Bioinformatics, 38(19), 2022, 4546–4553
https://doi.org/10.1093/bioinformatics/btac574
Advance Access Publication Date: 23 August 2022
Original Paper

OXFORD

Systems biology

Predicting cancer drug response using parallel heterogeneous graph convolutional networks with neighborhood interactions

Wei Peng [1,*], Hancheng Liu[1], Wei Dai[1], Ning Yu[2] and Jianxin Wang [3,4,*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650050, P.R. China, [2]Department of Computing Sciences, The College at Brockport, State University of New York, Brockport, NY 14422, USA, [3]School of Computer Science and Engineering, Central South University, Changsha 410083, P.R. China and [4]Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, P. R. China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 3, 2022; revised on July 26, 2022; editorial decision on August 15, 2022; accepted on August 22, 2022

Abstract

Motivation: Due to cancer heterogeneity, the therapeutic effect may not be the same when a cohort of patients of the same cancer type receive the same treatment. The anticancer drug response prediction may help develop personalized therapy regimens to increase survival and reduce patients' expenses. Recently, graph neural network-based methods have aroused widespread interest and achieved impressive results on the drug response prediction task. However, most of them apply graph convolution to process cell line-drug bipartite graphs while ignoring the intrinsic differences between cell lines and drug nodes. Moreover, most of these methods aggregate node-wise neighbor features but fail to consider the element-wise interaction between cell lines and drugs.

Results: This work proposes a neighborhood interaction (NI)-based heterogeneous graph convolution network method, namely NIHGCN, for anticancer drug response prediction in an end-to-end way. Firstly, it constructs a heterogeneous network consisting of drugs, cell lines and the known drug response information. Cell line gene expression and drug molecular fingerprints are linearly transformed and input as node attributes into an interaction model. The interaction module consists of a parallel graph convolution network layer and a NI layer, which aggregates node-level features from their neighbors through graph convolution operation and considers the element-level of interactions with their neighbors in the NI layer. Finally, the drug response predictions are made by calculating the linear correlation coefficients of feature representations of cell lines and drugs. We have conducted extensive experiments to assess the effectiveness of our model on Cancer Drug Sensitivity Data (GDSC) and Cancer Cell Line Encyclopedia (CCLE) datasets. It has achieved the best performance compared with the state-of-the-art algorithms, especially in predicting drug responses for new cell lines, new drugs and targeted drugs. Furthermore, our model that was well trained on the GDSC dataset can be successfully applied to predict samples of PDX and TCGA, which verified the transferability of our model from cell line *in vitro* to the datasets *in vivo*.

Availability and implementation: The source code can be obtained from https://github.com/weiba/NIHGCN.

Contact: weipeng1980@gmail.com or jxwang@mail.csu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancers have become one of the leading course of death, seriously threatening human health. Although some anticancer drugs have been developed, the therapeutic effect may not be the same when a cohort of patients of the same cancer type receive the same treatment because of their difference in genomic profiles (Lloyd *et al.*, 2021; Rubin, 2015). Hence, the therapy response prediction may help develop personalized therapy regimens to increase survival and reduce patients' expenses (Adam *et al.*, 2020; Menden *et al.*, 2019; Xia *et al.*, 2022). However, experimental testing of multiple therapeutic strategies for a patient is infeasible for practical and financial reasons. Recently, several large-scale anticancer drug screen projects have produced massive amounts of drug sensitivity profiles for thousands of cancer cell lines, and become available through public

repositories, such as Genomics of Drug Sensitivity in Cancer (GDSC) (Yang *et al.*, 2012), Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012), Cancer Therapeutics Response Portal (CTRP) (Seashore-Ludlow *et al.*, 2015) and NCI-60 (Ross *et al.*, 2000). These public datasets enable researchers to build computational methods that analyze the differences in the cancer cell line genomic and transcriptomic profiles to predict drug sensitivity or resistance.

Current anticancer drug response prediction algorithms can be divided into three broad categories from the perspective of the prediction task. One formulates the prediction task of cell–drug responses as regression where continuous values such as the half-maximal inhibitory concentration (IC50) values are predicted. These models usually leverage regression models, such as ridge regression (Geeleher *et al.*, 2014), LASSO (Tibshirani, 1996) and elastic network (Zou and Hastie, 2005), to infer a causal relationship between cell line expression profiles and drug response. However, these models cannot efficiently extract non-linear features of drugs and cell lines or fit their relationship correctly. The other kind of method is the classification-based method. These methods firstly extract features of cell lines or drugs and input these features into classifiers, such as support vector machine (SVM) (Huang *et al.*, 2017), Random Forest (Lind and Anderson, 2019; Su *et al.*, 2019) and convolutional neural network (CNN) (Liu *et al.*, 2019, 2020), to predict IC50 values or indicate whether the drug is sensitive or resistant. Considering the cell line omics data and the drug chemical profiles are high-dimensional and complex, some researchers frequently employ neural network-based methods, such as autoencoder (AE), stacked AE (Li *et al.*, 2019), variation AE (Rampášek *et al.*, 2019) and graph convolutional neural network (Liu *et al.*, 2020), to learn features of cell lines or drugs and predict drug response. Recently, some multi-omics data-based models have been proposed and achieved impressive performance in drug response prediction because of the complementarity in the multi-omics data. For example, Su *et al.* (2019) used a cascade forest model to stitch together high-dimensional feature vectors, including cell line gene expression and copy number variation for individual drug response prediction. Sharifi-Noghabi *et al.* (2019) proposed MOLI. This self-encoder-based multi-omics late-integration algorithm used a neural network to generate features of cell lines by encoding their gene expression, somatic mutation and copy number variation data. They then concatenated these features and used a neural network classifier to predict the drug response to a cell line. The shortcoming of the classification-based methods was that most of them fail to extensively use the known inter- and intra-associations of cell lines and drugs. The third class of models casts the drug response prediction as a link prediction problem. These models were proposed based on the assumption that chemically and structurally similar compounds may have a similar biological effect on known cell lines. Zhang *et al.* (2018) constructed a heterogeneous network including cell line, drug and target genes and their connections. They executed an information flow-based algorithm on the heterogeneous network to infer the potential drug responses to a cell line. Wang *et al.* (2017) considered similarity among cell lines, drugs and targets and used a similarity-regularization matrix decomposition (SRMF) approach to decompose known drug-cell line associations into drug features and cell line features. Then they utilized the drug and cell line features to reconstruct the drug-cell line response matrix. Peng *et al.* (2022) fused multiple cell line multi-omics data and drug chemical structure data to construct a cell line-drug heterogeneous network and implement graph convolution operations on the heterogeneous network to learn cell line and drug features. The drug responses were inferred from the values of the reconstructed drug-cell line response matrix. Liu *et al.* (2022) constructed graph neural networks for cancer drug response (CDR) prediction and used the contrast learning task as a regularizer in a multi-task learning paradigm. The graph convolutional network (GCN)-based methods can simultaneously consider the nodes' features and network structures to learn features of drugs and cells, which succeed in drug response prediction (Liu *et al.*, 2022; Peng *et al.*, 2022). However, most previous GCN-based methods run graph convolution on the cell line-drug bipartite graph (Liu *et al.*, 2022; Peng *et al.*, 2022). They collected information from neighborhoods with different node types (i.e. cell lines aggregate drugs' features, and drugs aggregate cell lines' features) and ignored the intrinsic differences between cell lines and drug nodes. Moreover, most GCN-based approaches implemented graph convolution operations as linear feature aggregations (i.e. weighted sums) of the target node's neighbors. However, the interactions within node features provided helpful signals for node feature encoding (Lian *et al.*, 2018).

Considering the above issues, we propose a heterogeneous GCN model for anticancer drug response prediction based on different levels of neighborhood interaction (NI), namely NIHGCN (see Fig. 1). We first construct a heterogeneous network where the nodes are drugs, and cell lines and the edges are composed of the known response information. After that, the cell line gene expression and drug molecular fingerprints are linearly transformed separately to obtain their feature vectors with the same dimension. We then input the feature vectors of cell lines and drugs as node attributes into an interaction module consisting of a parallel graph convolution network (PGCN) layer and a NI layer. The cell line and drugs not only aggregate node-level features from their neighbors in the heterogeneous network through graph convolution operation but also consider the element-level of interactions with their neighbors in the NI layer. Finally, we use the linear correlation coefficient to estimate the degree of drug response to the cell lines. Extensive comparative experiments were conducted to assess the effectiveness of our model. Our model achieved good performance on GDSC and CCLE datasets compared with the state-of-the-art algorithms. We also verified our model by training it on the *in vitro* dataset GDSC and predicting drug response for PDX and patients of the TCGA dataset. The results demonstrate the transportability of our model from *in vitro* data to *in vivo* data.

2 Materials

We adopted the following four datasets to build and test our model. They are Genomics of Drug Sensitivity in Cancer (GDSC) (Yang *et al.*, 2012), Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012), PDX Encyclopedia dataset (PDX) (Gao *et al.*, 2015) and The Cancer Genome Atlas(TCGA) (Ding *et al.*, 2016), which record gene expression data in cell lines and corresponding drug responses profiles in the cell lines.

The GDSC database provides two tables, i.e. Supplementary Table S4A (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/suppData/TableS4A.xlsx) and Supplementary Table S5A (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/suppData/TableS5C.xlsx), to help us infer drug sensitivity and resistance states. Supplementary Table S4A is a logarithmic matrix of half-maximal inhibitory concentration (IC50) values for all screened cell line/drug combinations, containing 990 cancer cell lines and 265 tested drugs. Supplementary Table S5C records the sensitivity thresholds for the 265 drugs. A drug is considered sensitive in a cell line if its value in Supplementary Table S4A does not exceed the threshold in Supplementary Table S5C. Otherwise, it is resistant to the cell line (Sharifi-Noghabi *et al.*, 2019). The CCLE database provides 11 670 records of cell line-drug trials. Each record reports experimental information such as drug target, dose, log(IC50) and effective area. According to the methods of Staunton *et al.* (2001) and Peng *et al.* (2022), we categorized the drug response into sensitivity and resistance by comparing its log(IC50) values with a predefined threshold.

Our approach involves gene expression and drug substructure fingerprinting. Drug substructure fingerprints came from the PubChem (Bolton *et al.*, 2008) database. Gene expression data were obtained from GDSC and CCLE databases, and the preprocessing method was similar to Sharifi-Noghabi *et al.* (2019). After preprocessing, we got 962 cell lines and 228 drugs in the GDSC database. A binary response matrix was constructed in our model according to thresholds, with 1 representing sensitivity, 0 representing resistance and missing values filled with 0. The missing values were masked and not involved in model optimization in our model.

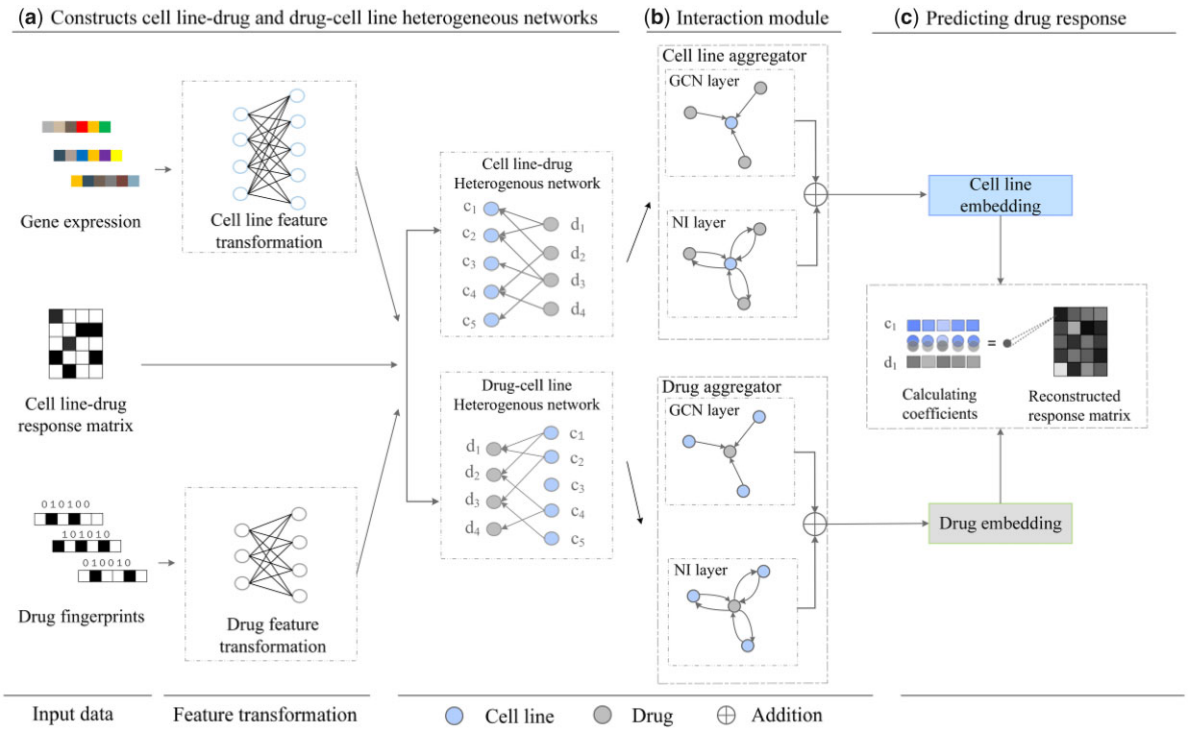

Fig. 1. The framework of the NIHGCN. (**a**) It constructs a heterogeneous bipartite network consisting of drugs, cell lines and the known drug response information. Drug fingerprint and gene expression went separately transformed to feature vectors. (**b**) It inputs the cell line and drug feature vectors into the heterogeneous network and employs an interaction module including a parallel GCN layer and a NI layer to learn feature representations of cell line and drug separately. (**c**) The drug response is predicted by calculating the correlation coefficient between cell line and drug embeddings

Similarly, we obtained 436 cancer cell lines and 24 drugs in the CCLE database.

PDX dataset is available in Supplementary File of Gao *et al.* (2015), containing gene expression profiles and drug responses values. We obtained six drugs shared by GDSC and PDX. PDX drug responses were divided into two groups, responses ('CR' and 'PR') and non-response ('SD' and 'PD'). After preprocessing the gene expression profiles (Sharifi-Noghabi *et al.*, 2019), we obtained 191 response records for 6 drugs from the PDX dataset. We retrieved clinical annotations of patients' drug responses in the TCGA dataset from the Supplementary Material of Ding *et al.* (2016) and obtained 22 drugs shared by GDSC and TCGA. TCGA drug responses were divided into two groups, responses ('Complete Response' and 'Partial Response') and non-response ('Stable Disease' and 'Progressive Disease'). The corresponding gene expression of the TCGA samples was from FirehoseBroadGDAC (http://gdac.broadinstitute.org/runs/stddata__2016_01_28/). After preprocessing the gene expression profiles (Sharifi-Noghabi *et al.*, 2019), we obtained 430 response records for 22 drugs from the TCGA dataset. Supplementary Table S1 summarizes the statistics of every dataset. The detailed processing and response recorder for drug and cell lines in every dataset, please refer to the Supplementary Files.

3 Methods

3.1 Overview

We cast the drug response prediction as the task of revealing the missing links between the drugs and cell lines. NIHGCN takes four steps to finish the predictions. Figure 1 shows the structure of NIHGCN. First, cell line gene expression and drug molecular fingerprints are linearly transformed to obtain drug and cell line properties with the same dimension. Then NIHGCN inputs the cell line and drug feature vectors into a heterogeneous network and employs an interaction module including a parallel graph convolution layer (GCN) and a NI layer to learn feature representations of cell line and drug separately. Finally, the missing links between cell lines and drugs are recovered by calculating the linear correlation coefficients of feature representations of cell lines and drugs.

3.2 Cell line/drug property extraction

The original features of the cell line are gene expression and the original features of the drug are drug molecular fingerprints, which have different dimensions. Therefore, we transform the original features of the cell line and the drug separately and then use these transformed features of the same dimension as properties of the drug and the cell line for further analysis.

At first, we preprocess the gene expression data in cell lines by calculating their $\overline{expr_i}$ (see Equation (1)):

$$\overline{expr_i} = \frac{(expr_i - u_i)}{\sigma_i} \tag{1}$$

where $expr_i$ is a column vector representing the expression value of the ith gene in all cell lines, and the mean u_i and standard deviation σ_i represent the expression abundance of the ith gene in all cell lines. After normalizing the cell line, we obtain the cell line features by Equation (2):

$$H_c = C \cdot \theta_c \tag{2}$$

where $C \in R^{m \times h_c}$ is the cell line gene expression matrix preprocessed by Equation (1), m is the numbers of cell lines, $H_c \in R^{m \times h}$ is the cell line projection feature matrix, $\theta_c \in R^{h_c \times h}$ is the set of parameters of the cell line linear transformation.

Drug properties are obtained directly by linear transformation using drug molecular fingerprints:

$$H_d = D \cdot \theta_d \tag{3}$$

where $D \in R^{n \times h_d}$ drug molecular fingerprint matrix, n is the numbers of drugs, $H_d \in R^{n \times h}$ is the drug projection feature matrix, $\theta_d \in R^{h_d \times h}$ is the set of parameters of the drug linear transformation.

3.3 Interaction module

This work aims to learn proper feature representations of cell lines and drugs to reconstruct the drug-cell line reaction matrix. We first build a bipartite heterogeneous network based on the known drug responses in cell lines, i.e. sensitive or resistant. Then the drugs and cell lines learnt their feature representations by transforming and aggregating the neighborhood feature information on the graph architecture using an interaction module. The interaction module consists of a PGCN layer and a NI layer, aggregating features from neighbors at the node and element level.

Let $G = (A \in \{0,1\}^{(m\times n)}, H_c \in R^{m\times h}, H_d \in R^{n\times h})$ represent the bipartite heterogeneous network. A is the adjacent matrix of the network, whose rows correspond to cell lines and columns to correspond drugs. A value of 1 in A_{cd} indicates that cell line c reacts with drug d. H_c and H_d are the attribute matrices of cell lines and drugs, calculated by Equations (2) and (3). $\mathcal{N}(c)$ and $\mathcal{N}(d)$ denote the neighbor set of the cell line c and drug d in the heterogeneous network, respectively. $\tilde{\mathcal{N}}(c) = \{c\} \cup \mathcal{N}(c)$ and $\tilde{\mathcal{N}}(d) = \{d\} \cup \mathcal{N}(d)$ after adding self-connections for cell line c and drug node d. $\tilde{q}_c = |\tilde{\mathcal{N}}(c)|$ and $\tilde{q}_d = |\tilde{\mathcal{N}}(d)|$ are the degree of cell line c and drug d.

3.3.1 Parallel graph convolution network layer

Considering the heterogeneous nature of the cell line-drug bipartite graph, we implemented two parallel graph convolution operations on the two-part heterogeneous network and independently aggregated node-wise features from neighbors for the cell lines and drugs. The process can be mathematically described as follows:

$$GCN\left(\left\{h_d^{(k-1)}\right\}_{d\in\tilde{\mathcal{N}}(c)}\right) = h_c^{(k-1)}W_c^{(k)} + \sum_{d\in\tilde{\mathcal{N}}(c)} \frac{1}{\sqrt{\tilde{q}_c\tilde{q}_d}} h_d^{(k-1)}W_c^{(k)} \tag{4}$$

$$GCN\left(\left\{h_c^{(k-1)}\right\}_{c\in\tilde{\mathcal{N}}(d)}\right) = h_d^{(k-1)}W_d^{(k)} + \sum_{c\in\tilde{\mathcal{N}}(d)} \frac{1}{\sqrt{\tilde{q}_c\tilde{q}_d}} h_c^{(k-1)}W_d^{(k)} \tag{5}$$

where $h_c^{(k-1)}$ and $h_d^{(k-1)}$ denote the representations of the cell line node c and the drug node d at the $(k-1)$th iteration, respectively, and the $h_c^{(0)}$ and $h_d^{(0)}$ are the initial cell line and drug attributes obtained from Equations (2) and (3). We calculated the Laplacian matrices $\mathcal{L}_c = D_c^{-\frac{1}{2}}AD_d^{-\frac{1}{2}}$ and $\mathcal{L}_d = D_d^{-\frac{1}{2}}A^TD_c^{-\frac{1}{2}}$ for cell line aggregation and drug aggregation, respectively, where $D_{c(ij)} = \sum_j A_{ij} + 1$ and $D_{d(ij)} = \sum_j A_{ji} + 1$. Considering the features of the nodes themselves, we introduced the cell line self-features $SC = \left(D_c^{-1} + I_m\right)H_c^{(k-1)}$ and the drug self-features $SD = \left(D_d^{-1} + I_n\right)H_d^{(k-1)}$. To numerically solve Equations (4) and (5), we rewrote them in the matrix-vector format as follows:

$$GCN\left(\left\{h_d^{(k-1)}\right\}_{d\in\tilde{\mathcal{N}}(c)}\right) = (SC + \mathcal{L}_cH_d^{(k-1)})W_c^{(k)} \tag{6}$$

$$GCN\left(\left\{h_c^{(k-1)}\right\}_{c\in\tilde{\mathcal{N}}(d)}\right) = (SD + \mathcal{L}_dH_c^{(k-1)})W_d^{(k)} \tag{7}$$

3.3.2 Neighborhood interaction layer

Dot product between two vectors can emphasize common properties and dilute divergent information. In the NI layer, we captured fine-grained neighbor features by element-wise dot product the target node with its neighbor node (see Equations (8) and (9)):

$$NI\left(\{h_d^{(k-1)}\}_{d\in\tilde{\mathcal{N}}(c)}\right) = h_c^{(k-1)} \odot h_c^{(k-1)}W_c^{(k)} + \sum_{d\in\tilde{\mathcal{N}}(c)} \frac{1}{\sqrt{\tilde{q}_c\tilde{q}_d}} \left(h_c^{(k-1)} \odot h_d^{(k-1)}\right)W_c^{(k)} \tag{8}$$

$$NI\left(\{h_c^{(k-1)}\}_{c\in\tilde{\mathcal{N}}(d)}\right) = h_d^{(k-1)} \odot h_d^{(k-1)}W_d^{(k)} + \sum_{c\in\tilde{\mathcal{N}}(d)} \frac{1}{\sqrt{\tilde{q}_c\tilde{q}_d}} \left(h_d^{(k-1)} \odot h_c^{(k-1)}\right)W_d^{(k)} \tag{9}$$

where $\odot$ stands for element-wise dot product, which can emphasize the common properties of a node pair. Similarly, Equations (10) and (11) can be written in the matrix-vector format as follows:

$$NI\left(\left\{h_d^{(k-1)}\right\}_{d\in\tilde{\mathcal{N}}(c)}\right) = (SC + \mathcal{L}_cH_d^{(k-1)}) \odot H_c^{(k-1)}W_c^{(k)} \tag{10}$$

$$NI\left(\left\{h_c^{(k-1)}\right\}_{c\in\tilde{\mathcal{N}}(d)}\right) = (SD + \mathcal{L}_dH_c^{(k-1)}) \odot H_d^{(k-1)}W_d^{(k)} \tag{11}$$

3.3.3 Heterogeneous aggregation module

By considering both the graph convolution layer and the NI layer, we obtained the cell line and drug aggregation functions, respectively (see Equations (12) and (13)):

$$H_c^{(k)} = \sigma\left((1-\alpha)(SC + \mathcal{L}_cH_d^{(k-1)})W_c^{(k)} + \alpha(SC + \mathcal{L}_cH_d^{(k-1)}) \odot H_c^{(k-1)}W_c^{(k)}\right) \tag{12}$$

$$H_d^{(k)} = \sigma\left((1-\alpha)(SD + \mathcal{L}_dH_c^{(k-1)})W_d^{(k)} + \alpha(SD + \mathcal{L}_dH_c^{(k-1)}) \odot H_d^{(k-1)}W_d^{(k)}\right) \tag{13}$$

where σ is the ReLU activation function and α is a hyper-parameter to balance the contribution of the two parts. $H_c^{(k)} \in R^{m\times f}$ and $H_d^{(k)} \in R^{n\times f}$ are the final representations of the cell lines and drugs obtained after k-step embedding propagation, respectively. $H_c^{(0)} \in R^{m\times h}$ and $H_d^{(0)} \in R^{n\times h}$ are the initial attributes H_c and H_d for cell lines and drugs, respectively. $W_c^{(k)} \in R^{h\times f}$ and $W_d^{(k)} \in R^{h\times f}$ are the weight parameters of the cell line and drug aggregators, respectively, where we used different weight matrices for cell lines and drugs to aggregate features separately.

3.4 Drug response prediction

Finally, we used the linear correlation coefficient to estimate the degree of drug response to the cell line. The linear correlation coefficient between the feature vectors of drugs and cell lines is defined as follows:

$$\mathrm{Corr}(h_i, h_j) = \frac{(h_i - \mu_i)(h_j - \mu_j)^T}{\sqrt{(h_i - \mu_i)(h_i - \mu_i)^T}\sqrt{(h_j - \mu_j)(h_j - \mu_j)^T}} \tag{14}$$

where $h_i \in H_c^k$ and $h_j \in H_d^k$ are the feature representation vectors of the ith cell line and jth drug, respectively, μ_i and μ_j are the means of h_i and h_j, respectively. Since the correlation coefficient takes values in the range $[-1,1]$, we used Equation (15) to activate the output:

$$\varphi(h) = \frac{1}{1 + e^{-\gamma h}} \tag{15}$$

Since most drug response data are negative samples, an appropriate parameter γ allows Equation (15) to have a more appropriate gradient for positive and negative samples than directly using the sigmoid activation function, facilitating model convergence and accelerating parameter updating. Finally, the cell line-drug association matrix is reconstructed as:

$$\hat{A} = \varphi\left(\mathrm{Corr}\left(H_c^k, H_d^k\right)\right) \tag{16}$$

We adopted the following loss function for the model constraint:

$$\mathcal{L}(A, \hat{A}) = -\frac{1}{m\times n}\sum_{i,j} M_{ij}\left[A_{ij}\ln\left(\hat{A}_{ij}\right) + (1 - A_{ij})\ln\left(1 - \hat{A}_{ij}\right)\right] \tag{17}$$

where m and n represent the number of cell lines and drugs respectively. M is an indicator matrix. When the association between the

ith cell line and the jth drug is in the training set, $M_{ij} = 1$, otherwise $M_{ij} = 0$. In summary, Supplementary Algorithm S1 gives the pseudo-code for NIHGCN (see Supplementary Files). We used the PyTorch framework to implement the model code and the Adam optimizer to optimize the loss function. For the parameter setting, please refer to Supplementary Files.

4 Experiments

4.1 Baseline

To evaluate the performance of our model, we compare our approach with the following baselines.

- HNMDRP (Zhang *et al.*, 2018) predicts novel cell line-drug associations by incorporating cell line genomic profile, drug chemical structure, drug-target and PPI information. It learns cell line features and drug features through network propagation.
- SRMF (Wang *et al.*, 2017) learns drug and cell line feature representations based on the similarity of the regularization matrix decomposition. It uses the feature representations to reconstruct the drug-cell line reaction matrix and infers drug response in cell lines.
- DeepDSC (Li *et al.*, 2019) uses a stacked deep self-encoder to extract cell features from gene expression data and then combines them with drug chemistry features for drug response prediction.
- DeepCDR (Liu *et al.*, 2020) is a hybrid GCN-based anticancer drug response predictor integrating multi-omics profiles of cell lines and drugs.
- MOFGCN (Peng *et al.*, 2022) calculates cell line similarity and drug similarity as the cell line and drug's input features. It runs a graph convolutional neural network to diffuse cell line and drug similarity and reveals potential associations between cell lines and drugs.
- GraphCDR (Liu *et al.*, 2022) is the latest method for drug response prediction using a graph convolution network and a contrast learning framework.

4.2 Experimental design

In this experiment, we used both datasets *in vitro* (i.e. GDSC and CCLE) and *in vivo* (i.e. PDX encyclopedia and TCGA) to build and test our model. All methods used gene expression for cell lines and molecular fingerprints for drugs for a fair comparison. Each model was evaluated using ROC curves and PR curves and their area under the curves, AUC and AUPRC. In our experiments, we tested our model and baselines under the following five settings:

- Test 1: Comparing our model with HNMDRP, SRMF, DeepCDR, DeepDSC, MOFGCN and GraphCDR for cell line-drug association matrix reconstruction when randomly zeroing values in the matrix.
- Test 2: Comparing our model with HNMDRP, SRMF, DeepCDR, DeepDSC, MOFGCN and GraphCDR for cell line-drug association matrix reconstruction when blinding to one row or column at random.
- Test 3: Comparing our model with HNMDRP, SRMF, DeepCDR, DeepDSC, MOFGCN and GraphCDR when training model *in vitro* data and predicting drugs *in vivo* data.
- Test 4: Doing ablation study for our model to investigate the performance contribution of the different components.
- Test 5: Doing case study of our model to see if it can detect novel drug responses in cell lines.

4.3 Experimental results

4.3.1 Randomly zeroing values cross-validation (Test 1)

The HNMDRP, SRMF, DeepCDR, DeepDSC, MOFGCN, GraphCDR and our model were link prediction-based methods that predict drug response by estimating the association probability between drugs and cell lines. In experimental design Test 1, we randomly divided known cell line-drug associations (positive samples) into five equal fractions and performed five times 5-fold cross-validation. In each round of validation, we selected 1/5 of the positive samples and an equal number of negative samples from the association matrix as test data. The remaining 4/5 positive and the remaining negative samples were training data. The main findings of the experimental results are as follows:

Table 1 shows that our model NIHGCN consistently outperformed all baselines on both datasets when classifying whether the drugs are sensitive or resistant to cell lines. The AUC and AUPRC values of our model achieved 87.60% and 88.03% on GDSC, and 88.06% and 88.03% on CCLE, which are 0.76% higher in AUC and 0.73% higher in AUPRC on GDSC and 1.98% higher in AUC and 2.14% higher in AUPRC on CCLE compared with the best baseline MOFGCN. MOFGCN is a GCN-based approach. The outperformance of our model shows the importance of capturing NI information and considering the heterogeneity of the network nodes. The GCN-based approaches (e.g. MOFGCN, GraphCDR) perform better on two datasets than deep learning-based baselines such as DeepCDR and DeepDSC and are also much better than the matrix decomposition-based method SRMF and the network propagation-based approach HTMDRP. An intuitive explanation is that the GCN-based approach can explicitly encode drugs and cell lines by aggregating information from neighbors with similar response signals.

To verify whether the output of Equation (14) correlates with IC50s, we added the IC50 regression test. We modified the output of our model as Equation (14) to estimate the drug's IC50 values and adopted three regression indicators, Pearson's correlation (PCC), Spearman's correlation (SCC) and Root Mean Square Error (RMSE) for evaluation. We conduct the regression test under the randomly zeroing values cross-validation on GDSC and CCLE datasets. Supplementary Table S2 reports that the predictions of our model show higher correlates with IC50s than other competing methods.

4.3.2 Predicting responses to new drugs or new cell lines (Test 2)

Test 2 aims to assess the ability of every method to predict responses to new drugs or new cell lines. The rows in the cell line-drug association matrix represent cell lines and the columns represent drugs. We cleared one row or column of the cell line-drug association matrix as the test set and selected the remaining rows or columns as the training set. To avoid being too general or too specific, we only tested rows or columns that contained at least ten positive samples (Staunton *et al.*, 2001). After screening, we obtained 227 of 228 drugs and 658 of 962 cell lines in the GDSC experiments and 20 of 24 drugs and 26 of 436 cell lines in the CCLE experiments. Table 2 reports the average AUC, AUPRC values of all models across all testing cell lines or drugs on the GDSC datasets, and the results on the CCLE dataset are presented in Supplementary Table S3. Our model still controls the best performance under this test set. Compared to the best baseline DeepDSC, our model achieved 4.36% and 3.52% improvement in AUC and AUPRC on the GDSC dataset and 1.49% and 1.13% improvement in AUC and AUPRC on the CCLE dataset in the single-row clearing test. In the single-column clearing test, the AUC and AUPRC of our model were 3.13% and 3.76% higher than the best base line DeepCDR on GDSC and are 2.31% and 1.83% higher than DeepCDR on the CCLE dataset.

When a new drug has no prior known associations with cell lines, it generates features from its initial attributes. However, our model can still successfully identify drug responses to cell lines, which may be attributed to the parallel GCN strategy effectively

Table 1. Results of randomly zeroing values cross-validation on GDSC and CCLE datasets

	GDSC		CCLE	
Algorithm	AUC	AUPRC	AUC	AUPRC
HNMDRP	$0.7258 \pm 3 \times 10^{-5}$	$0.7198 \pm 4 \times 10^{-5}$	$0.7104 \pm 1 \times 10^{-4}$	$0.6956 \pm 2 \times 10^{-4}$
SRMF	$0.6563 \pm 2 \times 10^{-4}$	$0.6605 \pm 5 \times 10^{-5}$	$0.7669 \pm 4 \times 10^{-5}$	$0.7418 \pm 2 \times 10^{-5}$
DeepCDR	$0.7849 \pm 5 \times 10^{-5}$	$0.7827 \pm 6 \times 10^{-5}$	$0.8289 \pm 1 \times 10^{-4}$	$0.8185 \pm 2 \times 10^{-4}$
DeepDSC	$0.8118 \pm 4 \times 10^{-4}$	$0.8311 \pm 1 \times 10^{-4}$	$0.8594 \pm 1 \times 10^{-4}$	$0.8607 \pm 1 \times 10^{-4}$
MOFGCN	$0.8684 \pm 7 \times 10^{-6}$	$0.8730 \pm 1 \times 10^{-5}$	$0.8608 \pm 1 \times 10^{-4}$	$0.8589 \pm 1 \times 10^{-4}$
GraphCDR	$0.8136 \pm 4 \times 10^{-5}$	$0.8193 \pm 3 \times 10^{-5}$	$0.8474 \pm 2 \times 10^{-4}$	$0.8495 \pm 2 \times 10^{-4}$
Ours[1]	$\mathbf{0.8760} \pm 1 \times 10^{-5}$	$\mathbf{0.8803} \pm 1 \times 10^{-5}$	$\mathbf{0.8806} \pm 1 \times 10^{-4}$	$\mathbf{0.8803} \pm 1 \times 10^{-4}$

[1]The best results are in bold font.

Table 2. Results of new drugs or new cell lines response prediction on GDSC datasets

	New cell lines		New drugs	
Algorithm	AUC	AUPRC	AUC	AUPRC
HNMDRP	—	—	$0.6951 \pm 1 \times 10^{-2}$	$0.6935 \pm 1 \times 10^{-2}$
SRMF	$0.5807 \pm 1 \times 10^{-2}$	$0.6153 \pm 1 \times 10^{-2}$	$0.6683 \pm 6 \times 10^{-3}$	$0.6757 \pm 6 \times 10^{-3}$
DeepCDR	$0.7526 \pm 8 \times 10^{-3}$	$0.7664 \pm 8 \times 10^{-3}$	$0.7605 \pm 9 \times 10^{-3}$	$0.7565 \pm 1 \times 10^{-2}$
DeepDSC	$0.7831 \pm 8 \times 10^{-3}$	$0.7994 \pm 8 \times 10^{-3}$	$0.7472 \pm 1 \times 10^{-2}$	$0.7514 \pm 1 \times 10^{-2}$
MOFGCN	$0.7190 \pm 5 \times 10^{-3}$	$0.7366 \pm 5 \times 10^{-3}$	$0.7601 \pm 7 \times 10^{-3}$	$0.7558 \pm 8 \times 10^{-3}$
GraphCDR	$0.7122 \pm 9 \times 10^{-3}$	$0.7061 \pm 9 \times 10^{-3}$	$0.7614 \pm 8 \times 10^{-3}$	$0.7501 \pm 9 \times 10^{-3}$
Ours[1]	$\mathbf{0.8267} \pm 7 \times 10^{-3}$	$\mathbf{0.8346} \pm 8 \times 10^{-3}$	$\mathbf{0.7927} \pm 6 \times 10^{-3}$	$\mathbf{0.7877} \pm 6 \times 10^{-3}$

[1]The best results are in bold font.

learning cell lines' features independently. Similarly, the new cell lines successfully detect their reactions to drugs even though they have no neighbors in the heterogeneous network. It may be the drugs learn feature representations independently through the parallel GCN. We also noticed that HNMDRP failed to produce predicted results when clearing a row. The HNMDRP only infers potential drug-cell line associations from drugs with similar chemical structures and similar drug targets and does not consider cell lines with similar gene expression profiles. Therefore, HNMDRP cannot predict drug response to new cell lines.

Targeted drugs are a group of drugs that interact with genes related to cancer progression. We selected drugs from the GDSC and CCLE databases that target CDK4 to investigate the performance of our model in predicting response to targeted drugs. CDK4 is a cyclin-dependent kinase that drives cell proliferation by pairing with other proteins that inhibit retinoblastoma (Staunton *et al.*, 2001). We screened three targeted drugs from the GDSC database, PD-0332991, AT-7519 and CGP-082996, and one targeted drug, PD-0332991, from the CCLE database. For one of the targeted drugs, we cleared its corresponding column in the cell line-drug association matrix as the test set and trained the model on the remaining columns to reveal responses to the targeted drug. Supplementary Table S4 reports that our model achieved a good performance in predicting responses to the targeted drugs. To further prove the validity of the model, we employed a k-means method to cluster the cell lines into two groups according to their embeddings from NIHGCN or their original features. The cell lines are represented as points on a 2D map via the UMAP tool. Supplementary Figure S1 illustrates that compared to the initial features, using their embeddings from NIHGCN can divide the cell lines into two groups well. Moreover, we also leveraged two internal indicators, Silhouette Coefficient (SC) and Davies-bouldin Index (DBI) and an external indicator, Normalized Mutual Information (NMI), to evaluate the quality of the clustering results. Supplementary Table S5 shows that the features learned by our model can successfully separate cell lines into two groups with the high compactness within clusters and the high separation among clusters. The higher NMI values indicate that our embedding features categorize the cell lines into a sensitive and resistant group that has a higher overlap with benchmark datasets.

4.3.3 Drug response predictions *in vivo* dataset (Test 3)

It is a challenge to transfer the drug response prediction model learned in cell line (*in vitro*) to clinical contexts *in vivo* (Ma *et al.*, 2021). In Test 3, we trained our model and other baselines on GDSC *in vitro* dataset and then applied them to predict drug response in two *in vivo* datasets, i.e. patient-derived xenograft (PDX) in mice dataset and TCGA patient data. There are different numbers of genes involved in GDSC, PDX and TCGA. The numbers of genes in the samples of GDSC, PDX and TCGA are 19 451, 22 378 and 20 531, respectively. Hence we focus on the common genes in the samples of GDSC and PDX, GDSC and TCGA. When predicting PDX samples, we chose the expressions of the 18 942 genes shared by GDSC and PDX as the inputting cell line features. Similarly, when predicting TCGA samples, we selected the expressions of the 18 948 genes shared by GDSC and TCGA as the input cell line features. The data from the two datasets may not follow the same distribution and appear to be batch effects. Before training and testing the models, we gathered the data from the two datasets. Then we scaled the gene expression data by the mean gene expression and its standard deviation over all cell lines and samples in Equation (1), which, to some extent, alleviates the batch effect. Supplementary Table S6 shows the performance comparison between our model and baselines when predicting 191 drug-cell line reactions in the PDX dataset and 430 drug-cell line reactions in the TCGA dataset. Our model had 1.15% and 2.93% higher AUC and AUPRC values than the best baseline, DeepCDR, on the PDX dataset. When predicting the drug response of the TCGA samples, our model still controlled the best performance in AUC values, except that its AUPRC value achieved 0.6356, a little lower than the two best baselines, DeepCDR and GraphCDR. These results confirm the transferability of our model from cell line *in vitro* to the datasets *in vivo*.

4.3.4 Ablation study (Test 4)

The NIHGCN model mainly consists of three components (i) type-based feature transformation, (ii) node and element level interaction

and (iii) heterogeneous aggregation. Firstly, we separately perform a linear transformation on the cell line gene expression and drug molecular fingerprints to obtain their feature vectors with the same dimensionality. After that, NIHGCN inputs the cell line and drug feature vectors into a heterogeneous network and employs an interaction module including a parallel GCN layer and a NI layer to learn node-level and element-level features of cell line and drug separately. Considering the difference of cell lines and drugs, we aggregate cell line features and drug features separately. To investigate which component contributes NIHGCN model's excellence, we set up the four model variations. (i) Without type-based feature transformation: it passes the cell line gene expression and drug molecular fingerprints through a linear transformation layer with the same parameters. (ii)Without NI layer: it sets $\alpha = 0$ in Equations (12) and (13) to remove the contribution of the NI layer. (iii) Without GCN layer: it sets $\alpha = 0$ in Equations (12) and (13) to remove the contribution of the graph convolution layer. (iv) Without heterogeneous aggregation: it replaces the parameters W_c and W_d in Equations (12) and (13) with a common parameter matrix. Moreover, we test two additional model variations to test the effectiveness of our processing strategies. (v)With rows normalization: it revises Equation (1) and scales each gene expression value by its corresponding mean and SV in each cell line. (vi) With similarity features: it considers cell line similarity and drug similarity as input attributes of our model.

Supplementary Table S7 compares the prediction performance between NIHGCN and its variants on the GDSC and CCLE datasets. We observed that removing either the graph convolution layer or the NI layer alone degraded NIHGCN performance, indicating the effectiveness of the interaction module. Moreover, we noticed that the NIHGCN without type-based feature transformation and the NIHGCN without heterogeneous aggregation have lower prediction performance than the original NIHGCN model in AUC and AUPRC values on the two datasets. The cell lines and drugs possess different molecular features generated by disparate biological techniques. Moreover, the heterogeneous network consists of different numbers of drugs and cell lines, usually having more cell line nodes than drug nodes. Hence, separately learning drug and cell line features will be beneficial to capturing the inherent mechanism of drug response in cell lines. Finally, the NIHGCN model combines multiple strategies and leads to higher performance than its variants, suggesting it successfully improves anticancer drug response prediction using NI-based heterogeneous graph convolution. Additionally, we noticed that normalizing cell line gene expressions by matrix rows resulted in 0.04% and 0.45% lower AUC values for GDSC and CCLE datasets, respectively, compared with the normalizing cell line gene expressions by matrix columns. It suggests that our data scale strategy can preserve gene original relative magnitude between different cell lines and alleviate batch effect to some extent. Using cell line similarity and drug similarity as input attributes of our model reduces its prediction performance. The similarity features may cause information loss due to the limited number of cell lines and drugs in the experimental datasets.

4.3.5 Case study (Test 5)

Previous research found that 20% of the responses were missing in the observed data (Liu *et al.*, 2022). To further assess whether our model can detect novel drug responses in cell lines, we trained the model with all known cell line drug responses in the GDSC dataset and then predicted unknown responses in the GDSC data. We focused on two clinically approved drugs, Dasatinib and GSK690693. Supplementary Table S8 shows the top 10 sensitive cell lines of Dasatinib and GSK690693 predicted by our model. We indicated two cell lines, A549 and Hey, are exposed to Dasatinib, which was reported in previous literature. Zhang *et al.* (2020) were the first to report that Dasatinib could induce pyroptosis in A549 tumor cells. Le *et al.* (2010) reported that Dasatinib caused a small amount of apoptosis and a large amount of autophagy to inhibit cell growth in HEY ovarian cancer cells. For the drug GSK690693, three cell lines, RCH-ACV, MOLT-16 and JEKO-1, were sensitive. Levy *et al.* (2009) investigated the response of pre-B cell RCH-ACV and T-cell line MOLT-16 to the pan-akt kinase inhibitor GSK690693. According to their study, GSK690693 can effectively inhibit the proliferation of RCH-ACV and MOLT-16. Liu *et al.* (2020) found that GSK690693 could effectively inhibit the proliferation of MCL cell line JeKo-1.

5 Conclusions

This work proposed a NI-based heterogeneous graph convolution network method for anticancer drug response prediction, namely NIHGCN. NIHGCN constructed a heterogeneous bipartite network consisting of drugs, cell lines and the known drug response information. Drug fingerprint and gene expression went separately transformed to feature vectors with the same dimension and were input into our model as initial drug and cell line attributes. Considering the heterogeneous nature of cell lines and drugs, NIHGCN ran two parallel convolutional operations combined with a network interaction layer on the bipartite networks. Thus, the cell lines and drugs can separately aggregate node-wise and element-wise features from their neighbors. We predicted the anticancer drug response by calculating the linear correlation coefficients of feature representations of cell lines and drugs. We conducted extensive experiments to assess the effectiveness of our model. Our model considers the difference between drugs and cell lines and separately aggregates node-wise and element-wise features from neighbors, which helps improve the anticancer drug response prediction to a higher level on two cell line datasets. Our model that was well trained on the GDSC dataset can be successfully applied to predict samples of PDX and TCGA, which verified the transferability of our model from cell line *in vitro* to the datasets *in vivo*. The case study proves that model can detect novel drug responses to cell lines. However, there is still room to improve our model for predicting the anticancer drug response on the *in vivo* dataset. It is important to mitigate batch effects between different datasets. Our future work will focus on designing advanced approaches to handle batch effects in transferring learning and integrating multi-biological data to improve the drug response prediction and identify the related drug targets.

Funding

This work was supported in part by the National Natural Science Foundation of China [61972185], the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization [U1909208], the Natural Science Foundation of Yunnan Province of China [2019FA024] and Yunnan Ten Thousand Talents Plan young.

Conflict of Interest: The authors declare that they do not have any conflict of interest.

Data availability

Some source data are available at Supplementary files online and the source code can be obtained from https://github.com/weiba/NIHGCN.

References

Adam,G. *et al.* (2020) Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.*, **4**, 1–10.

Barretina,J. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bolton,E.E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler,R.A and Spellmeyer,D.C. (eds) *Annual Reports in Computational Chemistry*. Elsevier, Amsterdam, pp. 217–241.

Ding,Z. *et al.* (2016) Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, **32**, 2891–2895.

Gao,H. *et al.* (2015) High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.*, **21**, 1318–1325.

Geeleher,P. *et al.* (2014) Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol.*, **15**, 1–12.

Huang,C. *et al.* (2017) Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. *PLoS One*, **12**, e0186906.
Le,X.F. *et al.* (2010) Dasatinib induces autophagic cell death in human ovarian cancer. *Cancer*, **116**, 4980–4990.
Levy,D.S. *et al.* (2009) AKT inhibitor, GSK690693, induces growth inhibition and apoptosis in acute lymphoblastic leukemia cell lines. *Blood*, **113**, 1723–1729.
Li,M. *et al.* (2019) DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **18**, 575–582.
Lian,J. *et al.* (2018) xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, London United Kingdom*, pp. 1754–1763.
Lind,A.P. and Anderson,P.C. (2019) Predicting drug activity against cancer cells by random Forest models based on minimal genomic information and chemical properties. *PLoS One*, **14**, e0219774.
Liu,P. *et al.* (2019) Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics*, **20**, 1–14.
Liu,Q. *et al.* (2020) DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, **36**, i911–i918.
Liu,X. *et al.* (2022) GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Brief. Bioinform.*, **23**, bbab457.
Liu,Y. *et al.* (2020) Extensive investigation of benzylic N-containing substituents on the pyrrolopyrimidine skeleton as Akt inhibitors with potent anticancer activity. *Bioorg. Chem.*, **97**, 103671.
Lloyd,J.P. *et al.* (2021) Impact of between-tissue differences on pan-cancer predictions of drug sensitivity. *PLoS Comput. Biol.*, **17**, e1008720.
Ma,J. *et al.* (2021) Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer.*, **2**, 233–244.
Menden,M.P. *et al.*; AstraZeneca-Sanger Drug Combination DREAM Consortium. (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, **10**, 1–17.
Peng,W. *et al.* (2022) Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J. Biomed. Health Inform.*, **26**, 1384–1393.
Rampášek,L. *et al.* (2019) Dr. VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, **35**, 3743–3751.
Ross,D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
Rubin,M.A. (2015) Health: make precision medicine work for cancer care. *Nature*, **520**, 290–291.
Seashore-Ludlow,B. *et al.* (2015) Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.*, **5**, 1210–1223.
Sharifi-Noghabi,H. *et al.* (2019) MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, **35**, i501–i509.
Staunton,J.E. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. USA*, **98**, 10787–10792.
Su,R. *et al.* (2019) Deep-Resp-Forest: a deep Forest model to predict anti-cancer drug response. *Methods*, **166**, 91–102.
Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*, **58**, 267–288.
Wang,L. *et al.* (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer*, **17**, 1–12.
Xia,F. *et al.* (2022) A cross-study analysis of drug response prediction in cancer cell lines. *Brief. Bioinform.*, **23**, bbab356.
Yang,W. *et al.* (2012) Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.
Zhang,F. *et al.* (2018) A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.*, **8**, 1–9.
Zhang,J. *et al.* (2020) Distinct characteristics of dasatinib-induced pyroptosis in gasdermin E-expressing human lung cancer A549 cells and neuroblastoma SH-SY5Y cells. *Oncol. Lett.*, **20**, 145–154. [Mismatch
Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.