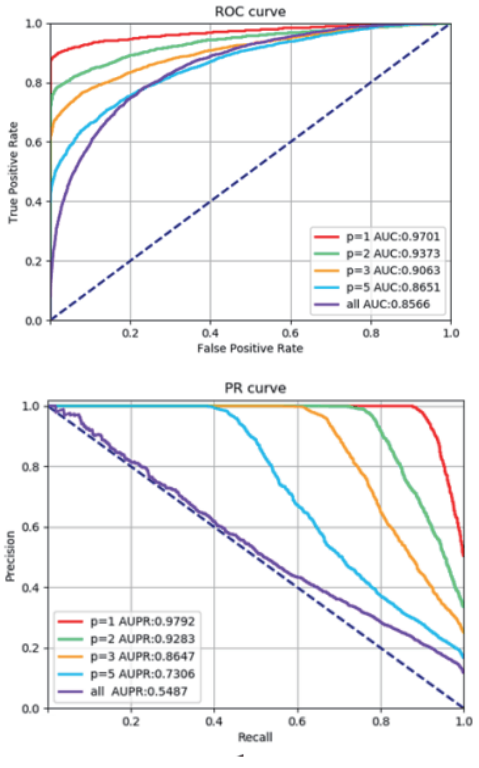



图 4 GDSC 数据集不同正负样本比例的 ROC 和 PR 曲线图

为了验证算法对新药物 - 细胞系响应关系的预测能力，本文对两种药物 Dasatinib 和 GSK690693 进行分析。首先从 GDSC 数据集中挑选与 Dasatinib 药物没有响应关系的细胞系，并逐一将其与该药物构成细胞系 - 药物对。之后将这些组合输入经过预训练的 MCGNN 模型中获取预测得分。最终预测情况如表 1 所示。

表 1 药物 Dasatinib 和 GSK690693 新细胞系的预测结果

| Drug | Rank | Cancer cell | PMID | Drug | Rank | Cancer | PMID |
|---|---|---|---|---|---|---|---|
| Dasatinib | 1 | ZR7530 | N/A | GSK690693 | 1 | ZR7530 | N/A |
| | 2 | YKG1 | N/A | | 2 | WSUDLCL2 | N/A |
| | 3 | YH13 | N/A | | 3 | UACC62 | N/A |
| | 4 | YAPC | 34 818 551 | | 4 | KATO-III | 28 860 825 |
| | 5 | WM793 | 23 251 610 | | 5 | RH18 | N/A |
| | 6 | WM115 | 30 587 121 | | 6 | SKM1 | N/A |
| | 7 | VMRCRC2 | N/A | | 7 | RCHACV | 19 064 730 |
| | 8 | VMRCRW | N/A | | 8 | NCIN87<br>NCIH929 | 34 760 336 |
| | 9 | UMUC3 | 31 554 791 | | 9 | NCIH929 | N/A |
| | 10 | UACC62 | 34 957 688 | | 10 | MELHO | 32 204 402<br>N/A |

## 3 结语

本文利用深度学习强大的特征学习能力，提出了一种新的药物 - 细胞系响应预测方法（MCGNN）。该方法主要包括特征提取和关系预测两部分。其中，前者对细胞系的多组学特征采用三通道深度学习模型处理，由药物的分子图特征出发，采用图神经网络模块提取药物多级子结构特征，引入共同注意力机制评价各通路特征组合对药物 - 细胞系敏感性响应的影响，并优化特征；后者串联药物细胞系对的特征向量，使用深度神经网络预测它们之间的作用关系。该模型在药物敏感性响应预测方面表现出了显著的预测性能，通过在 GDSC 和 CCLE 数据集上的实验以及与其他算法的比较，证实了 MCGNN 具有较高的预测精度。此外，该模型的通用性和可适应性较高，能够广泛应用于其他药物相关预测问题的研究。

## 参考文献：

[1]SEBAUGH J L.Guidelines for accurate EC50/IC50 estimation [J].Pharmaceutical statistics, 2011,10(2):128-134.

[2]BARRETINA J, CAPONIGRO G, STRANSKY N, et al.The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity[J].Nature, 2012,483(7391):603-607.

[3]YANG W, SOARES J, GRENINGER P, et al. Genomics of drug sensitivity in cancer (GDSC):a resource for therapeutic biomarker discovery in cancer cells[J]. Nucleic acids research, 2012, 41:D955-D961.

[4]IORIO F, KNIJNENBURG T A, VIS D J, et al.A landscape of pharmacogenomic interactions in cancer[J]. Cell, 2016, 166(3): 740-754.

[5]YAN X Y, ZHANG S W, YIU S M, et al. Interpretable prediction of drug-cell line response by triple matrix factorization[J]. 2021, 9(4): 426-439.

[6]YANG J, LI A, LI Y, et al.A novel approach for drug response prediction in cancer cell lines via network representation learning[J]. Bioinformatics, 2019, 35(9): 1527-1535.

[7]ZHU Y, OUYANG Z, CHEN W, et al.TGSA: protein–protein association-based twin graph neural networks for drug response prediction with similarity augmentation [J].Bioinformatics, 2022, 38(2): 461-468.

[8]CHOI J, PARK S, AHN J.RefDNN:a reference drug based neural network for more accurate prediction of anticancer drug resistance[J].Scientific reports, 2020,10(1):1861.

[9]LIU Q, HU Z, JIANG R, et al.DeepCDR: a hybrid graph convolutional network for predicting cancer drug response[J]. Bioinformatics, 2020,36:i911-i918.

[10]LIU X, SONG C, HUANG F, et al. GraphCDR:a graph neural network method with contrastive learning for cancer drug response prediction[J].Briefings in bioinformatics, 2022, 23(1): ARTN bbab457.

[11]HAN K, JENG E E, HESS G T, et al. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions[J].Nat biotechnol, 2017, 35(5): 463-477.

[12]XU K, HU W, LESKOVEC J, et al.How powerful are graph neural networks? [EB/OL].(2018-10-01)[2024-03-01].https://doi.org/10.48550/arXiv.1810.00826.

[13]NYAMABO A K, YU H, SHI J. SSI-DDI:substructure-substructure interactions for drug-drug interaction prediction[J].Briefings in bioinformatics,2021,22(6):bbab133-1-bbab133-10.

## 【作者简介】

李晴（2000—），女，河南开封人，硕士研究生，研究方向：生物信息学、深度学习。

闫效莺（1977—），女，山西阳泉人，博士，副教授，CCF 会员，研究方向：生物信息学、深度学习。

靳艳春（2000—），女，河南郑州人，硕士研究生，研究方向：生物信息学、深度学习。

（收稿日期：2024-03-15）