OXFORD

Briefings in Bioinformatics, 2022, 1–13

https://doi.org/10.1093/bib/bbac605

Review

A systematic assessment of deep learning methods for drug response prediction: from in vitro to clinical applications

Bihan Shen[†], Fangyoumin Feng[†], Kunshi Li, Ping Lin, Liangxiao Ma and Hong Li

Corresponding author: Hong Li, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. E-mail: lihong01@sibs.ac.cn

[†]The authors would like to declare that Bihan Shen and Fangyoumin Feng contributed the same to this work.

Abstract

Drug response prediction is an important problem in personalized cancer therapy. Among various newly developed models, significant improvement in prediction performance has been reported using deep learning methods. However, systematic comparisons of deep learning methods, especially of the transferability from preclinical models to clinical cohorts, are currently lacking. To provide a more rigorous assessment, the performance of six representative deep learning methods for drug response prediction using nine evaluation metrics, including the overall prediction accuracy, predictability of each drug, potential associated factors and transferability to clinical cohorts, in multiple application scenarios was benchmarked. Most methods show promising prediction within cell line datasets, and TGSA, with its lower time cost and better performance, is recommended. Although the performance metrics decrease when applying models trained on cell lines to patients, a certain amount of power to distinguish clinical response on some drugs can be maintained using CRDNN and TGSA. With these assessments, we provide a guidance for researchers to choose appropriate methods, as well as insights into future directions for the development of more effective methods in clinical scenarios.

Keywords: drug response prediction, personalized therapy, deep learning, graph embedding, benchmark

Introduction

The goal of precision oncology is to deliver therapies tailored to the molecular profile of an individual's tumor. Due to the paucity of clinical pharmacogenomics datasets, researchers often use preclinical models, especially cancer cell lines, as a proxy. From the great efforts made by the scientific community, large-scale pharmacogenomics resources, including the Cancer Cell Line Encyclopedia (CCLE) [1], Genomics of Drug Sensitivity in Cancer (GDSC) [2] and Cancer Therapeutics Response Portal (CTRP) [3, 4], are publicly available for further investigation. The development of computational methods to solve the drug response prediction problem has been expedited through the use of accumulating data.

Drug response prediction is an important and challenging problem in both bioinformatics and translational medicine. Various categories of methods, including kernel-based methods, network-based methods, regression models, traditional machine learning and deep learning (DL) models [5–9], have been developed. To identify optimal methods and provide suggestions for the improvement of new models, the National Cancer Institute (NCI) and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) launched a drug sensitivity prediction challenge as early as 2014 [5], and several other articles were later published for benchmarking model performance [6, 10–13]. The top-tier methodology usually models nonlinear relationships, utilizes prior biology knowledge and implements sophisticated preprocessing and feature selection [5, 6]. Deep learning (DL) methods are gaining popularity due to high capacity, flexibility and better generalizability across cell line datasets [11]. However, the evaluation of different DL methods is relatively lacking compared to other methodological categories.

In addition to the prediction methodology, other factors related to prediction performance have been explored: the experimental variability of drug responses across studies [11], the integration of multiple data sources [10], the incorporation of biological pathway or network information [5], pan-cancer or tissue-specific modelling [6], the predictive power of multi-omics or single omics

Bihan Shen is a PhD student in the Cancer Systems Biology group at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Fangyoumin Feng is a postdoctoral research fellow in the Cancer Systems Biology group at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Kunshi Li is a master student in the Cancer Systems Biology group at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Ping Lin is a postdoctoral research fellow in the Cancer Systems Biology group at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Liangxiao Ma is a technical expert in high-performance storage and computing in Bio-Med Big Data Center at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Hong Li is a Research Director and Group Leader of the Cancer Systems Biology group at Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.

Received: September 8, 2022. **Revised:** October 30, 2022. **Accepted:** December 9, 2022

data [5, 6], the importance of cell line or drug diversity on model generalizability [11] and the prediction accuracy for drugs that do not appear in the training datasets [13]. However, the systematic evaluation of the predictive ability of preclinical models for clinical patients is insufficient.

Here, we conduct a systematic and comprehensive assessment of representative DL drug response prediction methods, covering both single-drug learning (SDL) and multi-drug learning (MDL) paradigms. These methods were evaluated on cancer celllines using nine metrics. In previous studies, prediction models were only assessed by their overall performance on all drug-cell line pairs. Here, we also considered the model performance for each drug. More importantly, the ability of a model to predict drug response in clinical cohorts was estimated. Based on these results, we provide constructive suggestions and future development directions for the community of scientists interested in this issue.

Brief overview of deep learning methods for drug response prediction

There is a wide variety of DL methods for drug response prediction, including SDL or MDL paradigms that are based on whether drugs are predicted separately or together (Figure 1A and B) [9], regression or classification models that are based on whether the output variables are continuous or discrete response values and models characterized by the DL architectures.

The SDL paradigm is intended to independently predict drug response for a given drug, which makes a linkage between complex omics data and drug sensitivity measures. The nature of the drug response of cell lines is nonlinear; thus, DL methods, such as dense neural networks (DNNs) [14], autoencoders (AEs) [15] and variational autoencoders (VAEs) [16], which are powerful mathematical frameworks, have been employed to better capture nonlinear relationships.

The MDL paradigm digests information from both cell lines and drugs, which facilitates the integration of multiple drugs and scales up the training data. MDL models usually consist of three components: a cell line embedding branch (CEB), which is used to encode genomic profiles (e.g. expression profiles, mutation status and copy number variation); a drug embedding branch (DEB), which is used to encode drug features (e.g. simplified molecular-input line-entry system (SMILES) strings, molecular fingerprints and molecular structure graphs [17]); and a prediction module (PM), which is used to fuse the embedded vectors to predict drug response. Various architectures have been employed. For CEB, DNNs [18], convolutional neural networks (CNNs) [19–21], AEs [22] and attention mechanisms [23] have been used to encode cell line profiles. Additionally, prior biological knowledges were utilized via models like visible neural networks (VNNs) [24] and graph neural networks (GNNs) [25], which better present the features of cell lines. The architectures of DEB often depend on the format of the drug features. Molecular fingerprints and descriptors are usually handled by DNNs [24] and CNNs [19], SMILES strings (defined as a strings of characters) are addressed with CNNs [20, 23], and molecular structure graphs are processed by GNNs as a graph embedding task. With respect to PM, multi-modal fusion is the key to taking advantage of the complementarity of CEB and DEB. The late fusion strategy, where the extracted feature vectors from CEB and DEB are concatenated and fed into a DNN [20–22, 24, 25] or CNN [19, 26], and then transformed into an output neuron to predict drug response, has been widely adopted. Recently, an intermediate fusion strategy in the form of contextual attention has been implemented to boost the interactions of CEB and DEB [23].

Representative methods for assessments

We first queried PubMed using the keywords ‘drug response + prediction + deep learning’ and limited the publication date to those works published after 2019. Then, we manually removed methods without readily available source codes. To balance comprehensiveness and computational cost, we selected the representative state-of-the-art methods with various deep learning architectures. Finally, three MDL methods (DrugCell, PaccMann and TGSA) and three SDL methods (CRDNN, VAEN and MOLI) were included in this benchmark analysis (Table 1).

CRDNN is a DNN model that uses transcriptome as input, the hyperbolic tangent (tanh) as the activation function and the mean squared error (MSE) as the loss function [14]. Although its architecture is simple, better performance was achieved using this model than traditional machine learning methods when predicting drug response and survival in several clinical cohorts.

VAEN uses a VAE with a sigmoid activation function to obtain the embedding of the rank normalized transcriptome data and then uses an elastic net regression model to predict drug response [16]. Notably, the VAE loss function is composed of a reconstruction term (MSE) and a regularization term (Kullback–Leibler divergence) to retain the continuity and completeness in the latent space [27].

MOLI is a Multi-Omics (transcriptome, mutation and copy number variations) Late Integration method [15]. It employs three DNNs with an ReLU activation function to learn latent representations from different types of data and then regularizes the representations by introducing triplet loss, which enforces the distance between samples with the same labels smaller than those with different labels. As a classification model, its PM is a DNN with a sigmoid activation function and binary cross-entropy loss.

PaccMann takes full advantage of an attention mechanism to integrate SMILES strings and gene expression [23]. First, gene attention weights are generated using a softmax layer, and then gene expression values are filtered to ensure that the most informative genes are addressed. Second, SMILES strings are filtered using CNNs with various kernel sizes and then fed into a multi-head contextual attention layer using the filtered genes as a context. The so-called multiscale convolutional attentive encoder is used for the intermediate fusion of gene expression and SMILES strings as well as the extraction of local and long-range dependencies on drug structures.

DrugCell utilizes a visible neural network (VNN) that models the hierarchical organization of biological processes to enhance mechanistic interpretability on the CEB with the tanh activation function. Morgan fingerprints encoded by DNN are concatenated with the VNN-encoded latent vectors, and then fed into another DNN module for regression. Based on the interpretable structure, this method was validated for learning drug response mechanisms and was even extended for the design of synergistic drug combinations [24].

TGSA applies twin GNNs to represent cell lines and drugs. More specifically, a drug is represented as a molecular graph that takes atoms as nodes and chemical bonds as edges; a cell line is represented as a gene–gene interaction graph, where nodes are genes and edges are gene–gene interactions curated from the STRING database. The graph isomorphism network (GIN)

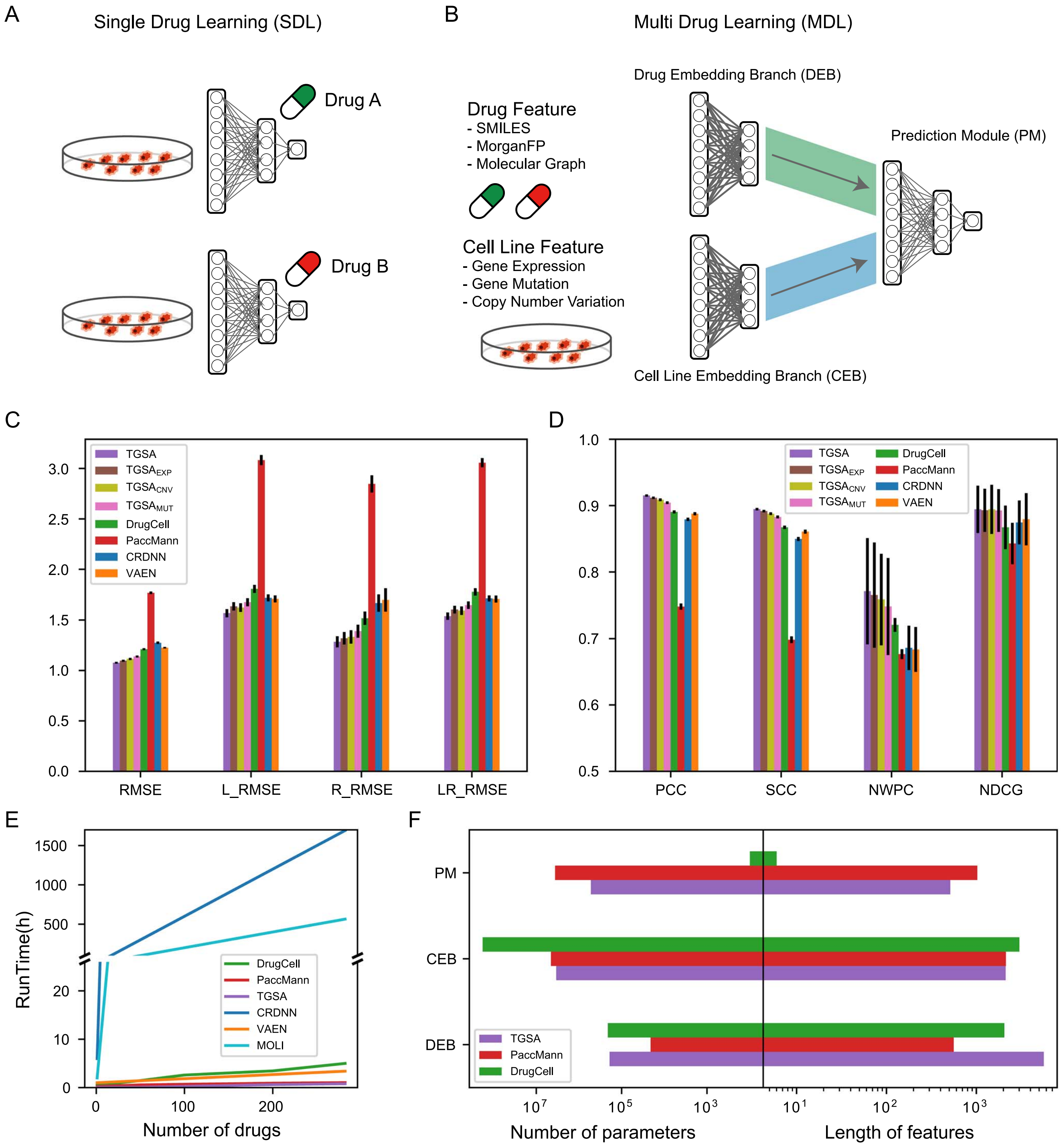

Figure 1. Overall performance of two typical deep learning paradigms for drug response prediction. (**A** and **B**) Illustration of two paradigms of DL drug response prediction models: (**A**) SDL, which constructs one model for each drug, and (**B**) DL, which builds one parameter-sharing model for multiple drugs, which generates embeddings of drug and cell line features and fuses them into the prediction module. (**C**-**D**) The overall performance evaluated by 5-fold cross validation using the entire GDSC dataset where (**C**) depicts the RMSE family metrics and (**D**) represents the correlation metric family. The error bars represent the 95% confidence intervals. (**E**) The runtime varied with the increase of sample size. The time consumed by parameter selection was taken into account. CRDNN was performed on the CPU servers, while the remaining models were assessed on a GPU server equipped with 32 threads, 376G RAM and Nvidia RTX 2080Ti. **(F)** The number of parameters assigned to each part of the MDL methods.

and graph attention network (GAT) are used to update the node features of the drug and cell line graph through the message passing procedure, respectively. Graph embeddings of drugs and cell lines undergo a similarity augmentation procedure implemented using GraphSAGE to fuse information at higher granularity levels [25]. In TGSA, gene expression profiles, mutation and copy number variations are taken as input; and $TGSA_{EXP}$, $TGSA_{CNV}$ and $TGSA_{MUT}$ corresponded to the single omics models used in our research.

We also compared DL methods with the conventional machine learning methods: support vector regression (SVM), Bayesian ridge regression (Bayes) and elastic net regression (Enet). The machine

Table 1. Summary of representative deep learning methods for drug response prediction

	Cell line feature	Cell line embedding	Drug feature	Drug embedding	Prediction module	Loss function	Reference
DrugCell	M	VNN	Morgan FP	DNN	DNN	MSE	[24]
PaccMann	E[a]	Self-attention	SMILES	CNN	Contextual-attention	MSE	[23]
TGSA	E[a], M, C	GAT	Molecular graph	GIN	DNN	MSE	[25]
CRDNN	E[a]	DNN	–	–	DNN	MSE	[14]
VAEN	E[b]	VAE	–	–	Elastic Net	MSE and MAE	[16]
MOLI	E[a], M, C	Auto-encoder	–	–	DNN	Cross entropy and triplet loss	[15]

E: expression profiles; M: mutation status; C: copy number variation; FP: fingerprint; DNN: dense neural network; CNN: convolutional neural network; VNN: visible neural network; GAT: graph attention network; VAE: variational autoencoder; GIN: graph isomorphism network; MSE: mean squared error; MAE: mean absolute error [a]z-score standardization [b]Rank normalization

learning models for each drug were built using the Python package scikit-learn.

Cancer cell line datasets

The GDSC dataset, which was downloaded from https://www.cancerrxgene.org/downloads (release 8.2, accessed 1 December 2021), includes multiple omics data from 988 cell lines involving 28 tissue types and 446,146 response readouts from 518 drugs. Gene expression profiles obtained from the microarray were RMA-normalized, log2-transformed and standardized accordingly (Table 1); only non-silent mutations were retained and coded as 0 for the wild type and 1 for the mutated; gene-level copy number variations were obtained for the cell line as GISTIC scores and binarized by assigning 0 for the copy-neutral and 1 for the deletions or amplifications. Drugs were searched through the PubChem website to obtain unified CIDs and canonical SMILES strings. Other molecular representations for drugs, e.g. Morgan fingerprints (nbits = 2048, radius = 2) and molecular structure graphs, were generated by the RDKit Python package (http://www.rdkit.org). Drugs with CIDs and cell lines with all types of the aforementioned omics data were included in our benchmark study. The CCLE dataset with 471 cell lines and 24 drugs was downloaded from https://depmap.org/portal/ccle (2021Q4) and processed in the same way as the GDSC dataset.

Drug response

The half-maximal inhibitory concentration (IC50) and the area under the dose–response curve (AUC) are commonly used measures of drug sensitivity. Although the performance of a prediction model may differ when using IC50 or AUC, compatible method comparison results are often obtained using these two metrics [6]. Therefore, we used IC50 to measure drug sensitivity in our benchmark work (evaluation using AUC yielded similar results).

For regression models, log transformation and max-min standardization were performed on the training and test datasets, respectively. For classification models, the binarization of IC50 for each drug was performed using the heuristic outlier procedure with four steps: upsampling IC50 to add samples, estimating the kernel density, modelling the population of resistant cell lines as a normal distribution, and evaluating the cumulative distribution to find the binarization threshold [28].

Variation partitioning [29] was conducted to estimate the portion of variation of IC50 values explained by cell lines or drugs using the ‘varpart()‘ function from the ‘vegan‘ R package, where cell line and drug were served as explanatory variables.

Three statistics were calculated to characterize the distribution of IC50. The standard deviation measures the dispersion of responses to a drug. The bimodality coefficient depicts the selective killing activity of the anti-cancer drug [30] and is defined as:

$$\text{bimodality coefficient}^{\mathrm{d}} = \frac{\mathrm{g}^2 + 1}{\mathbf{k} + 3(n-1)^2 / (n-2)\,(n-3)}, \qquad (1)$$

where n denotes the number of cell lines screened for the given drug **d**, **g** denotes the skewness and **k** denotes the excess kurtosis relative to the normal distribution. A higher bimodality coefficient denotes a distribution that is both strongly skewed (high absolute value of g) and light-tailed (small value of k).

The density coverage is used to calculate the cumulative distribution probability on the entire dataset spanning the 10th and 90th percentiles of each drug as follows:

$$\text{density coverage}^{\mathrm{d}} = CDF\left(\pi^{\mathrm{d}}_{0.9}\right) - CDF\left(\pi^{\mathrm{d}}_{0.1}\right), \qquad (2)$$

where $CDF(\cdot)$ is the cumulative distribution function of the IC50 of all the drug-cell line pairs and the 10th and 90th percentiles of the given drug d are denoted as $\pi^{\mathrm{d}}_{0.1}, \pi^{\mathrm{d}}_{0.9}$. A higher density coverage implies that the distribution of the drug is closer to that of the entire dataset.

The Mann–Whitney test was used to evaluate whether there was a significant difference in the above statistics among the different drug groups.

Evaluation on cancer cell lines

Five-fold cross validation was used for model evaluation on the GDSC dataset. The data were divided into training, validation and test sets at a ratio of 3:1:1. Cell lines were stratified to ensure an even proportion of tissue types in each set. For MDL methods, the training, validation and test sets contained 140 244, 46 747 and 46 747 drug-cell line pairs, respectively; for SDL methods, the median sizes of the training, validation and test sets were 540, 179 and 179, respectively. For each prediction method, 30 sets of hyperparameters were selected at random from the grid of recommended configurations in the original papers. For every fold, all 30 models were trained and evaluated, and the model with the highest Pearson correlation coefficient on the validation dataset was selected as the optimal model. Since each sample was tested, we can easily obtain the predicted results of all the cell line-drug pairs by concatenating each test fold. The performance of the entire dataset and individual drugs can then be measured based on the evaluation metrics.

The evaluation metrics quantify the performance of a predictive model by comparing the observed values, $y = (y_1, y_2, \ldots, y_n)$, with the predicted values, $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$, in different aspects.

We adapted nine metrics from [6] and classified them into two groups. One group, which we refer to as the RMSE family, was used to measure the average magnitude of residuals. RMSE is the root mean squared error between the ground-truth and predicted values among all cell lines; L_RMSE, R_RMSE and LR_RMSE only consider only sensitive, only resistant and both sensitive and resistant cell lines, respectively [6]. In the RMSE family, a lower value indicates better performance. The other group, which we refer to as the correlation metric family, was used to depict the coherence between the observed and predicted values and includes the Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SCC), normalized discounted cumulative gain (NDCG) [31], probabilistic c-index (PC)/normalized weighted probabilistic c-index (NWPC) [5], and area under the receiver operator characteristic curve (ROC_AUC). In the correlation metric family, a higher value indicates better performance. In particular, PC/NWPC and NDCG are used to evaluate the rank coherence between the predicted and observed values. NDCG measures the ability to rank highly relevant results (i.e. sensitive cell lines) at the top of the list:

$$\mathrm{DCG}\left(y,\hat{y}\right)=\sum\nolimits_{i=1}^{n}\frac{2^{-y_i}-1}{\log_2\left(r\left(\hat{y}_i\right)+1\right)},$$

$$\mathrm{NDCG}\left(y,\hat{y}\right)=\frac{\mathrm{DCG}\left(y,\hat{y}\right)}{\mathrm{DCG}\left(y,y\right)} \qquad (3)$$

where $r(\hat{y}_i)$ is the position of $\hat{y}_i$ on the sorted $\hat{y}$in ascending order. More sensitive cell lines have smaller y and lower $r(\hat{y}_i)$, thus a higher NDCG indicates that the model can identify the most sensitive cell lines.

PC is used to evaluate the rank coherence between the predicted and observed values, and is defined as the ratio of the concordant pairs to all possible combinations:

$$\mathbf{PC}\left(y,r\left(\hat{y}\right)\right)=\frac{2}{n\left(n-1\right)}\sum_{i<j}hp\left(y_i,y_j,r\left(\hat{y}_i\right),r\left(\hat{y}_j\right),\sigma\left(\mathbf{y}\right)\right) \qquad (4)$$

$$hp\left(y_i,y_j,r\left(\hat{y}_i\right),r\left(\hat{y}_j\right),\sigma\left(\mathbf{y}\right)\right)=\begin{cases}\frac{1}{2}\left(1+\mathbf{erf}\left(\frac{y_i-y_j}{2\sigma\left(\mathbf{y}\right)}\right)\right),r\left(\hat{y}_i\right)>r\left(\hat{y}_j\right);\\ 0.5,r\left(\hat{y}_i\right)=r\left(\hat{y}_j\right);\\ \frac{1}{2}\left(1+\mathbf{erf}\left(\frac{y_j-y_i}{2\sigma\left(\mathbf{y}\right)}\right)\right),r\left(\hat{y}_i\right)<r\left(\hat{y}_j\right)\end{cases}$$

$$\mathbf{erf}(a)=\frac{2}{\sqrt{\pi}}\int_0^a e^{-t^2}dt\,.$$

NWPC is the normalized weighted PC when considering all drugs:

$$\mathrm{WPC}(M)=\frac{\sum_d w_d\cdot\mathrm{PC}_d}{\sum_d w_d},$$

$$\mathrm{NWPC}=\frac{\mathrm{WPC}-\mathrm{WPC_{min}}}{\mathrm{WPC}-\mathrm{WPC_{max}}}, \qquad (5)$$

where the drug weight w_d is determined as $w_d=\left(\mathrm{PC}_d^*-\mu_d\right)/\sigma_d$. A random ranking of n items is generated for the given drug d, $\mathbf{R_d}$ and $\mathbf{PC_d}=\mathbf{PC}(y,R_d)$ is computed; the above process is repeated 1000 times to produce an empirical null distribution with a median and standard deviation $(\mu_\mathrm{d},\sigma_\mathrm{d})$ and the gold standard $\mathbf{PC_d^*}=\mathbf{PC}(y,r(y))$ is computed.

Since MOLI is a classification model, it was only evaluated by ROC_AUC to measure the ability of the model to discriminate between sensitive and resistant cell lines. The other eight metrics were used to evaluate regression models under two settings: pooling all drugs together and separately computing for each drug.

The five-fold cross-validation was randomly repeated ten times, 50 values were obtained for each evaluation metric, and then the standard deviations and 95% confidence intervals were calculated for statistical evaluation of the DL methods. Due to the high computational cost of training CRDNN models, 10 drugs were randomly selected to represent the confidence intervals of all drugs in CRDNN.

Ablation study

An ablation study was performed to estimate the importance of CEB and DEB via paralyzing each of them. More specifically, each element of molecular profiles was set to zero (CellZeros), and the identical SMILES string (i.e. CCCCCCCCCCCC) replaced the original (DrugZeros) to disable CEB and DEB, respectively.

Computation evaluation of the runtime

CRDNN was performed on the CPU servers with 32 threads and 128G RAM. The rest were performed on a GPU server equipped with 32 threads, 376G RAM and Nvidia RTX 2080Ti. The runtime of each method was captured using the time function available in the R or Python environments.

Evaluation on clinical cohorts

The Cancer Genome Atlas (TCGA) provides extensive molecular data together with the clinical information of patients spanning dozens of cancer types. The drug treatment information from TCGA was curated by Ding *et al.* [32]. We selected 16 drugs with sample sizes larger than 10 (Supplementary Table S2): cisplatin, paclitaxel, gemcitabine, 5-fluorouracil, temozolomide, docetaxel, doxorubicin, etoposide, bleomycin, pemetrexed, vinorelbine, tamoxifen, bicalutamide, sorafenib, vinblastine and methotrexate. To avoid the effects of multiple rounds of treatments, patients with responses to multiple drugs were excluded. The corresponding omics data were downloaded via the GDC portal (https://portal.gdc.cancer.gov/). RNA-seq gene expression data were converted from counts to TPM (transcripts per million) and log-transformed. ComBat [33] was used to remove batch effects between cell line and patient datasets. The 'standard ComBat' pooled gene expression matrix of two dataset and adjusted the technical artifacts using an empirical Bayes approach; the 'reference-batch ComBat' took GDSC cell line dataset as baseline to adjust patient data, which do not see the patient data before model training [34]. Mutations and copy number variations in TCGA were transformed to numerical vectors in the same manner as for cancer cell lines. An additional clinical trial containing 169 relapsed myeloma patients treated with bortezomib was collected to validate DL methods using the same procedure as TCGA dataset [35].

Hyperparameters for each model were determined by the above 5-fold cross validation on cancer cell lines. The optimal models were trained using the whole GDSC dataset and tested on the patient datasets. In the clinical cohorts, patients were categorized as follows: progressive disease (PD), stable disease (SD), partial response (PR) and complete response (CR). In our study, PD and SD patients were treated as non-responders, while PR and CR patients were treated as responders. Next, the discrimination of each prediction model between the non-responders and responders was checked by three metrics (effect size, *P*-value and ROC_AUC). The effect size was calculated as the mean difference between the predicted IC50 of responders and non-responders; the *P*-value was estimated via the Mann–Whitney test with the alternative hypothesis that responders have a lower mean IC50; ROC_AUC was computed on the predicted IC50 of patients to aggregate the performance across various thresholds.

Results

Overall performance of the prediction methods

We selected six typical and state-of-the-art methods of drug response prediction, including three SDL methods (CRDNN, VAEN, MOLI) and three MDL methods (DrugCell, PaccMann, TGSA). Detailed descriptions of each model are listed in Table 1. These methods were assessed on the GDSC dataset, which contains 233 738 drug sensitivity readouts involving 282 drugs and 966 cell lines.

Prediction accuracy was evaluated by cross validation. By pooling each test fold together, eight evaluation metrics were calculated among all cell-drug pairs in the GDSC dataset for the eight regression models including single omics variants of TGSA. The results of RMSE and its variants are shown in Figure 1C. The RMSEs of most models except PaccMann were below 0.15. If only sensitive and/or resistant cell lines were considered, the error slightly increased (L_RMSE, R_RMSE and LR_RMSE), which indicates that it is difficult to predict the extreme and rare responses. Regarding the correlation metrics shown in Figure 1D, PCC and SCC were used to evaluate linear and monotonic relationships based on raw and ranked values, respectively. Relatively poor performance on these two metrics, with a PCC of 0.74 and an SCC of 0.69, was observed using PaccMann, while values over 0.8 were obtained using the other models. DrugCell, TGSA, CRDNN and VAEN performed fairly well on all metrics. Additionally, the comparison between single-omics and multi-omics input on TGSA attested that gene expression was the most informative and that the integration of multi-omics data slightly enhanced the prediction performance (Figure 1C and D).

SDL models had a linear time complexity as the number of drugs increased, while MDL models tended to have a sublinear time cost due to the shared parameters among drugs (Figure 1E). VAEN only utilizes the DL strategy to obtain low-dimensional representations of expression profiles that can be quickly processed in elastic net; thus, it is more scalable than other SDL models. DrugCell was slower than the other MDL models due to the larger number of parameters. With closer scrutiny on the architecture of MDL models (Figure 1F), it is not surprising to find that more parameters are assigned to CEB compared with DEB and PM, which are subject to the length and complexity of cell line features. Considering the overall performance and time cost, TGSA is more recommended.

Previous publications have shown that deep learning methods have certain advantages compared with conventional machine learning methods in different scenarios of drug response prediction scenarios: within cell line datasets, cross cell line datasets and transfer to clinical cohorts [11, 13, 14]. To explore this, we compared the six DL methods with SVM, Bayes and Enet. Most of the DL models yielded better performance than those machine learning methods on all eight metrics (Supplementary Table S1). Data sources are another focus of discussion with respect to drug sensitivity estimation. We compared the overall performance of deep learning methods on another large cancer cell line dataset, CCLE. The performance of each DL model fluctuates between the GDSC and CCLE datasets, but the relative level of performance among the different models was consistent (Supplementary Figure S1). Such consistency of method comparison among different data sources have also been observed in a previous report [11].

Relative contributions of cell lines and drugs to MDL models

The measure of drug response is determined by both the cell line and drug, but the relative importance of these two factors is not fully understood. Therefore, we assessed the effects of the cell line and drug in terms of the ground-truth distribution and the contribution of different components of MDL methods.

Intuitively, there was a larger disparity in each drug's IC50 distribution than that of each cell line (Figure 2A). Then, the portions of variance of the response explained by cell lines or drugs were estimated by variation partitioning analysis. As Figure 2B shows, drugs accounted for 73.6% of the variance in IC50 values, while cell lines accounted for only 5.9%.

To understand the relative contributions of CEB and DEB to MDL methods, an ablation study was developed to evaluate the performance of a model in the CellZeros and DrugZeros settings, where models were blind to different cell lines and drugs, respectively. Generally, the overall performance decreased significantly on all three MDL methods in the DrugZeros setting, with much higher RMSE and lower SCC values, while CellZeros had minimal impact on the overall performance, indicating that DEB has a greater contribution (Figure 2C and D). This finding is consistent with the result of variation partitioning; i.e. the variance between cell lines is too subtle to be captured by MDL methods; therefore, it is crucial to improve CEB to better distinguish the differences among cell lines. Furthermore, we found that the prediction performance significantly decreased when calculating the RMSE and SCC values for each drug using CellZeros (Figure 2E and F). This finding suggests that the evaluation on single-drug level could better measure the influence of cell features compared to on all drugs together.

Considering the large difference among drugs and the outcome of the ablation study, a null model was introduced to represent the extreme scenario, in which IC50 variation was only determined by the drug, which means all cell lines have the same IC50 values for the given drug. Surprisingly, a relatively good overall performance (PCC 0.83, SCC 0.86 and RMSE 1.36) was obtained using the null model, although it only represents the variability among drugs (Figure 2G). Taken together, these results highlight the necessity of a separate evaluation of each drug for method comparison.

Evaluation on single-drug level

A common application scenario in precision medicine is the estimation of the response of a drug on different tumor samples; thus, we evaluated the performance of DL methods for each drug. The pipeline of single-drug level assessment is shown in Supplementary Figure S2. Nine evaluation metrics were not independent, and their associations were measured by the Spearman's rank correlation coefficient (Figure 3A-C, Supplementary Figure S3). As expected, the RMSE family (RMSE, L_RMSE, R_RMSE and LR_RMSE) showed high internal consistency, which was similarly observed in the correlation metric family (SCC, PCC, ROC_AUC, PC and NDCG). However, metrics from the RMSE family were poorly correlated with those from the correlation metric family. MSE (the square of RMSE) is a commonly used loss function in deep learning models (Table 1); however, but minimizing MSE cannot guarantee higher score on the correlation metrics.

Next, we compared the performance of DL models and the null model using nine evaluation metrics on single-drug level

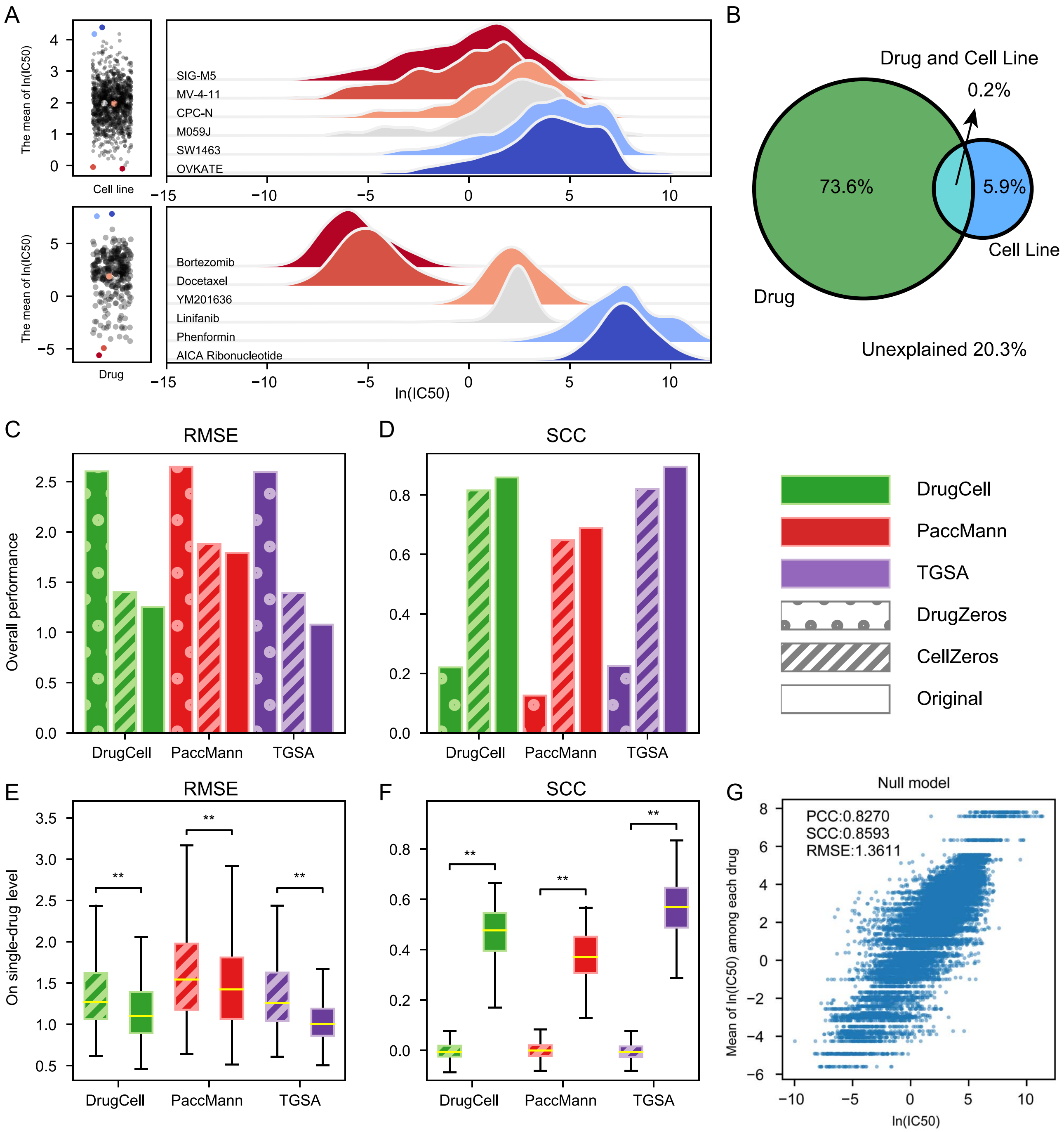

Figure 2. Variation partitioning and an ablation study assess the effects of drugs and cell lines on the prediction performance. (**A**) The distribution of IC50 values for the representative cell lines or drugs. (**B**) The proportion of IC50 variance explained by drugs and cell lines. (**C**-**F**) Delineation of the impact of drug and cell line embedding branches assessed in the ablation study. (**C**-**D**) and (**E**-**F**) illustrate the overall and single-drug level performance, respectively. Asterisks indicate the level of statistical significance by the Mann–Whitney test: * $P < 0.05$, ** $P < 0.01$. (G) The scatter plot depicts the correlation between the null model and observed data.

(Figure 3D-F, Supplementary Figure S4). Similar to the overall performance, TGSA outperformed the other models when evaluating for each drug. The ROC_AUC value of MOLI oscillated around 0.5, indicating that the predictions resembled random guesses, which might result from information loss during the discretization of IC50 values when treated it as a classification problem.

So far, DrugCell, TGSA, CRDNN and VAEN have performed relatively well on both overall and single-drug level assessments. The leading model, TGSA, achieved an overall RMSE of 1.07 and an SCC of 0.89; the RMSE values of each drug ranged from 0.50 to 2.66 with a median of 1.00, and the SCC values ranged from 0.18 to 0.83 with a median of 0.57.

Predictability of individual drugs

Based on the above results (Figure 3), the ability to correctly predict drug sensitivity varied among drugs. SCC between the observed and predicted IC50 values was used to measure the predictability of each drug. Then, we tested whether the predictability of the drugs was consistent among the different

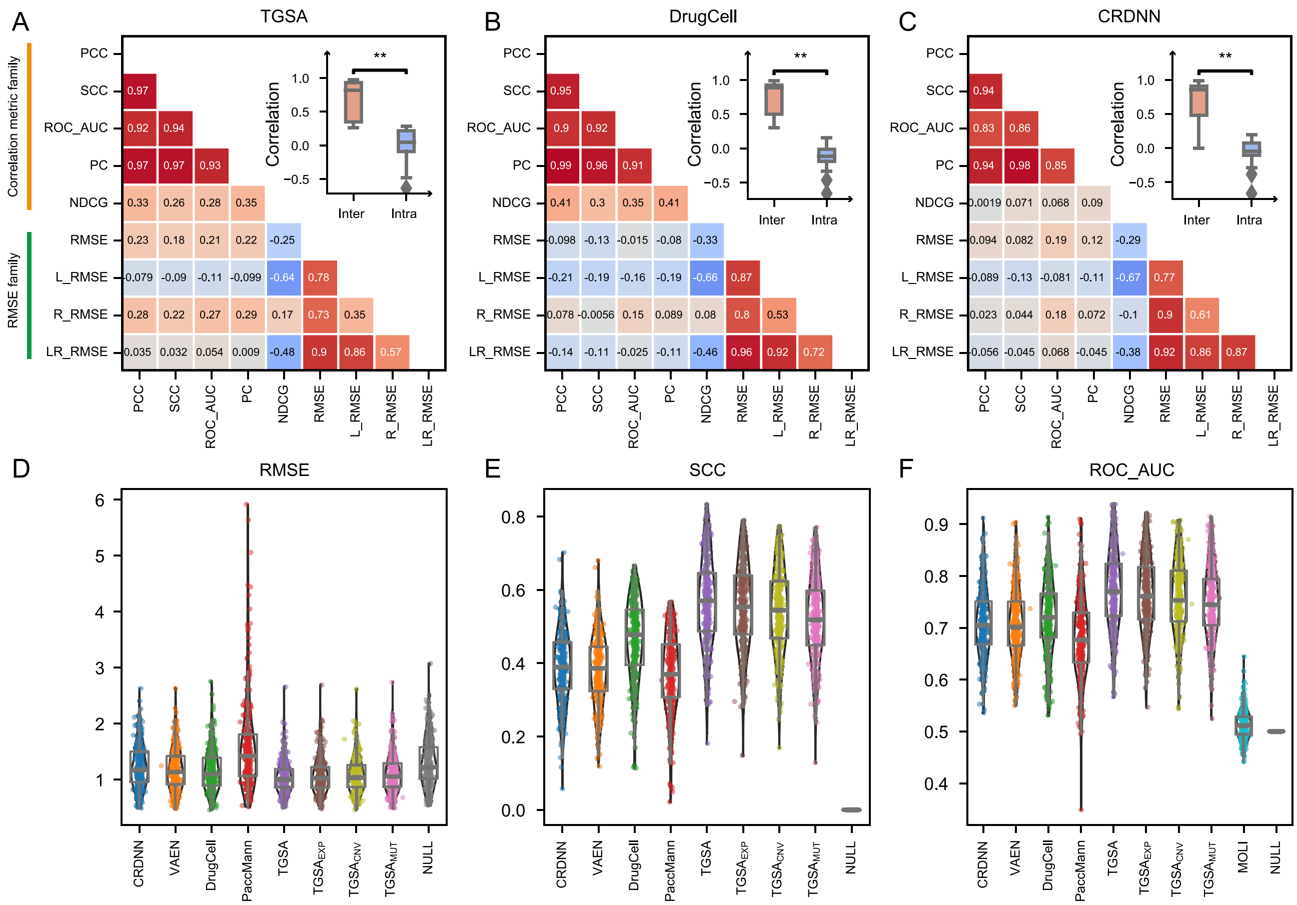

Figure 3. The prediction performance of DL methods for the response of each drug. (**A**-**C**) The consistency of nine evaluation metrics on TGSA (**A**), DrugCell (**B**), CRDNN (**C**) and others (Supplementary Figure S1). Nine metrics were classified them into two families: the RMSE family (RMSE, L_RMSE, R_RMSE, LR_RMSE) and the correlation metric family (PCC, SCC, NDCG, PC, ROC_AUC). Intra-and inter-metric-family consistency are defined as the Spearman's rank correlation coefficient within the metric family and across metric families, respectively. Intra-metric-family consistency is significantly higher than inter-metric-family consistency (Mann–Whitney test, ** $P < 0.01$). (**D**-**F**) The single-drug level performance of nine DL models and null model was assessed by RMSE (**D**), SCC (**E**), ROC_AUC (**F**) and other metrics (Supplementary Figure S2). The classification model MOLI was only evaluated by ROC_AUC.

models. The correlation of predictability between any two methods was significantly positive (Figure 4A), indicating that there are drugs that could be well predicted regardless of the methods and also drugs that are difficult to correctly predict. Models from the same learning paradigm showed significantly higher consistency than those from different learning paradigms (Figure 4A). It is implied that the predictability might be bound by the learning paradigm.

Furthermore, drugs were classified into four groups (Figure 4B): predictable drugs, with SCC values ranked in the top 50% of all methods (P group); unpredictable drugs with SCC values ranked in the bottom 50% of all methods (U group); difficult to predict using mutation status but easy to predict using expression profiles, with SCC values ranked in the bottom 50% for DrugCell and TGSA$_{\mathrm{MUT}}$ while ranked top 50% for TGSA$_{\mathrm{EXP}}$, PaccMann, CRDNN and VAEN (M group); and the remaining drugs (O group).

To comprehensively characterize factors that influence drug predictability, we defined three statistical measurements and compared them in different drug groups (Figure 4D-F). The standard deviation indicates the dispersion of IC50 among cancer cell lines within the given drug; the bimodality coefficient indicates the selectivity of the killing activity of each drug; and the density coverage reflects the consistency of the IC50 distribution between the given drug and the entire dataset. All these statistics significantly differed between the P and U groups, indicating that drugs that are easier to predict have in high variance, selective killing activity and high consistency with the overall distribution.

Next, pathway enrichment analysis based on the drug target pathway was conducted to investigate the relationship between the predictability and mechanism of the given drug using Fisher's exact test. As Figure 4C shows, drugs targeting at ERK MAPK signaling were enriched in the M group, DNA replication and chromatin histone acetylation were enriched in the P group, and IGF1R signaling was enriched in the U group.

In summary, the predictability of individual drugs might be bound by the learning paradigm, IC50 distribution and drug mechanism of action.

Transferability on patients

Research on preclinical models is aimed at facilitating the clinical applications. To further evaluate the inductive transferability on TCGA patients, batch effects between cell line and patient datasets were removed by the standard ComBat, and then DL methods trained from all cancer cell lines were applied to predict patients' response of 16 drugs (Supplementary Table S2,

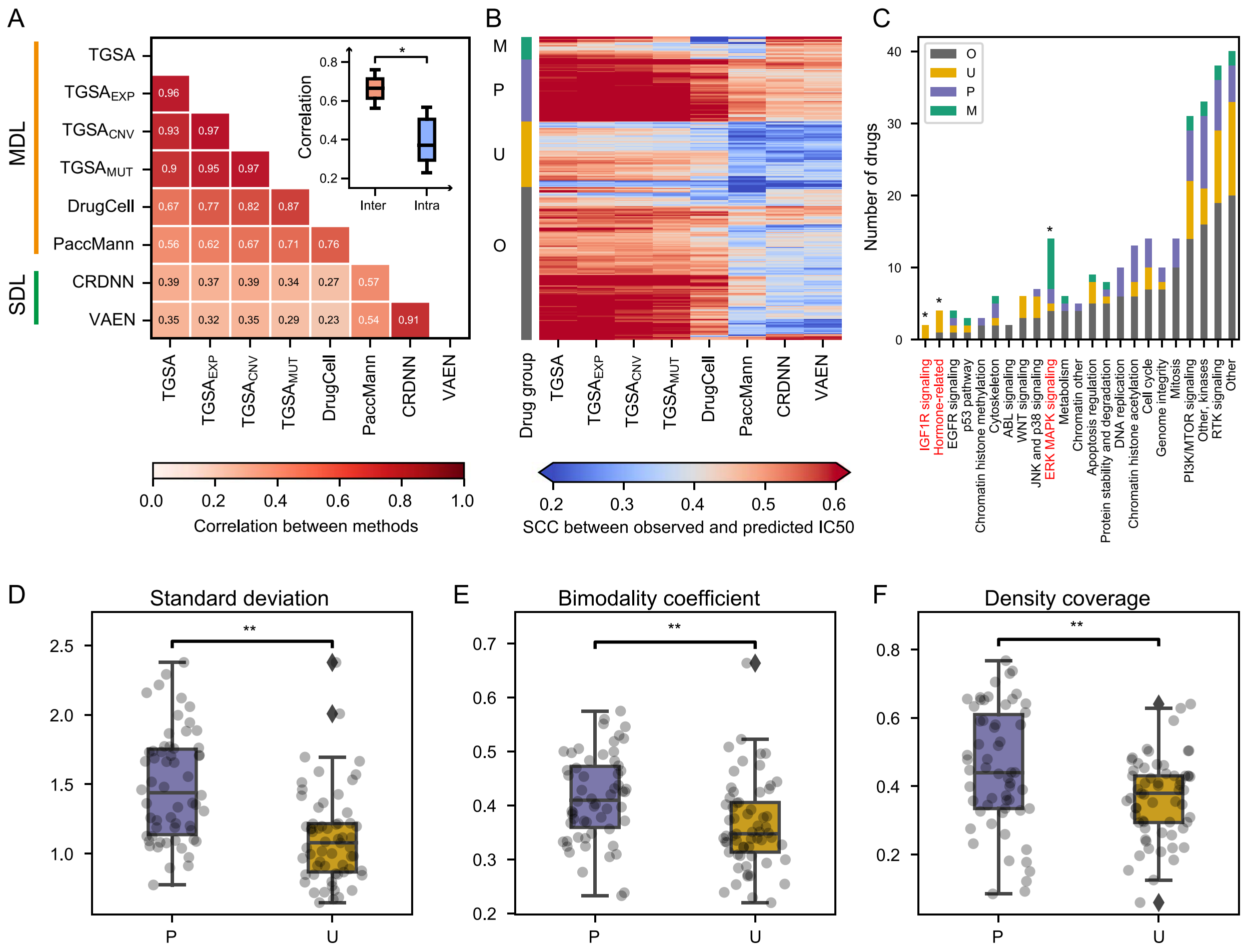

Figure 4. Predictability of individual drugs. (**A**) The consistency of DL methods. The colors represent the SCC values between any two methods. Intra-paradigm consistency was significantly higher than inter-paradigm consistency (Mann–Whitney test, * $P < 0.05$). (**B**) The heatmap of SCC values shows the different predictability among drugs. Drugs were categorized into four groups. M: drugs that were difficult to predict using mutation status but easy to predict by expression profiles, P: predictable drugs, U: unpredictable drugs, O: others. (**C**) Pathway enrichment was applied to assess the potential association between drug predictability and mechanism of action (Fisher's exact test, * $P < 0.05$). (**D**-**F**) depict the difference between the IC50 distribution of the P and U groups: standard deviation (**D**), bimodality coefficient (**E**) and density coverage (**F**) (Mann–Whitney test, ** $P < 0.01$).

Supplementary Figure S5). Regarding the effect size and P-value of predicted IC50 values between responders and non-responders, for some drugs, adequate performance (effect size >0, P-value <0.05) was obtained with some models (Figure 5B): PaccMann for doxorubicin; TGSA for cisplatin, etoposide and tamoxifen; TGSA$_{EXP}$ for bicalutamide, cisplatin and sorafenib; CRDNN for cisplatin, doxorubicin, etoposide, gemcitabine and vinorelbine. This implies that the performance of different models on different drugs may vary greatly as a result of the huge discrepancy between cell lines and patients caused by experimental bias, biological contexts and so on. Compared with other models, CRDNN surpassed the statistical significance threshold on 5 out of 16 drugs in the TCGA dataset, indicating that the end-to-end SDL model might be more powerful for discriminating tumor samples (Figure 5B). Unexpectedly, more than half of the ROC_AUC values of DrugCell and TGSA$_{MUT}$ were under 0.5, which means these models tended to mistake the responders and non-responders (Figure 5A). Given that these two models only used somatic mutation profiles, we compared the total number of mutations for cell lines and patients. As seen in Supplementary Figure S6, TCGA patients carried less mutations than GDSC cell lines, which may result from different mutation calling pipelines and control samples. Therefore, consistent data preprocessing and normalization are important when transferring models to other datasets.

Among the 16 drugs mentioned above, a statistically significant difference between response and non-response patients was observed for 9 drugs, including etoposide and cisplatin, while the predicted responses were undistinguishable for the other 7 drugs regardless of the methods used (Figure 5B). This phenomenon occurs even using the latest transfer learning technologies that were published recently [36–38] (Supplementary Table S3). There are a number of possible reasons for this outcome, such as differences between cell lines and patients, the lack of drug sensitivity related features and the reduction of statistical power due to small sample sizes. To this end, we built DNN classifiers directly from the expression profiles of patients instead of cell lines. Eight drugs with sample sizes larger than 80 were chosen, and the performance was measured by ROC_AUC (Figure 5C). Patient response to etoposide and cisplatin were well predicted by both cell line- and patient-trained models, while patient response to 5-fluorouracil and gemcitabine were difficult to predict whether using cell line- or patient-trained models. Better predictability for temozolomide, paclitaxel and docetaxel was

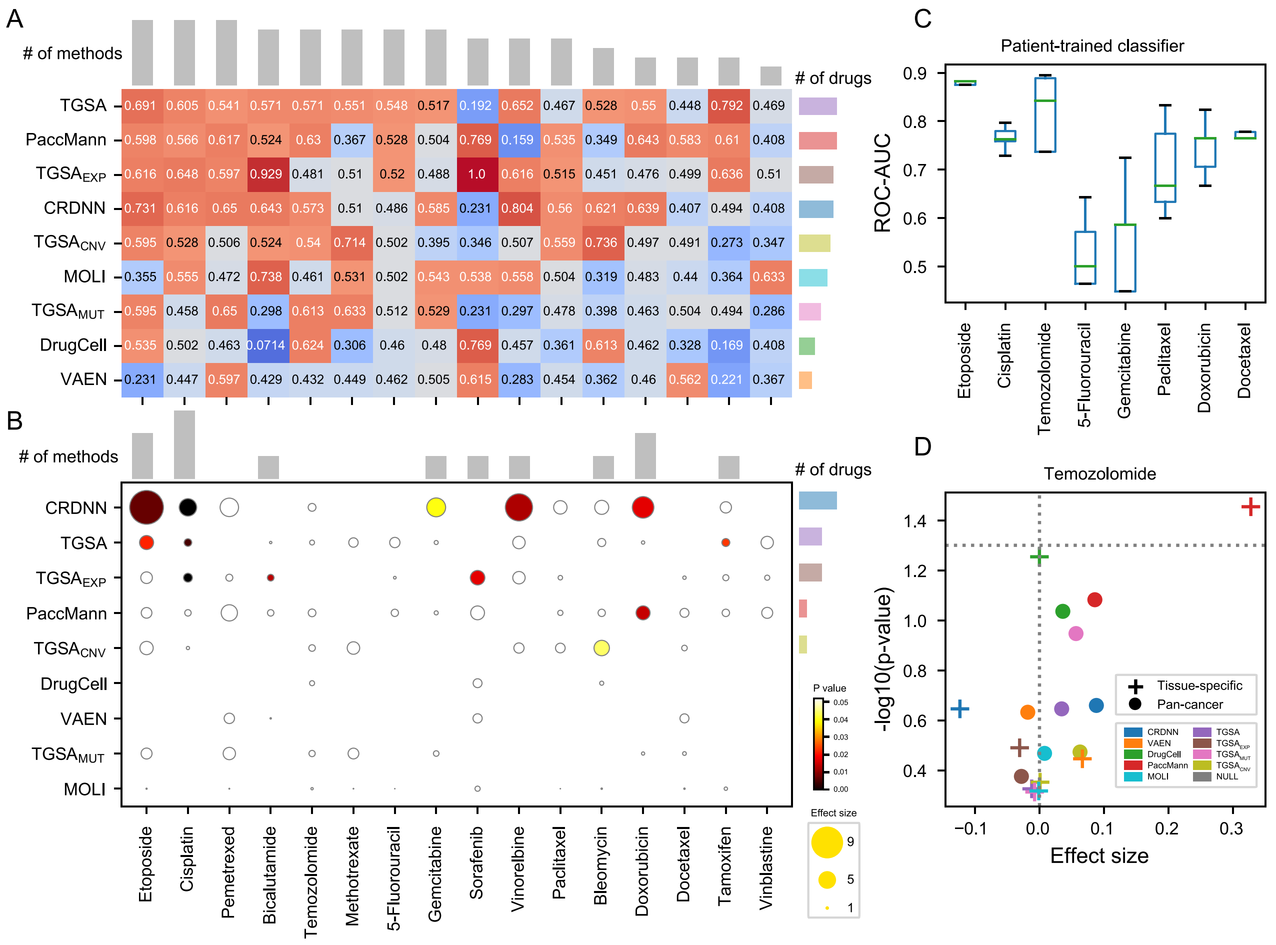

Figure 5. Assessment of model transferability on clinical cohorts. (**A**) and (**B**) depict the performance of 16 drugs trained on GDSC cancer cell lines and tested on TCGA patients. (**A**) The heatmap plots of the ROC_AUC values for each drug. The barplots on the upper and right panels depict the number of methods and drugs with the ROC_AUC values larger than 0.5, respectively. (**B**) The performance in terms of *P*-value and effect-size. The bar plots on the upper and right panels depict the number of methods and drugs with an effect size larger than 0 and *P*-value small than 0.05, respectively. (**C**) The boxplot shows the ROC_AUC values for each drug assessed by five-fold cross validation for models trained directly on patient data. (**D**) The volcano plot shows the prediction results of temozolomide using pan-cancer cell lines or cell lines of the central nervous system.

achieved using patient-trained models than by cell line-trained models; in particular, the mean ROC_AUC of temozolomide was above 0.8 for the patient-trained model. We noted that patients treated with temozolomide in TCGA mainly suffer brain tumor (Supplementary Figure S5). Since temozolomide is the leading compound for the treatment of brain cancer in clinical practice [39], we suppose that models built for specific tissue types might be more suitable than pan-cancer model. As shown in the volcano plot (Figure 5D), statistical significance was improved for some models when trained only on the cell lines cultured from central nervous system tumors, especially PaccMann, which yielded statistically significant results.

To improve the robustness of method comparison, another patient cohort treated with bortezomib was used to evaluate transferability of DL methods [35]. CRDNN was still the best model on this dataset. The predicted sensitivity of bortezomib by CRDNN are significantly different between responders and non-responders ($P = 0.047$). TGSA and PaccMann can distinguish responders and non-responders to some extent (*P*-values are 0.06 and 0.07, respectively).

Although batch correction by pooling training and validation datasets together is common when applying prediction models to external datasets [14, 40], there is potential risk of data leak. Another strategy called 'reference-batch ComBat' was used [34], which do not see the patient data before model training and there is no risk of data leak. The new results are largely consistent with previous results from the 'standard ComBat' (Supplementary Figure S7). CRDNN maintains the best performance no matter using 'reference-batch ComBat' or 'standard ComBat'. The number of drugs with significant difference between responders and non-responders decrease from 9 to 7. The slight performance drop may be due to no data leak of the stricter 'reference-batch ComBat'.

In summary, deep learning models trained from cancer cell lines hold promise for transferring to clinical patients, but the accuracy varies based on the drugs and prediction methods utilized. New approaches are needed to address the challenge from preclinical models to clinical applications.

Discussions

In this study, we systematically assessed DL methods for drug response prediction, including three SDL methods and three MDL methods, from various perspectives: overall and drug level performance, computational efficiency and transferability on TCGA clinical data (Table 2). Generally, MDL models achieved promising

Table 2. Summary of benchmark results: computational efficiency, method performance on cell lines and patients

	DrugCell	PaccMann	TGSA	CRDNN	VAEN	MOLI
Computational efficiency	**	****	*****	*	***	*
Accuracy	*****	**	*****	****	***	*
TCGA transferability	*	***	**	***	**	*

More asterisks indicate better on the addressed item.

performance on cancer cell lines with less time consumption, and could be used to predict new drugs that were unseen in the training datasets. However, current MDL models require the molecular representation of drugs; thus, they cannot be directly used for monoclonal antibody drugs. The SDL model is very slow when predicting a large number of drugs. However, in regard to clinical applications that estimate the response of a drug, the performance of CRDNN is relatively robust. Moreover, our study provides some directions for the development of more effective methods for drug response prediction.

First, there are large differences among drugs in terms of their IC50 values. According to the result of variation partitioning, 73.6% of the true IC50 labels is explained by the drug, and intuitively, it is difficult to learn the discrepancy among cell lines. Meanwhile, the ablation study demonstrated DEB that contributed more to the final performance of MDL models. Therefore, how to balance the information strengths from CEB and DEB through PM to enforce the ability to distinguish different cell lines may be a key point. Several concepts from intermediate fusion methods, such as attention-based and bilinear pooling-based fusion [41], might be beneficial.

Second, graph embedding of cell lines beyond the Euclidean space might enhance the expressive power of CEB. DrugCell directly mapped the neurons of a deep neural network into the Gene Ontology Biological Process hierarchy and constrained that the information only flowed from child subsystems to parent systems; TGSA adopted a typical message passing neural networks framework to gather information from neighbors. These two models utilized biomedical graphs to represent cell line status, breaking new ground for drug response prediction. Of note, graphs of different cell lines shared the same topology structure in either of the above methods. If variable-structure graphs are used for representing cell lines, the model may learn more information.

Third, the reproducibility and transferability of prediction methods are important for clinical applications. Data preprocessing, such as the normalization of omics data, tissue-specific modelling, and the computational alignment of patients and cell lines, should be considered for the predictive improvement. In previous DL methods, batch effects were simply removed before model construction, but researchers now make forays into transductive learning. Ma et al. designed a few-shot learning model aiming to identify applicable input features on both cell lines and patients through its 'pretraining' and 'few-shot learning' two-phase training process [42]; Velodrome adapted the cell line and patient domain by the object function, combining a supervised loss for accuracy, an alignment loss for generalization and a consistency loss for invariant latent space [38]. New transfer learning technologies will bring a brighter future for the prediction of clinical drug response.

Finally, single-cell sequencing makes it possible to explore the response of cell subpopulations. Although it is difficult to directly apply the existing drug sensitivity prediction methods to single-cell data, a few studies have already started to explore this topic. They predicted drug responses by generating a 'pseudo-bulk' expression profile from single-cells [43] or developed drug combinations to target drug-tolerant cell subpopulations [44]. In the future, better approaches to address single-cell drug responses may be developed under new computational frameworks such as multi-instance learning [45].

Key Points

- Representative DL models for drug response prediction were systematically assessed in terms of computational consumption, prediction performance and transferability on clinical cohorts. On large-scale cell line data, TGSA, which has a low time cost and high accuracy, is recommended; while for single-drug applications in clinical scenarios, CRDNN shows better performance.
- Variation partitioning and ablation studies revealed that it is more difficult to capture the differences among cell lines. Improved architectures such as graph embedding utilizing biomedical prior knowledge are important to represent cell lines for personalized response prediction.
- It is a challenging task to transfer drug response prediction models trained by cell line data to patients due to discrepancy in biology context, experimental conditions and other factors. New transfer learning techniques are urgently needed for clinical applications.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib. The codes and instruction for implementing and comparing DL methods are available at https://github.com/LihongLab/Suppl-data-Benchmark. The data used for training andvalidating the drug sensitivity prediction models are available at https://zenodo.org/record/7264573#.Y16Ed3ZByUl.

Acknowledgments

We thank Xin Li (Shanghai Institute of Nutrition and Health, CAS) and Wenju Cai (Guizhou science data center Gui'an Supercomputing Center) for providing computational resource.

Funding

This work is supported by National Natural Science Foundation of China (32170680, T2122018), Natural Science Foundation of Shanghai (21ZR1476000), National Key R&D Program of China (2021YFF1200902), CAS Youth Innovation Promotion Association (2018307) and the development fund for Shanghai talents.

References

1. Barretina J, Caponigro G, Stransky N, *et al.* The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.

2. Garnett MJ, Edelman EJ, Heidorn SJ, *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**:570–5.
3. Basu A, Bodycombe NE, Cheah JH, *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**:1151–61.
4. Seashore-Ludlow B, Rees MG, Cheah JH, *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**:1210–23.
5. Costello JC, Heiser LM, Georgii E, *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**:1202–12.
6. Chen J, Zhang L. A survey and systematic assessment of computational methods for drug response prediction. *Brief Bioinform* 2021;**22**(1):232–246.
7. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Brief Bioinform* 2020;**22**:360–79.
8. Fangyoumin F, Bihan S, Xiaoqin M, *et al.* Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. *J Genet Genomics* 2021;**48**:540–51.
9. Firoozbakht F, Yousefi B, Schwikowski B. An overview of machine learning methods for monotherapy drug response prediction. *Brief Bioinform* 2022;**23**(1):bbab408.
10. Guvenc Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Brief Bioinform* 2019;**22**:346–59.
11. Xia F, Allen J, Balaprakash P, *et al.* A cross-study analysis of drug response prediction in cancer cell lines. *Brief Bioinform* 2022;**23**(1):bbab356.
12. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, *et al.* Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief Bioinform* 2021;**22**(6):bbab294.
13. Chen Y, Zhang L. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? *Brief Bioinform* 2022;**23**(1):bbab378.
14. Sakellaropoulos T, Vougas K, Narang S, *et al.* A deep learning framework for predicting response to therapy in cancer. *Cell Rep* 2019;**29**:3367–3373.e4.
15. 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**:i501–9.
16. Jia P, Hu R, Pei G, *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* 2021;**12**:1740.
17. An X, Chen X, Yi D, *et al.* Representation of molecules for drug response prediction. *Brief Bioinform* 2022;**23**(1):bbab393.
18. Chiu YC, Chen HH, Zhang T, *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics* 2019;**12**:18.
19. Chang Y, Park H, Yang H-J, *et al.* Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep* 2018;**8**:8857.
20. 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**:408.
21. Nguyen TT, Nguyen GTT, Nguyen T, *et al.* Graph convolutional networks for drug response prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**19**(1):146–154.
22. 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* 2019;1–1.
23. Manica M, Oskooei A, Born J, *et al.* Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol Pharm* 2019;**16**:4797–806.
24. Kuenzi BM, Park J, Fong SH, *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020;**38**:672–684.e676.
25. Zhu Y, Ouyang Z, Chen W, *et al.* TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics* 2021;**38**(2):461–468.
26. Liu Q, Hu Z, Jiang R, *et al.* DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;**36**:i911–8.
27. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 2013.
28. Knijnenburg TA, Klau GW, Iorio F, *et al.* Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci Rep* 2016;**6**:36812.
29. Borcard D, Legendre P, Drapeau P. Partialling out the spatial component of ecological variation. *Ecology* 1992;**73**:1045–55.
30. Corsello SM, Nagari RT, Spangler RD, *et al.* Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nature cancer* 2020;**1**:235–48.
31. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 2002;**20**:422–46.
32. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* 2016;**32**:2891–5.
33. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
34. Zhang Y, Jenkins DF, Manimaran S, *et al.* Alternative empirical Bayes models for adjusting for batch effects in genomic studies. *BMC Bioinformatics* 2018;**19**:1–15.
35. Mulligan G, Mitsiades C, Bryant B, *et al.* Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* 2007;**109**:3177–88.
36. Peres da Silva R, Suphavilai C, Nagarajan N. TUGDA: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from in vitro to in vivo settings. *Bioinformatics* 2021;**37**(Suppl 1):i76–i83.
37. Mourragui SMC, Loog M, Vis DJ, *et al.* Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proc Natl Acad Sci U S A* 2021;**118**(49):e2106682118.
38. Sharifi-Noghabi H, Harjandi PA, Zolotareva O, *et al.* Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence* 2021;**3**:962–72.
39. *FDA Approved Drug Products: TEMODAR (temozolomide) capsules and injection*. https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/022277s013lbl.pdf (30 July 2022, date last accessed).
40. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines. *Genome Biol* 2014;**15**:R47.
41. Zhang C, Yang Z, He X, *et al.* Multimodal intelligence: representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing* 2020;**14**:478–93.
42. Ma J, Fong SH, Luo Y, *et al.* Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer* 2021;**2**:233–44.
43. Chawla S, Rockstroh A, Lehman M, *et al.* Gene expression based inference of cancer drug sensitivity. *Nat Commun* 2022;**13**:1–15.

44. Aissa AF, Islam AB, Ariss MM, *et al.* Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun* 2021;**12**:1–25.

45. Carbonneau M-A, Cheplygina V, Granger E, *et al.* Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recognition* 2018;**77**:329–53.