Briefings in Bioinformatics, 23(1), 2022, 1–9

https://doi.org/10.1093/bib/bbab457
Problem Solving Protocol

GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction

Xuan Liu, Congzhi Song, Feng Huang, Haitao Fu, Wenjie Xiao and Wen Zhang

Corresponding author. Wen Zhang, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. E-mail: zhangwen@mail.hzau.edu.cn

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

Abstract

Predicting the response of a cancer cell line to a therapeutic drug is an important topic in modern oncology that can help personalized treatment for cancers. Although numerous machine learning methods have been developed for cancer drug response (CDR) prediction, integrating diverse information about cancer cell lines, drugs and their known responses still remains a great challenge. In this paper, we propose a graph neural network method with contrastive learning for CDR prediction. GraphCDR constructs a graph neural network based on multi-omics profiles of cancer cell lines, the chemical structure of drugs and known cancer cell line-drug responses for CDR prediction, while a contrastive learning task is presented as a regularizer within a multi-task learning paradigm to enhance the generalization ability. In the computational experiments, GraphCDR outperforms state-of-the-art methods under different experimental configurations, and the ablation study reveals the key components of GraphCDR: biological features, known cancer cell line-drug responses and contrastive learning are important for the high-accuracy CDR prediction. The experimental analyses imply the predictive power of GraphCDR and its potential value in guiding anti-cancer drug selection.

Key words: Cancer drug response prediction; Graph neural network; Contrastive learning; Multi-omics; Drug structure

Introduction

Cancer is one of the most intractable diseases that cause millions of deaths each year over the world. Drug discovery plays a crucial role in cancer therapy and precision medicine. Traditional methods of anti-cancer drug discovery are mainly based on *in vivo* animal experiments and *in vitro* drug screening, but these methods are expensive and laborious [1]. Recent advances in pharmacogenomics have developed several databases, such as Cancer Cell Line Encyclopedia (CCLE) [2] and Genomics of Drug Sensitivity in Cancer (GDSC) [3], which provide genome-wide data about cancer cell lines and drug responses against these cell lines. These valuable resources enable researchers to investigate the drug response mechanism in cancer therapy and have been extensively utilized to establish machine learning methods for cancer drug response (CDR) prediction.

Over decades, a number of machine learning methods have been proposed for CDR prediction. The matrix factorization-based (MF) methods reconstruct the known CDRs by the product of decomposed factors that are usually constrained by side information of cancer cell lines and drugs [4–6]. Network-based methods construct networks with bio-entities (cancer cell lines, drugs, etc.) and their associations, then formulate the

Xuan Liu is a PhD candidate in the College of Informatics at Huazhong Agricultural University.
Congzhi Song is a postgraduate in the College of Informatics at Huazhong Agricultural University.
Feng Huang is a PhD candidate in the College of Informatics at Huazhong Agricultural University.
Haitao Fu Haitao Fu is a PhD candidate in the College of Informatics at Huazhong Agricultural University.
Wenjie Xiao is a student in the Information School at University of Washington.
Wen Zhang is a professor in the College of Informatics at Huazhong Agricultural University.
Submitted: 10 July 2021; **Received (in revised form):** 25 September 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

original problem as a link prediction task solved by using random walk [7, 8], information flow [9, 10] or subset connection [11]. Classification-based methods learn representation vectors from cancer cell line-drug responses to train classifiers with different models, such as support vector machine [12], random forest [13, 14], logistic regression [15] and integration model [16]. For the CDR prediction, deep learning-based methods learn the latent representation of cancer cell lines and drugs from biochemical data through different network architectures, such as multi-layer perception [17, 18], convolutional neural networks [19, 20] and recurrent neural networks [21].

Although the previous methods have led to significant progress in CDR prediction, there is still room for improvement. In recent years, graph neural networks (GNNs [22]), which apply deep learning to graphs, have shown good performance in lots of bioinformatics problems [23–25], and also motivated us to develop GNN-based CDR prediction models. Since the annotated cancer cell line-drug responses are scarce [3], the generalization capability of models are restricted. Self-supervised learning has emerged as a powerful technique for generating pseudo-label data from the data itself to relieve the data scarcity. Contrastive learning is a class of self-supervised methods, which aims to learn discriminative representations by maximizing agreement/disagreement between the similar/dissimilar instances [26–28]. Moreover, previous works [29, 30] indicate that the biochemical information (e.g. multi-omics data of cancer cell lines and SMILES structure of drugs) is helpful for CDR prediction. Thus, incorporating the biochemical information into the GNN with contrastive learning can learn more meaningful representation and boost the performance of CDR prediction.

In this study, we propose a GNN method with contrastive learning, namely GraphCDR, for CDR prediction. GraphCDR constructs a GNN framework to integrate the biochemical information of cancer cell lines and drugs as well as their known responses. First, multi-omics representations of cancer cell lines learned via DNNs and the molecular graph representations of drugs learned via GNNs are taken as attributes of nodes in a CDR graph, which treats cancer cell lines, drugs as nodes and their sensitive responses as edges. Second, we employ a GNN encoder to learn the latent embedding of cancer cell lines and drugs from the CDR graph for prediction. Further, a contrastive learning task is designed to improve discriminative expressiveness of the GNN encoder and generalize the prediction, which contrasts the embeddings from the CDR graph and the graph constructed based on resistant responses. All contributions are summarized as follows:

- GraphCDR integrates the biochemical features of cancer cell lines and drugs as well as known cancer cell line-drug responses under a GNN framework, which leverages diverse information to boost the performance of CDR prediction.
- By taking the domain knowledge into account, a contrastive learning task is designed and incorporated into GraphCDR as a regularizer to enhance the generalization ability.
- In the absence of known responses, GraphCDR can also utilize the biochemical information of cancer cell lines/drugs for CDR prediction, which ensures inductive predictive capability when given new cell lines/drugs.

Datasets

Multi-omics data for cancer cell lines. CCLE [2] provides genomic, transcriptomic and epigenomic profiles for more than 1000 cancer cell lines. Following DeepCDR [20], we downloaded genomic mutation, gene expression and DNA methylation by using DeMap portal (https://depmap.org/). Specifically, 34 673 unique mutation positions within the related genes (697 genes from COSMIC Cancer Gene Census [31]) were collected as genomic mutation data. The gene expression data were obtained by the log-normalized TPM value of gene expression. The DNA methylation data were directly obtained from the processed Bisulfite sequencing data of promoter 1kb upstream TSS region. The three omics data (i.e. genomic, transcriptomic and epigenomic) of a cancer cell line can be represented as 34 673-dimensional, 697-dimensional and 808-dimensional feature vectors, respectively.

Molecular graph data for drugs. PubChem [32] provides validated chemical structure information for 19 million unique compounds. We downloaded the SMILES strings of all drugs from PubChem. By leveraging the ConvMolFeaturizer method in DeepChem library [33], the SMILES string of each drug can be compiled into a molecular graph where the nodes and edges denote chemical atoms and bonds, respectively. The attribute of each atom node in a drug is represented as a 75-dimensional feature vector, described in [34].

Cancer cell line-drug responses. In this study, we collected the IC_{50} values (natural log-transformed) from GDSC database for measuring responses between cancer cell lines and drugs. We binarized IC_{50} values according to the threshold of each drug provided by the reported maximum screening concentration [3]. Furthermore, we removed cell lines that lacked any type of omics data and drugs that shared the same Compound ID (CID) in PubChem. Finally, we compiled a dataset containing 11 591 sensitive responses and 88 981 resistant responses across 561 cell lines and 222 drugs. Among all the 561 × 222 = 124 542 responses, approximately 19.2% (23 970) of responses (i.e. IC_{50} values) were missing/unknown. Beyond that, we downloaded the activity area value of responses from CCLE database, and categorized the responses by setting a threshold (sensitive if the z-score normalized active area > 0.8; otherwise resistant) according to [35]. We finally created another dataset with 7307 responses (1375 sensitive ones and 5932 resistant ones) between 317 cell lines and 24 drugs.

Methods

GNN encoder

GNN is a feed-forward neural network specifically designed for directly processing graph to generate node representations. Given a graph $\mathcal{G} = (\boldsymbol{X}, \boldsymbol{A})$, where $\boldsymbol{X}$, $\boldsymbol{A}$ are, respectively, node attributes and the adjacency matrix, the GNN encoder Ω takes $\mathcal{G}$ as input and outputs node embeddings $\boldsymbol{H}$ by repeated aggregation over local node neighborhoods. Such neighborhood aggregation can be abstracted as:

$$\boldsymbol{h}_i^{(k)} = \phi^{(k)}\left(\boldsymbol{h}_i^{(k-1)}, f^{(k)}\left(\left\{\boldsymbol{h}_j^{(k-1)} : \forall j \in \mathcal{N}(i)\right\}\right)\right) \quad (1)$$

Here, $\boldsymbol{h}_i^{(k)} = \boldsymbol{H}^{(k)}[i, :]$ is the embedding for node i at the k-th layer and $\boldsymbol{H}^{(0)} = \boldsymbol{X}$; $\phi^{(k)}$ is a combination function, $\mathcal{N}(i)$ denotes a set of nodes adjacent to i in $\boldsymbol{A}$, and $f^{(k)}$ is an aggregation function.

GraphCDR

As shown in Figure1, the framework of GraphCDR includes three following modules: (i) the node representation module extracts

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

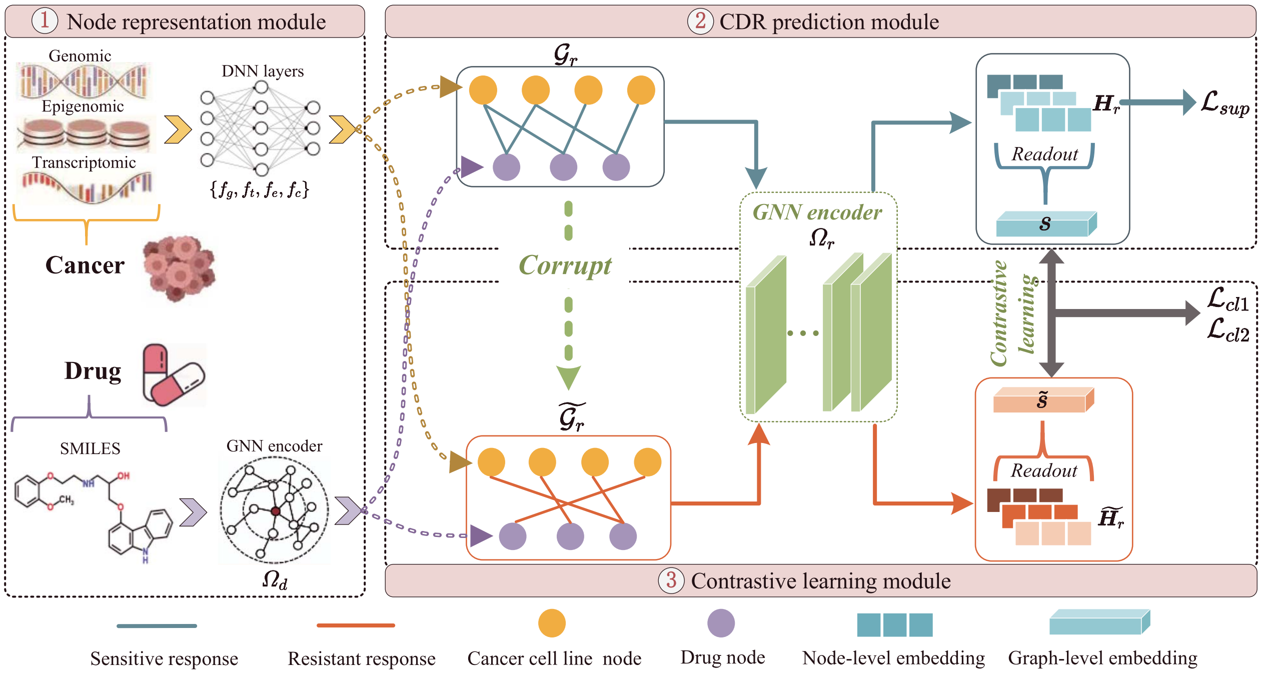

Figure 1. Overview of GraphCDR framework. ① Node representation module extracts representations from biochemical features of cancer cell lines/drugs via DNN layers $\{f_g, f_t, f_e, f_c\}$/GNN encoder Ω_d, then takes them as node attributes of the CDR graph $\mathcal{G}_r$ (cell lines and drugs represent nodes, and their sensitive responses represent edges). ② CDR prediction module employs a GNN encoder Ω_r to learn node-level embeddings $\mathbf{H}_r$ over $\mathcal{G}_r$ for the CDR prediction task. ③ Contrastive learning module learns corrupted node-level embeddings $\widetilde{\mathbf{H}_r}$ from a designed corrupted graph $\widetilde{\mathcal{G}_r}$ (using resistant responses) through Ω_r, then constructs a contrastive learning task to enhance the generalization ability of the prediction model by contrasting $\mathbf{H}_r$ and $\widetilde{\mathbf{H}_r}$.

cancer cell line/drug representations from the multi-omics profiles/the molecular graph via DNN layers/GNN encoder; (ii) the CDR prediction module formulates the known cancer cell line-drug responses as a CDR graph, and takes the preceding representations as the input attributes of nodes (i.e. cancer cell lines and drugs), and learns the latent embedding of nodes through a GNN encoder to predict novel CDRs; (iii) the contrastive learning module presents a contrastive learning task based on the CDR graph and its corrupted graph, and incorporate the task into GraphCDR as a regularizer.

Node representation module

Cancer cell line representation. Following previous study [20], omics-specific neural network layers are designed to integrate multi-omics information so as to obtain the representation for each cancer cell line. Here, we resort to the late-integration fashion in which each neural network layer will first learn a representation of a specific omics feature and then be concatenated together. Given a cancer cell line with its multi-omics features (i.e. a genomic feature vector $\mathbf{c}_g$, a transcriptomic feature vector $\mathbf{c}_t$ and an epigenomic feature vector $\mathbf{c}_e$), we can encode it into an F-dimensional representation by:

$$\mathbf{c} = f_c\left\{\left[f_g(\mathbf{c}_g)||f_t(\mathbf{c}_t)||f_e(\mathbf{c}_e)\right]\right\} \tag{2}$$

where, $\mathbf{c} \in \mathbb{R}^F$ is the representation of a cancer cell line, $||$ is a vector concatenation operator and $\{f_g, f_t, f_e, f_c\}$ are diverse neural network layers for feature transformation. Given a set of cancer cell lines $\mathcal{C} = \{c_i\}_{i=1}^{N_\mathcal{C}}$, we finally obtain representations $\mathbf{C} \in \mathbb{R}^{N_\mathcal{C} \times F}$ of all cancer cell lines for the follow-up modeling.

Drug representation. As described before, a drug can serve as a molecular graph where nodes represent atoms and edges are chemical bonds. We denote graph $\mathcal{G}_d = (\mathbf{X}_d, \mathbf{A}_d)$ as the molecular graph for drug d. $\mathbf{X}_d \in \mathbb{R}^{N_d \times F_d}$ is a matrix that records the attribute vectors ($F_d = 75$) of all atoms and $\mathbf{A}_d \in \mathbb{R}^{N_d \times N_d}$ is an adjacency matrix representing the bonds, where N_d is the number of atoms in the molecular graph of drug d. Here, we apply a GNN encoder Ω_d to capture the latent representation of atom nodes, denoted by $\mathbf{H}_d \in \mathbb{R}^{N_d \times F}$, where $\hat{h}_i = \mathbf{H}_d[i,:]$ is the latent representation of the node i. As different drug molecular graphs have different numbers of atom nodes, we apply a global max pooling (GMP) layer over all nodes to produce a summary representation of the entire graph as the representation for drug d: $\mathbf{d} \in \mathbb{R}^F \leftarrow \text{GMP}(\mathbf{H}_d)$. Given a set of drugs $\mathcal{D} = \{d_i\}_{i=1}^{N_\mathcal{D}}$, we utilize the GNN encoder Ω_d to obtain representations $\mathbf{D} \in \mathbb{R}^{N_\mathcal{D} \times F}$ of all drugs for the subsequent modeling.

CDR prediction module

The cancer cell line-drug responses can be formulated as an undirected heterogeneous graph (i.e. CDR graph) $\mathcal{G}_r = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes that contain two disjoint sets of entities (i.e. cancer cell lines $\mathcal{C}$ and drugs $\mathcal{D}$) and $|\mathcal{V}| = N_\mathcal{C} + N_\mathcal{D}$; $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes the set of edges representing cancer cell line-drug responses (i.e. sensitive responses). The goal of the CDR prediction is to learn a mapping function $\Theta(\omega) : \mathcal{E} \to [0, 1]$ from edges to scores, where ω is the learnable parameter of Θ, such that we can determine the probability of cancer cell line-drug pairs having sensitive responses. $\mathcal{G}_r$ can be further indicated by an adjacency matrix $\boldsymbol{A}_r \in \{1, 0\}^{|\mathcal{V}| \times |\mathcal{V}|}$ and node attributes $\mathbf{X}_r \in \mathbb{R}^{|\mathcal{V}| \times F}$, where $\mathbf{A}_r(c, d) = 1$ if cancer cell line c is sensitive to drug d and $\mathbf{A}_r(c, d) = 0$ otherwise, and $\mathbf{X}_r = \left[\begin{smallmatrix}\mathbf{C}\\ \mathbf{D}\end{smallmatrix}\right]$ (i.e. the cell line

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

representations **C** and drug representations ***D*** that are learned before, are taken as $\mathbf{X}_r$).

In this paper, we employ a GNN encoder Ω_r to the CDR graph $\mathcal{G}_r$, $\Omega_r : (\mathbf{X}_r, \boldsymbol{A}_r) \rightarrow \mathbf{H}_r \in \mathbb{R}^{|\mathcal{V}| \times F'}$, which learns latent embeddings of nodes. We denote $\vec{\boldsymbol{h}}_c = \mathbf{H}_r[c, :]$ and $\vec{\boldsymbol{h}}_d = \mathbf{H}_r[d, :]$, respectively, as final embeddings for cancer cell line node c and drug node d.

For the CDR prediction, we utilize the final embeddings of cancer cell line node c and drug node d to predict the probability of their sensitive response $\hat{p}_{cd}$ through a scoring function with the inner product:

$$\hat{p}_{cd} = \mathrm{Sigmoid}(\vec{\boldsymbol{h}}_c \vec{\boldsymbol{h}}_d^{\,T}) \tag{3}$$

Then the loss of supervised CDR prediction task can be formulated as:

$$\mathcal{L}_{sup} = -\frac{1}{|\mathcal{S}|} \sum_{(c,d) \in \mathcal{S}} \left(p_{cd} \log \hat{p}_{cd} + (1 - p_{cd}) \log (1 - \hat{p}_{cd})\right) \tag{4}$$

where $\mathcal{S}$ is the training set of responses and p_{cd} denotes true label for the response between nodes c and d.

Contrastive learning module

Inspired by deep graph infomax (DGI) [26], we present a contrastive learning task that contrasts embeddings from the CDR graph $\mathcal{G}_r$ and its corrupted graph $\widetilde{\mathcal{G}}_r$, to enhance the model's generalization ability. The procedure of the contrastive learning is as follows.

We construct a corrupted CDR graph $\widetilde{\mathcal{G}}_r = (\mathbf{X}_r, \widetilde{\mathbf{A}}_r)$ based on resistant responses between cancer cell lines and drugs (obtained from the training set $\mathcal{S}$). The reason behind this design is intuitive: resistant responses naturally imply opposite information against the sensitive responses, and hence it is believed that DGI can refine the embeddings learned from $\mathcal{G}_r$ via maximizing dissimilarities between them and their counterparts (the embeddings learned from $\widetilde{\mathcal{G}}_r$), thereby making the prediction model be more discriminative.

Following the vanilla DGI, we then obtain the corrupted node embeddings $\widetilde{\mathbf{H}_r}$ from the corrupted CDR graph through the same GNN encoder Ω_r: $(\mathbf{X}_r, \widetilde{\mathbf{A}_r}) \rightarrow \widetilde{\boldsymbol{H}_r} \in \mathbb{R}^{|\mathcal{V}| \times F'}$. The objective of our contrastive learning task is formulated as:

$$\mathcal{L}_{cl1} = -\frac{1}{2|\mathcal{V}|} \left(\sum_{v \in \mathcal{V}} \log \Gamma(\vec{\boldsymbol{h}}_v, \mathbf{s}) + \sum_{v \in \mathcal{V}} \log(1 - \Gamma(\widetilde{\vec{\boldsymbol{h}}}_v, \mathbf{s})) \right) \tag{5}$$

where **s** is the graph-level embedding obtained by a readout function, $R : \boldsymbol{H}_r \in \mathbb{R}^{|\mathcal{V}| \times F'} \rightarrow \mathbf{s} \in \mathbb{R}^{F'}$, and $\Gamma(\cdot, \cdot)$ is the contrastive discriminator constructed by a simple bilinear function $\sigma(\vec{\boldsymbol{h}}^{T} W\mathbf{s})$ that estimates similarities between the node-level embeddings and the graph-level embedding. W is a learnable scoring matrix and σ is the logistic sigmoid nonlinearity.

Finally, different from the vanilla DGI, we extend the contrastive learning mechanism from another perspective: maximizing disagreements between node-level embeddings $\boldsymbol{H}_r$ and the corrupted graph-level embedding $\widetilde{\mathbf{s}} = R(\widetilde{\boldsymbol{H}_r})$, which can be formulated as:

$$\mathcal{L}_{cl2} = -\frac{1}{2|\mathcal{V}|} \left(\sum_{v \in \mathcal{V}} \log \Gamma(\widetilde{\vec{\boldsymbol{h}}}_v, \widetilde{\mathbf{s}}) + \sum_{v \in \mathcal{V}} \log(1 - \Gamma(\vec{\boldsymbol{h}}_v, \widetilde{\mathbf{s}})) \right) \tag{6}$$

Optimization

To implement the CDR prediction task and the contrastive learning task simultaneously, we optimize the following objective function that combines Eq.4, Eq.5 and Eq.6:

$$\mathcal{L} = (1 - \alpha - \beta)\mathcal{L}_{\text{sup}} + \alpha\mathcal{L}_{\text{cl1}} + \beta\mathcal{L}_{\text{cl2}} \tag{7}$$

where α and β are hyper-parameters that balance the contributions of different tasks. The pseudo-codes of GraphCDR are illustrated in Algorithm 1.

Algorithm 1 GraphCDR

```
Input: multi-omics data of cancer cell lines $\mathcal{C}$; molecular graph of drugs $\mathcal{D}$; training set of responses $\mathcal{S}$; hyper-parameters $\alpha, \beta$.
Output: $\Theta(\omega)$
1: while GraphCDR not converge do
2:     /**** Cancer cell line representation
3:     for $c \in \mathcal{C}$ do
4:         Calculate $\mathbf{c}$ via Eq.2
5:     end for
6:     /**** Drug representation
7:     for $d \in \mathcal{D}$ do
8:         Calculate $\boldsymbol{H}_d \leftarrow \Omega_d(\boldsymbol{X}_d, \boldsymbol{A}_d)$
9:         $\boldsymbol{d} \leftarrow \mathrm{GMP}(\boldsymbol{H}_d)$
10:    end for
11:    /**** CDR prediction
12:    Initialize $\boldsymbol{X}_r$ via $\{\boldsymbol{c}_i\}_{i=1}^{|\mathcal{C}|}$ and $\{\boldsymbol{d}_i\}_{i=1}^{|\mathcal{D}|}$
13:    $\mathcal{G}_r \leftarrow (\boldsymbol{X}_r, \boldsymbol{A}_r)$
14:    Calculate $\boldsymbol{H}_r \leftarrow \Omega_r(\boldsymbol{X}_r, \boldsymbol{A}_r)$
15:    $\boldsymbol{s} \leftarrow R(\boldsymbol{H}_r)$
16:    for $(c, d) \in \mathcal{S}$ do
17:        Calculate $\hat{p}_{cd}$ via Eq.3
18:    end for
19:    Calculate $\mathcal{L}_{sup}$ via Eq.4
20:    /**** Contrastive learning
21:    $\widetilde{\mathcal{G}_r} \leftarrow (\boldsymbol{X}_r, \widetilde{\boldsymbol{A}_r})$
22:    Calculate $\widetilde{\boldsymbol{H}_r}, \widetilde{\boldsymbol{s}}$ like line 14-15
23:    Calculate $\mathcal{L}_{cl1}$ and $\mathcal{L}_{cl2}$ via Eq.5 and Eq.6
24:    $\nabla_\omega((1 - \alpha - \beta)\mathcal{L}_{\text{sup}} + \alpha\mathcal{L}_{cl1} + \beta\mathcal{L}_{cl2})$
25: end while
26: return $\Theta(\omega)$
```

Experiments

Model evaluation

In this study, we take the GDSC dataset as the main dataset to evaluate the performance of prediction models. We randomly select 90% all known cell line-drug responses from the GDSC dataset to compile the cross-validation set, and use the remaining 10% responses as the independent test set, ensuring no overlap between these two sets. The sensitive and resistant responses are taken as positive and negative samples, respectively. Then, we consider the following two experimental configurations.

- Cross-validation: the 5-fold cross-validation (5-CV) is implemented on the cross-validation set by randomly dividing responses into five equal parts. The hyper-parameters of GraphCDR are also set according to the 5-CV results and are then used for other experiments.
- Independent test: the prediction models are trained on the cross-validation set and tested on the independent test set. Furthermore, we also conduct an independent test on the

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

CCLE dataset, in which 90% and 10% responses are used as the train set and the test set, respectively.

We evaluate the experimental results using two metrics: the area under curve (AUC) and the area under the precision-recall (AUPR). Besides, accuracy and f1-score are also taken into account.

Experimental settings

In the node representation module for cancer cell line, we represent f_g, f_t and f_c by using three different fully connected layers. Considering the mutation positions are distributed linearly along the chromosome, we take a 1D convolutional layer as f_e. In the node representation module for drug, we employ the graph convolutional network (GCN) [22] layers as the GNN encoder Ω_d, therefore, the representation of atom node i can be updated by:

$$\hat{\boldsymbol{h}}_i^{(k_d)} = \mathrm{ReLU}^{(k_d)}\left(\sum_{j\in\mathcal{N}(i)\cup\{i\}} \frac{1}{\sqrt{q_i q_j}} \cdot \mathbf{w}_d^{(k_d)} \cdot \hat{\boldsymbol{h}}_j^{(k_d-1)}\right) \tag{8}$$

where $q_i = 1+|\mathcal{N}_i|$, $\mathbf{w}_d$ is the weight matrix parameter, and $k_d = 3$. The representation dimension of a cell line/drug is fixed to 100 ($F = 100$).

In the CDR prediction module, the GNN encoder Ω_r is set to a k_r-layer ($k_r = 1$) GCN with the PReLU [36] function, and the embedding of node $v \in \mathcal{V}$ can be formulated by:

$$\vec{\boldsymbol{h}}_v^{(k_r)} = \mathrm{PReLU}^{(k_r)}\left(\sum_{u\in\mathcal{N}(v)\cup\{v\}} \frac{1}{\sqrt{q_v q_u}} \cdot \mathbf{w}_r^{(k_r)} \cdot \vec{\boldsymbol{h}}_u^{(k_r-1)}\right) \tag{9}$$

The embedded dimension of node v is fixed to 256 ($F' = 256$). We design an attentive readout function as $R : \mathbf{s} = \sum_{v\in\mathcal{V}} a_v \cdot \vec{\boldsymbol{h}}_v^{(k_r)}$, a_v denotes the attention score of node v:

$$a_v = \frac{\exp\left(f_a([\vec{\boldsymbol{h}}_v^{(0)} \| \vec{\boldsymbol{h}}_v^{(k_r)}])\right)}{\sum_{v\in\mathcal{V}} \exp\left(f_a([\vec{\boldsymbol{h}}_v^{(0)} \| \vec{\boldsymbol{h}}_v^{(k_r)}])\right)} \tag{10}$$

where f_a is a fully connected layer for mapping embedding to a real number. In Eq.3, we concatenate cell line/drug node embeddings of different GCN layers as the final embedding: $\vec{\boldsymbol{h}}_c = [\vec{\boldsymbol{h}}_c^{(0)} || \vec{\boldsymbol{h}}_c^{(k_r)}]/\vec{\boldsymbol{h}}_d = [\vec{\boldsymbol{h}}_d^{(0)} || \vec{\boldsymbol{h}}_d^{(k_r)}]$. Furthermore, we employ Adam with a learning rate of 0.001 as the optimizer.

In addition to the above empirical settings, several hyper-parameters in GraphCDR need to be tuned: the coefficients α and β in the objective function. Here, we choose both α and β from $\{0.00, 0.05, ..., 0.50\}$ for the hyper-parameter optimization under the cross-validation on the GDSC dataset. According to the result of hyper-parameter optimization (Supplementary Table S1), we fix both α and β to 0.3 because they produced the best AUC and AUPR scores. More details are presented in Supplementary Table S2.

Method comparison

Table 1. Results of ablation study

Method	AUC	AUPR
GraphCDR	**0.8496**	**0.5237**
GraphCDR (w/o GO)	0.8430	0.5187
GraphCDR (w/o TO)	0.8437	0.5113
GraphCDR (w/o EO)	0.8466	0.5202
GraphCDR (w/o MG)	0.8317	0.4970
GraphCDR (w/o CL)	0.8428	0.5109
GraphCDR (w/o GNN)	0.7687	0.3719
GraphCDR (w CS)	0.8234	0.4436

We compare our method with the following baselines.

- **Random Forest** [13] is a random forest (RF)-based CDR predictor that utilizes oncogene mutational spectrum of cancer cell lines and fingerprint chemical descriptors of drugs.
- **HNMDRP** [10] constructs a heterogeneous network from multiple sub-networks related to drugs, proteins, cancer cell lines and targets, and then employs an information flow algorithm to predict CDRs.
- **NRL2DRP** [12] integrates cancer cell lines and drugs with the protein–protein interaction network, and learns the cancer cell line representations from the network, and then build support vector machine-based predictors for individual drugs.
- **tCNNs** [19] is a CNN-based CDR prediction method that uses SMILES sequences of drugs and genomic mutation data of cancer cell lines.
- **RefDNN** [18] is a deep neural network-based CDR prediction method that utilizes gene expression profiles of cancer cell lines and molecular structure similarity profiles of drugs.
- **DeepCDR** [20] is a hybrid GCN-based CDR predictor which integrates multi-omics profiles of cell lines and chemical structures of drugs.

The cross-validation results. We conduct the 5-CV to evaluate the performances of GraphCDR and baselines on the cross-validation set of GDSC, as described in Section Model evaluation. As shown in Figure2A, GraphCDR outperforms all the baselines, and exceeds two best baselines: DeepCDR and NRL2DRP by 2.02% and 3.27%, respectively, in AUC scores, and 4.39% and 8.25%, respectively, in AUPR scores. By using CCLE and GDSC annotation information, we group the cross-validation result of GraphCDR according to the tissue and target pathway types, and then calculate the AUC and AUPR scores of GraphCDR on each group for specific analysis. The results of 24 tissues and 23 target pathways are shown in Figure2B. The results show that GraphCDR performs differently on different tissues and different target pathways. GraphCDR achieves AUC scores higher than 0.8 on 11 tissues and 16 target pathways, respectively. The cross-validation results demonstrate the superiority of GraphCDR when performing on CDR prediction task.

The independent test results. We conduct the independent test to further assess the performances of prediction models on the GDSC and CCLE datasets, as described in Section Model evaluation. Since the hyper-parameter tuning and model training are irrelevant to the independent test set, the independent test can better measure generalization ability of GraphCDR to unseen data. As shown in Figure2C, GraphCDR achieves a higher AUC and AUPR scores of 0.8426 and 0.5148 than baselines on the GDSC dataset. On the CCLE dataset, the AUC and AUPR scores of GraphCDR are, respectively, 0.9563 and 0.8877, which are still superior to baselines (results are provided in Supplementary Table S3). The independent test results show the high generalization ability of GraphCDR.

The results of other metrics in the cross-validation and independent test are provided in Supplementary Tables S3–S5.

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

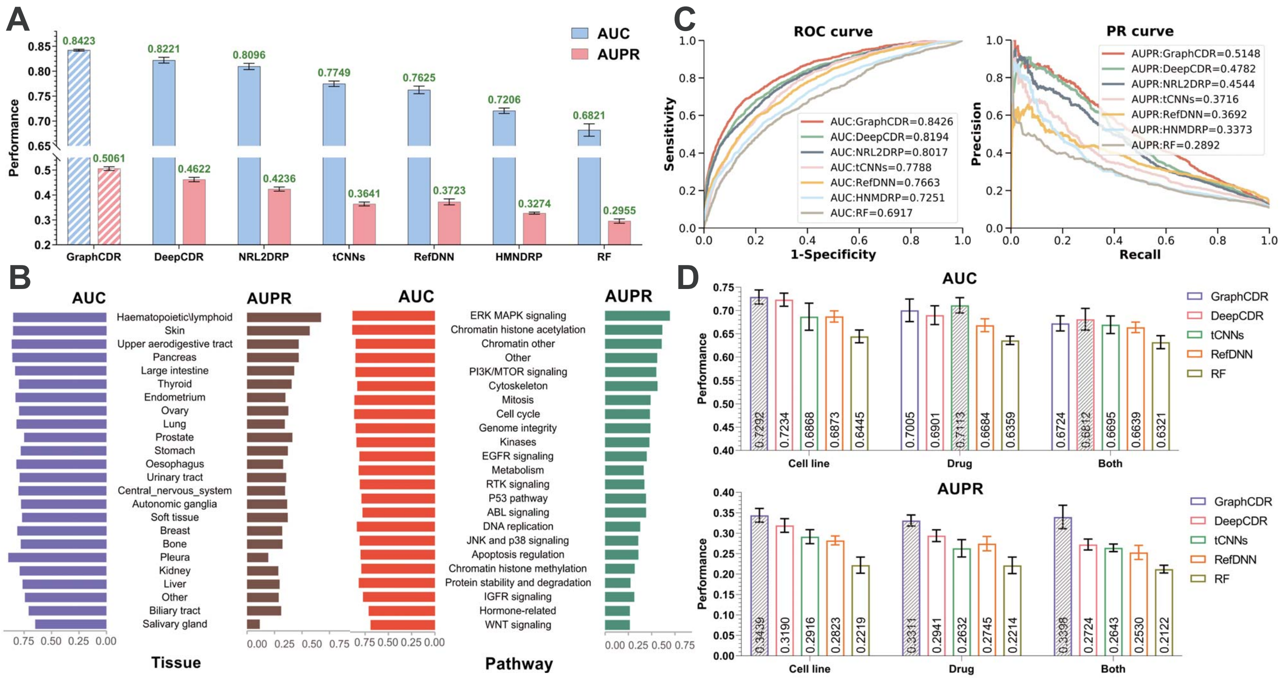

Figure 2. The performance of GraphCDR with different experimental configurations on the GDSC dataset. (A) The performance of GraphCDR and baselines on the cross-validation. (B) The performance of GraphCDR across 24 tissues (left) and 23 target pathways (right), respectively. (C) The receiver operating characteristic (ROC) and precision–recall (PR) curve of GraphCDR and baselines on the independent test. (D) The AUC (up) and AUPR (down) scores of GraphCDR and baselines in Inductive capability study.

Inductive capability study

To evaluate the inductive capability of GraphCDR on new cell lines/drugs, a challenging experiment is conducted on the GDSC dataset. We randomly split entities (cell lines/drugs/both types) into five equal parts. In each fold, one part of entities is used to simulate new ones. The prediction model is trained on the remaining four parts of entities and their related responses, and then makes predictions for the new ones.

We consider several baselines with inductive capability (i.e. DeepCDR, tCNNs, RefDNN and RF), and compare GraphCDR with them under three inductive configuration experiments (cell lines, drugs and both types), and the results of all models are shown in Figure2D. In the inductive capability study for cell lines, GraphCDR outperforms baselines by achieving a higher AUC score and AUPR score. In the inductive capability study for drugs, the performances of all methods drop slightly, because cell lines may share similar genetic information while chemical structures of drugs can be diversified, but GraphCDR also produces the best results. It is worth mentioning that the inductive capability study for both types (cell lines and drugs) is a more strict evaluation, and the performances of baselines decreased significantly, but GraphCDR still achieves comparable AUC scores and higher AUPR scores when compared with baselines.

The above studies show that GraphCDR outperforms the state-of-the-art inductive CDR prediction methods and has good adaptability in inductive learning.

Ablation study

To further investigate the importance of components, multi-omic data of cancer cell lines, molecular graph data of drugs, the GNN framework and the contrastive learning task, we design the following variants of GraphCDR:

- **GraphCDR without genomics** (w/o GO) removes the genomics omics data.
- **GraphCDR without transcriptomics** (w/o TO) removes the transcriptomics omics data.
- **GraphCDR without epigenomics** (w/o EO) removes the epigenomics omics data.
- **GraphCDR without molecular graph representation** (w/o MG) uses randomized drug representations instead of drug representations learned from molecular graphs.
- **GraphCDR without contrastive learning task** (w/o CL) removes the contrastive learning task.
- **GraphCDR without the GNN framework** (w/o GNN) removes the CDR graph and GNN encoder Ω_r, and directly uses cell line representations **C** and drug representations ***D*** through the inner product for prediction.
- **GraphCDR with the contrastive strategy from the vanilla DGI** (w CS) adopts the corrupted CDR graph with row-wise shuffled node attributes.

The ablation study is conducted on the GDSC dataset. We randomly split all responses into five equal parts to implement 5-CV, and the results are shown in Table 1. When removing omics data, the AUC scores of GraphCDR variants (w/o GO, w/o TO and w/o EO) range from 0.8430 to 0.8466, and the AUPR scores range from 0.5113 to 0.5202, indicating the usefulness of all individual omics profiles. The results of the variant (w/o MG) demonstrate that the prediction performance is significantly boosted with the drug molecular topology, compared to removing it. As expected, the performances of the variant (w/o CL) show that the contrastive learning task makes a contribution to the prediction. The experiment on the variant (w/o GNN) demonstrates that our GNN framework does enhance performance and performs better than directly using biochemical representations. Besides, our contrastive strategy is superior to the vanilla DGI (i.e. the

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

Table 2. Top 10 predicted cancer cell lines for two drugs

Drug	Rank	Cancer cell line	PMID
Dasatinib	1	EFM-192A	N/A
	2	TT2609-C02	N/A
	3	HSC-2	N/A
	4	**MCF7**	22306341
	5	8505C	N/A
	6	SNG-M	N/A
	7	SW-1710	N/A
	8	**786-0**	26984511
	9	**CAL-12T**	22649091
	10	LCLC-103H	N/A
GSK690693	1	GA-10	N/A
	2	HGC-27	N/A
	3	**RCH-ACV**	19064730
	4	IGROV1	N/A
	5	NCI-H929	N/A
	6	RH-18	N/A
	7	NCI-H1650	N/A
	8	**JeKo-1**	32120074
	9	**HCC202**	26181325
	10	**MOLT-16**	19064730

node attribute shuffling) according to the results of the variant (w CS).

In general, GraphCDR leverages the biological information, the GNN framework and the contrastive learning for the high-accuracy CDR prediction, while the removal of components will undermine the predictive capacity.

Case study

In this section, we conduct case studies to verify whether GraphCDR could predict novel cancer cell line-drug responses. We train the GraphCDR model with all known cancer cell line-drug responses in the GDSC dataset, and then predict novel ones (i.e. unknown responses in the GDSC dataset, described in Section Datasets). The prediction results are provided in Supplementary Table S6.

Here, we take two clinically approved drugs Dasatinib and GSK690693 for analysis. The top 10 cancer cell lines that have responses with two drugs predicted by GraphCDR are illustrated in Table 2. Three out of 10 predicted cell lines could be proved to be sensitive to Dasatinib. For breast cancer cell line MCF7, its response against Dasatinib was identified as moderately sensitive [37]. Dasatinib showed a significant dose-dependent influence on early non-metastatic 786-0 ccRCC cell lines, increasing their apoptosis while decreasing proliferation [38]. According to Sen et al.'s trials [39], they classified the response of lung cancer cell line CAL-12T to Dasatinib as sensitive. For GSK690693, four cell lines could be confirmed to be sensitive. GSK690693 restrained the proliferation of cells from both T-cell and B-/pre-B-cell origin within the ALL cell panel, with RCH-ACV and MOLT-16 found to be sensitive to GSK690693 [40]. Liu et al. [41] discovered that GSK690693 is effective in inhibiting the proliferation of MCL cell line JeKo-1. In Korkola et al.'s study [42], they measured responses for GSK690693 in breast cancer cell line HCC202, which was identified to be sensitive according to the threshold [3].

Therefore, the case studies demonstrate that GraphCDR could help to find out the novel cancer cell line-drug responses.

Conclusion

In this work, we present a GNN-based method with contrastive learning namely GraphCDR for CDR prediction. GraphCDR integrates multi-source information of bio-entities under a GNN frame. Moreover, by leveraging information derived from data itself, GraphCDR designs a contrastive learning task as a regularizer within a multi-task learning paradigm to boost the prediction performance. GraphCDR outperforms state-of-the-art CDR prediction models under various experimental configurations.

In future work, we provide two directions for improving CDR prediction: (i) Resistant responses have an inherently different semantic meaning as compared to sensitive responses, and it motivates us to employ the signed GCN model [43] on the signed graph (i.e. the CDR graph having both sensitive and resistant responses). (ii) We obtain the node embeddings from the CDR graph having only cell lines and drugs, and diverse biological association information (e.g. drug–target interactions, different similarity networks of bio-entities) has not been well exploited. Incorporating more associations for CDR prediction deserves consideration.

Key Points

- GraphCDR integrates the biochemical features of cancer cell lines and drugs as well as known cancer cell line-drug responses under a graph neural network framework, which leverages diverse information to boost the performance of CDR prediction.
- By taking the domain knowledge into account, a contrastive learning task is designed and incorporated into GraphCDR as a regularizer to enhance the generalization ability.
- In the absence of known responses, GraphCDR can also utilize the biochemical information of cancer cell lines/drugs for CDR prediction, which ensures inductive predictive capability when given new cell lines/drugs.

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank anonymous reviewers for valuable suggestions.

Funding

This work was supported by the National Natural Science Foundation of China (62072206, 61772381); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662021JC008). The funders have no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Data availability

The data sets and source code can be freely downloaded from https://github.com/BioMedicalBigDataMiningLab/GraphCDR.

References

1. Li K, Du Y, Li L, *et al*. Bioinformatics approaches for anti-cancer drug discovery. *Curr Drug Targets* 2020;**21**(1):3–17.
2. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**(7391):603–7.
3. Iorio F, Knijnenburg TA, Vis DJ, *et al*. A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**(3):740–54.
4. Ammad-Ud-Din M, Khan SA, Malani D, *et al*. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* 2016;**32**(17):i455–63.
5. Suphavilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. *Bioinformatics* 2018;**34**(22):3907–14.
6. Wang L, Li X, Zhang L, *et al*. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 2017;**17**(1):1–12.
7. Stanfield Z, Coşkun M, Koyutürk M. Drug response prediction as a link prediction problem. *Sci Rep* 2017;**7**(1):1–13.
8. Turki T, Wei Z. A link prediction approach to cancer drug sensitivity prediction. *BMC Syst Biol* 2017;**11**(5):1–14.
9. Zhang N, Wang H, Fang Y, *et al*. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol* 2015;**11**(9):e1004498.
10. Zhang F, Wang M, Xi J, *et al*. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**(1):1–9.
11. Meybodi FY, Eslahchi C. Predicting anti-cancer drug response by finding optimal subset of drugs. *Bioinformatics* 2021; btab466. doi: 10.1093/bioinformatics/btab466.
12. Yang J, Li A, Li Y, *et al*. A novel approach for drug response prediction in cancer cell lines via network representation learning. *Bioinformatics* 2019;**35**(9):1527–35.
13. Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS One* 2019;**14**(7):e0219774.
14. Su R, Liu X, Wei L, *et al*. Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 2019;**166**:91–102.
15. Yu L, Zhou D, Gao L, *et al*. Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods* 2021;**192**:85–92.
16. Gerdes H, Casado P, Dokal A, *et al*. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nat Commun* 2021;**12**(1):1–15.
17. Li M, Wang Y, Zheng R, *et al*. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**(2):575–582.
18. Choi J, Park S, Ahn J. RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci Rep* 2020;**10**(1):1–11.
19. Liu P, Li H, Li S, *et al*. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics* 2019;**20**(1):1–14.
20. Liu Q, Hu Z, Jiang R, *et al*. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;**36**(Supplement_2):i911–8.
21. Li Q, Huang J, Zhu H, *et al*. Prediction of Cancer Drug Effectiveness Based on Multi-Fusion Deep Learning Model. In: *Annual Computing and Communication Workshop and Conference (CCWC)*. Las Vegas, NV, USA: IEEE, 2020:0634–0639. doi: 10.1109/CCWC47524.2020.9031163
22. Xu K, Hu W, Leskovec J, *et al*. How powerful are graph neural networks? In: *International Conference on Learning Representations (ICLR)*. New Orleans, Louisiana, United States: OpenReview.net, 2019.
23. Li J, Zhang S, Liu T, *et al*. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020;**36**(8):2538–46.
24. Yu Z, Huang F, Zhao X, *et al*. Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics* 2021;**22**(4):bbaa243. doi: 10.1093/bib/bbaa243.
25. Nyamabo AK, Yu H, Shi JY. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics* 2021;bbab133. doi: 10.1093/bib/bbab133.
26. Velickovic P, Fedus W, Hamilton WL, *et al*. Deep Graph Infomax. In: *nternational Conference on Learning Representations (ICLR)*. New Orleans, Louisiana, United States: OpenReview.net, 2019.
27. Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning (ICML)*. Virtual Conference: PMLR, 2020;**119**:1597–607.
28. Qiu J, Chen Q, Dong Y, *et al*. Gcc: Graph contrastive coding for graph neural network pre-training. In: *International Conference on Knowledge Discovery & Data Mining (KDD)*. Virtual Event CA USA: ACM, 2020, 1150–60. doi: 10.1145/3394486.3403168.
29. Kearnes SM, McCloskey K, Berndl M, *et al*. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016;**30**(8):595–608.
30. Sharifi-Noghabi H, Zolotareva O, Collins CC, *et al*. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;**35**(14):i501–9.
31. Sondka Z, Bamford S, Cole CG, *et al*. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 2018;**18**(11):696–705.

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025

32. Kim S, Chen J, Cheng T, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**(D1):D1102–9.
33. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al*. Convolutional networks on graphs for learning molecular fingerprints In: *Conference on Neural Information Processing Systems (NeurIPS)*. Montreal Canada: Curran Associates, Inc., 2015;**2**:2224–2232.
34. Ramsundar B, Eastman P, Walters P, *et al*. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. Sebastopol, California: O'Reilly Media, Inc., 2019.
35. Dong Z, Zhang N, Li C, *et al*. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer* 2015;**15**(1): 1–12.
36. He K, Zhang X, Ren S, *et al*. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *International conference on computer vision (ICCV)*. Santiago, Chile: IEEE, 2015, 1026–34.
37. Park BJ, Whichard ZL, Corey SJ. Dasatinib synergizes with both cytotoxic and signal transduction inhibitors in heterogeneous breast cancer cell lines–lessons for design of combination targeted therapy. *Cancer Lett* 2012;**320**(1):104–10.
38. Roseweir AK, Qayyum T, Lim Z, *et al*. Nuclear expression of Lyn, a Src family kinase member, is associated with poor prognosis in renal cancer patients. *BMC Cancer* 2016;**16**(1):1–10.
39. Sen B, Peng S, Tang X, *et al*. Kinase impaired BRAF mutations confer lung cancer sensitivity to Dasatinib. *Sci Transl Med* 2012;**4**(136):136ra70. doi: 10.1126/scitranslmed.3003513.
40. Levy DS, Kahana JA, Kumar R. AKT inhibitor, GSK690693, induces growth inhibition and apoptosis in acute lymphoblastic leukemia cell lines. *Blood, The Journal of the American Society of Hematology* 2009;**113**(8):1723–9.
41. Liu Y, Zhang Z, Ran F, *et al*. Extensive investigation of benzylic N-containing substituents on the pyrrolopyrimidine skeleton as Akt inhibitors with potent anticancer activity. *Bioorg Chem* 2020;**97**:103671.
42. Korkola JE, Collisson EA, Heiser L, *et al*. Decoupling of the PI3K pathway via mutation necessitates combinatorial treatment in HER2+ breast cancer. *PLoS One* 2015;**10**(7):e0133219.
43. Derr T, Ma Y, Tang J. Signed graph convolutional networks. In: *International Conference on Data Mining (ICDM)*. IEEE, 2018, 929–34. doi: 10.1109/ICDM.2018.00113.

Downloaded from https://academic.oup.com/bib/article/23/1/bbab457/6415314 by guest on 04 August 2025