文章编号：2095-2163(2023)05-0098-05 中图分类号：TP301.6 文献标志码：A

基于异构图卷积网络的药物-细胞系响应预测

郭帅旗，闫效莺
(西安石油大学 计算机学院，西安 710065)

摘 要：准确预测药物敏感性响应是当前个性化治疗的关键，然而，如何高效融合药物、细胞系以及已观察到的药物-细胞系作用关系数据，仍面临着很大挑战。本文基于药物结构分子指纹特征数据、细胞系基因表达谱数据和已知的药物-细胞系作用关系数据，提出一种融合异构图卷积网络和深度神经网络的药物-细胞系预测模型 HGCNDCP。首先分别计算药物相似性和细胞系相似性，构建异构网络；通过异构图卷积模型对药物和细胞系进行特征表示学习；使用 DNN 模型预测药物-细胞系响应关系；最后在 GDSC 数据集中测试，并与 TMF、HNMDRP、NRL2DRP 和 HRWR 算法进行比较。测试结果表明，本文算法具有较高的预测性能。

关键词：图卷积；异构网络；药物；细胞系；深度神经网络；k-CV 交叉验证

Prediction of drug-cell response by heterogeneous graph convolution network

GUO Shuaiqi, YAN Xiaoying
(College of Computer Science, Xi′an Shiyou University, Xi′an 710065, China)

【Abstract】 Predicting drug sensitivity response accurately is critical for current personalized treatment. But, how to integrate diverse information about cancer cell lines, drugs and their observed responses efficiently, still remains a great challenge. In this paper, we proposed a heterogeneous graph convolution network method with deep neural network (DNN) for drug-cell response prediction, based on the chemical structure features, gene expression profiles of cell lines, and drug responses across cells (HGCNDCP). First, we calculate the cell line similarity and drug similarity separately, and construct the heterogeneous network; Then use the heterogeneous graph convolution model to learn the characteristics of the drug and cell line, and use the DNN model to predict the drug-cell line response; Finally, based on GDSC datasets, testing and comparison with TMF, HNMDRP, NRL2DRP and HRWR algorithms. The results show that the algorithm in this paper has high predictive performance.

【Key words】 graph convolution; heterogeneous network; drug; cell line; deep neural network; k-CV cross validation

0 引 言

近年来，卷积神经网络(Convolution Neural Network，CNN)在图像处理、机器视觉和自然语言处理等领域均取得了较好的应用。但随着大量非结构化图数据，如交通网络、社交网络和生物信息学领域中蛋白质作用关系网络(PPI)等的出现，图卷积神经网络(Graph Convolution Network，GCN)应运而生，并迅速发展，已成为图表示学习的重要方法之一。

随着健康理念的发展，精准医疗已成为疾病诊疗的方向，如何在分子水平上精确测定个体病人对药物治疗的响应情况，是精准医疗的基础和关键。然而，对个体病人进行大量药物临床试验是不可行的，因此需要建立模型，预测药物对个体疾病的敏感性响应关系。随着高通量测序技术的发展，药物敏感性数据库，如 NCI-60 数据[1]、癌症细胞系百科全书(CCLE)[2]和癌症药物敏感性基因组学数据库(GDSC)[3]等相继发布，这些数据库中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，近些年国内外学者提出了许多药物-细胞系响应的预测方法。这些方法大致可分为基于回归模型[4]、基于网络推断[5]、基于矩阵分解[6]和基于深度学习[7]的预测方法等等。如：Iorio 等人[4]通过弹

基金项目：西安石油大学研究生创新与实践能力培养计划项目(YCS21112078)；西安石油大学博士科研启动基金(134890003)。
作者简介：郭帅旗(1996-)，男，硕士研究生，主要研究方向：图卷积神经网络、机器学习；闫效莺(1977-)，女，博士，副教授，CCF 会员，主要研究方向：深度学习、生物信息学。
通讯作者：闫效莺 Email：xiaoying_yan@126.com
收稿日期：2022-06-24

性网络和 LASSO 构建基因表达值与响应值之间的回归预测模型，但其缺点是忽略了药物的相关信息；Yang 等人[5]利用网络表示学习提取特征，并采用 SVM 预测药物响应关系（NRL2DRP）；Stanfield 等人[8]提出基于异构网络的带重启随机游走算法预测药物敏感性响应（HRWR）；Yan 等人[6]提出一种具有可解释性的三矩阵分解方法，预测药物敏感性响应（TMF）；Li 等人[9]提出一种应用堆叠的深度自动编码器方法预测药物敏感性响应（DeepDSC）。

虽然上述算法已极大推动了药物-细胞系作用关系预测的研究，但预测精度仍有很大的提升空间。特别是近几年图神经网络的出现，已成功应用于药物-靶蛋白，药物-药物作用关系预测等生物信息学相关问题研究[10-11]之中。考虑药物-细胞系敏感关系中网络节点的异质性，本文对适用于同构网络的 GCN 算法进行改进，提出了一种新的基于异构图卷积网络和深度神经网络的药物-细胞系响应预测方法（HGCNDCP）。该方法首先分别计算药物相似性和细胞系相似性，并融合药物相似性特征、细胞系相似性特征以及药物-细胞系响应关系，构建异构网络；然后在异构网络上使用图卷积操作，通过不断聚合邻居节点特征，可同时捕获异构图网络的拓扑结构特征和节点特征，得到药物和细胞系两类对象的特征表示数据，最后使用深度神经网络（DNN）预测药物-细胞系响应关系，并在 GDSC 数据集中对算法进行验证。

1　基于异构图卷积网络的预测方法 HGCNDCP

1.1　图卷积网络 GCN

GCN 是 CNN 算法在图数据领域应用的产物。GCN 模型可用于捕获非欧氏空间中存在的复杂网络关系及对象（或实体）间的各种依赖关系。该模型通过不断聚合邻居节点信息，来更新自身节点特征，可同时获捕图结构拓扑特征和节点特征。因此，GCN 可从原始图数据和节点特征中，更好地进行特征表示与学习。基于同构图的 GCN 模型定义如下：

给定图 $G=(V,E)$，其中 V 是 n 个节点的集合，E 是节点之间边的集合；对应的邻接矩阵记作 $\boldsymbol{A}$，其元素 a_{ij} 代表节点 v_i 与 v_j 之间的连接关系，节点特征矩阵记为 $\boldsymbol{X}\in R^{n\times p}$；一阶 GCN 模型定义为 $H^{l+1}=f(\hat{\boldsymbol{A}}H^l\boldsymbol{W}^l)$，$H^0=\boldsymbol{X}$，其中 $\hat{\boldsymbol{A}}$ 为归一化的邻接矩阵，H^l、$\boldsymbol{W}^l$ 分别为 l 层的节点表示和映射权重矩阵。

1.2　异构网络构建

1.2.1　构建药物相似性矩阵

药物特征描述符主要包括化学结构描述符、分子指纹等。本文采用 Pubchem 数据库中记录的药物分子指纹描述符[6]，将药物表示为 881 维的子结构特征。因此，药物特征矩阵可表示为 $\boldsymbol{F}_d\in R^{N\times p}$，$p=881$，$N$ 为药物数目。其中，药物 d_i 的特征记为 $f_{d_i}=[s_{i1},\cdots,s_{il},\cdots,s_{ip}]$。若第 l 个子结构特征在药物 d_i 中存在 $s_{il}=1$，否则 $s_{il}=0$。由于具有相似化学结构的药物在细胞系中表现出相似的反应，在此使用 Jaccard 系数计算药物结构相似性，其公式如下：

$$S_d(d_i,d_j)=\frac{|f_{d_i}\cap f_{d_j}|}{|f_{d_i}\cup f_{d_j}|}=\frac{|f_{d_i}\cap f_{d_j}|}{|f_{d_i}|+|f_{d_j}|-|f_{d_i}\cap f_{d_j}|} \tag{1}$$

1.2.2　构建细胞系相似性矩阵

对于细胞系的基因表达谱数据来说，每个细胞系均包含16 383个基因的表达值，因此细胞系的特征矩阵可表示为 $\boldsymbol{F}_c\in R^{M\times q}$、$q=16\,383$，$M$ 为细胞系数目，其中细胞系 c_i 的特征记为 $f_{c_i}=[g_{i1},\cdots,g_{il},\cdots,g_{iq}]$，$g_{il}$ 表示第 l 个基因在细胞系 c_i 中的表达值。由于具有相似基因表达谱的细胞系会表现出相似的药物反应。本文使用皮尔逊相关系数计算细胞系之间的相似性，公式如下：

$$S_c(c_i,c_j)=\frac{\sum(\boldsymbol{f}_{ci}-\bar{\boldsymbol{f}}_{ci})(\boldsymbol{f}_{cj}-\bar{\boldsymbol{f}}_{cj})}{\sqrt{\sum(\boldsymbol{f}_{ci}-\bar{\boldsymbol{f}}_{ci})^2\sum(\boldsymbol{f}_{cj}-\bar{\boldsymbol{f}}_{cj})^2}} \tag{2}$$

其中，$\boldsymbol{f}_{ci}$ $\boldsymbol{f}_{cj}$ 分别为细胞系 c_i 和 c_j 的特征列向量，$\bar{\boldsymbol{f}}_{ci}$ $\bar{\boldsymbol{f}}_{cj}$ 分别表示其对应列向量的均值。

1.2.3　药物-细胞系响应关系网络

已知的药物-细胞系响应关系可以表示为二分图 $G=(V,E)$，其中 $V=\{V_c,V_d\}$ 表示药物和细胞系两类节点，E 表示药物与细胞系之间已知的 IC50 响应值。本文使用 Iorio 等人[12]提供的阈值，将已观察响应值划分为敏感性和耐药性两类。其中，敏感响应 16 804 个，耐药响应 125 647 个，未知响应 33 595个，由此构建邻接矩阵为 $\boldsymbol{A}_{cd}$。

由药物相似性矩阵、细胞系相似性矩阵和药物-细胞系响应关系构建形成的异构网络模型如图 1 所示。

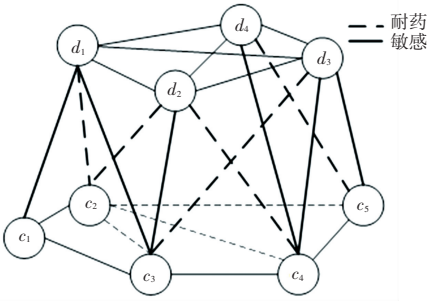

图 1 药物-细胞系响应异构网络

Fig. 1 Drug-cell line heterogeneous network model

1.3 基于异构图卷积的药物-细胞系响应预测算法

基于异构图卷积的药物-细胞系响应预测模型如图 2 所示,算法实现步骤如下:

Input:药物相似性网络 S_d、细胞系相似性网络 S_c、药物-细胞系二分图网络 A_{cd}(边权重"1"和"0",分别表示敏感性和耐药性响应类别);

Output:药物-细胞系响应关系预测得分。

Step 1 对 S_d、S_c 以及 A_{cd} 按如下方式重构,得到异构网络邻接矩阵 $\boldsymbol{A}$ 和特征矩阵 $\boldsymbol{S}$。

$$\boldsymbol{A}=\begin{bmatrix}0 & A_{cd}\\ A_{cd}^{T} & 0\end{bmatrix} \tag{3}$$

$$\boldsymbol{S}=\begin{bmatrix}S_c & 0\\ 0 & S_d\end{bmatrix} \tag{4}$$

Step 2 矩阵归一化

邻接矩阵归一化:$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$,其中 $\boldsymbol{D}$ 为对角矩阵,对角元素为 $D_{ii}=\sum_j a_{ij}$。

特征矩阵归一化:$\hat{\boldsymbol{S}}=\boldsymbol{D}_S^{-1}\boldsymbol{S}$,其中 $\boldsymbol{D}_S$ 为对角矩阵,对角元素为 $D_{S,ii}=\sum_j S_{ij}$。

Step 3 基于异构网络的图卷积操作,得到药物和细胞系嵌入特征表示为 F':

$$F'=\mathrm{ReLU}(\hat{\boldsymbol{S}}(I+\hat{\boldsymbol{A}})\boldsymbol{W}_e+B) \tag{5}$$

其中,$\boldsymbol{W}_e$ 为可训练的权值矩阵;B 为偏置;F' 为异构图卷积操作的输出,包括映射后的药物嵌入特征矩阵 $\boldsymbol{F}_d'(N\times n_e)$ 和细胞系嵌入特征矩阵 $\boldsymbol{F}_c'(M\times n_e)$。

Step 4 特征向量聚合,将药物嵌入特征和细胞系嵌入特征拼接形成药物-细胞系对的特征 $\boldsymbol{X}\in R^{K\times P}$,$K$ 为样本数,P 为样本特征的维度,见公式(6)。

$$\boldsymbol{X}(c_i,d_j)=\{F_{c_i}'\,\|\,F_{d_j}'\} \tag{6}$$

Step 5 构建预测器。使用深度神经网络(DNN)作为 HGCNDCP 的预测器。

$$\Pr(y\mid X,\theta)=f(Z_{out}W_{out}+b_{out}) \tag{7}$$

$$Z_{out}=f(Z_kW_k+b_k) \tag{8}$$

$$Z_{k+1}=f(Z_kW_k+b_k) \tag{9}$$

其中,Z_{out} 和 $Z_k(k=0,\cdots,l)$ 是 DNN 模型中对应权重 W_{out}、W_k 和偏置 b_{out}、b_k 的隐层神经元,$Z_0=\boldsymbol{X}$。$y\in\Re^{K\times t}$ 为 K 个药物-细胞系样本对的预测值。

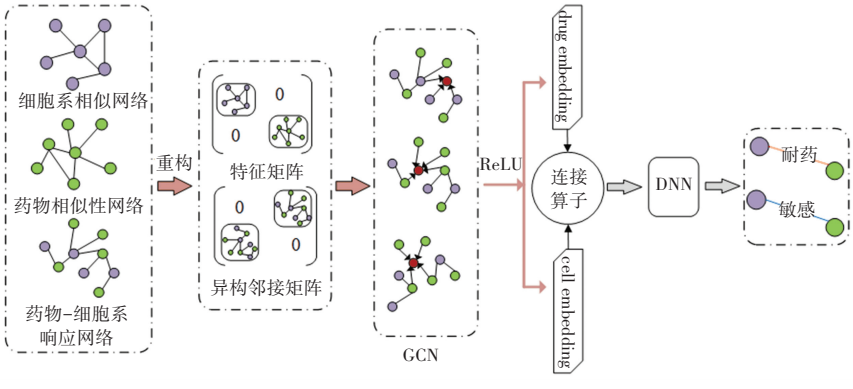

图 2 HGCNDCP 网络模型图

Fig. 2 The flowchart of HGCNDCP pipeline

2 数据来源与实验方法

本文使用开源的 GDSC[3] 作为基准数据集,网址为 http://www.cancerrxgene.org/,其中包括 256 个药物、1 001 个癌症细胞系,以及药物和细胞系的对数变换半抑制浓度值 IC50[13]。该值代表要使 50% 的细胞生长受到抑制所需的药物浓度,是药物-细胞系响应的测量值。考虑到实验的具体进行,需要

对 GDSC 基准数据集进行筛选、清洗等预处理。经预处理后，本文得到 183 种同时具有化学结构特征和药物反应数据的药物，962 种同时具有基因组特征和药物反应数据的细胞系。在这些药物与细胞系之间，药物-细胞系响应总共有 176 046 个，其中敏感响应 16 804 个，耐药响应 125 647，未知响应 33 595个。

使用 PyCharm 集成开发环境，Pytorch 1.7.1 作为框架。采用文献[5]的验证方法，基于上述预处理后的数据集，采用 5-CV 交叉验证方法进行实验，即将数据随机分成大致相等的 5 份，每一份轮流作为测试样本，其余 4 份做训练集。对测试集中每个药物-细胞系对样本进行预测，并将预测结果与实际标签进行对比。使用 *ROC* 曲线下面积 *AUC* 表征模型预测性能。*AUC* 值越大，表示算法性能越好。

3　实验结果与分析

3.1　损失函数与参数设置

本文算法包括基于异构 GCN 的特征表示和 DNN 预测器两部分，其中第一部分的损失函数采用二元加权交叉熵，第二部分采用二元交叉熵，见公式(10)、公式(11)。

$$L_f(p,q) = -\sum_{i,j} p(a_{ij})\log(q(a_{ij})) * W_{pos} + p(a_{ij})\log(q(a_{ij})) \tag{10}$$

$$L_p = \frac{1}{T}\sum_i -\log s(a_{ij}) + \beta\left(\sum_l \frac{1}{2}\| W^l \|^2 + \sum_l \frac{1}{2}\| b^l \|^2\right) \tag{11}$$

公式(10)中，$p(a_{ij})$ 为 a_{ij} 的真实标签，$q(a_{ij}) = \sigma(F_{c_i}' \cdot F_{d_j}'^{\mathrm{T}})$ 是由异构 GCN 生成的两类节点嵌入特征向量内积计算出的预测概率，W_{pos} 为负样本与正样本数目比的权重；公式(11)中，$s(a_{ij})$ 是 DNN 预测的得分值，其值越大表示该样本呈现敏感性的概率越大，T 为类别数量，β 为权值衰减系数。式子前一项旨在对所有类别计算平均损失，后一项旨在为权值矩阵和偏置矩阵提供 L2 范数约束。两部分均使用 Adam 优化器[14]、ReLU 激活函数和批量归一化[15]处理。通过寻找损失函数的最小值和最佳精度对参数进行网格搜索。其中，基于异构 GCN 的特征表示部分，分别设置嵌入特征数 $n_e = \{5,25,50,75,100\}$，并根据式(10)计算训练损失。由图 3 可见，不同数量的嵌入特征，训练过程相似，在 200 轮训练之前，损失快速下降，在 1 000 轮之后，陆续趋于收敛状态。其中潜在因子数 n_e 为 75 和 100 时，损失值相差无几，误差可缩小至 10^{-3} 数量级。本文设定 $n_e = 100$，学习率 $lr = 0.01$。DNN 预测器的各层维度分别为[200,128,96,64,2]，lr = 3.25e-5，衰减系数 β = 1e-5。

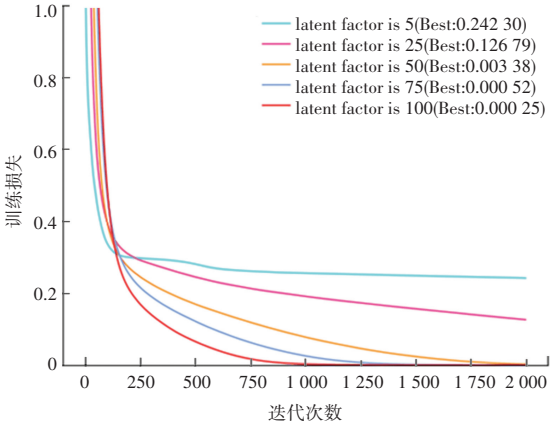

图 3　隐层节点数对异构图 GCN 特征提取的影响

Fig. 3　Influence of the number of latent factors on the heterogeneous GCN feature extractor

3.2　方法比较

采用 5-CV 交叉验证方法，将本文算法 HGCNDCP 与 HNMDRP[16]、HRWR[8]、NRL2DRP[5]和 TMF[6]算法进行比较。数据集中敏感性数据为正样本，耐药性为负样本。*ROC* 曲线下面积 *AUC* 结果详见表 1。

表 1　算法性能比较结果

Tab. 1　Performance comparison

Methods	*AUC*
HNMDRP	0.739 1
HRWR	0.847 4
NRL2DRP	0.788 2
TMF	0.825 6
HGCNDCP	**0.948 8**

由表 1 可见，HGCNDCP 的 *AUC* 值比 HNMDRP 提高了 20.97%、比 HRWR[8]算法提高了 10.14%、比 NRL2DRP 提高了 16.06%、比 TMF 提高了 12.32%，证明 HGCNDCP 具有更优的预测性能。

3.3　k-CV 交叉验证对模型性能的影响

为了评估不同 k-CV 对模型性能的影响，本文分别进行 2-CV、5-CV 和 10-CV 交叉验证，其对应 *AUC* 结果如图 4 所示。

结果表明，预测精度随着训练数据集的增多而增加，10-CV 验证的训练数据集大于 2-CV 验证和 5-CV 验证，其 *AUC* 值也高于两者。

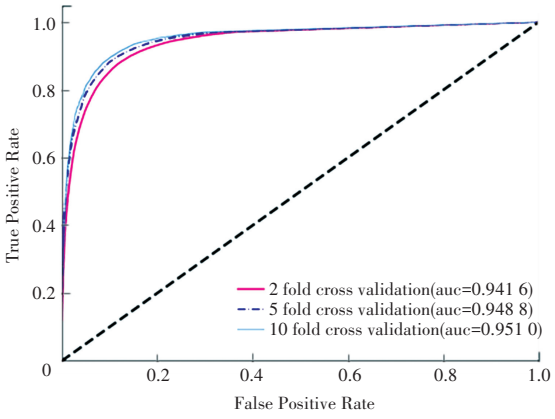

图 4 不同交叉验证方法下预测性能

Fig. 4 Prediction performance under different cross validation methods

4 结束语

为了更高效地预测药物-细胞系之间的敏感性响应关系,本文在图卷积神经网络的基础上提出了基于异构图卷积网络的药物-细胞系响应预测方法(HGCNDCP)。研究表明:

(1)使用药物结构分子指纹特征数据、细胞系的基因表达谱数据和药物-细胞系作用关系数据,对学习药物、细胞系的特征提取有重要影响。

(2)通过构建异构网络,使用图卷积神经网络,能够有效地聚合邻居特征信息,得到较好的药物和细胞系的表征。

(3)通过使用 GDSC 数据集,并与其它算法的一系列实验比对,HGCNDCP 具有较高的预测精度,能够较好地预测药物细胞系响应,从而为药物细胞系响应预测提供有效的思路和方法。

参考文献

[1] GHOLAMI A, HAHNE H, WU Z, et al. Global Proteome Analysis of the NCI-60 Cell Line Panel[J]. Cell Reports,2013, 4 (3): 609-620.

[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.

[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(D1): 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] 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.

[6] KORAS K, KIZLING E, JURAEVA D, et al. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines[J]. Scientific Reports, 2021, 11 (1): 1-16.

[7] LIU Q, HU Z, JIANG R, et al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response[J]. Bioinformatics, 2020, 36(Supplement_2): i911-i918.

[8] STANFIELD Z, COŞKUN M, KOYUTÜRK M. Drug response prediction as a link prediction problem[J]. Scientific Reports, 2017, 7(1): 40321.

[9] LI M, WANG Y, ZHENG R, et al. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines[J]. IEEE/ACM Trans Comput Biol Bioinform,2021, 18 (2): 575-582.

[10]徐国保, 陈媛晓, 王骥. 基于图卷积网络的药物靶标关联预测算法[J]. 计算机应用,2021, 41 (5): 5.

[11]FENG Y H, ZHANG S W, SHI J Y. DPDDI: a deep predictor for drug-drug interactions[J]. BMC bioinformatics, 2020, 21(1): 1-15.

[12]IORIO F, KNIJNENBURG T A, VIS D J, et al. A Landscape of Pharmacogenomic Interactions in Cancer - ScienceDirect[J]. Cell,2016, 166 (3): 740-754.

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

[14] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[15]IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// International conference on machine learning. pmlr, 2015: 448-456.

[16] ZHANG F, WANG M, XI J, et al. A novel heterogeneous network-based method for drug response prediction in cancer cell lines[J]. Scientific Reports,2018, 8 (1): 3355.