Applied Intelligence
https://doi.org/10.1007/s10489-022-03294-w

A multi-view multi-omics model for cancer drug response prediction

Zhijin Wang[1] · Ziyang Wang[1] · Yaohui Huang[2] · Longquan Lu[3] · Yonggang Fu[1]

Accepted: 23 January 2022

Abstract

Cancer drug response prediction is the fundamental task in precision medicine, which provides opportunities for cancer therapy. Several methods have been proposed to screen drugs, via building computational models on multi-omics data. However, the view value missing problem caused by unknown cancers or tumors has not been addressed. For this reason, a multi-view multi-omics (MvMo) model is proposed to predict cancer drug response values. The proposed MvMo model first represents the input heterogeneous data in different kinds of embeddings and features, such as token embeddings and latent features. Then several views are generated to observe interconnections among those representations. Finally, the predictions are generated based on the outputs of these views. Experimental results on the collected real data show the efficiency of the proposed method in terms of speed and accuracy.

Keywords Cancer drug response · Prediction · Multi-view learning · Multi-omics data

1 Introduction

Precision medicine provides opportunities to therapy life-threatening diseases, such as cancer or tumor. The cancer heterogeneity among patients causes the challenge of drug screening, which means a common type of two patient-specific cancers usually has huge differences of effects on a drug [3]. Meanwhile, patient-specific cancer can not be tested on many drugs in clinical settings.

Recently, the non-clinical and clinical databases have been publicly available, the predictive model becomes a promising way to screen drugs for specific cancer cell lines. The databases include but are not limited to cell line databases, cancer genome databases, chemical compounds databases, and cancer-drug sensitivity databases. Hence, several models have been proposed to predict the cancer-drug-response (CDR) value, via leveraging cell line information and drug molecular information [28].

Several methods have been proposed to predict CDR value, via feeding cancer information and drug information into machine learning model [7]. Such as, Ridge Regression [13], MOLI [20], and DeepCDR [17]. However, there are interactions between cancers and drugs, which can not be directly observed. For this reason, the cell line data and drug information data can not be treated as two independent parts of model inputs. Moreover, the CDR values are usually missing in unknown cancers or tumors, which is also known as view value missing problem in multi-view learning [6]. A toy example of the view value missing problem is compared with the random value missing problem in Fig. 1.

This paper proposes a multi-view multi-omics (MvMo) model to deal with the two problems, and make more accurate predictions. The proposed MvMo consists of three

This article belongs to the Topical Collection: *Special Issue on Multi-view Learning*
Guest Editors: Guoqing Chao, Xingquan Zhu, Weiping Ding, Jinbo Bi and Shiliang Sun

✉ Yonggang Fu
yonggangfu@jmu.edu.cn

Zhijin Wang
zhijinecnu@gmail.com

Ziyang Wang
ndwang123@gmail.com

Yaohui Huang
yhhuang5212@gmail.com

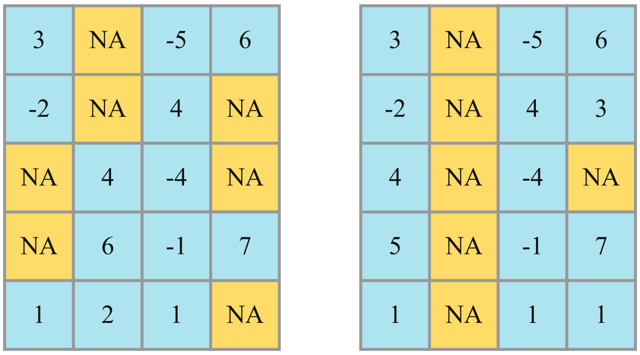
Longquan Lu
longquanlus@gmail.com

[1] Computer Engineering College, Jimei University, Yinjiang Road 185, Xiamen, 361021, China

[2] College of Electronic Information, Guangxi University for Nationalities, Daxue East Road 188, Nanning, 530006, China

[3] School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China

Published online: 11 March 2022



3	NA	-5	6
-2	NA	4	NA
NA	4	-4	NA
NA	6	-1	7
1	2	1	NA

(a) Random missing

3	NA	-5	6
-2	NA	4	3
4	NA	-4	NA
5	NA	-1	7
1	NA	1	1

(b) View missing

Fig. 1 Two value-missing examples. The left part is a toy example of random value missing. The right part is another toy example of view value missing

stages. Firstly, the multi-omics data and drug molecular data are represented embeddings or latent features. All the represented embeddings or features share a common dimension. Secondly, five views are generated to represent interactions among embeddings and features. Finally, the outputs of views are gathered to generate CDR predictions.

The remainder of this paper is organized as follows. Section 2 addresses this research. Section 3 gives problem formulation on CDR prediction using multi-omics data. Section 4 illustrates the proposed MvMo. Section 5 configures the experiment environment. Section 6 shows the experimental results and analyses. Finally, we conclude in Section 7.

2 Related work

This section addresses the research by recalling several CDR prediction studies. According to the nature of methods, they are divided into three categories as follows.

2.1 Traditional machine learning methods

Due to the growth of computational power, and the ability of high-dimensional data processing, machine learning methods had been applied to CDR prediction by leveraging omics data integrative analyses [1, 2].

Regression-based methods mapped the inputs to outputs in a linear or non-linear manner. ENet [31] combined both L_1 and L_2 regularization terms to the loss function of linear regression. GELnet [22] extended the ENet by incorporating a gene similarity matrix into a regularizor. RWEN [4] used an iterative weighted elastic regression network based on gene expression information to predict drug resistance response. Sotudian and Paschalidis [23] was a regression model constructed based on ENet, which abandoned the idea of directly predicting drug response and instead ranked drug sensitivity through a paired comparison of drug sensitivity between cell lines. PairwiseMKL [8] constructed a practical pairwise learning framework based on kernels, which integrated heterogeneous data sources into a single model, by learning the weights of kernels with different information sources and combining the input kernels. KRR [14] learned associations between drugs and mapped gene expression information of individual cell lines to a high-dimensional feature space to predict and rank the response of each drug to that cell line.

Random Forest (RF) is a classifier that ultimately votes on the output of multiple decision trees and accomplishes the prediction by capturing the nonlinear relationships among the data. Riddick et al. [19] incorporated feature filtering and outlier filtering frameworks into the RF model. HARF [18] assigns weights to the tree based on the type of cancer in the sample and achieves a performance lead compared to the traditional RF method in the case of uneven drug response. Costello et al. [9] designed an integrated model that integrates Support Vector Machine (SVM) to process multi-omics information separately. Finally, the output of each vector machine is weighted and averaged to derive the prediction value.

Methods in this category are difficult to capture the connection between heterogeneous data sources, and the integration difficulties caused by data heterogeneity have not been fundamentally solved.

2.2 Recommendation system methods

Several researchers assumed that similar cell lines or drugs would have similar effects in drug response [29]. Therefore, recommendation systems were applied to predict CDR values by incorporating other known CDR values of similar cell lines and drugs. In a recommender framework [26], a CDR value is regarded as a user-item-preference value. Hence, lots of recommendation algorithms can be used to predict CDR values.

Owning to the success of matrix factorization in rating prediction [27], this technique had been popular in predicting values within a sparse matrix. CaDRReS [24] used the matrix factorization to learn the potential associations between cell lines and drugs. DualNet [30], SRMF [25], and HNMDRP [29] regularized cell line-cell line similarity matrix, and drug-drug similarity matrix in their loss functions to achieve a relative good learning process. BMTMKL [9] used multiple kernel functions to learn multi-omics features, and the output of each kernel was weighted to generate predictions in consequence of the last kernel function. KBMF [10] presented a personalized ranking drug recommender, by incorporating drug features and omics data of unknown cell lines. cwKBMF

[11] was built on the KBMF by extending path-based features.

Most of the recommendation-based methods extended the matrix factorization framework by adding similarity regularization terms, which usually neglected the inherent features of drugs and cell lines. Meanwhile, the drugs and cell lines may interact with each other in the molecular level, which leads to poor explanations [21] on the prediction results.

2.3 Deep learning methods

Deep learning methods were generally based on combinations or stacks of Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and attention mechanisms. MVLR [12] proposed a multi-view multi-task model based on functional link networks, which treated different data sources as independent feature sets. CDRScan [5] used stacked CNNs to map inputted cell line data and molecular drug data to their corresponding CDR value. MOLI [20] used triplet loss to receive the integrated outputs of several different neural networks, where a network was applied to a kind of input data. DeepCDR [17] utilized graph convolution network (GCN) to process drug molecular information, and merged it with cell line multi-omics data to predict CDR values. DeepDSC [16] used stacked encoders to extract genomics from gene expression data, and then chemical characteristics of drugs were jointed to generate response values.

There are still relatively few deep learning methods in drug response prediction. The main reason is that the deep learning methods are straightforward to overfitting and usually can not generalize in different data sets due to their high degree of parameterization.

In summary, the major challenge in drug response prediction lies in the fusion of heterogeneous data from multiple sources. Technically, the significant differences in the data structure, dimensionality, signal-to-noise ratio, and complexity of multi-omics data pose a great challenge in representation learning. Moreover, these methods commonly ignore the view value missing problem from unknown cell lines of cancers or tumors. In this paper, the proposed MvMo fuses multi-omics data in a multi-view framework to alleviate the view missing problem and generate accurate predictions.

3 Problem formulation

Let N_c be the number of cell lines, N_d be the number of drugs, and symbol $\boldsymbol{R} \in \mathbb{R}^{N_c \times N_d}$ be the observed cancer-drug-response (CDR) matrix. Symbol $R_{c,d} \in \boldsymbol{R}$ is the entry of $\boldsymbol{R}$. Symbol $\hat{\boldsymbol{R}} \in \mathbb{R}^{N_c \times N_d}$ denotes the predicted CDR matrix.

The task of CDR prediction is to predict missing values in the CDR matrix. The in-matrix missing values is denoted by symbol $\hat{\boldsymbol{R}}$. In recommendation systems [24], the CDR prediction task is formulated as:

$$\hat{\boldsymbol{R}} = F(\boldsymbol{R}), \tag{1}$$

where $\boldsymbol{R}$ presents observed CDR values, and $\hat{\boldsymbol{R}}$ denotes unknown CDR values. In reality, new cancers or tumors are unknown, and are not tested with any drugs. Hence, we subject (1) to $\{c_1 \in \boldsymbol{R} \neq c_2 \in \hat{\boldsymbol{R}}\}$, where c_1 and c_2 are cell lines happens in $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$, respectively.

Due to the enrichment of cell line information and drug molecular information, the CDR prediction can be formulated as a machine learning problem [5, 29] as well. Let symbol $\boldsymbol{P} \in \mathbb{R}^{N_c \times N_p}$ be the genetic expression of cell lines, symbol $\boldsymbol{M} \in \mathbb{R}^{N_c \times N_m}$ be the mutation positions of cell lines, symbol $\boldsymbol{J} \in \mathbb{R}^{N_c \times N_j}$ be the genetic methylation of cell lines, symbol $\boldsymbol{F} \in \mathbb{R}^{N_d \times N_a \times N_f}$ be the molecular features of drugs. Hence, the task is defined as:

$$\hat{\boldsymbol{R}} = F(\boldsymbol{P}, \boldsymbol{M}, \boldsymbol{J}, \boldsymbol{F}). \tag{2}$$

Not only the observed CDR values, but also the multi-omics data should be considered while predicting unknown values in this paper. Therefore, the CDR prediction problem is re-formulated as follows:

$$\hat{\boldsymbol{R}} = F(\boldsymbol{R}, \boldsymbol{P}, \boldsymbol{M}, \boldsymbol{J}, \boldsymbol{F}), \tag{3}$$

where $\boldsymbol{R}$ is the observed cancer-drug interactions, and $\boldsymbol{P}, \boldsymbol{M}, \boldsymbol{J}, \boldsymbol{F}$ are cell line information and drug features.

Table 1 lists the main notations used in this paper.

4 The proposed MvMo

This section illustrates the proposed MvMo. The proposed MvMo consists of three stages: input data representation, view generation, and view combination. Figure 2 gives the graphical illustration of MvMo.

Firstly, all the input data are represented to a latent space using embeddings of transformation, since those multi-omics data and drug feature data are heterogeneous, which can not be operated with each other to connect with their corresponding response values. The data representation operations are shown in the left part of Fig. 2.

Secondly, several views are generated to receive the embeddings and transformed features, and then the latent interactions are calculated. The target of those views is to observe the potential reactions between cell lines and drugs in terms of embeddings or features. Those views are displayed in the middle part of Fig. 2.

Finally, those views are combined to connect with the CDR values. And the backward propagation algorithms are applied to the model.

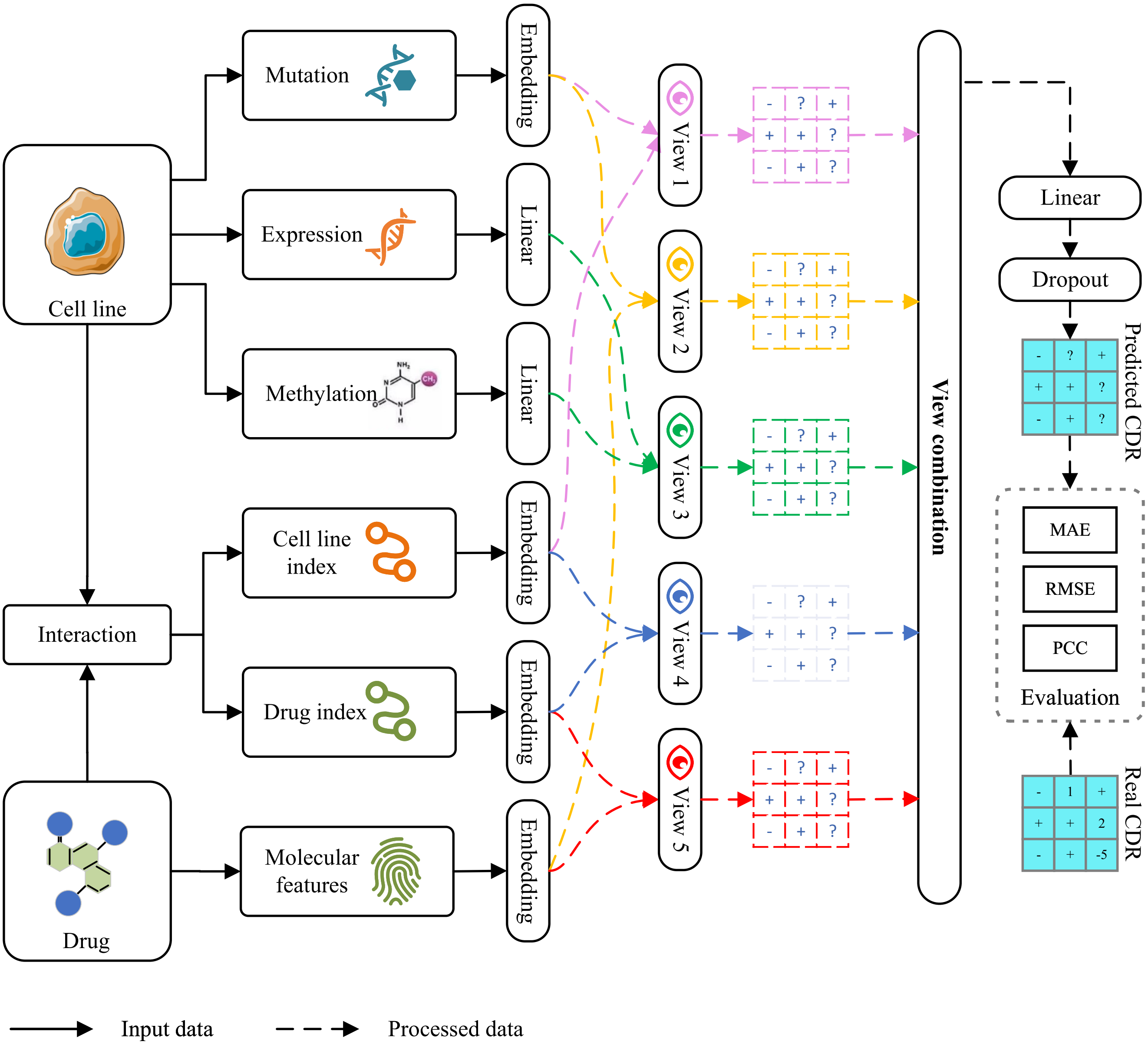


Fig. 2 The graphical illustration of the proposed Multi-view Multi-Omics (MvMo) model

4.1 Molecular data representation

Mutation embedding The mutation data is a sequence of $\{0, 1\}$ symbols, and the mutation positions are marked with symbol 1. Motivated by natural language processing (NLP), a mutation sequence is regarded as a text, and can be represented by embeddings. The mutation embedding process is formulated as follows:

$$\boldsymbol{E}^{m} = F(\boldsymbol{M}), \tag{4}$$

where N_e denotes the embedding size, N_x is the maximum number of mutated positions in a cell line, $\boldsymbol{M} \in \mathbb{R}^{N_c \times N_m}$ is the mutation sequences, and $\boldsymbol{E}^{m} \in \mathbb{R}^{N_c \times N_x \times N_e}$ denotes mutation embeddings.

Genetic expression features The genetic expression features consist of a series of continuous values. To operate the features with other embeddings or features, the size of these features is condensed to the size of embeddings. The condense operation is formulated as follows:

$$\boldsymbol{E}^{p} = F(\boldsymbol{P}), \tag{5}$$

where N_e denotes the feature size, $\boldsymbol{P} \in \mathbb{R}^{N_c \times N_p}$ is the mutation sequences, and $\boldsymbol{E}^{p} \in \mathbb{R}^{N_c \times N_e}$ denotes condensed expression features.

Methylation features The methylation features consist of a series of continuous variables as well. The condense operation is formulated as follows:

Table 1 Major notations

Symbol	Notation
N_c	The number of cell lines
N_d	The number of drugs
$\boldsymbol{R}$	Observed CDR matrix, $\boldsymbol{R} \in \mathbb{R}^{N_c \times N_d}$
$R_{c,d}$	Entry of matrix $\boldsymbol{R}$
$\hat{\boldsymbol{R}}$	Predicted CDR matrix, $\hat{\boldsymbol{R}} \in \mathbb{R}^{N_c \times N_d}$
$\hat{R}_{c,d}$	Entry of matrix $\hat{\boldsymbol{R}}$
N_m	The number of total mutated positions
$\boldsymbol{M}$	Mutation matrix, $\boldsymbol{M} \in \mathbb{R}^{N_c \times N_m}$
N_p	The number of genes of a cell line
$\boldsymbol{P}$	Expression matrix, $\boldsymbol{P} \in \mathbb{R}^{N_c \times N_p}$
N_j	The number of methylation positions
$\boldsymbol{J}$	Methylation matrix, $\boldsymbol{J} \in \mathbb{R}^{N_c \times N_j}$
N_a	The maximum number of atoms in a molecular
N_f	The number atom features
$\boldsymbol{F}$	Drug molecular feature tensor, $\boldsymbol{F} \in \mathbb{R}^{N_d \times N_a \times N_f}$
$F(\cdot)$	A mapping from inputs to outputs
N_e	The global embedding size
N_x	The maximum number of mutated positions
$\boldsymbol{E}^m$	Mutation embedding matrix $\boldsymbol{E}^m \in \mathbb{R}^{N_c \times N_x \times N_e}$
$\boldsymbol{E}^p$	Expression feature matrix $\boldsymbol{E}^p \in \mathbb{R}^{N_c \times N_e}$
$\boldsymbol{E}^j$	Methylation feature matrix $\boldsymbol{E}^j \in \mathbb{R}^{N_c \times N_e}$
$\boldsymbol{E}^c$	Cell line latent feature matrix $\boldsymbol{E}^c \in \mathbb{R}^{N_c \times N_e}$
$\boldsymbol{E}^d$	Drug latent feature matrix $\boldsymbol{E}^d \in \mathbb{R}^{N_d \times N_e}$
$\boldsymbol{E}^f$	Molecular embeddings $\boldsymbol{E}^f \in \mathbb{R}^{N_d \times N_a \times N_f \times N_e}$

$$\boldsymbol{E}^j = F(\boldsymbol{J}), \tag{6}$$

where N_e denotes the feature size, $\boldsymbol{J} \in \mathbb{R}^{N_c \times N_j}$ is the mutation sequences, and $\boldsymbol{E}^j \in \mathbb{R}^{N_c \times N_e}$ denotes condensed methylation features.

Cell line latent features and drug latent features Motivated by matrix factorization techniques, which factorize the CDR matrix into cell line latent features and drug latent features. The size of latent features is set the same as the size of embeddings. The operation is formulated as:

$$\boldsymbol{E}^c, \boldsymbol{E}^d = F(\boldsymbol{R}), \tag{7}$$

where $\boldsymbol{R}$ is the observed CDR values, $\boldsymbol{E}^c \in \mathbb{R}^{N_c \times N_e}$ is cell line latent features, and $\boldsymbol{E}^d \in \mathbb{R}^{N_d \times N_e}$ is drug latent features.

Drug molecular embeddings The drug molecular consists of several atoms, and each atom has similarity features, which is also a sequence of {0, 1} symbols. The molecules are also represented by embeddings, and are formulated as follows:

$$\boldsymbol{E}^f = F(\boldsymbol{F}), \tag{8}$$

where $\boldsymbol{F} \in \mathbb{R}^{N_d \times N_a \times N_f}$ is the molecular feature tensor, and $\boldsymbol{E}^f \in \mathbb{R}^{N_d \times N_a \times N_f \times N_e}$ denotes mutation embeddings.

Many methods can be applied to condense the above inputs to a vector/matrix/tensor with a given size, such as CNN, RNN, and Transformer. However, this paper focuses on the multi-view framework. Hence, we abandon all those excellent networks to learn the advantages of the proposed MvMo model. For (4), (7) and (8), we regard each input as a list of tokens, and use the tokens to lookup embeddings. For (5) and (6), a linear layer to used to reduce the dimension to the embedding size. These simple representations are prepared to witness the MvMo framework.

4.2 View generation

To observed the interactions between cell line information and drug information, five views are generated to learn latent response values.

View 1 observes the interconnections between cell line and it's mutation. Given a cell line c, the observations from view 1 is formulated as follows:

$$\boldsymbol{O}^1 \leftarrow V1\left(\boldsymbol{E}^m_{c,:,:}, \boldsymbol{E}^c_{c,:}\right), \tag{9}$$

where $\boldsymbol{O}^1 \in \mathbb{R}^{N_x}$ denotes the output of view 1, $\boldsymbol{E}^m \in \mathbb{R}^{N_c \times N_x \times N_e}$ is mutation embeddings, and $\boldsymbol{E}^c \in \mathbb{R}^{N_c \times N_m}$ is the cell line latent features.

View 2 observes the relationship between genome mutation and drug molecular feature. Given a cancer-drug pair (c, d), the observations from view 2 is formulated as follows:

$$\boldsymbol{O}^2 \leftarrow V2\left(\boldsymbol{E}^m_{c,:}, \boldsymbol{E}^f_{d,:,:,:}\right), \tag{10}$$

where $\boldsymbol{O}^2 \in \mathbb{R}^{N_a \times N_f}$ denotes the output of view 2.

View 3 observes the connections between expression and methylation. Given a cell line c, the observations from view 3 is formulated as follows:

$$O^3 \leftarrow V3\left(\boldsymbol{E}^p_{c,:}, \boldsymbol{E}^j_{c,:}\right), \tag{11}$$

where $O^3 \in \mathbb{R}$ denotes the output of view 3.

View 4 observes the factorized cell line latent matrix and drug latent matrix. Given a cancer-drug pair (c, d), the product of the two latent features could be regarded as a predicted CDR value.

$$O^4 = \boldsymbol{E}^c_{c,:} \cdot \boldsymbol{E}^d_{d,:}, \tag{12}$$

where $O^4 \in \mathbb{R}$ denotes the output of view 4.

View 5 observes the interconnections between drug latent feature and drug molecular features. The observations are represented by:

$$\boldsymbol{O}^5 \leftarrow V5\left(\boldsymbol{E}^d_{d,:}, \boldsymbol{E}^f_{d,:,:,:}\right), \tag{13}$$

where $\boldsymbol{O}^5 \in \mathbb{R}^{N_a \times N_f}$ denotes the output of view 5.

To alleviate the burden of calculation, we use inner product to represent their interconnections in (9), (10), (11) and (13).

4.3 View combination

We gather the outputs from the above five views, and get:

$$\boldsymbol{O} = \left[\boldsymbol{O}^1, \boldsymbol{O}^2, O^3, O^4, \boldsymbol{O}^5\right], \tag{14}$$

where $\boldsymbol{O} \in \mathbb{R}^{1\times(N_x+N_a*N_f*2+2)}$ denotes the flatten concatenated view outputs.

The combination task is to link the outputs $\boldsymbol{O}$ to the CDR value $R_{c,d}$. Since this paper focuses on designing a multi-view framework, two consecutive linear layers receive the view outputs $\boldsymbol{O}$ to generate predictions.

5 Experimental configuration

This section introduces datasets, performance measurements, comparable methods and model configurations.

5.1 Datasets

The basic statistics of relevant datasets used in this paper are listed in Table 2.

To investigate the benefits of the proposed MvMo, we follow the collection procedures in previous studies, such as NCI-DREAM [9], and DeepCDR [17]. The original data were downloaded and aggregated from the well-known bioinformatics databases as follows:

(1) *Genomics of Drug Sensitivity in Cancer*[1] (GDSC) is the largest public resource for information on drug sensitivity in cancer cells and molecular markers of drug response. It provides a large amount of IC_{50} values matching with cell and drug pairs.

(2) *Cancer Cell Line Encyclopedia*[2] (CCLE) is a cancer cell line database, which contains large-scale, robust, well-defined cancer cell line models. We focus on three cell line multi-omics data: expression, methylation, and mutation.

(3) *The Cancer Genome Atlas*[3] (TCGA) is a large database of human cancer genetic information, including mutations, mRNA expression, miRNA expression and methylation data.

(4) *PubChem*[4] is the world largest public database of chemical information. It provides hundreds of drug structure data.

These datasets contain 494 cell lines, 237 drugs, and 94314 observed response values. About 29% of them are missing values. The cell line data contains genetic expression with shape 494 × 697, methylation with shape 494 × 808, and gene mutation with shape 494 × 34673. The cell line data contains 24 cancer types. The drug data contains molecular graph data and drug feature data. The molecular in the drug data has 96 atoms at most and has 75 features at most.

To reduce the biases from model and data, cross-validation and train/test data split methods were applied on all the comparable models. Another reason may be the drug response for a specific-patient is complex, especially at the molecular level, deep learning models may easily be stuck in extreme values for a specific dataset. Here, 10-fold cross-validation were used to split the train/test CDR values. For each group of cross-validation, three kinds of split methods were employed. The train/test data were splitted according to non-overlapping cell lines, non-overlapping drugs, and random. The TCGA types are used to guarantee no interacted cell lines between train data and test data. One of the ten parts of data is left out as test data in turn, which means each model will be running ten times.

5.2 Performance measurements

Many measurements can be applied to assess model performance. The common used mean absolute error (MAE), root mean squared error (RMSE), and Pearson Coefficient Correlation (PCC) [17] are adopted to evaluate the prediction performance of $ln(IC_{50})$ values. They are formulated as follows:

(1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{|\boldsymbol{T}|}\sum_{(c,d)\in \boldsymbol{T}} |R_{c,d} - \hat{R}_{c,d}|, \tag{15}$$

(2) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{|\boldsymbol{T}|}\sum_{(c,d)\in \boldsymbol{T}} \left(R_{c,d} - \hat{R}_{c,d}\right)^2}, \tag{16}$$

[1] https://www.cancerrxgene.org/

[2] https://sites.broadinstitute.org/ccle

[3] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[4] https://pubchem.ncbi.nlm.nih.gov/

Table 2 The basic statistics of the four datasets used in experiments

	IC_{50}	Expression	Methylation	Mutation	Drug feature	Molecular graph
#Cell lines	494	494	494	494	–	–
#Drugs	237	–	–	–	237	237
#Interactions	94314	–	–	–	–	–
#Genes	–	697	808	–	–	–
#Atoms	–	–	–	–	–	96 (Max)
#Features	–	–	–	–	75 (Max)	–
#Genetic locus	–	–	–	34673 (Max)	–	–
#TCGA type	24	24	24	24	–	–
Data source	GDSC	CCLE	CCLE	CCLE	PubChem	PubChem

(3) Pearson Correlation Coefficient (PCC)

$$PCC = \frac{\sum_{(c,d)\in \boldsymbol{T}} \left(R_{c,d} - \bar{R}\right) * \left(\hat{R}_{c,d} - \bar{\hat{R}}\right)}{\sqrt{(\sum_{(c,d)\in \boldsymbol{T}} \left(R_{c,d} - \bar{R}\right)^2} \sqrt{\sum_{(c,d)\in \boldsymbol{T}} \left(\hat{R}_{c,d} - \bar{\hat{R}}\right)^2}}, \tag{17}$$

where $\boldsymbol{T}$ is the testing set, $R_{c,d}$ is a real value, $\hat{R}_{c,d}$ is a predicted value, $\bar{R}$ is the mean value in the testing set, and $\bar{\hat{R}}$ is the mean value of predicted CDR values.

The number of trainable parameters, GPU memory consumption, and training time per batch are used to compare the model size and training speed.

5.3 Competition models

(1) Matrix Factorization (MF) [27] is a famous technique in predicting user-item-rating value in recommender systems. Here, it's used to predict CDR values, i.e., $ln(IC_{50})$.

(2) Multiple Linear Regression (MLR) [13] linearly makes connections between every input elements with the output elements. The input elements consist of genetic expression, methylation, gene mutation and drug features.

(3) CDRScan [5] published several versions.[5] The three versions with relative good performance were employed as comparable methods. CDRScan-Master applies two stacked CNNs on molecular fingerprints and genomic mutations to represent drugs and cancer profiles, respectively. The representations are fed into the third stacked CNN. Compared with CDRScan-Master, CDRScan-Shallow has few CNN layers in the third stacked CNN, but has more linear layers operations. CDRScan-FullConnected replaced the third stacked CNN using full connected layers. For convenience, CDRScan-Master, CDRScan-Shallow, and CDRScan-FullConnected are presented by CDRScan-M, CDRScan-S, and CDRScan-FC, respectively.

(4) DeepCDR [17] leverages CNNs and graph convolutional networks (GCNs) to process multi-omics data of cell lines and chemical features of drugs, respectively. The codes are available at github.com.[6]

5.4 Model configuration

For fair competition on all models, the batch size is set to 64. The Adam optimizer [15] is adopted to train all the models. All the compared models are implemented in PyTorch 1.8.2 (LTS), and are ran four graphics processing units of NVIDIA Tesla V100. Moreover, the comparable models have achieved the same accuracy as their corresponding literature. All the size of latent features and the size of embeddings are set to 20.

The MF is solely based on interaction data, i.e., IC_{50}. The MLR, DeepCDR, CDRScan use all the molecular level data as input and corresponding observed IC_{50} as target. All these methods are running on the same 10-fold cross-validation.

6 Results and analyses

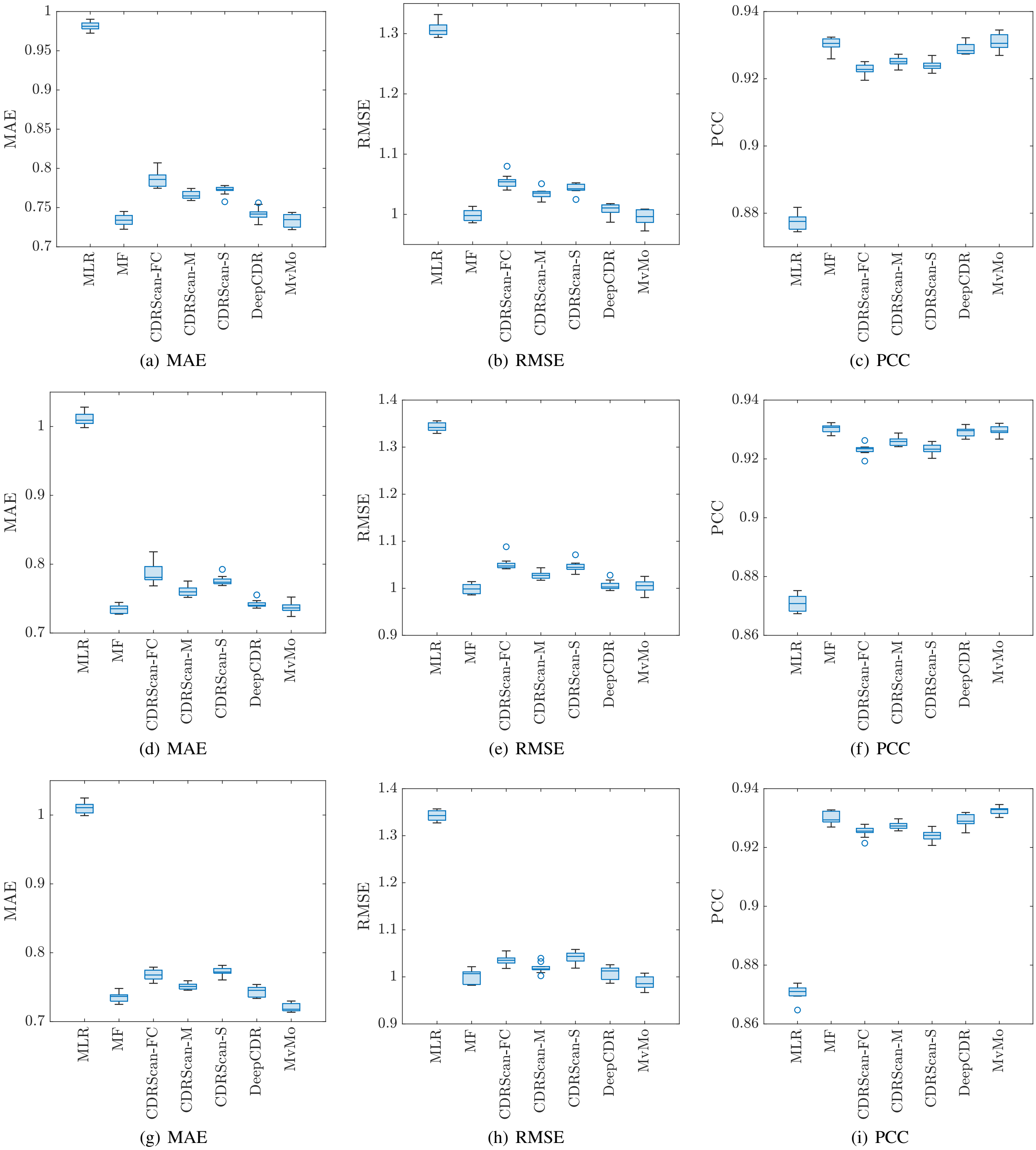
This section gives the results and provides insights from the experimental findings. The performance competition is provided in the perspective of prediction performance and model complexity.

6.1 Comparisons

For fair competition, all the comparable methods were ran on common tasks. The experimental results are shown in Fig. 3. The major observations from the results are listed below:

[5] http://github.com/summatic/CDRScan

[6] http://github.com/kimmo1019/DeepCDR



(a) MAE

(b) RMSE

(c) PCC

(d) MAE

(e) RMSE

(f) PCC

(g) MAE

(h) RMSE

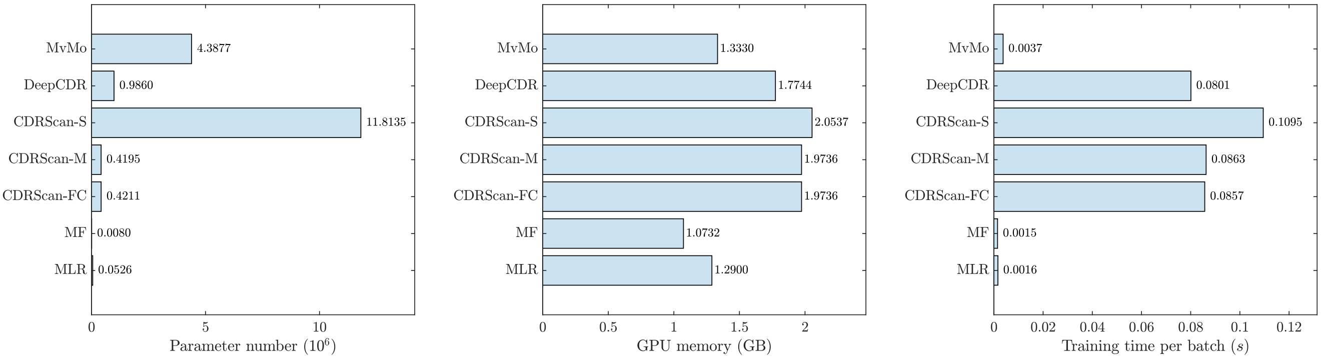
(i) PCC

Fig. 3 Box-plots of seven methods in terms of MAE, RMSE and PCC. (a-c) split based on non-overlapping cell lines. (d-f) split based on non-overlapping drugs. (g-i) random split

(1) When observing at the three metrics on three groups of data, the proposed MvMo has the best performance for all metrics when compared with other methods.

(2) When observing at the three metrics on three groups of data, MLR gets the worst performance.

(3) According to the box results of each method on non-overlapping cell line dataset and non-overlapping drug



(a) Trainable parameters (b) Memory usage (c) Training time

Fig. 4 Comparisons of seven methods in terms of trainable parameters, memory usage and training time. The small values mean low time complexity or space complexity. The MvMo has relative few parameters and training time

dataset, MLR, MF, DeepCDR and MvMo are stable, but CDRScan methods fiercely shakes.

(4) The CDRScan methods gain relative stable results on the random split dataset.

There are several potential reasons for the poor performance of MLR:

(1) The relationship between multiple characteristic information and drug response is not linear.
(2) MF has the second-best prediction performance and has excellent training time and space complexity.
(3) There is a strong correlation between multi-omics and drug structure data. MLR cannot capture this relationship.
(4) Due to the high degree of heterogeneity of those molecules, it is difficult to find intrinsic patterns solely using linear functions.

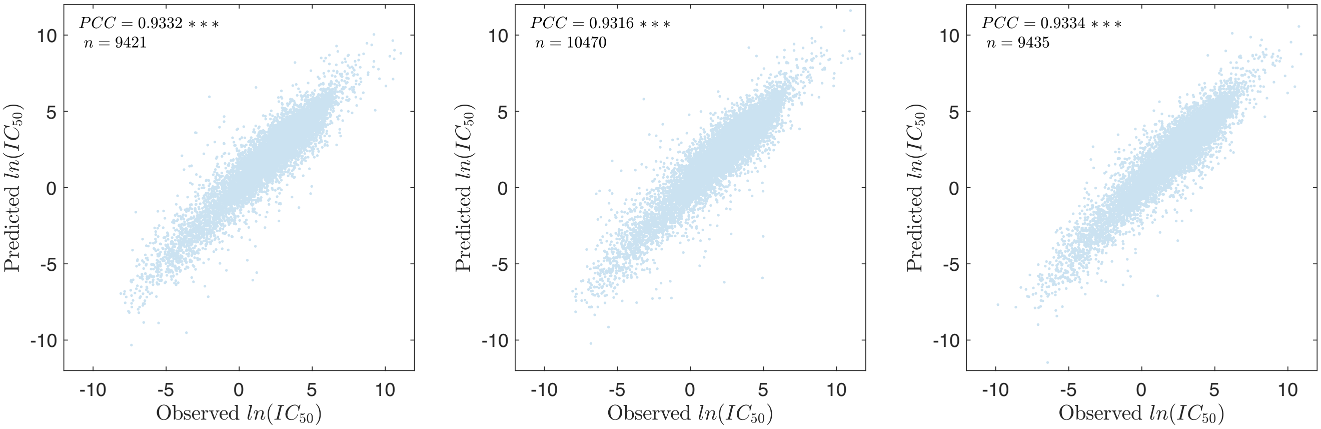
Compared with MvMo, MF ignores the intrinsic characteristics of drugs and cell lines, the predictions are generated based on the learned latent features. This reveals the impact of molecular data.

For CDRScan-FC, more linear layers bring better stability but may affect the model's ability to capture nonlinear features. For CDRScan-S, fewer convolutional units reduce the accuracy. Compared with CDRScan, DeepCDR simplifies the processing of omics data, which can fuse more omics data. DeepCDR achieves third-best performance, somehow owning to the benefits from GCN on drug molecular representation. There is still a gap between DeepCDR and MF. A possible reason for this improvement is that latent interactions are much more important than feature representation only.

MvMo shows the best performance in all three metrics. It uses an embedding component to encode the input data, which compresses the high-dimensional input data into a low-dimensional feature space and filters out fluctuations. This allows MvMo to quickly and efficiently complete the extraction of low- and high-order latent interactions.

6.2 Complexity analyses

To investigate the complexity of comparable methods, we demonstrate the indicators of parameter number, memory



(a) non-overlapping cell line split. (b) non-overlapping drug split. (c) random split.

Fig. 5 The visualized correlations between observed CDR values and predicted CDR values in terms of three train/test split methods

usage, and training time per batch of methods in Fig. 4, the batch size of them to be consistent.

As shown in Fig. 4, MF and MLR significantly reduce the number of parameters. Compared with MF and MLR, the inevitable traversal operation in the CNN layer leads to higher time complexity on DeepCDR and CDRScan. CDRScan-S has the most trainable parameters due to the reduction of convolutional units. The data mapping of Embedding increases the parameters of MvMo, but the dimensionality reduction it brings allows MvMo to fuse more data with low time and space complexity and achieve more accurate predictions.

6.3 Prediction analyses

Three groups of validation experiments on the proposed MvMo were visualized using scatter plots, see Fig. 5. The CDR values of new cell lines are visualized in Fig. 5(a). The CDR values of new drug are visualized in Fig. 5(b). The CDR values of known cell lines and known drugs are visualized in Fig. 5(c). Since the plots of 10-fold predictions make the presentation so long, we randomly chose one result from each group of 10-fold datasets.

When observing the PCC values in the three subfigures, they are very close. The proposed MvMo has an ability in overcoming view value missing problem. The PCC values show significant correlations between real values and predicted values, which reflects the high predictive accuracy of the MvMo method. A possible reason is that higher concentrations indicate poor performance. There are more ineffective responses than effective ones for all the datasets.

7 Conclusions

This paper focused on the cancer-drug-response prediction using multi-omics data in a multi-view framework, which was named MvMo. The response values on new cancers or new tumors were predicted, via the observations from several views on multi-omics data. Technically, the proposed MvMo represented those heterogeneous input data into equal-dimensionality embeddings or features, so the latent interaction between cell line and drug representations can be made. Experimental results on the real CDR datasets reveal the proposed MvMo in alleviating the view value missing problem and show the benefits of the proposed MvMo in terms of prediction accuracy as well.

In the future, we would like to investigate the detailed combination positions of drug molecules and proteins, which could provide the interpretability of cancer-drug reactions at the molecular level.

Acknowledgements This work was supported in part by Jimei University (no. ZP2021013), the Science project of Xiamen City (no. 3502Z20193048), the Education Department of Fujian Province (CN) (no. JAT200277), and the Natural Science Foundation of Fujian Province (CN) (no. 2021J01859).

We would like to thank the editor and anonymous reviewers for their helpful comments in improving the manuscript quality.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests None.

References

1. Adam G, Rampasek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A (2020) Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 4(1):19
2. Ali M, Aittokallio T (2019) Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev 11(1):31–39
3. Baptista D, Ferreira PG, Rocha M (2021) Deep learning for drug response prediction in cancer. Brief Bioinform 22(1):360–379
4. Basu A, Mitra R, Liu H, Schreiber SL, Clemons PA (2018) RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. Bioinformatics 34(19):3332–3339
5. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, Jung J, Shin JM (2018) Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep 8(1):8857
6. Chao G, Sun S (2018) Semi-supervised multi-view maximum entropy discrimination with expectation laplacian regularization. Inf Fusion 45(2):296–306
7. Chen J, Zhang L (2021) A survey and systematic assessment of computational methods for drug response prediction. Brief Bioinform 22(1):232–246
8. Cichonska A, Pahikkala T, Szedmak S, Julkunen H, Airola A, Heinonen M, Aittokallio T, Rousu J (2018) Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics 34:509–518
9. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Hintsanen P, Khan SA, Mpindi JP et al (2014) A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 32(12):1202–1212
10. Ammad-ud din M, Georgii E, Gonen M, Kallioniemi O, Wennerberg K, Poso A (2014) Integrative and personalized QSAR analysis in cancer by kernelized bayesian matrix factorization. J Chem Inf Model 54:35
11. Ammad-ud din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, Kaski S (2016) Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics 32(17):i455–i463
12. Ammad-ud din M, Khan SA, Wennerberg K, Aittokallio T (2017) Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics 33(14):i359–i368

13. Geeleher P, Cox NJ, Huang R (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol 15(3):R47
14. He X, Folkman L, Borgwardt K (2018) Kernelized rank learning for personalized drug recommendation. Bioinformatics 34(16):2808–2816
15. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations, San Diego, CA, USA. OpenReview.net
16. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, Wang J (2021) Deepdsc: A deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform 18(2):575–582
17. Liu Q, Hu Z, Jiang R, Zhou M (2020) DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 36(Supplement_2):911–918
18. Rahman R, Matlock K, Ghosh S, Pal R (2017) Heterogeneity aware random forest for drug sensitivity prediction. Sci Rep 7(1):11347
19. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, Fine HA (2010) Predicting in vitro drug sensitivity using random forests. Bioinformatics 27(2):220–224
20. Sharifi-Noghabi H, Zolotareva O, Collins C, Ester M (2019) Moli: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 35:501–509
21. Snow O, Sharifi-Noghabi H, Lu J, Zolotareva O, Lee M, Ester M (2021) Interpretable drug response prediction using a knowledge-based neural network. In: Proceedings of the 27th international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 3558–3568
22. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM (2016) Pathway-based genomics prediction using generalized elastic net. PLoS Comput Biol 12(3):1–23
23. Sotudian S, Paschalidis IC (2021) Machine learning for pharmacogenomics and personalized medicine: A ranking model for drug sensitivity prediction. IEEE/ACM Trans Comput Biol Bioinform:1–1
24. Suphavilai C, Bertrand D, Nagarajan N (2018) Predicting cancer drug response using a recommender system. Bioinformatics 34(22):3907–3914
25. Wang L, Li X, Zhang L, Gao Q (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 17(1):513
26. Wang Z, He L (2016) User identification for enhancing IP-TV recommendation. Knowl-Based Syst 98:68–75
27. Wang Z, Chen K, He L (2018) Asysim: modeling asymmetric social influence for rating prediction. Data Sci Pattern Recog 2:25–40
28. Ye Q, Hsieh CY, Yang Z, Kang Y, Chen J, Cao D, He S, Hou T (2021) A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun 12(1):6775
29. Zhang F, Wang M, Xi J, Yang J, Li A (2018) A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep 8(1):3355
30. Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS (2015) Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS Comput Biol 11(9):18
31. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc 67(2):301–320

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Zhijin Wang received the Ph.D. degree from the Department of Computer Science and Technology, East China Normal University, Shanghai, China, in 2016. He is currently with the Computer Engineering College, Jimei University, Xiamen, China. His current research interests include recommendation system, data mining, and artificial intelligence in health/healthcare.

Ziyang Wang received the B.S. degree from the College of Software, Taiyuan University of Technology, Taiyuan, China, in 2020. He is currently pursuing a Master's Degree at the Computer Engineering College, Jimei University, Xiamen, China. His current research interests include time series analysis and bioinformatics.

Yaohui Huang received the B.S. degree in computer science from Chengyi University College, Jimei University, Xiamen, China, in 2020, and is currently pursuing the master's degree with Guangxi University for Nationalities, Nanning, China. His research interests include data mining, time series data processing and analysis, deep learning and data fusion.

Longquan Lu received the B.S. degree from the College of Computer Engineering, Jimei University, Xiamen, China, in 2021. He is currently pursuing a Master's Degree at the School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China. His current research interests include time-series data analysis and bioinformatics.

Yonggang Fu received his bachelor and master degree from Xi'an Jiaotong University in Computational Mathematics in 1995 and 1998 respectively. He received his doctor degree in Computer Software and Theory in Shanghai Jiaotong University in 2005. Now he is a professor in Jimei University. His main research concentrated in multimedia security, and artificial intelligence.