Decoding Drug Response With Structurized Gridding Map-Based Cell Representation

Jiayi Yin, Hanyu Zhang, Xiuna Sun, Nanxin You, Minjie Mou, Mingkun Lu, Ziqi Pan, Fengcheng Li, Honglin Li, Su Zeng, and Feng Zhu

***Abstract*—A thorough understanding of cell-line drug response mechanisms is crucial for drug development, repurposing, and resistance reversal. While targeted anticancer therapies have shown promise, not all cancers have well-established biomarkers to stratify drug response. Single-gene associations only explain a small fraction of the observed drug sensitivity, so a more comprehensive method is needed. However, while deep learning models have shown promise in predicting drug response in cell lines, they still face significant challenges when it comes to their application in clinical applications. Therefore, this study proposed a new strategy called DD-Response for cell-line drug response prediction. First, a limitation of narrow modeling horizons was overcome to expand the model training domain by integrating multiple datasets through source-specific label binarization. Second, a modified representation based on a two-dimensional structurized gridding map (SGM) was developed for cell lines & drugs, avoiding feature correlation neglect and potential information loss. Third, a dual-branch, multi-channel convolutional neural network-based model for pairwise response prediction was constructed, enabling accurate outcomes and improved exploration of underlying mechanisms. As a result, the DD-Response demonstrated superior performance, captured cell-line characteristic variations, and provided insights into key factors impacting cell-line drug response. In addition, DD-Response exhibited scalability in predicting clinical patient responses to drug therapy. Overall, because of DD-response's excellent ability to predict drug response and capture key molecules behind them, DD-response is expected to greatly facilitate drug discovery, repurposing, resistance reversal, and therapeutic optimization.**

***Index Terms*—Drug response prediction, cell lines representation, precision medicine, structurized gridding map.**

Received 6 October 2023; revised 26 November 2023; accepted 7 December 2023. Date of publication 13 December 2023; date of current version 7 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grants 82373790, 22220102001, 81872798, and U1909208, in part by the Natural Science Foundation of Zhejiang Province under Grant LR21H300001, in part by the National Key R&D Program of China under Grant 2022YFC3400501, in part by the Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Support Plan of China, in part by the Fundamental Research Fund for Central University under Grant 2018QNA7023, in part by "Double Top-Class" University Project under Grant181201∗194232101, in part by the Key R&D Program of Zhejiang Province under Grant 2020C03010, in part by the Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine), in part by the Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, in part by Alibaba Cloud, and in part by the Information Technology Center of Zhejiang University. *(Jiayi Yin and Hanyu Zhang are co-first authors.) (Corresponding authors: Feng Zhu; Su Zeng; Honglin Li.)*

Jiayi Yin, Hanyu Zhang, Xiuna Sun, Nanxin You, Minjie Mou, Mingkun Lu, Ziqi Pan, and Su Zeng are with the College of Pharmaceutical Sciences, Department of Clinical Pharmacy, The First Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Zhejiang 310027, China (e-mail: yinjiayi@zju.edu.cn; hanyu_zhang@zju.edu.cn; sunxiuna@zju.edu.cn; younanxin@zju.edu.cn; moumj@zju.edu.cn; mingkun@zju.edu.cn; panziqi@zju.edu.cn; zengsu@zju.edu.cn).

Fengcheng Li was with the College of Pharmaceutical Sciences, Department of Clinical Pharmacy, The First Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Zhejiang 310027, China. He is now with the Children's Hospital, Zhejiang University School of Medicine, Zhejiang University, Zhejiang 310027, China (e-mail: lifengcheng@zju.edu.cn).

Honglin Li is with the Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200050, China (e-mail: hlli@ecust.edu.cn).

Feng Zhu is with the College of Pharmaceutical Sciences, Department of Clinical Pharmacy, The First Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Zhejiang 310027, China, and also with the Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Zhejiang 310027, China (e-mail: zhufeng@zju.edu.cn).

Digital Object Identifier 10.1109/JBHI.2023.3342280

I. Introduction

A COMPREHENSIVE understanding of the mechanisms governing drug response is vital for drug discovery [1], drug repurposing [2], and drug resistance reversal [3]. Contemporary cancer treatment strategies often rely on low-toxicity anticancer drugs that specifically target proteins with abnormal expression or mutations [4], such as BCR-ABL fusion [5], EGFR mutations [6], HER2 overexpression [7], and KRAS mutations [8]. Unfortunately, not all types of cancer exhibit well-established biomarkers that are causally linked to drug response stratification [9], [10], and in most cases, the response of cancer to drugs is likely to depend on a combination of genomic variables [11]. In fact, single gene-drug associations can only account for a small fraction of the observed range of drug sensitivity across cancer for a given drug [12]. Therefore, considering only the simple relationship between drug targets or their mutation status is not sufficient for predicting the therapeutic efficacy of specific anticancer drugs [13]. Instead, early inference of drug response based on the pre-treatment cancer molecular profile and understanding the molecular mechanisms underlying drug response will greatly contribute to the development of precision medicine [1], [14], [15].

Recently, a certain amount of research has been reported on the prediction of drug cell line responses. At the early stage of the study, classical machine learning (ML) algorithms such as lasso regression [16], elastic net regression [17], support vector machines [18], and random forests [19] were adopted by many methods [20], [21]. With the rapid development of artificial intelligence (AI), deep learning (DL) algorithms have been widely used in contemporary model construction, and significant achievements have been made. For instance, many methods were presented based on deep neural networks, such as *RefDNN* [22], *DeepDRK* [23], and *Precily* [24]. Some methods were developed based on the graph neural network and the contrastive learning technique such as *DeepCDR* [25] and *GraphCDR* [26]. In addition, some methods integrated their model with cellular biology hierarchy, such as *DrugCell* [1] with its developed visible neural network.

However, despite the existence of numerous established methods, there remains a dearth of clear guidelines regarding their clinical implementation for drug therapy stratification [10], [27]. Several key concerns contribute to this issue: ***First***, limitations arise from the narrow horizon of the model, stemming from the single-source and unilateral modeling [28], [29], [30]. Due to inconsistent data annotations and standards from different sources [28], current methods, especially regression models, are hampered by multi-data source availability. The restricted training domain [24] and the overuse of unilateral architectures for opportunistic promotion [31] would reduce the functionality and representativeness of the model [29]. ***Second***, limitations arise from the defective representation of cell lines, stemming from the feature correlation neglect and potential information loss [31]. On the one hand, without considering the feature dependency, direct adoption of high-dimensional cell-line gene expression profiles in existing methods can lead to the "curse of dimensionality" and model overfitting [10], [22], [32]. On the other hand, rough engineering or over-simplification of the high-dimensional vectors by prior methods can result in information loss or imprecise characterization [23], [33], [34]. ***Last***, limitations arise from the low efficiency of knowledge extraction [35], stemming from the utilization of inappropriate models [36]. Most existing methods employed ordinary DL models which are unsatisfactory and incompetent in some aspects, and are usually optimized as "black boxes" that seek only predictive accuracy with little consideration of the underlying biological mechanisms [37]. This limitation severely restricts the identification of drug action mechanisms and hinders the development of precision medicine [1]. Therefore, there is an urgent need for novel drug response prediction models that accurately forecast outcomes and facilitate the exploration of the underlying mechanisms driving drug response.

In this study, a novel strategy, named DD-Response, for cell-line drug response prediction was therefore constructed. ***Firstly***, to overcome the limitations of narrow horizons, DD-Response harmonized multiple datasets by source-specific label binarization, enlarging the data space for modeling [38]. ***Secondly***, to avoid the defective representations of cell lines, DD-Response developed modified representations based on the created two-dimensional (2D) structurized gridding map (SGM), where the previous unordered feature vectors were rearranged into the organized grids for subsequent model learning [39]. ***Lastly***, to address the obstacles of inefficient knowledge extraction, DD-Response developed a dual-branch and multichannel convolutional neural network-based model for pairwise response prediction, which is considered to be highly capable of learning on image-like grid data [39]. Benefiting from the broad training domain, correlation-based SGM representations, and well-designed neural network, DD-Response achieved better cell-line drug response prediction performance compared with existing comparable methods. Furthermore, new insights into the cell-line characteristic variations from gene expression to drug response were also captured by DD-Response. DD-Response scored key factors affecting drug response in cell lines to gain an understanding of underlying mechanisms, and their involved functional pathways provided important clues to drug resistance reversal and repurposing. More importantly, DD-Response showed its scalability to predict the response of clinical patients to cancer drug therapy, which offers a valuable data reference for the rational utilization of drugs. All in all, DD-response is expected to greatly facilitate drug discovery, repurposing, resistance reversal, and therapeutic optimization. The source code and datasets of DD-Response are now available at https://github.com/idrblab/DD-Response.

II. Materials And Methods

A. Feature Encoding With Cancer Cell-Line Gene Expression and Drug Fingerprint

In this study, drug molecular fingerprint profiles and cell-line gene expression profiles were used as original representations, respectively. For drugs, their ISOSMILES were obtained from the PubChem database [40] via the open-source software RDkit (https://www.rdkit.org) for calculating molecular fingerprints (FPs), which consisted of 441 PharmacoErG FPs based on pharmacophore profiles [41], 881 PubChem FPs and 167 MACCS FPs based on key substructural features [42], [43]. For cell lines, 37,607 gene RNA-seq data (gene TPM values per million transcripts) from 1,432 cell lines were gathered from the GDSC, and log2 transformed to TPM values with pseudo-count 1, i.e., log2 (TPM+1) [44], [45]. Next, a comprehensive literature review of 37,607 genes was performed and combined with drug transporter and metabolizing enzyme data from VARIDT [46], INTEDE [47], and DrugMAP [48], which ultimately resulted in approximately 15,000 genes that were regarded as key features for cancer development and drug ADME. Then, the information on gene-related pathways was incorporated into this study, with a total of 9,497 genes having associated pathway information from the KEGG [49] and Reactome [50] databases. Ultimately, the expression profiles of the 9,497 genes were used as the original feature vectors of the cell lines.

Furthermore, optimal filtering was achievable by removing the lowest variance features [51]. The filtering of genes was guided by the variance of its expression across the 1,432 cell lines in GDSC. The filtering of FPs was guided by the variance of its manifestation across the 8,506,208 drug molecules in PubChem. The removal variance threshold value of < 0.0001 was set as the results-oriented selection, finally, 9,481 genes and 1,303 FPs

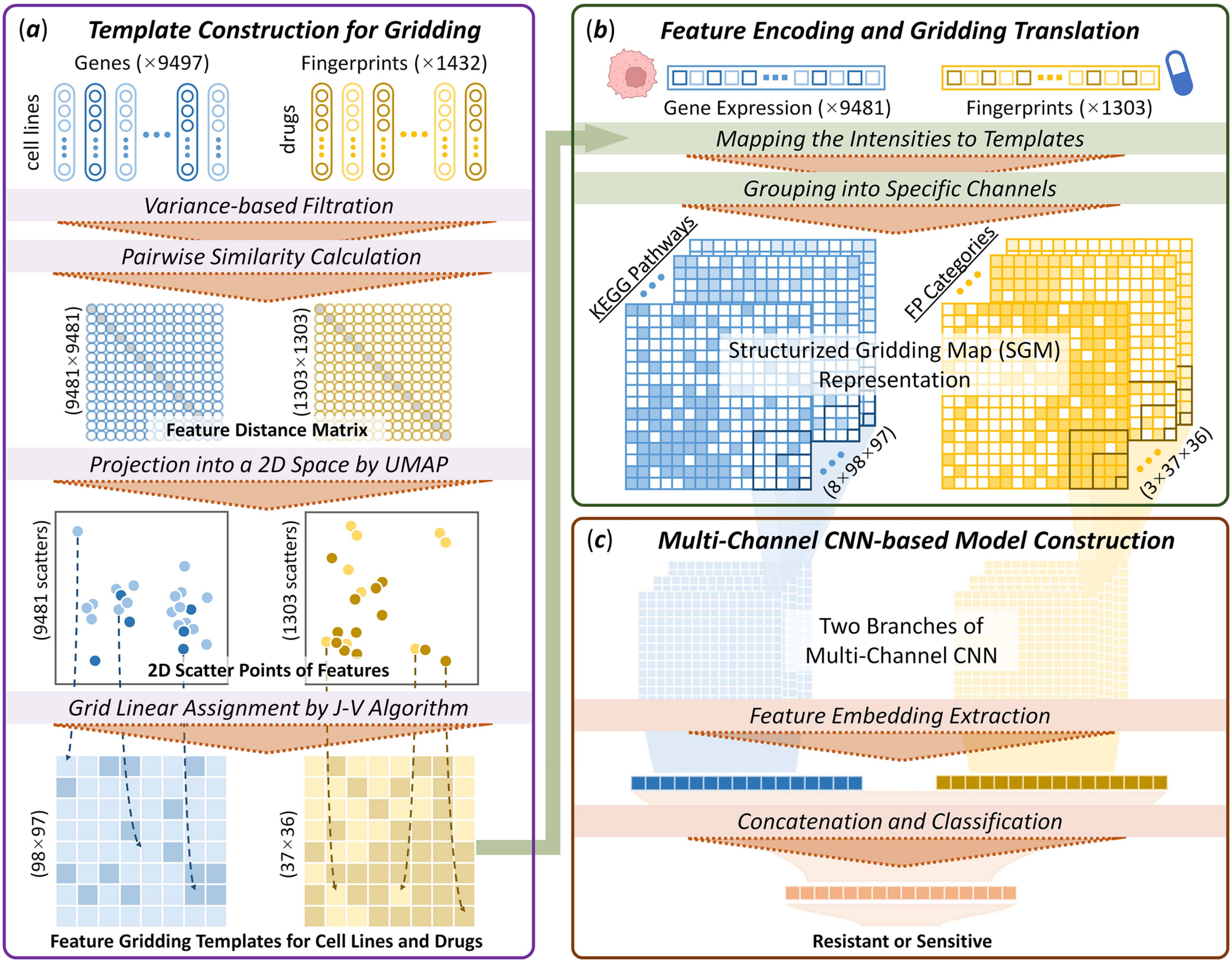
(a) Template Construction for Gridding
Genes (×9497)
Fingerprints (×1432)
cell lines
drugs
Variance-based Filtration
Pairwise Similarity Calculation
(9481×9481)
(1303×1303)
Feature Distance Matrix
Projection into a 2D Space by UMAP
(9481 scatters)
(1303 scatters)
2D Scatter Points of Features
Grid Linear Assignment by J-V Algorithm
(98×97)
(37×36)
Feature Gridding Templates for Cell Lines and Drugs
(b) Feature Encoding and Gridding Translation
Gene Expression (×9481)
Fingerprints (×1303)
Mapping the Intensities to Templates
Grouping into Specific Channels
KEGG Pathways
FP Categories
Structurized Gridding Map (SGM) Representation
(8×98×97)
(3×37×36)
(c) Multi-Channel CNN-based Model Construction
Two Branches of Multi-Channel CNN
Feature Embedding Extraction
Concatenation and Classification
Resistant or Sensitive

Fig. 1 Framework of DD-Response. (a) Template Construction for Gridding: Based on the pairwise similarity of features (represented by the distance matrix), the feature objects (genes for cell lines and FPs for drugs) were rearranged into a 2D grid using UMAP and J-V Algorithm step by step, to construct the templates for SGM representations of cell lines and drugs, respectively. (b) Feature Encoding and Gridding Translation: the feature value assignments and multilayer mapping transformations were performed for drug molecular fingerprints and cell line gene expression profiles, respectively. (c) Multi-Channel CNN-based Model Construction: a dual-branch and multi-channel convolutional neural network-based model, consisting of two CNN branches for feature embedding extraction and an FC-NN module for concatenating and classifying, was subsequently designed to learn the feature profiles of drugs and cell lines, respectively, and to predict the cell-line drug response patterns.

were used to constitute original feature vectors of cell lines and drugs.

B. Feature Correlation-Based Template Construction for Structurized Gridding

Based on the correlation of the feature manifestation in cell lines and in drugs (9,481×1,432 and 1,303×8,506,208 respectively), the pairwise distances among 9,481 genes and among 1,303 FPs were respectively calculated using the cosine correlation function (1), where fea_a and fea_b are both vectors of either genes (1,432 length) or FPs (8,506,208 length), then stored in the distance matrix of genes (9,481×9,481) and distance matrix of FPs (1,303×1,303) [39].

$$distance\left(fea_a, fea_b\right) = 1 - \frac{fea_a \cdot fea_b}{\|fea_a\| \times \|fea_b\|} \quad (1)$$

For either genes or FPs, each feature was projected into a 2D space as a scattered feature point based on their distance correlation. The embedding of these scatters was found by the closest possible equivalent fuzzy topological structure searching using Uniform Manifold Approximation and Projection (UMAP) [52]. These scatters were further assigned to a grid map, and structurized by minimizing the cost-squared distance matrix between feature scatters and feature grids (2). This linear assignment was implemented using the Jonker–Volgenant (J–V) algorithm [53]. Ultimately, the constructed grid maps for cell lines and drugs maintained the broadly learned correlation relationships and then were regarded as templates for original feature vector rearrangement (Fig. 1(a)).

$$distance\left(U^{scatter}, G^{grid}\right) = \left\|U^{scatter} - G^{grid}\right\|^2 \quad (2)$$

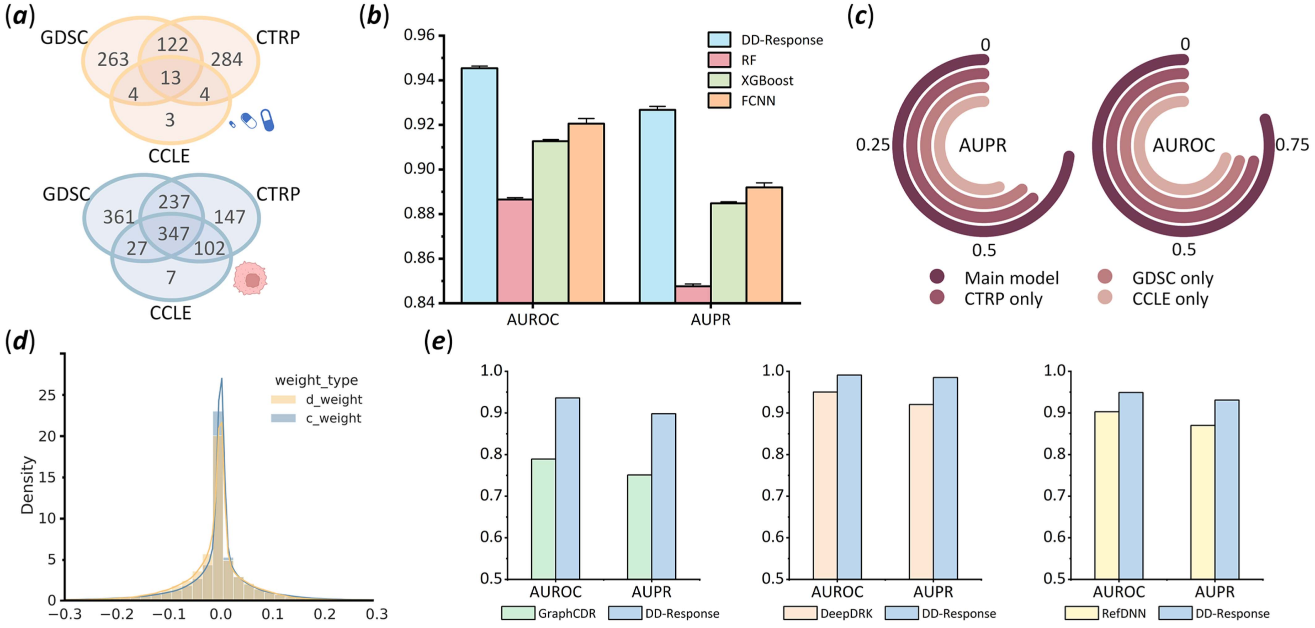

Fig. 2 Performance evaluation of DD-Response. (a) Cell lines and drugs from different data sources. Orange circles: interlacement of drugs; Blue circles: interlacement of cell lines. (b) The AUROC and AUPR testing results of DD-Response, RF, XGBoost, and FCNN models, DD-Response significantly outperformed the other classic ML models. Blue bar: DD-Response; Red bar: RF; Green bar: XGBoost; Orange bar: FCNN. (c) The performance of the DD-Response main model (Main model, colored in darker purple) and single-source trained models (CTRP only, GDSC only, and CCLE only, colored in three lighter purples in order) on the independent gCSI testing dataset. (d) The density distribution of neural network weights for processing cell-line (colored in blue) and drug (colored in orange) latent vectors. (e) Comparison with three existing methods, DD-Response outperformed the competing method. Green bar: GraphCDR; Orange bar: DeepDRK; Yellow bar: RefDNN; Blue bars: DD-Response.

As described above, a newly structurized gridding map (SGM) representation for each cell line and drug could be translated by mapping the intensities of all features to their corresponding locations in the relevant template (98×97 for cell lines and 37×36 for drugs). Moreover, additional grouping was employed to further enhance the representation of cell lines and drugs. For the cell lines, their SGM representations were grouped into 8 different channels based on the KEGG-defined biological pathway of the genes [49]. These categories included cellular processes, developmental biology, drug development, environmental information processing, genetic information processing, human diseases, metabolism, and organismal systems. For the drugs, their SGM representations were grouped into 3 channels based on the FP sources and properties. In summary, for the input of the model, each cell line would be represented by an 8×98×97 feature map, and each drug would be represented by a 3×37×36 feature map (Fig. 1(b)).

C. Preprocessing of Cell-Line Drug Response Labels From Diverse Data Sources

The fusion dataset utilized in this study was sourced from three prominent databases, namely Cancer Therapeutics Response Portal (CTRP) [54], [55], Genomics of Drug Sensitivity in Cancer (GDSC) [44], [45], and Cancer Cell Line Encyclopedia (CCLE) [38]. The information on drug high-throughput screening and cell line gene expression levels was extracted from the public websites of these databases. To elaborate, the GDSC dataset provided half maximal inhibitory concentration (IC50) values and area under the dose-response curve (AUC) values for 978 cancer cell lines and 543 drugs. The CCLE dataset contributed IC50 values and activity area values for 504 cancer cell lines and 24 drugs. As for the CTRP dataset, AUC values were obtained for 860 cancer cell lines and 547 drugs. Since the inconsistencies across the databases hindered comprehensive data analysis, this study initially normalized the compounds and cell lines across data sources. Compounds were standardized using Compound IDs (CID) from the PubChem database, while cell lines were standardized using cell line IDs from the GDSC database [45]. Ultimately, a total of 689 drugs and 1,227 cancer cell lines were included in the subsequent analysis. (as shown in Fig. 2(a) and Table I).

To standardize response labels across different data sources, the *Waterfall* method was employed to categorize the response values of drug-related cell lines, thereby determining the sensitivity (1) or resistance (0) [38]. The *Waterfall* method follows a sequential procedure, initially arranging the AUC or IC50 values of the drug against the cell lines in descending order to generate a curve. This curve represents the AUC/IC50 values on the Y-axis and the corresponding cell lines on the X-axis. To establish a cutoff value for the AUC/IC50 values, two distinct strategies were employed: Firstly, for linear curves exhibiting a Pearson correlation coefficient greater than 0.95, the median AUC/IC50 value across the cell lines was chosen as the cutoff point distinguishing sensitive from resistant AUC/IC50 values. Secondly, for non-linear curves with the Pearson correlation coefficient lower than 0.95, the cutoff point was determined based on the AUC/IC50 value of a specific boundary data point.

TABLE I
FUSION DATASET COLLECTION AND PREPROCESSING RESULTS

	No. of drugs	No. of drugs shared	No. of cell lines	No. of cell lines shared	No. of drug-cell line response	No. of sensitive data after binarization	No. of resistance data after binarization
GDSC	398	139	972	611	129 215	54 127	75 088
CCLE	24	21	483	476	6551	1041	5510
CTRP	423	139	832	686	74 317	28 166	46 151
Total	689	-	1227	-	210 083	83 334	126 749

This data point was selected as the maximum distance from the line connecting the highest and lowest AUC/IC50 values. The *Waterfall* method was initially proposed in a notable study focusing on the CCLE in 2012 and has since been adopted in various subsequent studies [38], [56]. Its efficacy in segregating data about drug sensitivity and resistance has been demonstrated [57].

D. Neural Network Model Construction, Training, Evaluation, Comparison, and Implementation

The DD-Response was composed of a multi-layer and multi-channel convolutional neural network (CNN) branches module for feature extraction and another fully connected layers module for concatenation and classification (Fig. 1(c)) [58], [59]. Cell lines and drugs were first learned by the two CNN branches. Both branches consisted of eight layers. The first layer was a convolutional layer (72 of 17×17 sized convolutional kernels for cell lines, and 48 of 9×9 sized kernels for drugs), with a stride of 1. The second layer was a max-pooling layer (72 of 5×5 sized pooling kernels for cell lines, and 48 of 3×3 sized pooling kernels for drugs), with a stride of 2. The third layer was a queue of three parallel convolutional layers (48 of 9×9, 5×5, and 1×1 sized convolutional kernels for cell lines, and 32 of 5×5, 3×3, and 1×1 sized kernels for drugs), with a stride of 1. The fourth layer separately combined the three-feature embedding obtained in parallel, and subsequently put into a max-pooling layer of the same size as the second layer for downsampling. The sixth layer was a queue of three parallel convolutional layers (96 of 9×9, 5×5, and 1×1 sized kernels for cell lines, and 64 of 5×5, 3×3, and 1×1 sized kernels for drugs), with a stride of 1. After completing an analogous combination in the seventh layer, the eighth layer performed a global max-pooling to obtain the final embedding. As a result, a 98 × 97 cell-line SGM representation and a 37 × 36 drug SGM representation would be learned to turn into a 288-length latent vector and a 192-length latent vector, respectively. After that, the cell-line latent vector and the drug latent vector were concatenated, and three fully connected layers (FC-layers) with 512, 256, and 64 neurons were stacked to learn the cell-line drug pairs and implement response prediction. Except that the Softmax function was used in the final discriminant layer, the ReLU function was used as the activation function [60].

For model training, the fusion dataset was split randomly, with 10% reserved for testing and the remaining 90% used as the training dataset. Then the stratified 5-fold cross-validation (CV) was employed for hyperparameter optimization while training. In stratified 5-fold CV, the training datasets are divided into five groups with the distribution of positive and negative samples as close as possible. In each CV round, one group is treated as the validation set in rotation, while the other four are collected as the training set [60]. As a result, the learning rate of 0.0005, batch size of 128, drop out of 0.0005, weight decay of 0.0005, early-stop steps of 16, and optimizer of Adam, were decided for model training in this study.

Model performance was evaluated according to the average values of area under the receiver operating characteristic curve (AUROC) and precision-recall curve (AUPR) on the prepared testing dataset among 5-fold CV models [60]. For further assessment, an independent testing dataset with 398 samples was collected from gCSI [61] after clearing away the duplicates already in the fusion dataset and binarizing.

DD-Response was further analyzed by comparing it with different classic machine-learning models and other existing methods. Here, three classic ML models, including Random Forrest (RF), Extreme Gradient Boosting (XGBoost), and a three-layer fully connected Neural Network (FCNN), were trained from scratch based on the same fusion dataset, feature engineering, hyperparameter optimization, and training strategy. As for the comparison with other existing methods, DD-Response took turns competing with the methods and used the dataset established by the rivals for model training. This is due to limitations in the availability of multi-omics data, where many cell lines in the fusion dataset cannot be not correctly encoded and do not fulfill the requirements of other existing methods for complete annotation of fusion datasets. The competing methods mainly included *GraphCDR* [26], *DeepDRK* [23], and *RefDNN* [22]. They all employed classification models and were published after 2020, which should be fully reproduced. In each paired comparison, the DD-Response model was trained from scratch based on the same strategy, while the model for comparison was trained from scratch based on the strategy reported in its original publication.

The whole framework was developed mainly based on Pytorch 1.8.1 and employed on the platform with Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz, NVIDIA(R) Tesla(R) P100 16GB GPU and 263GB RAM on CentOS Linux release 7.9.2009 (Core). All the scripts were written in Python 3.6.8.

E. Visualization of Characteristic Variation Before and After Model Learning

To reflect the characteristics of cell lines and drugs that the model had learned, their relative distribution variation before and after infusing the model with cell-line drug response information was compared. Based on UMAP, cell lines with

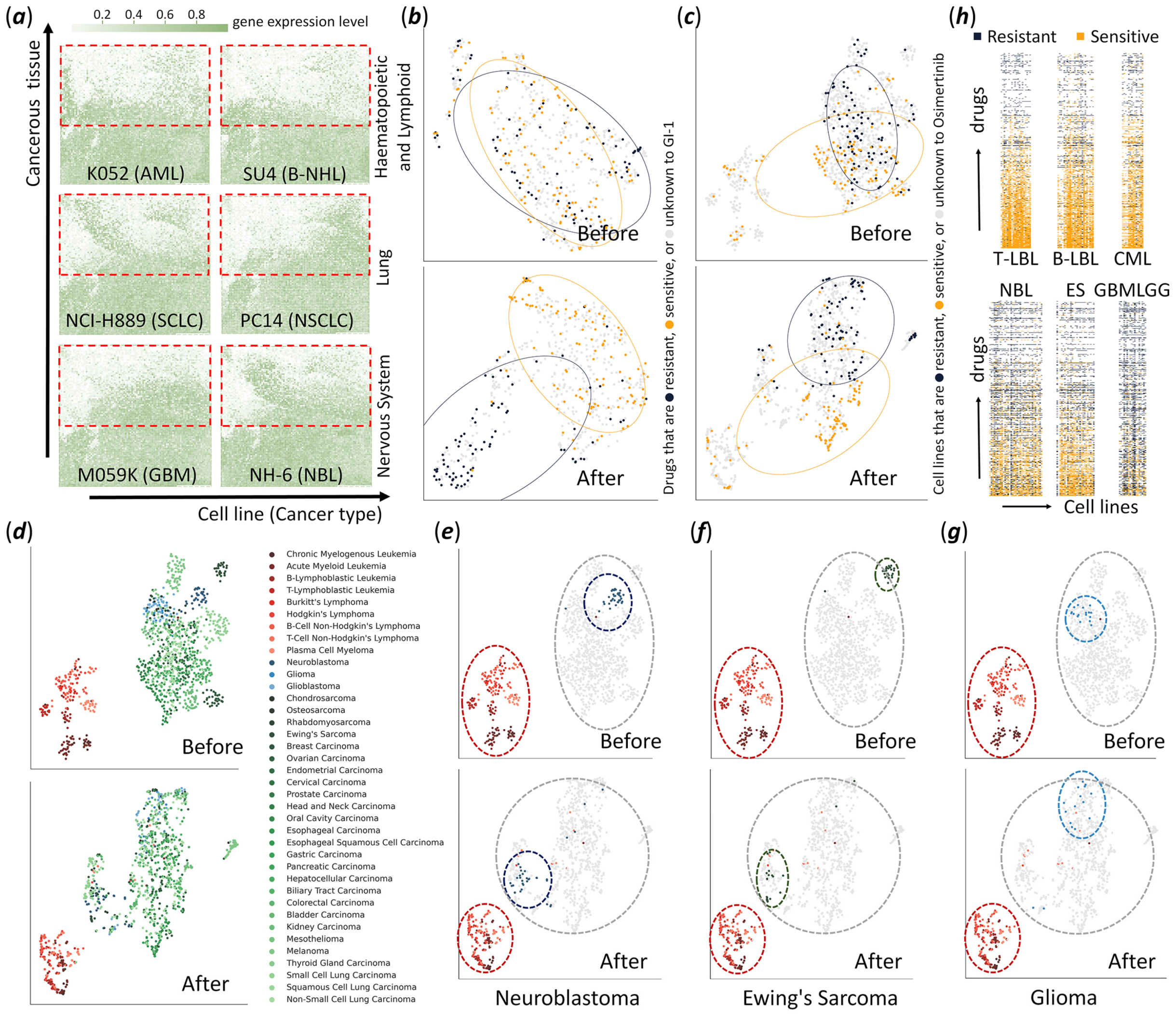

Fig. 3 Cell-line characterization and variations captured by DD-response. (a) The SGM representations of six different cancer cell lines (K052, SU4, NCI-H889, PC14, M059K, NH-6). Each grid in SGM represented a gene, and the intensity represented its expression. Specific clustered DEGs were framed by the red box. It could be observed that DEG clustering varied for different cancerous tissues and cancer types. (b) Drugs that were projected in a 2D space, each scatter represented a drug, and their relative distribution before and after model learning (upper panel and lower panel respectively). The sensitive and resistant drugs for the GI-1 cell line were more significantly separated after infusing cell-line drug response information. Black ones: resistant drugs to GI-1; Yellow ones: sensitive drugs to GI-1; Gray ones: drugs with unknown response to GI-1. (c) Cell lines that were projected in a 2D space, each scatter represented a cell line, and their relative distribution before and after model learning (upper panel and lower panel respectively). The sensitive and resistant cell lines for Osimertinib were more significantly separated after infusing cell-line drug response information. Black ones: resistant cell lines to Osimertinib; Yellow ones: sensitive cell lines to Osimertinib; Gray ones: cell lines with unknown response to Osimertinib. (d) The relative distribution of cancer types in cell lines before and after model learning (upper panel and lower panel respectively). Each scatter represented a cell line. (e)–(g) The relative distribution of Neuroblastoma (NBL), Ewing's sarcoma (ES), and Glioma (GBMLGG) cell lines before and after model learning. Red section: hematologic malignancy cell lines; Dark blue: Neuroblastoma cell lines; Green: Ewing's sarcoma cell lines; azure blue: Glioma cell lines; Gray: other solid tumor cell lines. (h) The responses of diverse drugs to T-cell lymphoblastic lymphoma (T-LBL), B-cell lymphoblastic lymphoma (B-LBL), chronic myelogenous leukemia (CML), NBL, ES, and GBMLGG cell lines.

original gene expression vectors and post-learned cell lines with latent embedding vectors were reduced to two dimensions and projected into two two-dimensional spaces, respectively, and a similar manipulation was carried out for the drugs with original fingerprint vectors and drugs with latent embedding vectors after learning (as shown in Fig. 3(b)–(g)). The distribution of their projection still preserved the high-dimensional global structure and therefore the relative positions of cell lines in the projected space could reveal their correlation [62].

F. Key Gene Identification With Integrated Gradients and Functional Enrichment

To identify the key genes (KGs) that mediate cell-line drug response, the *Integrated-Gradient* (IG) score was applied to measure the impact of each gene in model reasoning [57]. For any individual gene$_{ij}$ on the feature map, its IG score was defined by the integral of gradients along the path from the baseline to the input expression level [63]. The integral could be approximated using the Riemann rule and be calculated through (3), where

$IG_{ij}(x)$ value measured the importance of ij-th gene expression of the cell line in the current input x, x'_{ij} was the baseline expression level of ij-th gene (zero as a default), α was the scaling coefficient, $\partial F(x)/\partial x_{ij}$ was gradient of $F(x)$ along the ij-th dimension. Deployment of IG score calculation was realized mainly based on *Captum* python library (https://captum.ai/).

$$IG_{ij}\left(x\right) ::= \left(x_{ij} - x_{ij}^{'}\right) \times \int_{\alpha\,=\,0}^{1} \frac{\partial F\left(x' + \alpha \times \left(x - x'\right)\right)}{\partial x_{ij}} d\alpha \tag{3}$$

As a binary classification system, the model $F(x)$ outputted two types of response probabilities (resistance or sensitivity) for one input. Therefore, each gene obtained the IG score for resistance and the IG score for sensitivity, in anticipation of revealing the factors for resistant response and sensitive response respectively.

Based on the IG scores from all input samples, genes observed the following rules were selected as the KGs for a specific drug's resistance or sensitivity: (i) the gene's IG scores in specific drug correlative samples are significantly higher than its IG scores in other drugs correlative samples, which was manifested by Mann-Whitney U-test between the two groups (p-value $<$ 0.05) [64]; (ii) Median value of the gene's IG scores in specific drug correlative samples ranked top 200.

In addition, *Gene Set Enrichment Analysis* (GSEA) was performed on these KGs using the DAVID functional annotation tool [65] to further explore the biological patterns of cell line responses to drugs.

III. Results And Discussion

A. The Framework of DD-Response for Response Prediction Between Cell Lines and Drugs

DD-Response was designed to capture the determinants of cell-line drug response by considering cellular bioinformatics and medicinal chemistry factors. After considering factors such as data availability, modeling feasibility, and information effectiveness, the expression levels of a total of 9,481 genes, which are highly correlated with cancer development as well as drug transport and metabolism [46], [47], [48], were applied as the original representation of cancer cell lines (see **Methods A** for details). In terms of drug chemistry characterization, a total of 1,303 molecular features were used as original representations of the molecules [39], including PharmacoErG fingerprints (FPs) based on pharmacological characteristics [41], as well as PubChem FPs and MACCS FPs based on sub-structural features [66] (see **Methods A** for details). These original representations were provided with comprehensive biological information on cell lines and drug molecules.

To efficiently investigate high-dimensional original feature vectors while avoiding the loss of information that may result from rough engineering or oversimplification, a modified representation based on a two-dimensional (2D) structurized gridding map (SGM) was developed by DD-Response for both cell lines and drugs (as depicted in Fig. 1(a) and (b)). The method first rearranged high-dimensional feature vectors (genes for cell lines and FPs for drugs) into a grid map by considering the correlation between individual features. Subsequently, SGM templates were constructed based on the feature location in the prepared grid map (Fig. 1(a), see **Methods B** for details) to translate original representations into modified representations. Following this, feature values were assigned to the SGM according to the corresponding template, generating the modified two-dimensional SGM representations for model inputs, with dimensions of 98×97 for cell lines and 37×36 for drugs (as shown in Fig. 1(b), and see **Methods B** for details). Moreover, incorporating prior biological knowledge guidance, the cell-line 2D SGM representation was grouped into 8 layers based on KEGG-defined biological pathways [49], and the drug 2D SGM representation was grouped into 3 layers based on diverse molecular properties [39] to enhance the multi-frame or multiclass information density [67] (as shown in Fig. 1(b), and see **Methods B** for details). This novel two-dimensional structurized gridding representation approach reconstructed the intrinsic correlation between the features, providing valuable insights into the associations between different molecular features and gene expression.

With the aim of accurate cell-line drug response prediction and conducting mechanism research, a dual-branch and multi-channel convolutional neural network (CNN) based deep learning model was constructed [58], [68]. Specifically, the model was composed of a feature embedding extraction module as well as a concatenation and classification module, as illustrated in Fig. 1(c). Specifically, in the former module, SGM representations of cell lines and drugs were fed into two multi-channel CNN branches respectively. Then the latent vectors of both outputs were concatenated and passed through the latter module with fully connected layers (FC-layers), for learning and predicting the response patterns of cell-lines drug pairs. The architecture of the model was detailed in **Methods D**. These well-designed methods were designed to improve the cell-line drug response prediction, making DD-Response an important tool for drug discovery and development.

B. Credible Response Prediction Through Broad-Sighted and Unbiased Deep Learning

As current methods are restricted by their limited training domain, an intuitionistic solution is to expand the data space for cell-line and drug modeling [56]. The complementary strengths and interconnections in extant datasets provide the potential for valuable cross-cutting studies and comprehensive analyses, where the total value of the data is greater than the sum of its parts [28]. Therefore, DD-Response performed source-specific binarization for cell-line drug response data from GDSC, CTRP, and CCLE, respectively, as well as integration and model training based on labels (see **Methods C** for details, data statistics are shown in Table I). In this case, the amount of modeled cell lines and drugs was expanded by at least 25% compared with single-source data, incorporating additional information and data points substantially (as shown in Fig. 2(a)). The beforehand modeled additional cell lines and drugs could enable the model to perform accurate inferences of unknown responses on a larger scale [69]. The fusion dataset from several sources was used to train the model during 5-fold cross-validation (CV) (see **Methods D** for details). As a result, DD-Response achieved an average AUROC of 0.945 and AUPR of 0.927 at testing, which significantly outperformed Random Forrest (RF), Extreme Gradient Boosting (XGBoost), and fully-connected Neural Network (FCNN) (as shown in Fig. 2(b)).

In addition, an independent test was conducted on a cleaned dataset from Genentech Cell Line Screening Initiative(gCSI) [61] to compare the main model with models trained solely on data from one source (see **Methods D** for details). As a result, the main model showed significant advantages in AUROC and AUPR by at least 15% and 20%, respectively (as depicted in Fig. 2(c)). These results indicate that the larger training domain by fusing a wide range of data sources helps enhance the model's representativeness and generalization ability.

Moreover, it was crucial for the DD-Response model to equally consider both cell-line and drug information to reliably predict pairwise responses. Upon closer examination of the first FC-layer that implements pairwise embedding inference after latent vector concatenation [70], it was observed that there were no notable disparities in the distribution of model parameters for processing cell lines and drugs (as depicted in Fig. 2(d)). This finding suggested that the model took into account information from both cell lines and drugs in an unbiased manner when making response predictions [70]. Furthermore, DD-Response was compared with other existing cell-line drug response classification methods, which were fully reproducible and published after 2020. All models were trained from scratch during a 5-fold CV using identical datasets in each paired comparison (see **Methods D** for details). Notably, DD-Response outperformed the competing method by achieving a performance advantage of at least 5% in AUROC and AUPR on the test set, as demonstrated in Fig. 2(e). All the compelling results highlight the strong representative capacity of DD- Response for cell lines and drugs, as well as its unbiased and high-performing pairwise response prediction.

C. New Insights By Characterizing Cell-Line Variation From Gene Expression to Drug Response

The existing methods for predicting drug response to cancer cell lines usually depend on transcriptomic data for cell-line characterization [13], [24]. However, a naive application of omics data might lead to inaccurate results due to high-dimensional catastrophe and potential information loss [31], [34]. Thus, DD-Response rearranged the originally unordered gene expression profile into a two-dimensional structurized gridding map (SGM) based on expression dependence and correlation, achieving an organized layout and precise characterization of cell lines [39]. This superiority could be firstly reflected in the specific clustering of differentially expressed genes (DEGs) for specific-type cell lines, presented in its SGM representation (as illustrated in Fig. 3(a)). Meanwhile, the DEG clustering of different cell lines had intuitive variations from the perspectives of both 'cancerous tissue' and 'cancer type' (as shown in Fig. 3(a)), which could visualize the differences and similarities between cell lines at the gene expression level. On the other hand, the highly correlated gene clusters could increase the information density and help the CNN model to achieve more efficient regional feature extraction [71].

Moreover, it is important to note that the variations in the gene expression of cell lines are not necessarily equivalent to the variations in their response to drugs [72]. This implied that for accurate cell-line drug response prediction, deep learning models need to be able to identify and analyze the cell-line characteristic variation from gene expression to drug response. Thus, DD-Response incorporated potential patterns into the model by learning from diverse cell-line drug responses. By projecting the cell lines into 2D spaces respectively based on their original gene expression profiles and their output latent vectors of CNN branch (see **Methods E** for details), relative distribution variation of cell lines could be observed before and after infusing the model with response data. For example, the trained model effectively discriminated between the sensitive and resistant cell lines for Osimertinib, while it was improbable to tell the difference based simply on original gene expression (as shown in Fig. 3(c)). In addition, a similar phenomenon could also be observed from the drug perspective: the trained model was able to effectively discriminate between sensitive and resistant drugs to the GI-1 cell line (as shown in Fig. 3(b)).

In addition, the relative distribution variation of cancer types in cell lines further reflected the drug-response characteristics of different cancers. Cell lines from the same cancer type clustered together in terms of original gene expression, while distinct boundaries were observed between different cancer types, as shown in the upper panel of Fig. 3(d). Nevertheless, the infusing of drug-response information decreased the distinction between cancer types and shifted relative relationships, while their inherent differences persisted. For example, hematologic malignancies (cell lines represented in various shades of red) still had different distributions than solid tumors (cell lines represented in various shades of green), as shown in the lower panel of Fig. 3(d). Specifically by way of instance, cell lines of Neuroblastoma (NBL) and Ewing's sarcoma (ES) showed a stronger correlation with hematologic malignancies than other solid tumors (represented in grey) after being infused with drug-response information (as illustrated in Fig. 3(e) and (f)).

Meanwhile, cell lines of certain cancer types such as Glioma (GBMLGG) exhibited an opposite trend (as shown in Fig. 3(g)). By presenting drug responses in cell lines from different cancer types, the knowledge revealed by the model could correspond to the ground truth. As shown in Fig. 3(h), the responses of diverse drugs in NBL and ES cell lines resembled those in hematologic malignancies cell lines, such as T-cell lymphoblastic lymphoma (T-LBL), B-cell lymphoblastic lymphoma (B-LBL) and chronic myelogenous leukemia (CML), while GBMLGG displayed differently. In summary, DD-Response proved to be able to capture underlying feature patterns associated with drug response, which highlighted its immense potential as a reliable tool with an accurate understanding of the cell-line drug response.

D. Revealing Biological Mechanisms of Cell Lines Involving Their Response to Drugs

DD-Response has proven to achieve accurate cell-line drug response prediction and capture cell-line characteristic variation

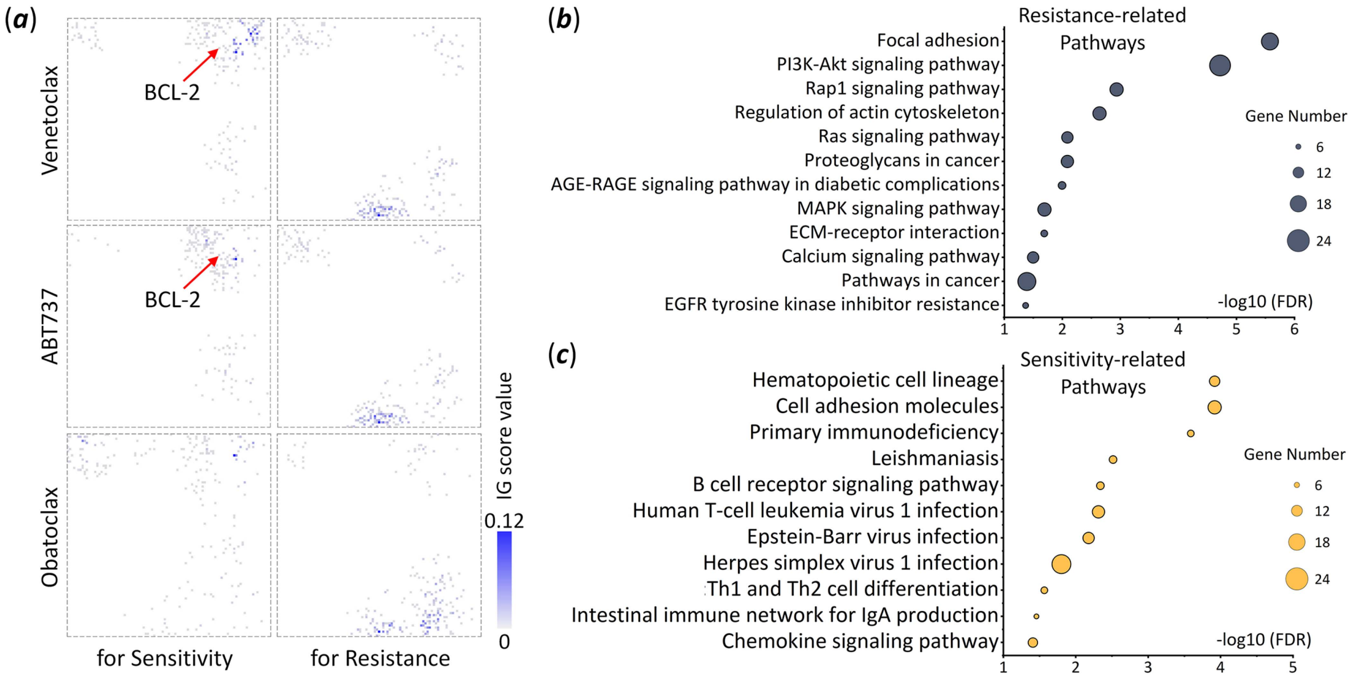

Fig. 4 (a) The KGs that were identified by DD-Response were shown in the SGMs for Venetoclax, ABT737, and Obatoclax. The intensity of each KG grid reflects its IG score value. (b)–(c) The KGs-based functional enrichment analysis revealed resistance-related and sensitivity-related pathways for Venetoclax. The horizontal coordinate represents the enrichment degree, and the bubble size represents the number of genes enriched.

from gene expression to drug response. Meanwhile, an in-depth understanding of the key genes (KGs) that induce sensitivity or resistance to cell lines is vital for drug resistance reversal, optimization, and repurposing. Thus, the *Integrated-Gradient* (IG) score for resistance and sensitivity was employed to identify KGs, which measured the gene impact on the model's perception of a sample as either sensitive or resistant [57]. In this study, three representative BCL-2 inhibitors: Venetoclax, ABT737, and Obatoclax were analyzed for their response-related KGs. For a specific drug, the influence of one gene was measured by its IG scores in all drug response samples. Based on the IG scores from all input samples, 200 KGs were selected for a specific drug's resistance or sensitivity based on concrete rules described in **Methods F**.

Specifically, the KGs identified for the three drugs were visualized in the cell-line SGMs based on the gene's median IG scores, as shown in Fig. 4(a). The visualization revealed that the KGs formed specific clustering in the SGMs for different drugs and different IG score types. The KG clustering region for Venetoclax and ABT737 were found to be quite similar (as shown in Fig. 4(a)), suggesting a potential similarity in their mechanisms of action. However, the KG clustering region for Obatoclax was significantly different from the other two drugs (as shown in Fig. 4(a)), indicating a distinct mechanism of action. Interestingly, the analysis also revealed that the gene BCL2, known to be strongly correlated with the sensitivity of Venetoclax and ABT737, was not found to be strongly correlated with that of Obatoclax. In fact, BCL2 did not appear among the 200 KGs for Obatoclax sensitivity. This result was supported by a previous study [13], which suggested that factors independent of high BCL2 expression may determine the sensitivity to Obatoclax [73]. Taken together, these findings indicated that different BCL-2 inhibitors may have differences in their mechanisms of action [74]. Moreover, the researchers also observed that the clustering region of KGs for sensitivity and resistance differed significantly, which suggested that focusing solely on factors inducing drug resistance may not be sufficient for drug reversal. It was crucial to also consider the factors that sensitize cells to the drug. All these findings emphasized the great importance of a comprehensive understanding of the mechanisms underlying drug response to enable the development of effective drug reversal strategies.

Moreover, Gene Set Enrichment Analysis (GSEA) was performed on the identified KGs for Venetoclax resistance and sensitivity, respectively. The pathways with enrichment degree (FDR) < 0.05 were demonstrated in Fig. 4(b) and (c) (see **Methods F** for details). For Venetoclax resistance-related pathways, in addition to some common pathways leading to cancer progression (e.g., Focal adhesion), DD-Response also identified pathways such as PI3K/AKT, MAPK, Ras, Calcium, Rap1, *etc.* (as shown in Fig. 4(b)). Most of these enriched pathways have been confirmed to affect Venetoclax resistance. For example, a study has utilized untargeted metabolomics to identify a significant effect of the PI3K/AKT pathway on Venetoclax resistance [75], and other studies have indicated that the activation of MAPK signaling pathway and RAS/MAPK pathway could lead to acquired resistance to Venetoclax [76], [77]. This implied that these identified pathways might be involved in the mechanism of Venetoclax resistance, and provided valuable insights into potential targets for resistance reversal. Among the Venetoclax sensitivity-related pathways, hematoma-related pathways appeared more frequently, such as B cell receptor signaling pathway [78], cell adhesion molecules [79], and Human T-cell leukemia virus 1 infection [80], which coincided with the scenario that this drug was mainly adapted for malignant lymphoid tumors (as shown in Fig. 4(c)). Some studies have demonstrated synergistic antitumor effects of Venetoclax in combination with drugs targeting these pathways [81]. More importantly, some of those identified pathways associated with different diseases suggest additional applications of this drug. For example, Venetoclax has been used in Epstein-Barr virus infection according to a recent report [82]. In conclusion, the ability of DD-Response to capture key factors influencing drug response in cell lines contributed to understanding the potential mechanisms and functional pathways of drug response, providing important clues for drug resistance reversal and repurposing.

E. Scalability of DD-Response for Drug Response Inference in Clinical Patients

Different from the traditional "one-size-fits-all" treatment of cancer drugs, precision medicine adopts more accurate targeted therapies, which predict the effects of drugs based on the characteristics of the patient's genetic information and provide more targeted and effective treatments [83], [84]. DD-Response achieved accurate prediction of cell-line drug response, precise characterization of drugs and cell lines, as well as effective identification of key biomarkers and signaling pathways. It was expected to conduct anticancer drug response prediction in patients, thus expanding the prospect of its clinical application.

Specifically, by applying Smriti Chawla's methodology to clean TCGA data [24], a total of 3,027 patient-drug combinations with recorded clinical responses were used for modeling. These patient-drug combinations involved 1,405 different patients and 133 unique drugs covering 29 cancer types, and raw genomic expression data of patients were obtained from the NCI Genomic Data Commons portal [85]. Following consistent procedures illustrated in **MATERIALS AND METHODS B-D**, an SGM representation template of the patient was constructed for mapping the intensity of all features, and the model for drug response inference in clinical patients was trained using TCGA data. As a result, it achieved a great performance with an average AUROC of 0.916 and AUPR of 0.936 on the testing set (as shown in Fig. 5), which was better than the other method based on the same dataset [24]. In conclusion, the precise characterization and effective feature extraction enabled DD-Response to accurately predict the clinical patients' response to cancer drug therapy. Therefore, DD-Response can be applied in the cancer drug treatment of clinical patients, promoting the development of personalized medicine.

IV. Conclusion and Perspectives

Here, DD-Response, a novel strategy consisting of the two-dimensional structurized gridding map-based representation and dual-branch multi-channel CNN-based DL model, was constructed for precise cell-line and drug characterization as well as accurate pairwise response prediction. Specifically, DD-Response excelled in cell-line drug response prediction through broad-sighted and unbiased deep learning. Secondly, benefiting from correlation-based SGM representation, new insights into cell lines from gene expression to drug response were captured by DD-Response, which allowed for a more comprehensive cognition of cell-line characteristics. Thirdly, the key factors that impact cell-line drug response were scored by DD-Response to reveal underlying mechanisms and their functional pathways, offering important clues for drug resistance reversal. Lastly, the DD-Response framework also demonstrated its ability to predict the response of clinical patients to cancer drug therapy.

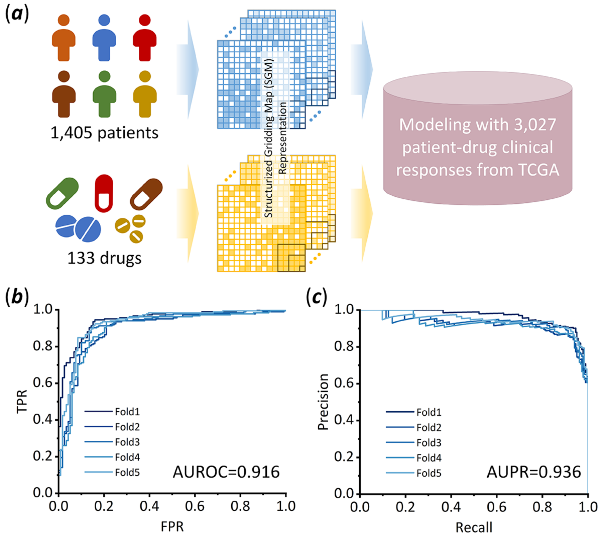

Fig. 5 Application of the DD-Response strategy to drug response prediction in TCGA patients. (a) The model construction was carried out by reconstructing patient's SGM representation template and applying DD-Response modeling strategy with the collected TCGA data. (b) The receiver operating characteristic curve (ROC) of the model on the testing set, with an average AUROC of 0.916. (c) The precision-recall curve (PR) of the model on the testing set, with an average AUPR of 0.936.

All in all, with the superior ability to predict drug responses and capture the key facts behind them, DD-response is expected to bring advancements in drug discovery, repurposing, resistance reversal, and therapeutic optimization. Specifically, by comprehensively analyzing the captured key facts involved in drug response, DD-response enables researchers to identify potential drug targets, design more effective therapies, and discover new strategies to overcome drug resistance. In addition, *in silico* drug response screening also saves time and resources by helping to repurpose existing drugs for new indications. DD-response's ability to predict how different patient groups will respond to a drug opens up opportunities for personalized medicine and patient-specific treatments. In essence, DD-response can hold great promise for accelerating the development of new therapies, improving patient prognosis, and advancing precision medicine. It will be constructed into open-service software to facilitate the use of researchers and clinical healthcare professionals. Future enhancements to DD-Response will be selected for integrating multi-modal data to further improve the drug response prediction for patients.

References

[1] B. M. Kuenzi et al., "Predicting drug response and synergy using a deep learning model of human cancer cells," *Cancer Cell*, vol. 38, no. 5, pp. 672–684, Nov. 2020.

[2] Z. Wu, P. J. Lawrence, A. Ma, J. Zhu, D. Xu, and Q. Ma, "Single-cell techniques and deep learning in predicting drug response," *Trends Pharmacological Sci.*, vol. 41, no. 12, pp. 1050–1065, Dec. 2020.

[3] W. Zhang et al., "Computational drug discovery for castration-resistant prostate cancers through in vitro drug response modeling," *Proc. Nat. Acad. Sci.*, vol. 120, no. 17, Apr. 2023, Art. no. e2218522120.

[4] P. L. Bedard, D. M. Hyman, M. S. Davids, and L. L. Siu, "Small molecules, big impact: 20 years of targeted therapy in oncology," *Lancet*, vol. 395, no. 10229, pp. 1078–1088, Mar. 2020.

[5] T. P. Braun, C. A. Eide, and B. J. Druker, "Response and resistance to BCR-ABL1-targeted therapies," *Cancer Cell*, vol. 37, no. 4, pp. 530–542, Apr. 2020.

[6] A. J. Cooper, L. V. Sequist, and J. J. Lin, "Third-generation EGFR and ALK inhibitors: Mechanisms of resistance and management," *Nature Rev. Clin. Oncol.*, vol. 19, no. 8, pp. 499–514, Aug. 2022.

[7] D. Y. Oh and Y. J. Bang, "HER2-targeted therapies - a role beyond breast cancer," *Nature Rev. Clin. Oncol.*, vol. 17, no. 1, pp. 33–48, Jan. 2020.

[8] L. Huang, Z. Guo, F. Wang, and L. Fu, "KRAS mutation: From undruggable to druggable in cancer," *Signal Transduct. Target. Ther.*, vol. 6, no. 1, Nov. 2021, Art. no. 386.

[9] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," *Nature Rev. Clin. Oncol.*, vol. 15, no. 2, pp. 81–94, Feb. 2018.

[10] G. Adam, L. Rampasek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: Challenges and recent progress," *NPJ Precis. Oncol.*, vol. 4, no. 1, 2020, Art. no. 19.

[11] E. Guney, J. Menche, M. Vidal, and A. L. Barabasi, "Network-based in silico drug efficacy screening," *Nature Commun.*, vol. 7, no. 1, Feb. 2016, Art. no. 10331.

[12] Z. Mazerska, A. Mroz, M. Pawlowska, and E. Augustin, "The role of glucuronidation in drug resistance," *Pharmacol. Therapeutics*, vol. 159, no. 1, pp. 35–55, Mar. 2016.

[13] M. G. Rees et al., "Correlating chemical sensitivity and basal gene expression reveals mechanism of action," *Nature Chem. Biol.*, vol. 12, no. 2, pp. 109–116, Feb. 2016.

[14] G. Zhang and D. W. Nebert, "Personalized medicine: Genetic risk prediction of drug response," *Pharmacol. Therapeutics*, vol. 175, no. 1, pp. 75–90, Jul. 2017.

[15] M. Byra, K. Dobruch-Sobczak, Z. Klimonda, H. Piotrzkowska-Wroblewska, and J. Litniewski, "Early prediction of response to neoadjuvant chemotherapy in breast cancer sonography using Siamese convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 3, pp. 797–805, Mar. 2021.

[16] S. Papillon-Cavanagh et al., "Comparison and validation of genomic predictors for anticancer drug sensitivity," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 597–602, Jul./Aug. 2013.

[17] L. Parca et al., "Modeling cancer drug response through drug-specific informative genes," *Sci. Rep.*, vol. 9, no. 1, Oct. 2019, Art. no. 15222.

[18] A. Daemen et al., "Modeling precision treatment of breast cancer," *Genome Biol.*, vol. 14, no. 10, 2013, Art. no. R110.

[19] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, no. 1, pp. 91–102, Aug. 2019.

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

[21] K. Vougas et al., "Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining," *Pharmacol. Therapeutics*, vol. 203, no. 1, Nov. 2019, Art. no. 107395.

[22] J. Choi, S. Park, and J. Ahn, "RefDNN: A reference drug based neural network for more accurate prediction of anticancer drug resistance," *Sci. Rep.*, vol. 10, no. 1, Feb. 2020, Art. no. 1861.

[23] J. Xu et al., "Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides," *Brief Bioinform*, vol. 22, no. 5, Sep. 2021, Art. no. bbab083.

[24] S. Chawla et al., "Gene expression based inference of cancer drug sensitivity," *Nature Commun.*, vol. 13, no. 1, Sep. 2022, Art. no. 5680.

[25] 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. i911–i918, Dec. 2020.

[26] X. Liu, C. Song, F. Huang, H. Fu, W. Xiao, and W. Zhang, "GraphCDR: A graph neural network method with contrastive learning for cancer drug response prediction," *Brief Bioinform.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab457.

[27] M. J. Chenoweth et al., "Global pharmacogenomics within precision medicine: Challenges and opportunities," *Clin. Pharmacol. Therapeutics*, vol. 107, no. 1, pp. 57–61, Jan. 2020.

[28] F. Xia et al., "A cross-study analysis of drug response prediction in cancer cell lines," *Brief Bioinform.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab356.

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

[30] X. Chen and L. Huang, "Computational model for disease research," *Brief Bioinform.*, vol. 24, no. 1, Jan. 2023, Art. no. bbac615.

[31] O. Bazgir, R. Zhang, S. R. Dhruba, R. Rahman, S. Ghosh, and R. Pal, "Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks," *Nature Commun.*, vol. 11, no. 1, Sep. 2020, Art. no. 4391.

[32] J. C. Costello et al., "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014.

[33] T. Ching et al., "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc., Interface*, vol. 15, no. 141, Apr. 2018, Art. no. 20170387.

[34] P. Geeleher et al., "Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies," *Genome Res.*, vol. 27, no. 10, pp. 1743–1751, Oct. 2017.

[35] T. H. Li, C. C. Wang, L. Zhang, and X. Chen, "SNRMPACDC: Computational model focused on siamese network and random matrix projection for anticancer synergistic drug combination prediction," *Brief Bioinform.*, vol. 24, no. 1, Jan. 2023, Art. no. bbac503.

[36] M. Mou et al., "A transformer-based ensemble framework for the prediction of protein-protein interaction sites," *Research*, vol. 6, no. 1, 2023, Art. no. 0240.

[37] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," *Brief Bioinform.*, vol. 22, no. 1, pp. 360–379, Jan. 18 2021.

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

[39] W. X. Shen et al., "Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations," *Nature Mach. Intell.*, vol. 3, no. 4, Apr. 2021, Art. no. 334.

[40] S. Kim et al., "PubChem in 2021: New data content and improved web interfaces," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1388–D1395, Jan. 2021.

[41] N. Stiefl, I. A. Watson, K. Baumann, and A. Zaliani, "ErG: 2D pharmacophore descriptions for scaffold hopping," *J. Chem. Inf. Model.*, vol. 46, no. 1, pp. 208–220, Jan./Feb. 2006.

[42] S. Kim et al., "PubChem 2023 update," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1373–D1380, Jan. 2023.

[43] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 6, pp. 1273–1280, Nov./Dec. 2002.

[44] 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, Jan. 2013.

[45] F. Iorio et al., "A landscape of pharmacogenomic interactions in cancer," *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 28 2016.

[46] J. Yin et al., "VARIDT 3.0: The phenotypic and regulatory variability of drug transporter," *Nucleic Acids Res.*, vol. 52, no. D1, Oct. 2023, Art. no. gkad818.

[47] J. Yin et al., "INTEDE: Interactome of drug-metabolizing enzymes," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1233–D1243, Jan. 2021.

[48] F. Li et al., "DrugMAP: Molecular atlas and pharma-information of all drugs," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1288–D1299, Jan. 2023.

[49] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, and M. Ishiguro-Watanabe, "KEGG for taxonomy-based analysis of pathways and genomes," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D587–D592, Jan. 2023.

[50] M. Gillespie et al., "The reactome pathway knowledgebase 2022," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D687–D692, Jan. 2022.

[51] Y. Sha, J. H. Phan, and M. D. Wang, “Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data,” in *Proc. IEEE 37th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 6461–6464.

[52] E. Becht et al., “Dimensionality reduction for visualizing single-cell data using UMAP,” *Nature Biotechnol.*, vol. 37, no. 1, Dec. 2018, pp. 38–44.

[53] R. Jonker and A. Volgenant, “A shortest augmenting path algorithm for dense and sparse linear assignment problems,” *Computing*, vol. 38, no. 4, pp. 325–340, 1987.

[54] A. Basu et al., “An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules,” *Cell*, vol. 154, no. 5, pp. 1151–1161, Aug. 2013.

[55] B. Seashore-Ludlow et al., “Harnessing connectivity in a large-scale small-molecule sensitivity dataset,” *Cancer Discov.*, vol. 5, no. 11, pp. 1210–1223, Nov. 2015.

[56] Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium, “Pharmacogenomic agreement between two cancer cell line data sets,” *Nature*, vol. 528, no. 7580, pp. 84–87, Dec. 2015.

[57] J. Chen et al., “Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data,” *Nature Commun.*, vol. 13, no. 1, Oct. 2022, Art. no. 6494.

[58] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[59] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters – improve semantic segmentation by global convolutional network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1743–1751.

[60] H. Zhang et al., “ncRNAInter: A novel strategy based on graph neural network to discover interactions between lncRNA and miRNA,” *Brief Bioinform.*, vol. 23, no. 6, Nov. 2022, Art. no. bbac411.

[61] P. M. Haverty et al., “Reproducible pharmacogenomic profiling of cancer cell line panels,” *Nature*, vol. 533, no. 7603, pp. 333–337, May 2016.

[62] L. McInnes, J. Healy, and J. J. a. e.-p. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *arXiv:1802.03426*. Accessed: Feb. 1, 2018. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180203426M

[63] N. Kokhlikyan et al.., “Captum: A unified and generic model interpretability library for PyTorch,” *arXiv:2009.07896*. Accessed: Sep. 1, 2020. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200907896K

[64] M. Fallahi-Sichani, S. Honarnejad, L. M. Heiser, J. W. Gray, and P. K. Sorger, “Metrics other than potency reveal systematic variation in responses to cancer drugs,” *Nature Chem. Biol.*, vol. 9, no. 11, pp. 708–714, Nov. 2013.

[65] B. T. Sherman et al., “DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update),” *Nucleic Acids Res.*, vol. 50, no. W1, pp. W216–W221, Jul. 2022.

[66] L. H. Hall and L. B. Kier, “Electrotopological state indexes for atom types - a novel combination of electronic, topological, and valence state information,” *J. Chem. Inf. Comput. Sci.*, vol. 35, no. 6, pp. 1039–1045, Nov./Dec. 1995.

[67] N. Schaworonkow and B. Voytek, “Enhancing oscillations in intracranial electrophysiological recordings with data-driven spatial filters,” *Plos Comput. Biol.*, vol. 17, no. 8, Aug. 2021, Art. no. e1009298.

[68] L. Liu, D. Li, and W. H. Wong, “Convergence rates of a partition based bayesian multivariate density estimation method,” in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, vol. 30, pp. 4738–4746.

[69] R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger, “Contrastive learning in protein language space predicts interactions between drugs and protein targets,” *Proc. Nat. Acad. Sci.*, vol. 120, no. 24, Jun. 2023, Art. no. e2220778120.

[70] L. Chen et al., “TransformerCPI: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments,” *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, Aug. 2020.

[71] C. Ivan, “Convolutional neural networks on randomized data,” in *Proc. CVPR Workshops*, 2019, pp. 1–8.

[72] L. K. Schatzle, A. Hadizadeh Esfahani, and A. Schuppert, “Methodological challenges in translational drug response modeling in cancer: A systematic analysis with FORESEE,” *Plos Comput. Biol.*, vol. 16, no. 4, Apr. 2020, Art. no. e1007803.

[73] C. R. Or, C. W. Huang, C. C. Chang, Y. C. Lai, Y. J. Chen, and C. C. Chang, “Obatoclax, a pan-BCL-2 inhibitor, downregulates survivin to induce apoptosis in human colorectal carcinoma cells via suppressing WNT/beta-catenin signaling,” *Int. J. Mol. Sci.*, vol. 21, no. 5, Mar. 2020, Art. no. 1773.

[74] G. Melo, C. A. B. Silva, A. Hague, E. K. Parkinson, and E. R. C. Rivero, “Anticancer effects of putative and validated BH3-mimetic drugs in head and neck squamous cell carcinomas: An overview of current knowledge,” *Oral Oncol.*, vol. 132, no. 1, Sep. 2022, Art. no. 105979.

[75] H. A. Alkhatabi et al., “Venetoclax-resistant MV4-11 leukemic cells activate PI3K/AKT pathway for metabolic reprogramming and redox adaptation for survival,” *Antioxidants*, vol. 11, no. 3, Feb. 2022, Art. no. 461.

[76] L. Han et al., “Concomitant targeting of BCL2 with venetoclax and MAPK signaling with cobimetinib in acute myeloid leukemia models,” *Haematologica*, vol. 105, no. 3, pp. 697–707, Mar. 2020.

[77] Q. Zhang et al., “Activation of RAS/MAPK pathway confers MCL-1 mediated acquired resistance to BCL-2 inhibitor venetoclax in acute myeloid leukemia,” *Signal Transduct. Target. Ther.*, vol. 7, no. 1, Feb. 2022, Art. no. 51.

[78] E. Ten Hacken, M. Gounari, P. Ghia, and J. A. Burger, “The importance of B cell receptor isotypes and stereotypes in chronic lymphocytic leukemia,” *Leukemia*, vol. 33, no. 2, pp. 287–298, Feb. 2019.

[79] P. Qin et al., “Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication,” *Cell Discov.*, vol. 7, no. 1, Jan. 2021, Art. no. 2.

[80] M. Matsuoka and K. T. Jeang, “Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation,” *Nature Rev. Cancer*, vol. 7, no. 4, pp. 270–280, Apr. 2007.

[81] K. Kielbassa et al., “Ibrutinib sensitizes CLL cells to venetoclax by interrupting TLR9-induced CD40 upregulation and protein translation,” *Leukemia*, vol. 37, no. 6, pp. 1268–1276, Jun. 2023.

[82] A. K. S. Chiang, K. P. Tam, and R. K. H. Au-Yeung, “Combination of bortezomib and venetoclax induces synergistic killing of epstein-barr virus-driven lymphoproliferative diseases by targeting the pro-survival function of latent membrane protein-1 and epstein-barr nuclear antigen-3C,” *Blood*, vol. 136, no. 1, pp. 12–13, Nov. 2020.

[83] F. Xie et al., “Multifactorial deep learning reveals pan-cancer genomic tumor clusters with distinct immunogenomic landscape and response to immunotherapy,” *Clin. Cancer Res.*, vol. 26, no. 12, pp. 2908–2920, Jun. 2020.

[84] S. Xiao, H. Lin, C. Wang, S. Wang, and J. C. Rajapakse, “Graph neural networks with multiple prior knowledge for multi-omics data analysis,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 9, pp. 4591–4600, Sep. 2023.

[85] R. L. Grossman et al., “Toward a shared vision for cancer genomic data,” *New England J. Med.*, vol. 375, no. 12, pp. 1109–1112, Sep. 2016.