Predicting Drug Response Based on Multi-Omics Fusion and Graph Convolution

Wei Peng, Tielin Chen, and Wei Dai

***Abstract*—Different cancer patients may respond differently to cancer treatment due to the heterogeneity of cancer. It is an urgent task to develop an efficient computational method to identify drug responses in different cell lines, which guides us to design personalized therapy for an individual patient. Hence, we propose an end-to-end algorithm, namely MOFGCN, to predict drug response in cell lines based on Multi-Omics Fusion and Graph Convolution Network. MOFGCN first fuses multiple omics data to calculate the cell line similarity and then constructs a heterogeneous network by combining the cell line similarity, drug similarity, and the known cell line-drug associations. Secondly, it learns the latent features for cancer cell lines and drugs by performing graph convolution operations on the heterogeneous network. Finally, MOFGCN applies the linear correlation coefficient to reconstruct the cancer cell line-drug correlation matrix to predict drug sensitivity. To our knowledge, this is the first attempt to combine graph convolutional neural network and linear correlation coefficient for this significant task. We performed extensive evaluation experiments on the Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) databases to validate MOFGCN's performance. The experimental results show that MOFGCN is superior to the state-of-the-art algorithms in predicting missing drug responses. It also leads to higher performance in predicting drug responses for new cell lines, new drugs, and targeted drugs.**

***Index Terms*—Drug response prediction, multi-omics fusion, graph convolutional neural network.**

Manuscript received February 9, 2021; revised May 22, 2021 and June 30, 2021; accepted July 28, 2021. Date of publication August 4, 2021; date of current version March 7, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61972185, in part by the Natural Science Foundation of Yunnan Province of China under Grant 2019FA024, in part by Yunnan Key Research and Development Program under Grant 2018IA054, and in part by Yunnan Ten Thousand Talents Plan Young. "Availability: https://github.com/weiba/MOFGCN". *(Corresponding author: Wei Peng.)*

Wei Peng is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650093, P. R. China, and also with the Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, Yunnan 650093, P. R. China (e-mail: weipeng1980@gmail.com).

Tielin Chen is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650093, P. R. China (e-mail: chentielin95@gmail.com).

Wei Dai is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China, and also with the Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, Yunnan 650500, China (e-mail: dw@cnlab.net).

Digital Object Identifier 10.1109/JBHI.2021.3102186

I. Introduction

CANCER is a complex and heterogeneous disease. Different cancer patients may respond differently to cancer treatment due to their genomic profiles [1]–[3]. Hence, identifying the drug response to different cell lines will guide us to design personalized treatments for individual patients, and it is a goal of cancer research in precision medicine. Recent advances in pharmacogenomics have given rise to many genomic data, including somatic mutation, copy number aberrations, methylate data, and hundreds of drug sensitivity and resistance in different cancer cell lines [4]–[6]. These data provide a wealth of information to build computational models that can utilize multi-omics data to predict the drug's effect on different cancer cell lines [7].

The computational algorithms for predicting drug response in cancer cell lines can be roughly divided into two categories from the view of using omics data. The first category algorithm only considers cell line gene expression when building predictive models [8], [9]. Geeleher *et al.* use the ridge regression model to correlate genome-wide expression with drug sensitivity to predict four clinical trial drugs, including Docetaxel and Cisplatin for breast cancer, Bortezomib for myeloma, and Erlotinib for non-small cell lung cancer [8]. The second type of model considers multi-omics data, including gene expression, somatic mutations, copy number variation, methylate data of the cell line, and the drug chemical structure features [3], [10]. Some of these models extract features for cells and drugs from these multi-omics profiles, then inputs the features into a classifier to predict drug sensitivity [11]–[17]. For example, Su *et al.* [12] transform cell line gene expression and copy number variation into high-dimensional feature vectors, then use a cascading forest model to splice the high-dimensional features to predict drug response in corresponding cell lines. Hossein *et al.* [13] propose a multi-omics late integration algorithm based on an auto-encoder, which uses a neural network to generate features for cell lines by encoding their gene expression, somatic mutations, and copy number variation data. Then they concatenate these features and utilize a neural network classifier to predict drug response to the cell lines. Besides the features of cell lines, Liu *et al.* [17] adopt a multi-layer graph convolutional neural network to encode medicinal chemistry features for drugs. Then, they concatenate the cell line features and drug features into a multi-layer convolutional neural network to predict drug effect on the cell lines. Based on the fact that similar cell lines and similar drugs exhibit similar drug responses [18], [19],

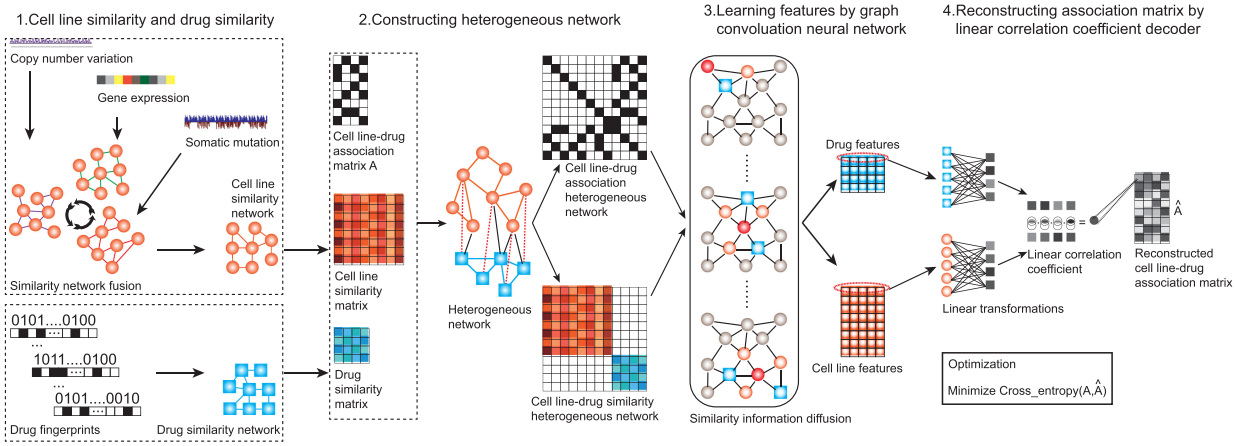

Fig. 1. The workflow of MOFGCN. The orange circle and blue square represent the cell line and drug, respectively. The solid orange line (between orange circles), solid blue lines (between blue squares), and solid black line (between orange circles and blue squares) represent the similarity between cell lines, the similarity between the drugs, and the true cell line-drug associations. The dotted line represents the potential cell line-drug associations.

Zhang *et al.* [18] construct a heterogeneous network model by calculating the Pearson correlation coefficient between cell line genomic profiles, drug chemical structures, and target gene. Subsequently, they perform an information flow-based algorithm on this network and obtain the prediction of drug response. Wang *et al.* [19] adopt a similarity-regularized matrix factorization (SRMF) method to decompose the known drug-cell line associations into the drug features and the cell line features. Meanwhile, they incorporate the similarity of drugs and gene expression profile similarity of cell lines into the drug-cell line matrix factorization model.

Although the previous great works have achieved promising results, most of them fail to consider both the drug/cell line attributes and the drug-cell line associations when learning features for drugs and cell lines. This paper develops an end-to-end algorithm, namely MOFGCN, to predict drug response based on Multi-Omics Data Fusion and Graph Convolution Network. We first construct a heterogeneous network, where nodes are drugs or cell lines and the edges are the known drug-cell line associations or self-loops of drugs or cell lines. Then we calculate the drug fingerprint similarity and cell line similarity network as the attributes of drugs and cell lines. The cell line similarity is measured by fusing cell line multi-omics data, including gene expression, copy number variation, and somatic mutation data [20]. After that, the graph convolution operation [21], [22] is employed to perform convolution operation on the heterogeneous network to extract features for drugs and cell lines from the drug/cell line attributes and the drug-cell line associations via message passing between the nodes of graphs. Finally, we reconstruct the cell line-drug association matrix by a linear correlation coefficient decoder to predict drug sensitivity or resistance in the cell lines. Fig. 1 shows the workflow of MOFGCN. We test our model on two databases, Genomics of Drug Sensitivity in Cancer (GDSC) [5] and Cancer Cell Line Encyclopedia(CCLE) [23]. Compared with the other five existing algorithms, the experimental results show that the MOFGCN algorithm achieves outstanding performance on several evaluation indexes such as AUC, ACC, F1 score, and MCC [24].

Overall, our main contributions are summarized as follows:

- We proposed an end-to-end deep learning algorithm based on multi-omics data fusion and graph convolution network, namely MOFGCN, to predict the drug response in cell lines.
- We constructed a heterogeneous network that consists of a cell line similarity network, a drug similarity network, and known drug-cell line associations. The cell line similarity network is built through fusing gene expression, copy number variation, and somatic mutation. Our work amply gathers the vital information of cell lines and drugs for drug response prediction.
- We introduced a linear correlation coefficient decoder to reconstruct the cell line-drug association matrix, which is highly interpretable. To our knowledge, MOFGCN is the first algorithm that combines the graph convolutional neural network and linear correlation coefficient decoder for drug response prediction.
- We performed extensive comparative experiments with the five state-of-the-art algorithms. The experimental results illustrate that our algorithm, MOFGCN, significantly outperforms the five state-of-the-art algorithms on the GDSC and CCLE datasets.

II. Materials and Methods

A. Materials

This study mainly involves two databases, GDSC and CCLE. In the GDSC database (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html), we downloaded two tables, TableS4A and TableS5C. TableS4A 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 test drugs. TableS5C includes the sensitivity thresholds of the 265 drugs. A drug will be considered

sensitive in a cell line if its value in TableS4A is not more than its threshold in TableS5C. Otherwise, it is resistant in the cell line. In the CCLE database (https://portals.broadinstitute.org/ccle/data), we obtain 11670 cell line-drug trial records. Each record contains trial information such as drug target, dose, logarithmic IC50, and active area. In subsequent experiments, we used logarithmic IC50 as a measure of sensitivity. Our method involves somatic mutations, gene expressions, copy number mutations, and drug substructure fingerprints. The drug substructure fingerprints were from the PubChem [6] database. The gene expression, copy number variation, and somatic mutation data came from the GDSC and CCLE databases, which were preprocessed similarly [13]. We only consider the cell lines for which all three omics data types are available.

After preprocessing, we obtained 962 cell lines and 228 drugs in the GDSC database. We categorized the cell line-drug response matrix as a binary matrix according to the threshold, with 1 representing sensitivity, 0 representing resistance, and missing values filled with 0. In our model and all comparing methods, miss values will be masked and will not participate in the model testing. For our model, SRMF, DeepForest, MOLI, and DeepCDR, the missing values will not join the calculation of the model loss. Similarly, we got 436 cancer cell lines and 24 drugs in the CCLE database. Finally, we obtained 1696 sensitive samples and 8768 resistant samples in the CCLE dataset; 20851 sensitive samples and 156512 resistant samples in the GDSC dataset.

B. Predicting Drug Response in Cell Lines

MOFGCN predicts drug sensitivity in cells based on multi-omics similarity network fusion and graph convolution network algorithms. MOFGCN first calculates the cell line similarity based on gene expression profiles, copy number variation, somatic mutation information, respectively. Then, it fuses the three similarities as the cell line similarity. The drug similarity is calculated based on the drug substructure fingerprints. After that, MOFGCN combines the cell line similarity, drug similarity, and the known drug cell associations to construct a heterogeneous network, where its nodes are drugs and cells, and its edges are the known drug-cell associations. We assumed that the cell line and the drug were self-connected in the heterogeneous network. The drug similarity and cell similarity are regarded as the drug attributes and cell line attributes. MOFGCN utilizes the graph convolutional neural network algorithm to diffuse the associations and similarity information among the heterogeneous network and learn features for drugs and cells in an end-to-end pattern. Finally, we reconstruct the cell line-drug association matrix by a linear correlation coefficient decoder to predict the drug's response in the cell line.

1) Cell Similarity and Drug Similarity: Based on the view that similar cell lines have similar drug sensitivity [3], [15], [25], we consider the similarity between cells or drugs as the cell line attributes or the drug attributes to infer the drug sensitivity. The cell line similarity can be measured in different ways, such as co-expression, co-variation, and co-mutations. Hence, this paper will use gene expression, copy number variation, and somatic mutations to calculate the cell similarity. Firstly, we use (1) to calculate the Gaussian regularization of gene expression:

$$\bar{expr}_i = \frac{(expr_i - \mu_i)}{\sigma_i} \tag{1}$$

where $expr_i$ is a column vector representing the expression values of the ith gene in all cell lines, μ_i and σ_i represent the mean and standard deviation of the expression abundance of the ith gene across all cell lines. After that, we use (2) to calculate cell-line gene expression similarity, copy number variation similarity, and somatic mutation similarity separately.

$$E_{ij} = e^{-\frac{\|x_i - x_j\|_2}{2\varepsilon^2}} \tag{2}$$

where x_i represents the gene expression values or copy number variation or somatic mutation of the ith cell line, and ε is the scale hyperparameters. Similar to [20], to fuse the three omics data of the cell line, we use (3) and (4) to calculate the full-kernel matrix and sparse-kernel matrix, respectively. The full-kernel matrix measures the distribution of omics similarity over the whole network. The sparse-kernel matrix measures the similarity information transferring within a certain distance, and the parameter N controls the length of information spread.

$$F_{ij} = \begin{cases} \frac{E_{ij}}{2\sum_j E_{ij}} & i \neq j \\ \frac{1}{2} & i = j \end{cases} \tag{3}$$

$$S_{ij} = \begin{cases} \frac{E_{ij}}{\sum_j E_{ij}} & j \in N_i \\ 0 & j \notin N_i \end{cases} \tag{4}$$

where E_{ij} is the similarity between cell lines i and j based on specific omics data. N_i is a proximity parameter, representing the top N_i of cell lines most similar to the ith cell line.

Let GF, CF, and MF represent the full-kernel matrix of gene expression, the full-kernel matrix of copy number variation, and the full-kernel matrix of somatic mutations, respectively. Let GS, CS, and MS denotes gene expression sparse-kernel matrix, copy number variation sparse-kernel matrix, and somatic mutation sparse-kernel matrix, respectively. Given $GF_0 = GF$, $CF_0 = CF$, $MF_0 = MF$, the full-kernel matrixes are updated iteratively as follows:

$$\begin{aligned} GF_t &= GS \cdot \frac{(CF_{t-1} + MF_{t-1})}{2} \cdot GS^T \\ CF_t &= CS \cdot \frac{(GF_{t-1} + MF_{t-1})}{2} \cdot CS^T \\ MF_t &= MS \cdot \frac{(CF_{t-1} + GF_{t-1})}{2} \cdot MS^T \end{aligned} \tag{5}$$

where $*^T$ represents transpose, $\cdot$ represents matrix multiplication, and t is the number of iterations. Finally, the similarity fusion matrix (RC) of cell lines can be obtained by (6):

$$RC = \frac{GF_t + CF_t + MF_t}{3} \tag{6}$$

The drug similarity is measured by the drug Jaccard similarity of their substructure fingerprints (see (7)).

$$RD_{ij} = \frac{|\, x_i \cap x_j \,|}{|\, x_i \cup x_j \,|} \tag{7}$$

where x_i and x_j represent the substructure fingerprints of the ith and jth drugs, respectively. After that, we introduce matrix R with $m+n$ rows and $m+n$ columns to store the cell line-similarity and drug similarity [26]:

$$R = \begin{vmatrix} RC & 0 \\ 0 & RD \end{vmatrix} \tag{8}$$

where RC represents the cell line similarity matrix, RD represents the drug similarity matrix.

2) Graph Convolutional Neural Network: Graph convolutional neural network is a network embedding technique that learns representation vectors for the nodes in the network while simultaneously considering both network structure and node attributes. Hence, we employ a graph convolutional neural network model to learn latent feature vectors for drugs and cell lines. Firstly, a heterogeneous network is constructed, where drugs and cells are nodes and the known drug-cell associations are edges. We assumed that the cell line and the drug were self-connected in the heterogeneous network. Let $\bar{A}$ be the adjacent matrix of the heterogeneous. It is defined as follows [26].

$$\bar{A} = \begin{vmatrix} I_m & A \\ A^T & I_n \end{vmatrix} \tag{9}$$

where m and n is the number of cell lines and drugs, respectively. I_m is the m-dimensional identity matrix. A is the cell line-drug association matrix, whose value is one means the drug affects the cell line, zero otherwise. For the data from the GDSC database, we binarize the cell line-drug association matrix according to logarithmic IC50 values (see (10)):

$$A_{ij} = \begin{cases} 1 & \text{response}_{ij} \le \text{threshold}_j \\ 0 & \text{other} \end{cases} \tag{10}$$

where response_{ij} represents the logarithmic IC50 value of the jth drug in the ith cell, and threshold_j represents the sensitivity threshold of the jth drug. According to [27], we use (11) to binarize the cell line-drug association matrix for data from the CCLE database:

$$A_{ij} = \begin{cases} 1 & normalize(\text{response}_j)_{ij} \le -0.8 \\ 0 & \text{other} \end{cases} \tag{11}$$

where $normalize(\cdot)$ represents the Gaussian regularization function.

Set $D_{(1)ii} = \sum_j A_{ij} + 1, D_{(2)ii} = \sum_j A_{ji} + 1$, and normalized adjacency matrix $L = I_{m+n} + D^{-\frac{1}{2}} \bar{A} D^{-\frac{1}{2}}$, where $D = \begin{vmatrix} D_{(1)} & 0 \\ 0 & D_{(2)} \end{vmatrix}$, $D_{ii} = \sum_j \bar{A}_{ij}$. Accordingly, the matrix L can be rewritten as:

$$L = \begin{vmatrix} D_{(1)}^{-1} + I_m & D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}} \\ \left(D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}}\right)^T & D_{(2)}^{-1} + I_n \end{vmatrix} \tag{12}$$

We regard the drug similarity and cell similarity as the node feature of the graph convolutional neural network. Hence, the single-layer graph convolutional neural network can be expressed as [21]:

$$\begin{aligned} H_{(1)} &= \sigma(L \cdot R \cdot W) \\ &= \sigma\left(\begin{vmatrix} \left(D_{(1)}^{-1} + I_m\right) \cdot RC & \left(D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}}\right) \cdot RD \\ \left(D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}}\right)^T \cdot RC & \left(D_{(2)}^{-1} + I_n\right) \cdot RD \end{vmatrix} \cdot \begin{vmatrix} W_{(1)} \\ W_{(2)} \end{vmatrix} \right) \end{aligned} \tag{13}$$

where $W \in R^{(m+n)\times h}$ is a weight parameter matrix of the neural network. h means the hidden layer dimension. σ represents the activation function. The ReLU activation function is used here. $W_{(1)} \in R^{m\times h}$ and $W_{(2)} \in R^{n\times h}$ are the weight parameter matrix of cell lines and drugs, respectively. The hidden layer $H_{(1)}$ is further divided into two parts:

$$\begin{aligned} H_{(1)} &= \begin{vmatrix} H_{(1-c)} \\ H_{(1-d)} \end{vmatrix} \\ &= \begin{vmatrix} \sigma((D_{(1)}^{-1} + I_m) \cdot RC \cdot W_{(1)} + (D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}}) \cdot RD \cdot W_{(2)}) \\ \sigma((D_{(1)}^{-\frac{1}{2}} A D_{(2)}^{-\frac{1}{2}})^T \cdot RC \cdot W_{(1)} + (D_{(2)}^{-1} + I_n) \cdot RD \cdot W_{(2)}) \end{vmatrix} \end{aligned} \tag{14}$$

where $H_{(1-c)} \in R^{m\times h}$ and $H_{(1-d)} \in R^{n\times h}$ represent the cell line and drug features embedded by the graph convolution layer, respectively.

3) Linear Correlation Coefficient Decoder: After obtaining feature representations for cell lines and drugs, we introduce linear correlation coefficients as a decoder to reconstruct the cell line-drug association matrix. Before decoding, we input the cell line features and the drug features into a single-layer neural network to do linear transformations, respectively:

$$\begin{aligned} H_{(2-c)} &= H_{(1-c)} \Theta_{(2-c)} \\ H_{(2-d)} &= H_{(1-d)} \Theta_{(2-d)} \end{aligned} \tag{15}$$

where $\Theta_{(2-c)} \in R^{h\times k}$ and $\Theta_{(2-d)} \in R^{h\times k}$ are the weight parameters of the neural network for cell lines and drugs, respectively. $H_{(2-c)} \in R^{m\times k}$ and $H_{(2-d)} \in R^{n\times k}$ are the k-size final feature presentations of cell lines and drugs respectively. The linear correlation coefficient between the two features is defined as follows:

$$Corr(h_i, h_j) = \frac{(h_i - \mu_i)(h_j - \mu_j)^T}{\sqrt{(h_i - \mu_i)(h_i - \mu_i)^T}\sqrt{(h_j - \mu_j)(h_j - \mu_j)^T}} \tag{16}$$

where $h_i \in H_{(2-c)}$ and $h_j \in H_{(2-d)}$ are the feature vectors with k-dimensional for the ith cell line and the jth drug, respectively. μ_i and μ_j are the mean values of h_i and h_j, respectively. Since the value range of the correlation coefficient is $[-1, 1]$, we use functions (17) and (18) to activate the output:

$$f(h) = \frac{e^{-\alpha h}}{1 + e^{-\alpha h}} \tag{17}$$

$$g(h) = \frac{h - \min(h)}{\max(h) - \min(h)}, h \in [-1, 1] \tag{18}$$

where α is the scaling paramete. In (18), $max(h)$ and $\min(h)$ represent the maximum and minimum values of the variable h in the interval $[-1, 1]$, respectively. Compared with the direct use of linear transformation, the appropriate parameter α can make (17) have a more appropriate gradient in the interval $[-1, 1]$, which promotes model convergence and accelerates parameter update. Finally, the cell line-drug association matrix $\hat{A}$ can be reconstructed as:

$$\hat{A} = g(f(Corr(H_{(2-c)}, H_{(2-d)}))) \tag{19}$$

The loss function for model constraints is as follows:

$$\ell(A, \hat{A}) = -\frac{1}{m \times n} \sum_{i,j} M_{ij}[A_{ij} ln(\hat{A}_{ij}) + (1 - A_{ij}) ln(1 - \hat{A}_{ij})] \tag{20}$$

where m and n represent the number of cell lines and drugs, respectively. M is a indicate matrix. $M_{ij} = 1$ when the association between the ith cell line and the jth drug is in the training set, otherwise $M_{ij} = 0$. Algorithm 1 lists the pseudocode for running our model.

We use the PyTorch framework to implement the model codes and the Adam optimizer to optimize the loss function. The hyperparameters are tuned by grid searching. Here, we set the scale parameter $\varepsilon = 2$, the proximity parameter $N = 11$, the number of iterations $t = 3$, the embedding layer dimension $h = 192$, the correlation information dimension $k = 36$, the scaling parameters $\alpha = 5.74$, the learning rate $lr = 5 \times 10^{-4}$, and the number of neural network iterations $epoch = 1000$.

Algorithm 1: MOFGCN

Input: Gene expression matrix, copy number variation matrix, somatic mutation matrix, drug fingerprint matrix, cell line-drug association matrix A, parameters ε, N, t, h, k, α, lr, $epoch$.

Output: Reconstructed cell line-drug association matrix.

Using (2) to calculate cell line similarities based on gene expression data, copy number variation, and somatic mutation, respectively.
Calculating gene expression full-kernel matrix (GF), copy number variation full-kernel matrix (CF), and somatic mutation full-kernel matrix (MF) with (3).
Calculating gene expression sparse-kernel matrix (GS), copy number variation sparse-kernel matrix (CS), somatic mutation sparse-kernel matrix (MS) with (4).
$GF_0 = GF, CF_0 = CF, MF_0 = MF$.
for $i = 1$ to t **do**
  $GF_i = GS \cdot \frac{(CF_{i-1} + MF_{i-1})}{2} \cdot GS^T$
  $CF_i = CS \cdot \frac{(GC_{i-1} + MF_{i-1})}{2} \cdot CS^T$
  $MF_i = MS \cdot \frac{(CF_{I-1} + GF_{i-1})}{2} \cdot MS^T$
end for
Calculating cell line similarity (RC) with Eqation (6).
Calculating drug similarity (RD) with (7).
Constructing cell line-drug similarity matrix (R) with (8).
Calculating adjacent matrix ($\bar{A}$) of the cell line-drug heterogeneous network with (9).
Calculating matrix L with (12).
for $i = 1$ to $epoch$ **do**
  $H_{(1)} = \sigma(L \cdot R \cdot W) = \begin{vmatrix} H_{(1-c)} \\ H_{(1-d)} \end{vmatrix}$
  $H_{(2-c)} = H_{(1-c)}\Theta_{(2-c)}$
  $H_{(2-d)} = H_{(1-d)}\Theta_{(2-d)}$
  Reconstructing the cell line-drug association matrix $\hat{A}$ with (19).
  $\ell(A, \hat{A}) = -\frac{1}{m \times n} \sum_{i,j} M_{ij}[A_{ij} ln(\hat{A}_{ij}) + (1 - A_{ij}) ln(1 - \hat{A}_{ij})]$
  Updating W, $\Theta_{(2-c)}$, and $\Theta_{(2-d)}$ by gradient descent and backpropagation.
end for
return $\hat{A}$.

III. Result

A. Experimental Design

To verify the effectiveness of our model MOFGCN, we compared it with other five state-of-the-art algorithms, including HNMDRP [18], SRMF [19], MOLI [13], DeepForest [12], and DeepCDR [17]. HNMDRP and SRMF introduce the known cell line-drug associations, gene expression, and drug chemical structure to construct a heterogeneous network. They predict the drug response by inferring new cell line-drug associations under the consideration of the gene similarity, drug similarity, and the known associations between cell lines and drugs. HNMDRP predicts novel cell line-drug associations by network propagation. Meanwhile, it also considers the interactions between drug targets and proteins. SRMF indicates drug sensitivity based on the similarity of regularization matrix factorization. MOLI, DeepForest, and DeepCDR input the drug features and cell line features into a classifier and classify whether the drugs are sensitive or resistant to cells. MOLI is the latest integration method based on deep neural networks, which uses type-specific encoders to learn features for different omics data types, including somatic mutation, copy number aberration, and gene expression data. It concatenates these features for drug response prediction. DeepForest inputs the cell line gene expression and copy number variation and then transforms the basic features into high-dimensional feature vectors through multi-grained scanning. It designs a cascade forest method with a feature optimization operation to classify the anti-disease drug response as sensitive or resistant. DeepCDR integrates cell line gene expression, methylation, somatic mutation, and drug features for drug response prediction. It leverages the graph convolutional neural network to learn drug features. The hyperparameters of all comparison algorithms adopt the settings recommended in the original text.

TABLE I
COMPARISON OF EVERY METHOD UNDER RANDOMLY ZEROING CROSS-VALIDATION ON TWO DATASETS

Dataset	Algorithm	AUC	ACC	Precision	Recall	F1 Score	MCC
GDSC	HNMDRP	$0.7258 \pm 3 \times 10^{-5}$	$0.6302 \pm 1 \times 10^{-4}$	$0.5890 \pm 1 \times 10^{-4}$	$0.8662 \pm 5 \times 10^{-4}$	$0.7008 \pm 1 \times 10^{-5}$	$0.2959 \pm 2 \times 10^{-4}$
	HNMDRP+Fusion	$0.7242 \pm 3 \times 10^{-5}$	$0.6267 \pm 2 \times 10^{-4}$	$0.5859 \pm 2 \times 10^{-4}$	$0.8714 \pm 8 \times 10^{-4}$	$0.7001 \pm 2 \times 10^{-5}$	$0.2915 \pm 4 \times 10^{-4}$
	SRMF	$0.6563 \pm 2 \times 10^{-4}$	$0.5587 \pm 6 \times 10^{-4}$	$0.5358 \pm 3 \times 10^{-4}$	$0.9078 \pm 8 \times 10^{-4}$	$0.6731 \pm 3 \times 10^{-5}$	$0.1615 \pm 1 \times 10^{-3}$
	SRMF+Fusion	$0.6500 \pm 3 \times 10^{-5}$	$0.5468 \pm 1 \times 10^{-4}$	$0.5268 \pm 1 \times 10^{-4}$	**0.9210** $\pm 5 \times 10^{-4}$	$0.6702 \pm 1 \times 10^{-5}$	$0.1410 \pm 2 \times 10^{-4}$
	MOFGCN+GE	$0.8018 \pm 1 \times 10^{-4}$	$0.7047 \pm 1 \times 10^{-4}$	$0.6578 \pm 3 \times 10^{-4}$	$0.8564 \pm 2 \times 10^{-4}$	$0.7437 \pm 4 \times 10^{-5}$	$0.4301 \pm 5 \times 10^{-4}$
	MOFGCN	**0.8622** $\pm 1 \times 10^{-5}$	**0.7726** $\pm 3 \times 10^{-5}$	**0.7402** $\pm 1 \times 10^{-4}$	$0.8411 \pm 3 \times 10^{-4}$	**0.7872** $\pm 1 \times 10^{-5}$	**0.5509** $\pm 7 \times 10^{-5}$
CCLE	HNMDRP	$0.7104 \pm 1 \times 10^{-4}$	$0.6328 \pm 4 \times 10^{-4}$	$0.5856 \pm 4 \times 10^{-4}$	$0.9219 \pm 1 \times 10^{-3}$	$0.7153 \pm 4 \times 10^{-5}$	$0.3281 \pm 7 \times 10^{-4}$
	HNMDRP+Fusion	$0.7106 \pm 2 \times 10^{-4}$	$0.6313 \pm 3 \times 10^{-4}$	$0.5836 \pm 2 \times 10^{-4}$	$0.9253 \pm 1 \times 10^{-3}$	$0.7151 \pm 3 \times 10^{-5}$	$0.3267 \pm 5 \times 10^{-4}$
	SRMF	$0.7669 \pm 4 \times 10^{-5}$	$0.6792 \pm 3 \times 10^{-5}$	$0.6190 \pm 3 \times 10^{-5}$	$0.9336 \pm 9 \times 10^{-4}$	$0.7441 \pm 5 \times 10^{-5}$	$0.4182 \pm 4 \times 10^{-4}$
	SRMF+Fusion	$0.7695 \pm 5 \times 10^{-5}$	$0.6824 \pm 8 \times 10^{-5}$	$0.6214 \pm 5 \times 10^{-5}$	**0.9342** $\pm 4 \times 10^{-4}$	$0.7462 \pm 6 \times 10^{-5}$	$0.4231 \pm 4 \times 10^{-4}$
	MOFGCN+GE	$0.8039 \pm 2 \times 10^{-4}$	$0.7281 \pm 4 \times 10^{-4}$	$0.6728 \pm 7 \times 10^{-4}$	$0.8961 \pm 1 \times 10^{-3}$	$0.7673 \pm 1 \times 10^{-4}$	$0.4872 \pm 1 \times 10^{-3}$
	MOFGCN	**0.8591** $\pm 1 \times 10^{-4}$	**0.7836** $\pm 2 \times 10^{-4}$	**0.7600** $\pm 8 \times 10^{-4}$	$0.8335 \pm 1 \times 10^{-3}$	**0.7940** $\pm 1 \times 10^{-4}$	**0.5719** $\pm 7 \times 10^{-4}$

We tested our method and all the baseline methods on the two databases: GDSC and CCLE. The receiver operating characteristic (ROC) curve and its area underneath (AUC) will be adopted to evaluate every model. We also show their best F1 score and corresponding accuracy (ACC), precision, recall, and Matthews correlation coefficient(MCC) values for comparison.

In our experiments, we conduct the test under the following four settings:

- Test 1: Comparing our model with HNMDRP and SRMF for the cell line-drug association matrix reconstruction when randomly zeroing the values in the matrix.
- Test 2: Comparing our model with HNMDRP and SRMF for the cell line-drug association matrix reconstruction when randomly blinding a row or a column.
- Test 3: Comparing our method with HNMDRP, SRMF, MOLI, DeepForest, and DeepCDR when predicting a single drug's response.
- Test 4: Comparing our model with HNMDRP, SRMF, MOLI, DeepForest, and DeepCDR when predicting response for some targeted drugs.

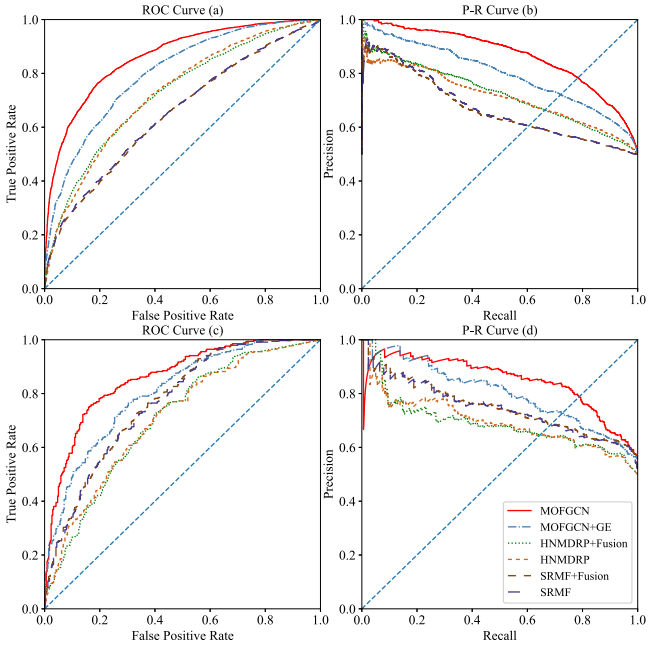

Fig. 2. ROC and P-R curves obtained under the setting Test 1 on GDSC (a)–(b) and CCLE (c)–(d).

B. Randomly Zeroing Cross-Validation

HNMDRP, SRMF, and our model reconstruct cell line-drug association matrix to predict drug response. We test their performance in two settings. In setting Test 1, we randomly partition the known cell line-drug associations (positive samples) into five equal parts and conduct five times five-fold cross-validations. In each round of validation, 1/5 positive samples are cleaned, and an equal amount of negative samples are randomly selected from the association matrix as the test data. The remaining 4/5 positive samples and the remaining negative samples are selected as the training data. MOFGCN measured the cell line similarity by fusing multi-omics data, including gene expression, copy number variation, and somatic mutation data. MOFGCN, HNMDRP, and SRMF all construct a heterogeneous network involving drug similarity, cell similarity, and known drug-cell line associations. The original HNMDRP and SRMF calculate the cell line similarity based on gene expression profiles. To test whether fusing multi-omics data improves the algorithm's performance, we compare the three methods that calculate the cell line similarity using gene expression data or fusion multi-omics data. For a fair comparison, all the three methods take the same way mentioned in the section "Cell Similarity and Drug Similarity" to calculate the cell line gene similarity and fuse multi-omics data.

Table I and Fig. 2 show the experimental results of three algorithms on the GDSC and CCLE databases, taking two strategies to construct the cell line similarity matrix. MOFGCN+GE, HNMDRP, and SRMF employ gene expression data to calculate the cell line similarity. MOFGCN, HNMDRP+Fusion, and SRMF+Fusion measure the cell line similarity by fusing multi-omics data. Our methods significantly outperform HNMDRP and SRMF on the GDSC and CCLE datasets in AUC, ACC, Precision, F1 score, and MCC when using gene expression profiles to construct cell line similarity matrix. We observe that our model fusing multi-omics data reaches 0.8622 and 0.8591 for AUC values on the GDSC and CCLE datasets, respectively. It performs best among all comparing methods on the two datasets. However, we also notice that fusing multi-omics data to construct the cell line similarity matrix cannot significantly improve the HNMDRP and SRMF on the GDSC and CCLE datasets. The observed improvement in our method compared

TABLE II
COMPARISON OF EVERY METHOD PREDICTING RESPONSE FOR A NEW DRUG OR A NEW CELL LINEON TWO DATASETS

Dataset	Algorithm		AUC	ACC	Precision	Recall	F1 Score	MCC
GDSC	Single row	SRMF	$0.5807 \pm 1 \times 10^{-2}$	$0.5859 \pm 7 \times 10^{-3}$	$0.5642 \pm 6 \times 10^{-3}$	$\mathbf{0.9458} \pm 6 \times 10^{-3}$	$0.6990 \pm 1 \times 10^{-3}$	$0.2258 \pm 3 \times 10^{-2}$
		HNMDRP	-	-	-	-	-	-
		MOFGCN	$\mathbf{0.7317} \pm 9 \times 10^{-3}$	$\mathbf{0.7002} \pm 8 \times 10^{-3}$	$\mathbf{0.6682} \pm 1 \times 10^{-2}$	$0.9026 \pm 2 \times 10^{-3}$	$\mathbf{0.7569} \pm 3 \times 10^{-3}$	$\mathbf{0.4444} \pm 2 \times 10^{-2}$
	Single column	SRMF	$0.6683 \pm 6 \times 10^{-3}$	$0.6033 \pm 6 \times 10^{-3}$	$0.5739 \pm 4 \times 10^{-3}$	$\mathbf{0.9281} \pm 5 \times 10^{-3}$	$0.7033 \pm 1 \times 10^{-3}$	$0.2651 \pm 1 \times 10^{-2}$
		HNMDRP	$0.6963 \pm 1 \times 10^{-2}$	$0.6505 \pm 1 \times 10^{-2}$	$0.6211 \pm 9 \times 10^{-3}$	$0.9143 \pm 3 \times 10^{-3}$	$0.7300 \pm 3 \times 10^{-3}$	$0.3325 \pm 4 \times 10^{-2}$
		MOFGCN	$\mathbf{0.7450} \pm 7 \times 10^{-3}$	$\mathbf{0.6879} \pm 6 \times 10^{-3}$	$\mathbf{0.6565} \pm 7 \times 10^{-3}$	$0.8759 \pm 2 \times 10^{-3}$	$\mathbf{0.7421} \pm 2 \times 10^{-3}$	$\mathbf{0.4076} \pm 2 \times 10^{-2}$
CCLE	Single row	SRMF	$0.6138 \pm 8 \times 10^{-3}$	$0.7055 \pm 3 \times 10^{-3}$	$0.6486 \pm 4 \times 10^{-3}$	$\mathbf{0.9699} \pm 1 \times 10^{-3}$	$0.7746 \pm 1 \times 10^{-3}$	$0.4741 \pm 7 \times 10^{-3}$
		HNMDRP	-	-	-	-	-	-
		MOFGCN	$\mathbf{0.8257} \pm 7 \times 10^{-3}$	$\mathbf{0.8104} \pm 5 \times 10^{-3}$	$\mathbf{0.7667} \pm 6 \times 10^{-3}$	$0.9524 \pm 1 \times 10^{-3}$	$\mathbf{0.8423} \pm 2 \times 10^{-3}$	$\mathbf{0.6536} \pm 1 \times 10^{-2}$
	Single column	SRMF	$0.4873 \pm 9 \times 10^{-3}$	$0.5204 \pm 1 \times 10^{-3}$	$0.5126 \pm 5 \times 10^{-4}$	$\mathbf{0.9782} \pm 2 \times 10^{-3}$	$0.6713 \pm 4 \times 10^{-5}$	$0.0788 \pm 6 \times 10^{-3}$
		HNMDRP	$0.6947 \pm 6 \times 10^{-3}$	$0.6498 \pm 4 \times 10^{-3}$	$0.6078 \pm 3 \times 10^{-3}$	$0.9103 \pm 3 \times 10^{-3}$	$\mathbf{0.7243} \pm 9 \times 10^{-4}$	$0.3428 \pm 1 \times 10^{-2}$
		MOFGCN	$\mathbf{0.7087} \pm 1 \times 10^{-3}$	$\mathbf{0.6507} \pm 1 \times 10^{-3}$	$\mathbf{0.6089} \pm 1 \times 10^{-3}$	$0.9038 \pm 5 \times 10^{-4}$	$0.7230 \pm 3 \times 10^{-4}$	$\mathbf{0.3536} \pm 4 \times 10^{-3}$

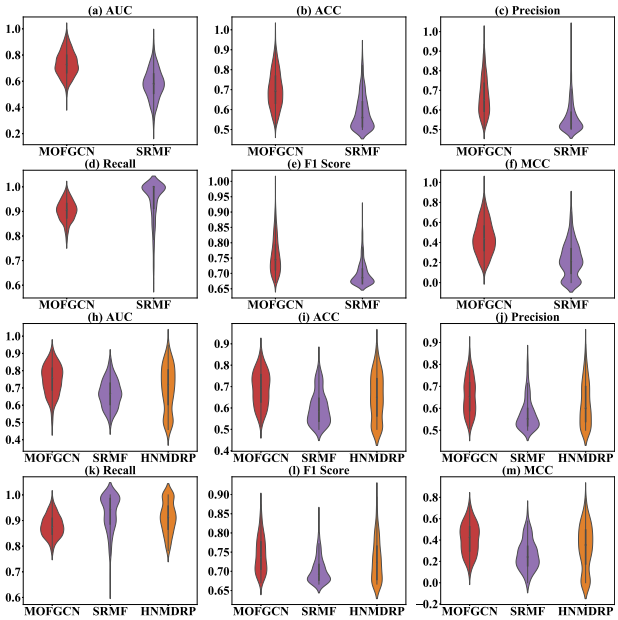

Fig. 3. Violin diagrams of single-row clearing (a)–(f) and single-column clearing (h)–(m) on GDSC dataset.

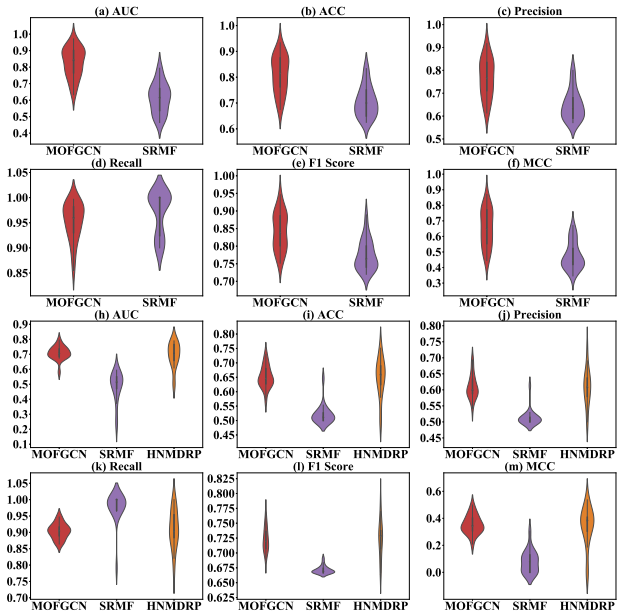

Fig. 4. Violin diagrams of single-row clearing (a)–(f) and single-column clearing (h)–(m) on CCLE dataset.

to HNMDRP and SRMF can be partially attributed to the high quality of cell line features and drug features learned by our method based on multi-omics data fusion and graph convolution.

Fig. 2 shows the ROC and P-R curves of all six comparing methods on GDSC and CCLE datasets. It can be seen intuitively from the figures that our method fusing multi-omics data to construct the cell line similarity matrix shows outstanding performance.

C. Predicting Response for a New Drug or a New Cell Line

Each row in the cell line-drug association matrix represents the cell line, and each column represents the drug. In Test 2, we clear a row or a column in the cell line-drug association matrix as the testing set and the remaining rows or columns as the training set. This setting aims to test every method's ability to predict response for a new drug or a new cell line. To avoid too general or too special, we only choose the row or column that contains at least ten positive samples for testing [27]. After screening, 227 of 228 drugs or 658 of 962 cell lines participate in the GDSC dataset experiments. 20 of 24 drugs and 26 of 436 cell lines participate in the CCLE dataset experiments. When we choose a row or a column as the testing set and clear it in the cell line-drug association matrix, the remaining rows or columns will take as the training set. In this process, our method will update the cell line-drug association matrix and will not update the similarity matrices in this testing. We plot violin diagrams (Fig. 3 and 4) to illustrate the AUC, ACC, precision, recall, and F1 score values of HNMDRP, SRMF, and our model on testing rows or columns of the GDSC and CCLE datasets. Table II reports the means and variances of AUC, ACC, precision, recall, F1 score, and MCC across all testing rows or columns on the GDSC and CCLE datasets. From the two figures and the table, we observe that our method leads to higher performance than HNMDRP and SRMF in all evaluation measures except in recall. We calculated the ACC, precision, recall, and MCC when the methods reach their best F1 scores. Different approaches obtain their best F1 scores at different precision and recall values. Our method gets its best F1 scores when predicting a small number of drugs or cells, showing relatively low recall values. However, our method has the highest

TABLE III
COMPARISON OF EVERY METHOD FOR PREDICTING A SINGLE DRUG'S RESPONSE ON TWO DATASETS

Dataset	Algorithm	AUC	ACC	Precision	Recall	F1 Score	MCC
GDSC	DeepForest	$0.6214 \pm 1 \times 10^{-2}$	$0.6403 \pm 5 \times 10^{-3}$	$0.6198 \pm 5 \times 10^{-3}$	$0.9099 \pm 2 \times 10^{-3}$	$0.7025 \pm 1 \times 10^{-3}$	$0.3280 \pm 1 \times 10^{-2}$
	MOLI	$0.6841 \pm 8 \times 10^{-3}$	$0.6817 \pm 4 \times 10^{-3}$	$0.6577 \pm 5 \times 10^{-3}$	$0.8920 \pm 1 \times 10^{-3}$	$0.7427 \pm 1 \times 10^{-3}$	$0.4072 \pm 1 \times 10^{-2}$
	DeepCDR	$0.7234 \pm 3 \times 10^{-3}$	$0.7119 \pm 2 \times 10^{-3}$	$0.6823 \pm 3 \times 10^{-3}$	$0.8842 \pm 8 \times 10^{-4}$	$0.7584 \pm 1 \times 10^{-3}$	$0.4646 \pm 9 \times 10^{-4}$
	SRMF	$0.7271 \pm 2 \times 10^{-2}$	$0.7229 \pm 1 \times 10^{-2}$	$0.6938 \pm 1 \times 10^{-2}$	$\mathbf{0.9121} \pm 3 \times 10^{-3}$	$0.7749 \pm 4 \times 10^{-3}$	$0.4877 \pm 4 \times 10^{-2}$
	HNMDRP	$0.7411 \pm 9 \times 10^{-3}$	$0.7268 \pm 6 \times 10^{-3}$	$0.6960 \pm 7 \times 10^{-3}$	$0.9035 \pm 1 \times 10^{-3}$	$0.7740 \pm 1 \times 10^{-3}$	$0.4954 \pm 2 \times 10^{-2}$
	MOFGCN	$\mathbf{0.8409} \pm 5 \times 10^{-3}$	$\mathbf{0.8143} \pm 4 \times 10^{-3}$	$\mathbf{0.7892} \pm 5 \times 10^{-3}$	$0.9082 \pm 1 \times 10^{-3}$	$\mathbf{0.8356} \pm 2 \times 10^{-3}$	$\mathbf{0.6522} \pm 1 \times 10^{-2}$
CCLE	DeepForest	$0.5924 \pm 4 \times 10^{-3}$	$0.6183 \pm 3 \times 10^{-3}$	$0.6002 \pm 4 \times 10^{-3}$	$\mathbf{0.9224} \pm 1 \times 10^{-3}$	$0.7122 \pm 9 \times 10^{-4}$	$0.2911 \pm 1 \times 10^{-2}$
	MOLI	$0.6899 \pm 5 \times 10^{-3}$	$0.6830 \pm 4 \times 10^{-3}$	$0.6597 \pm 5 \times 10^{-3}$	$0.8889 \pm 1 \times 10^{-3}$	$0.7430 \pm 1 \times 10^{-3}$	$0.4113 \pm 1 \times 10^{-2}$
	DeepCDR	$0.7010 \pm 2 \times 10^{-3}$	$0.6915 \pm 3 \times 10^{-3}$	$0.6636 \pm 4 \times 10^{-3}$	$0.8840 \pm 3 \times 10^{-4}$	$0.7453 \pm 1 \times 10^{-3}$	$0.4288 \pm 1 \times 10^{-2}$
	SRMF	$0.7403 \pm 1 \times 10^{-2}$	$0.7331 \pm 5 \times 10^{-3}$	$0.6968 \pm 6 \times 10^{-3}$	$0.9117 \pm 1 \times 10^{-3}$	$0.7786 \pm 2 \times 10^{-3}$	$0.5112 \pm 1 \times 10^{-2}$
	HNMDRP	$0.7202 \pm 2 \times 10^{-3}$	$0.7105 \pm 2 \times 10^{-3}$	$0.6753 \pm 3 \times 10^{-3}$	$0.9036 \pm 1 \times 10^{-3}$	$0.7612 \pm 1 \times 10^{-3}$	$0.4702 \pm 9 \times 10^{-2}$
	MOFGCN	$\mathbf{0.8464} \pm 3 \times 10^{-3}$	$\mathbf{0.8180} \pm 3 \times 10^{-3}$	$\mathbf{0.7965} \pm 3 \times 10^{-3}$	$0.8966 \pm 2 \times 10^{-3}$	$\mathbf{0.8345} \pm 2 \times 10^{-3}$	$\mathbf{0.6581} \pm 1 \times 10^{-2}$

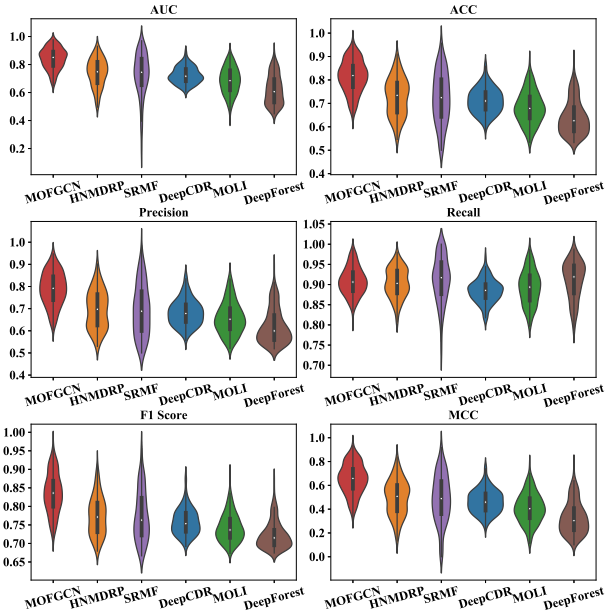

Fig. 5. Comparison of predicting a single drug's response on GDSC database.

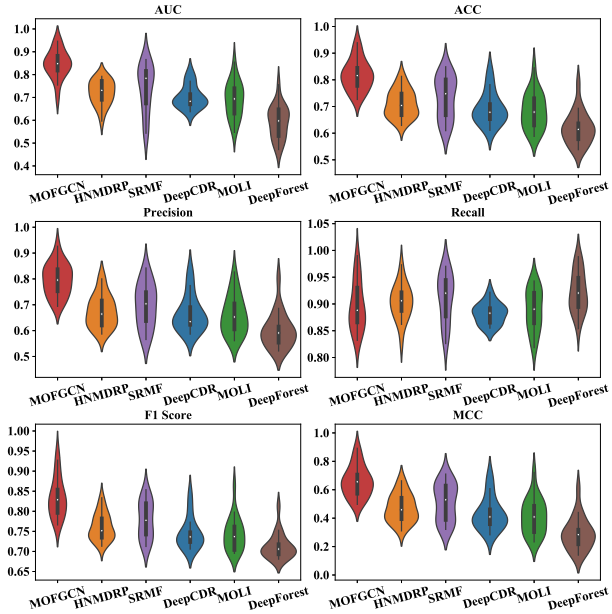

Fig. 6. Comparison of predicting a single drug's response on CCLE database.

F1 scores among all comparing methods, suggesting that our method can predict the drug response more correctly. We also notice that HNMDRP cannot produce prediction results when clearing a single row. Because HNMDRP only infers potential drug-cell line associations from the drugs with similar chemical structures and similar drug targets but fails to consider the cell lines with similar gene expression profiles. Hence, HNMDRP cannot predict drug response for new cell lines.

D. Predicting a Single Drug Response

MOLI, DeepCDR, and DeepForest are designed to extract features of single or several drugs to predict their response to different cell lines. We conduct Test 3 to comparing our method with HNMDRP, SRMF, MOLI, DeepForest, and DeepCDR when predicting a single drug response. We repeat five-fold cross-validations five times on the drugs that are sensitive to at least ten cell lines [27]. For a drug, we randomly choose 1/5 positive samples and the same amount of negative samples as the test data, and the other 4/5 positive samples and the remaining negative samples as the training data. To compare

TABLE IV
COMPARISON OF RUNNING TIME OF EVERY METHOD (SECONDS)

Algorithm	MOFGCN	HNMDRP	SRMF	MOLI	DeepCDR	DeepForest
GDSC	6.04	0.18	82.09	74.17	703.9	1132.9
CCLE	3.88	0.06	22.97	18.72	121.9	493.60

the performance of various algorithms clearly, we calculate the mean and variance of AUC, ACC, precision, recall, F1 score, and MCC across all testing drugs on the GDSC and CCLE datasets. Table III shows the experimental results. We notice that our method that integrates multi-omics data and adopts a graph convolution algorithm to predict the cell line-drug associations in the heterogeneous network achieves outstanding performance in all evaluation measures except in recall. We also observe that HNMDRP, SRMF, and our method considering the inter-association and intra-associations between cell lines and drugs lead to higher prediction performance than MOLI, DeepCDR, and DeepForest that focus on features of one drug. It suggests that combining other drug information can significantly improve the performance of the algorithm. Fig. 5 and Fig. 6 illustrate

TABLE V
COMPARISON OF EVERY METHOD FOR PREDICTING TARGETED DRUGS' RESPONSE ON TWO DATASETS

Dataset	Algorithm	AUC	ACC	Precision	Recall	F1 Score	MCC
GDSC	DeepForest	$0.6985 \pm 1 \times 10^{-3}$	$0.6433 \pm 4 \times 10^{-3}$	$0.6209 \pm 5 \times 10^{-3}$	$0.8274 \pm 9 \times 10^{-3}$	$0.7003 \pm 3 \times 10^{-4}$	$0.3175 \pm 1 \times 10^{-2}$
	MOLI	$0.7475 \pm 1 \times 10^{-3}$	$0.6821 \pm 2 \times 10^{-3}$	$0.6471 \pm 3 \times 10^{-3}$	$0.8431 \pm 6 \times 10^{-3}$	$0.7269 \pm 8 \times 10^{-4}$	$0.3919 \pm 7 \times 10^{-3}$
	DeepCDR	$0.7954 \pm 2 \times 10^{-3}$	$0.7156 \pm 4 \times 10^{-3}$	$0.6923 \pm 7 \times 10^{-3}$	$0.8355 \pm 3 \times 10^{-3}$	$0.7484 \pm 3 \times 10^{-4}$	$0.4554 \pm 3 \times 10^{-3}$
	SRMF	$0.7490 \pm 7 \times 10^{-4}$	$0.6404 \pm 1 \times 10^{-3}$	$0.5977 \pm 1 \times 10^{-3}$	$\mathbf{0.8911} \pm 4 \times 10^{-3}$	$0.7128 \pm 3 \times 10^{-5}$	$0.3342 \pm 6 \times 10^{-4}$
	HNMDRP	$0.7189 \pm 1 \times 10^{-3}$	$0.6613 \pm 3 \times 10^{-3}$	$0.6267 \pm 4 \times 10^{-3}$	$0.8651 \pm 7 \times 10^{-3}$	$0.7201 \pm 6 \times 10^{-4}$	$0.3643 \pm 7 \times 10^{-3}$
	MOFGCN	$\mathbf{0.8918} \pm 3 \times 10^{-4}$	$\mathbf{0.8291} \pm 6 \times 10^{-4}$	$\mathbf{0.8080} \pm 2 \times 10^{-3}$	$0.8704 \pm 1 \times 10^{-3}$	$\mathbf{0.8362} \pm 4 \times 10^{-4}$	$\mathbf{0.6637} \pm 2 \times 10^{-3}$
CCLE	DeepForest	$0.6218 \pm 4 \times 10^{-3}$	$0.5704 \pm 3 \times 10^{-3}$	$0.5457 \pm 1 \times 10^{-3}$	$\mathbf{0.9552} \pm 4 \times 10^{-3}$	$0.6910 \pm 4 \times 10^{-4}$	$0.2045 \pm 1 \times 10^{-2}$
	MOLI	$0.6737 \pm 4 \times 10^{-3}$	$0.6599 \pm 7 \times 10^{-3}$	$0.6331 \pm 7 \times 10^{-3}$	$0.8640 \pm 9 \times 10^{-3}$	$0.7210 \pm 1 \times 10^{-3}$	$0.3464 \pm 2 \times 10^{-2}$
	DeepCDR	$0.6643 \pm 4 \times 10^{-3}$	$0.6319 \pm 4 \times 10^{-3}$	$0.6047 \pm 5 \times 10^{-3}$	$0.8768 \pm 1 \times 10^{-2}$	$0.7061 \pm 6 \times 10^{-4}$	$0.3127 \pm 1 \times 10^{-2}$
	SRMF	$0.6884 \pm 1 \times 10^{-3}$	$0.6392 \pm 2 \times 10^{-3}$	$0.5930 \pm 2 \times 10^{-3}$	$0.9344 \pm 1 \times 10^{-3}$	$0.7230 \pm 6 \times 10^{-4}$	$0.3489 \pm 6 \times 10^{-3}$
	HNMDRP	$0.7623 \pm 3 \times 10^{-3}$	$0.7336 \pm 2 \times 10^{-3}$	$0.6901 \pm 4 \times 10^{-3}$	$0.8816 \pm 5 \times 10^{-3}$	$0.7689 \pm 1 \times 10^{-3}$	$0.4973 \pm 8 \times 10^{-3}$
	MOFGCN	$\mathbf{0.8436} \pm 4 \times 10^{-3}$	$\mathbf{0.8096} \pm 3 \times 10^{-3}$	$\mathbf{0.7889} \pm 6 \times 10^{-3}$	$0.8672 \pm 8 \times 10^{-3}$	$\mathbf{0.8201} \pm 3 \times 10^{-3}$	$\mathbf{0.6339} \pm 1 \times 10^{-2}$

the violin diagrams of all algorithms on the GDSC and CCLE databases, respectively. Compared with other algorithms, our algorithm has higher quartile values for all evaluation measures except recalls, and their distributions are more concentrated. Based on the above analysis, we believe that our method has better performance for single drug sensitivity prediction.

We also investigate the runtime cost of all comparing methods on the same computer. The computer configuration used in our experiment is 4 CPU cores and 16 GB memory. Table IV reports the average time each algorithm predicting a single drug on the GDSC and CCLE databases. As we can see from the table, our method runs much faster than others on the GDSC and CCLE datasets.

E. Predicting Response for Some Targeted Drugs

Targeted drugs are a group of drugs targeting the genes that help cancer to grow. There are many different types of targeted drugs according to the effect they have. Here, we chose the drugs that target ALK in the GDSC and CCLE databases to study our model's performance for predicting the response of a group of drugs. ALK is a tyrosine kinase receptor that is a clinical target of major cancer interest [28]. We screened four targeted drugs from the GDSC database, namely Crizotinib, NVP-TAE684, CH5424802, and XMD14-99, and one targeted drug NVP-TAE684 from the CCLE database. We used five-fold cross-validation to train and test our model and other comparing models on these targeted drugs. We randomly chose 1/5 of the positive samples and equal negative samples from the targeted drugs as the testing data, and 4/5 positive samples and remaining negative samples as training data. The process was repeated five times.

Table V shows experimental result comparison for all algorithms when predicting the targeted drug on the GDSC database and CCLE database, respectively. We observe that our method controls the best performance compared to all baseline methods on the two datasets in terms of AUC, ACC, precision, F1 score, and MCC values except in recall. Fig. 7 shows the ROC and P-R curves of all algorithms on the GDSC and CCLE datasets, respectively. For the ROC curve, when the false positive rate is low, MOFGCN reaches a higher true positive rate, which confirms that the algorithm can distinguish most positive samples from the negative samples. The P-R curves of our method are over others most of the time.

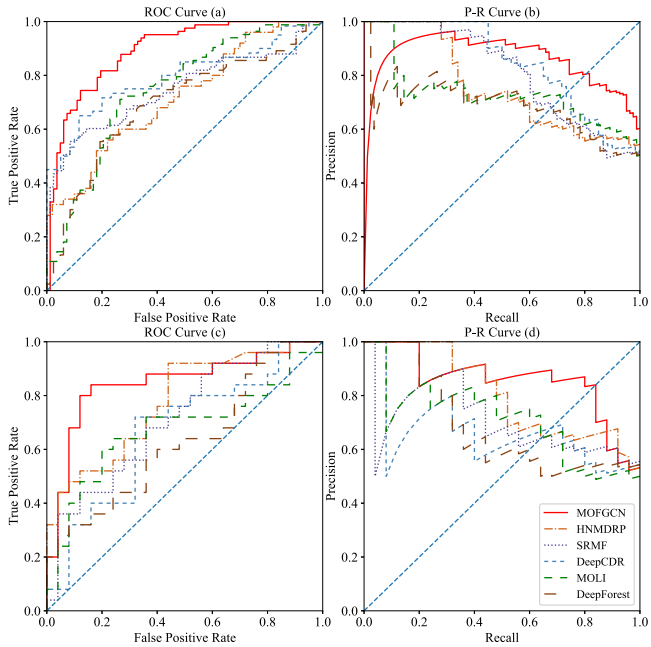

Fig. 7. ROC and P-R curve of predicting targeted drug's response on GDSC (a)–(b) and CCLE (c)–(d).

IV. CONCLUSION

This paper proposes an algorithm to predict drug response in cancer cell lines based on multi-omics fusion and graph convolution. The algorithm extracts useful information from cell line gene expression, copy number variation, and somatic mutation to generate the cell line similarity matrix. It then constructs a heterogeneous network, where drugs and cells are nodes and the known drug-cell associations are edges. After that, the algorithm uses the graph convolutional neural network to diffuse the similarity information and calculate the cell line features and the drug features. Finally, the algorithm uses a linear correlation coefficient decoder to reconstruct the cell line-drug association matrix for drug sensitivity prediction. We designed various experiments to verify the effectiveness of the algorithm. By analyzing the experimental results, we found that:

1) Compared with MOLI, DeepCDR, and DeepForest, inferring information from similar drugs improves the prediction performance.

2) Compared with HNMDRP and SRMF, fusing multi-omics data and the graph convolution neural network can effectively extract features for drugs and cell lines.
3) HNMDRP can hardly work for new drug prediction (single row clearing test). But our method can still predict drug sensitivity effectively, which shows that our algorithm is robust.

In the future, we will further improve the algorithm to predict drug sensitivity. Inspired by [29] and [30], we believe adding an attention mechanism to the graph convolutional neural network to extract higher-order information for nodes will improve the algorithm. Besides, introducing more omics data, such as methylation omics, will also enhance the algorithm.

REFERENCES

[1] Q. Li, R. Shi, and F. Liang, "Drug sensitivity prediction with high-dimensional mixture regression," *PLoS One*, vol. 14, no. 2, p.e0212108, 2019.

[2] M. J. Garnett *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, 2012.

[3] J. Sheng, F. Li, and S. T. Wong, "Optimal drug prediction from personal genomics profiles," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 4, pp. 1264–1270, Jul. 2015.

[4] T. LaFramboise, "Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances," *Nucleic Acids Res.*, vol. 37, no. 13, pp. 4181–4193, 2009.

[5] W. Yang *et al.*, "Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D955–D961, 2012.

[6] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "PubChem: Integrated platform of small molecules and biological activities," *Annu. Reports Comput. Chem.*, Elsevier, vol. 4, pp. 217–241, 2008.

[7] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, pp. 1–15, 2017.

[8] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome Biol.*, vol. 15, no. 3, pp. 1–12, 2014.

[9] L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Dr. VAE: Improving drug response prediction via modeling of drug perturbation effects," *Bioinformatics*, vol. 35, no. 19, pp. 3743–3751, 2019.

[10] S. Daoud, A. Mdhaffar, M. Jmaiel, and B. Freisleben, "Q-Rank: Reinforcement learning for recommending algorithms to predict drug sensitivity to cancer therapy," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3154–3161, Nov. 2020.

[11] R. Rahman, J. Otridge, and R. Pal, "IntegratedMRF: Random forest-based framework for integrating prediction from different data types," *Bioinformatics*, vol. 33, no. 9, pp. 1407–1410, 2017.

[12] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-Forest: A deep forest model to predict anti-cancer drug ressponse," *Methods*, vol. 166, no. 8, pp. 91–102, 2019.

[13] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, "MOLI: Multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, 2019.

[14] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young, and X. Lu, "Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Mol. Cancer Res.*, vol. 16, no. 2, pp. 269–278, 2018.

[15] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS Comput. Biol.*, vol. 11, no. 9, p. e1004498, 2015.

[16] T. T. Nguyen, G. T. T. Nguyen, and D.-H. Le, "Graph convolutional networks for drug response prediction," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, doi: 10.1109/TCBB.2021.3060430.

[17] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, "DeepCDR: A hybrid graph convolutional network for predicting cancer drug response," *Bioinformatics*, vol. 36, no. Suppl._2, pp. i 911–i918, 2020.

[18] F. Zhang, M. Wang, J. Xi, J. Yang, and A. Li, "A novel heterogeneous network-based method for drug response prediction in cancer cell lines," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018.

[19] L. Wang, X. Li, L. Zhang, and Q. Gao, "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization," *BMC Cancer*, vol. 17, no. 1, pp. 1–12, 2017.

[20] B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *2017 Inter. Conf. Learn. Representations*, Toulon, France, Apr. 2017.

[22] Y. Wang, Y. Yao, H. Tong, F. Xu, and J. Lu, "A brief review of network embedding," *Big Data Mining Analytics*, vol. 2, no. 1, pp. 35–47, 2018.

[23] J. Barretina *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[24] S. L. Westcott and P. D. Schloss, "OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units," *MSphere*, vol. 2, no. 2, 2017, pp. e00073–17.

[25] N.-N. Guan, Y. Zhao, C.-C. Wang, J.-Q. Li, X. Chen, and X. Piao, "Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization," *Mol. Therapy-Nucleic Acids*, vol. 17, no. 9, pp. 164–174, 2019.

[26] Y.-a. Huang, P. Hu, K. C. Chan, and Z.-H. You, "Graph convolution for predicting associations between mirna and drug resistance," *Bioinformatics*, vol. 36, no. 3, pp. 851–858, 2020.

[27] J. E. Staunton *et al.*, "Chemosensitivity prediction by transcriptional profiling," *Proc. Nat. Acad. Sci.*, vol. 98, no. 19, pp. 10787–10792, 2001.

[28] J. Van den Eynden *et al.*, "Phosphoproteome and gene expression profiling of alk inhibition in neuroblastoma cell lines reveals conserved oncogenic pathways," *Sci. Signaling*, vol. 11, no. 557, pp.eaar5680, 2018.

[29] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug-disease associations through layer attention graph convolutional network," *Brief. Bioinf.*, vol. 22, no. 4, p. bbaa243, 2021.

[30] P. Xuan, L. Gao, N. Sheng, T. Zhang, and T. Nakaguchi, "Graph convolutional autoencoder and fully-connected autoencoder with attention mechanism based method for predicting drug-disease associations," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1793–1804, May 2021.