OXFORD

Briefings in Bioinformatics, 2022, **23**(3), 1–10
https://doi.org/10.1093/bib/bbac100
Problem Solving Protocol

DeepTTA: a transformer-based model for predicting cancer drug response

Likun Jiang[†], Changzhi Jiang[†], Xinyu Yu, Rao Fu, Shuting Jin and Xiangrong Liu
Corresponding author: Xiangrong Liu, Department of Computer Science, Xiamen University, Information School, No.1 Zengcuo'an West Rod, Xiamen City 361005, Fujian Province, China. Fax: +86-592-2580258; E-mail: xrliu@xmu.edu.cn
[†]These authors contributed equally to this work.

Abstract

Identifying new lead molecules to treat cancer requires more than a decade of dedicated effort. Before selected drug candidates are used in the clinic, their anti-cancer activity is generally validated by *in vitro* cellular experiments. Therefore, accurate prediction of cancer drug response is a critical and challenging task for anti-cancer drugs design and precision medicine. With the development of pharmacogenomics, the combination of efficient drug feature extraction methods and omics data has made it possible to use computational models to assist in drug response prediction. In this study, we propose DeepTTA, a novel end-to-end deep learning model that utilizes transformer for drug representation learning and a multilayer neural network for transcriptomic data prediction of the anti-cancer drug responses. Specifically, DeepTTA uses transcriptomic gene expression data and chemical substructures of drugs for drug response prediction. Compared to existing methods, DeepTTA achieved higher performance in terms of root mean square error, Pearson correlation coefficient and Spearman's rank correlation coefficient on multiple test sets. Moreover, we discovered that anti-cancer drugs bortezomib and dactinomycin provide a potential therapeutic option with multiple clinical indications. With its excellent performance, DeepTTA is expected to be an effective method in cancer drug design.

Keywords: drug response prediction, cancer cell line, IC50, transformer, gene expression, transcriptome

Introduction

Drug development based on target proteins has been a successful approach over the past decades, but these methods cannot address diseases that lack well-defined protein targets [1]. Especially for cancer, although there are various mutations in tumor tissues, many drugs still have no specific target. Moreover, the intra- and inter-tumoral heterogeneity result in diverse anti-cancer drug responses among patients [2, 3], which illustrated the complexity of genomics. Recently, the accumulation of cancer multi-omics data provides an opportunity to understand how a tumor's omics characteristics can affect its responses to drugs [4]. For example, genomic, transcriptomic, proteomic and methylomic data have been shown to be successful in predicting drug response [5–8]. In the multi-omics data, transcriptome profiling can detect changes in gene activity and regulation by capturing quantitative expression patterns and has the capacity to describe the underlying phenotypes in great detail [9]. Changes in gene activity can be regarded as surrogates for many phenotypes, such as inflammation, vascularization, apoptosis [10], proliferation [11] and genomic instability [12]. Tumor transcriptome data can successfully acquire these traits influences important clinical variables, such as growth rate, metastatic potential [13] and response to drugs, and ultimately determines clinical progression and outcomes [14]. In the previous study, researchers have already used the transcriptome for drug discovery and generation [1, 15, 16].

Driven by advances in high-throughput technologies, sequencing technology has promoted freely available datasets for cancer cell drug sensitivity information and molecular markers of drug responses [17]. Genomics of Drug Sensitivity in Cancer (GDSC) are the most popular datasets in this field [18], providing researchers with multi-omics profiles including genomic, transcriptomic, proteomic and methylomic data. Besides, the half-maximal inhibitory concentration (IC50) is a common indicator reflecting drug response across cancer cell lines, and drug information is also recorded in GDSC. With the free and available big data, a handful of computational models have integrated omics and chemical descriptors to predict cell line drug sensitivity using a variety of methods including but not limited to matrix factorization [19], trace norm regularization

Likun Jiang is a PhD student at Xiamen University. His research interest is the classification of proteins in bioinformatics.
Changzhi Jiang is a PhD student at Xiamen University. Her research interest is information mining of biological networks in bioinformatics and drug discovery.
Xinyu Yu is a graduate student at Xiamen University. His research interest is bioinformatics.
Rao Fu is a graduate student at Xiamen University. His research interest is information mining of biological networks in bioinformatics.
Shuting Jin is a PhD student at Xiamen University. Her research interest is information mining of biological networks in bioinformatics and drug discovery.
Xiangrong Liu is a professor at Xiamen University. His research interests include bioinformatics and data mining.
Received: November 23, 2021. **Revised:** February 8, 2022. **Accepted:** February 27, 2022

[20], link predictions [21], collaborative filtering [22, 23], logistic regression [24], multi-layer neural networks and random forests [25], kernelized Bayesian matrix factorization [25], Pearson correlation-based similarity networks [26], Kronecker product kernel combined with support vector machines (SVMs) [27], autoencoders with elastic network and SVMs [28]. Those studies have shown that transcript data are the most useful data type for cancer drug response (CDR) prediction [24, 29, 30]. Besides, previous studies have used simple fingerprint descriptors [7, 25] to represent drugs, and there may be no information available for drug response prediction [15]. Some researchers use genomic mutation data by sparse matrix features [5, 25, 31, 32] for drug response prediction, but there was no significant effect. There are several deep learning models for CDR prediction tasks, such as DeepCDR [5], tCNNS [8], MOLI [6] and CDRscan [7]. CDRscan is the first model that adopts deep learning methods to predict the drug responses. However, only using genomic mutation as a cancer cell profile is not enough to predict the drug responses. tCNNs adopt two convolution networks to learn the representations of drugs and genomic mutation data; its performance has room for improvement. MOLI integrates the multi-omics data and introduces a triplet loss. DeepCDR is the state-of-the-art model in CDR predictions, which adopts a uniform graph conventional network (UGCN) to learn the representations of drugs and integrates the multi-omics data to improve the performance of predicting IC50. However, it is limited to capturing the structures of drugs using the UGCN modules and the model that used mutation features is too sparse.

We envisage an effective model that can extract and integrate drug and cancer cell transcriptome information, which makes it easier to find effective drugs for cancer treatment. Inspired by the natural language processing model, the transformer consists of a series of self-attention modules, which are effective in natural language representation and have been proved in various NLP models [33–36]. We use transformer architecture mining drug characteristics through compounds substructure, by which capture more structured information of SMILES and learn better representations for drugs. Finally, we construct an end-to-end deep learning model DeepTTA, which is composed of a drug feature extraction part based on transformer [36] and four-layer neural network transcriptome feature extraction sub-model. By comparing with existing advanced models such as DeepCDR, tCNNS, MOLI and CDRscan, our model DeepTTA achieved state-of-the-art performance in CDR prediction. In addition, prediction when training on binary datasets compared to other models, we observed a significant improvement in the area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPR) drug sensitivity prediction of DeepTTA.

Methods

Data

In this study, we use the GDSC database (www.cancerRxgene.org), which is a large public resource for information on drug sensitivity in cancer cells and molecular markers of drug response [17]. According to the official recommendation, we use the new version GDSC2, containing 135 242 pairs with IC50 consisting of 809 cell lines and 198 compounds. All the cell lines belong to 31 cancer types, an average of 167 drugs were used per cell line and each drug was tested on an average of 683 cell lines, respectively. We use the Pubchem ID recorded in GDSC to get the line notation of simplified molecular-input line entry system (SMILES) from Pubchem with a python package 'pubchempy' (https://github.com/mcs07/PubChemPy), which was initially used to extract the structural and chemical features of each drug. However, 44 drugs were not registered with PubchemID, leaving 154 drugs.

For the omics data, we selected transcriptional data from GDSC2, which was analyzed as described on the Human Genome U219 96-Array Plate using the Gene Titan MC instrument (Affymetrix). The robust multi-array analysis algorithm [32] was used to establish intensity values for each locus. Finally, there are 17 777 gene expression values for each cell line. With removing unrecorded instances, 805 cell lines remained. Each pair of cell line and compound as an instance, a total of 103 492 instances were employed to develop the deep learning models. To compare different models under uniform standards, as the data segmentation method of DeepCDR [5], we use the following methods for data segmentation:

- 'Random' segmentation strategy: to be consistent with the comparison model, for the 103 492 instances across 154 drugs and 805 cell lines recorded in GDSC, we randomly select 95% as train dataset and 5% as test dataset.
- 'According Cancer' segmentation strategy: to explore our model more suitable for prediction which cancer type or drug, we randomly selected 80% of 103 492 instances spanning 30 cancer types to train the model, the left 20% for case prediction across multiple TCGA cancer or different drugs.
- 'Missing' segmentation strategy: theoretically, 154 drugs and 805 cell lines would be performed 123 970 ($154 * 805$) experiments, but GDSC only contains 103 492 CDRs, with about 17% of CDR pairs missing. Therefore, we use the existing data as the training set to predict the unknown CDR pairs IC50.
- 'Independent' segmentation strategy: as described in the 'Random' segmentation strategy, the drug or cell lines in the test set may be included in the training set. To verify the comprehensive strength of the DeepTTA model, we carried out leave-drug-out and leave-cell-line-out data segmentation strategy.

For leave-drug-out strategy, we randomly selected 80% (123) drugs associated CDRs (82 795) as train sets and the remaining 20 679 CDRs associated with 20% (31) drugs as test sets. And we do the same for leave-cell-line-out strategy, the details are shown in Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

Model structure

We proposed an end-to-end deep learning model to predict the anti-cancer drug response. The structure of the prediction model is shown in Figure 1, named DeepTTA. The two main parts of the model are drug feature encoding by transformer model and multiple neural network extract cell line information. In the first part, the drug's SMILES was treated as a sequence divided into substructure, we fixed the length of the drug numerical vector to 50 based on the max substructure number of the drug, and it will be filled with 0 when it is shorter than 50. Then, substructure sequence vectors are fed into the neural network based on a transformer encoder to get representation vectors of a drug. For the cell line, we extracted features from gene expression data through four-layer neural network, of which three hidden layers include 1024, 256 and 64 neural units, separately. Finally, the high-level features of drugs and transcript data were concatenated together and fed into a classifier network with four fully connected layers. We used the Adam optimizer with a batch size of 128 and learning rate of 1e−3. Model constructs were implemented in Pytorch (https://pytorch.org/). We trained the model 300 epochs with 2 GeForce RTX 2080 Ti GPU and Intel Xeon CPU E5-2620-v4 2.10GHz. For details, please see the website: https://github.com/jianglikun/DeepTTC.

Drug feature mining by transformer architecture

To avoid the disadvantages of using molecular descriptors, we use Explainable Substructure Partition Fingerprint (ESPF) [37], which can decompose drugs into a discrete set of moderate-sized substructures that have strong predictive values. According to the principle of similarity [38, 39], molecules that induce similar effects in cell line gene expression will have similar molecular structures or share some pharmacophore signatures to some extent. Inspired by the sub-word units [40] in the language processing domain, ESPF is based on the Byte Pair Encoding algorithm [41]. By using the results of this work [37], ∼2700 substructures were obtained through ∼2 million ChEMBL drug SMILES, which has already proved good predictive performance.

After segment SMILES expression of the drug, the sequence information was involved to substructures by different granularities. To get inspiration from natural language processing, we use transformer model, the state-of-the-art deep learning architecture [36], for molecular representation learning to extract contextual information from the SMILES sequence. And we refer to the approach of handing drugs, which has proven to be useful in drug target interaction problems [42]. We first constructed a vocabulary set D that included different SMILES strings characters and tokenized the entire drug corpus. We defined the tokenized set as T. Then, the tokenized set T was updated with the new token that was the most frequent consecutive tokens. We repeated the update operation until no frequent token exceeded the threshold μ or the size of D reached the maximum length δ. After that, we could obtain a substructural sequence $\mathbf{S} = \{S_1, \ldots, \mathbf{S_l}\}$ of a drug with size l, where $S_i \in \mathbf{T}$. To capture the contextual chemical semantics information, we adopt the transformer encoders [36] to learn the representations of substructures. Therefore, for each input drug, we defined a matrix $M^s \in \mathcal{R}^{l\times\zeta}$ to represent the substructural sequence S, where l was the size of substructures or the cardinality of the vocabulary set D, and ζ was the maximum length of substructure sequence for the drug. The ith column M_i^s was a one-hot vector that represented the substructure index for the ith substructure of the drug sequence. We generate the content representation C_i for each drug using a learnable dictionary lookup matrix $W_c \in \mathcal{R}^{\gamma\times l}$:

$$C_i = W_c M_i^s \tag{1}$$

where γ was the size of representation for each substructure. To capture the positional information of substructures for drug, a positional representation P_i was calculated via a lookup dictionary [36] $W_{pos} \in R^{\gamma\times\zeta}$:

$$P_i = W_{pos} I_i \tag{2}$$

where $I_i \in R^{\zeta}$ was a one-hot vector where the ith position is 1. Therefore, we generated the new representation $\mathbf{e}_i$ via the sum of content and positional representations:

$$\mathbf{e}_i = C_i + P_i \tag{3}$$

After that, we could obtain representations of all substructure. To learn the chemical relationships between these substructures, we introduced a transformer encoder layer to encode representations of substructures. A transformer encoder layer included two sublayers, i.e. a multi-attention layer and a fully connected feed-forward network. The multi-attention layer was ensembled by several different self-attention layers. The new representation $\mathbf{e}_i$ was input into the multi-attention layer as follows:

$$\text{Attention}(\mathbf{e}_i) = \text{softmax}\left(\frac{(\mathbf{e}_i W^Q)(\mathbf{e}_i W^K)}{\sqrt{d}}\right) \times (\mathbf{e}_i W^V) \tag{4}$$

where $W^Q \in R^{\zeta\times\eta}$, $W^K \in R^{\zeta\times\eta}$ and $W^V \in R^{\zeta\times\eta}$ were learnable weight parameters, and $\frac{1}{\sqrt{d}}$ was a scaled factor.

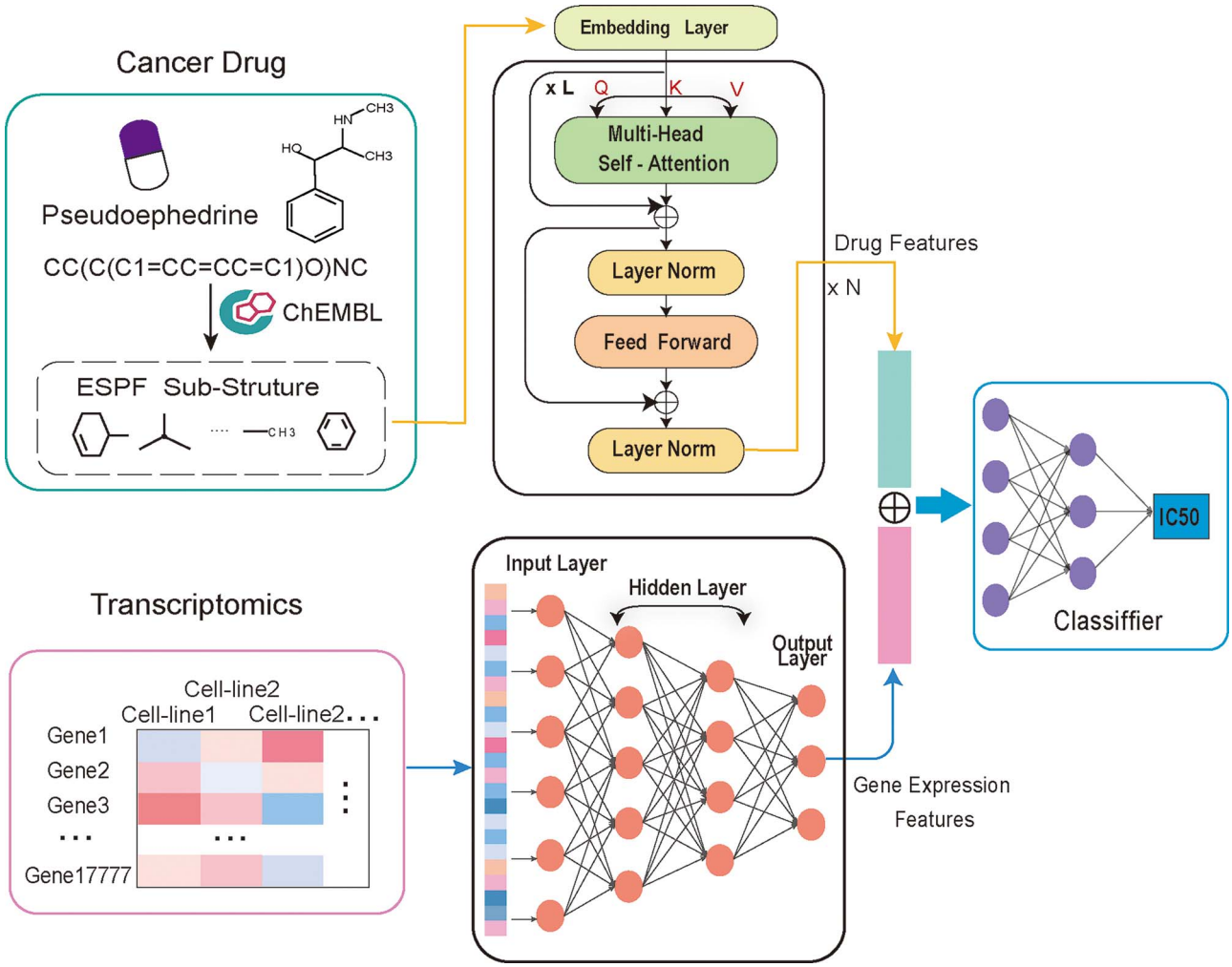

Figure 1. The structure of DeepTTA. It consists of three main parts, including drug feature mining, gene expression features extracted and classifier. In drug feature mining, we segment the drug into a substructural sequence S = {S_1, … .,S_l} by ESPF. After that, we introduced a transformer encoder layer to encode representations of substructures. There, we employ $L = 8$ parallel attention layers or heads, and composed of a stack of $N = 6$ identical layers. And new representation was achieved by multi-attention layers. For the cell line, we extract features from gene expression data by a four-layer neural network where three hidden layers include. Finally, the high-level features of drugs and transcript data were concatenated together and fed into a classifier network with four fully connected layers.

Then, the output of the multi-attention layer was input into the fully connected feed-forward layer as follows:

$$E = \max\left(0, \text{Attention}\left(e_i\right) W_1 + b_1\right) W_2 + b_2 \tag{5}$$

where $W_1 \in R^{\eta \times \varphi}$, $b_1 \in R^{\varphi}$, $W_2 \in R^{\varphi \times \phi}$, $b_1 \in R^{\phi}$ were learnable parameters. Hence, we obtained the final representation of each drug.

Evaluation criteria

The ground truth IC50 value between drugs and cancer cell lines was experimentally obtained and the natural logarithm conversion was carried out. For the model evaluation of regression prediction, we use the traditional measure root mean square error (RMSE) values to calculate the level of accuracy.

To measure the linear correlation between true and prediction IC50, we used the Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC). The calculation formula of the indicators is shown as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{6}$$

$$\text{PCC} = \frac{\text{cov}\left(y_i, \hat{y}_i\right)}{\sigma_{y_i}\sigma_{\hat{y}_i}} \tag{7}$$

$$\text{SRCC} = \frac{\sum_{i=1}^{m}\left(\text{rg}\left(y_i\right) - \overline{\text{rg}(y)}\right)\left(\text{rg}\left(\hat{y}_i\right) - \overline{\text{rg}\left(\hat{y}\right)}\right)}{\sqrt{\sum_{i=1}^{m}\left(\text{rg}\left(y_i\right) - \overline{\text{rg}(y)}\right)^2 \sum_{i=1}^{m}\left(\text{rg}\left(\hat{y}_i\right) - \overline{\text{rg}\left(\hat{y}\right)}\right)^2}} \tag{8}$$

where y_i and $\hat{y}_i$ were the true value of IC50 and predicted value of IC50, respectively. And rg demonstrated the rank, $\overline{\text{rg}(y)}$, $\overline{\text{rg}(\hat{y})}$ were the mean value of the $\text{rg}(y)$, $\text{rg}(\hat{y})$, respectively. And m represented the number of samples. We use the binarization of IC50 for drug sensitivity analysis and chose the AUROC, the AUPR and F1 score, three widely accepted measures of prediction accuracy in machine learning classifiers.

Table 1. Performance comparison of our method and existing methods

Model	Year	Instance	Pearson	Spearman	RMSE
MOLI	2019	—	0.813 ± 0.007	0.782 ± 0.005	2.282 ± 0.008
CDRscan	2018	152 549	0.871 ± 0.004	0.852 ± 0.003	1.982 ± 0.005
tCNNS	2019	172 114	0.910 ± 0.009	0.889 ± 0.008	1.228 ± 0.013
DeepCDR	2020	107 446	0.923 ± 0.005	0.898 ± 0.008	1.060 ± 0.033
DeepTTA	2021	103 492	**0.941 ± 0.003**	**0.914 ± 0.004**	**0.952 ± 0.007**

The best performance achieved by the model is shown in bold; MOLI and CDRscan did not provide the data; the model results come from comparative experiment [30]. The instance of MOLI was not recorded. We reproduced the model tCCNS and DeepCDR.

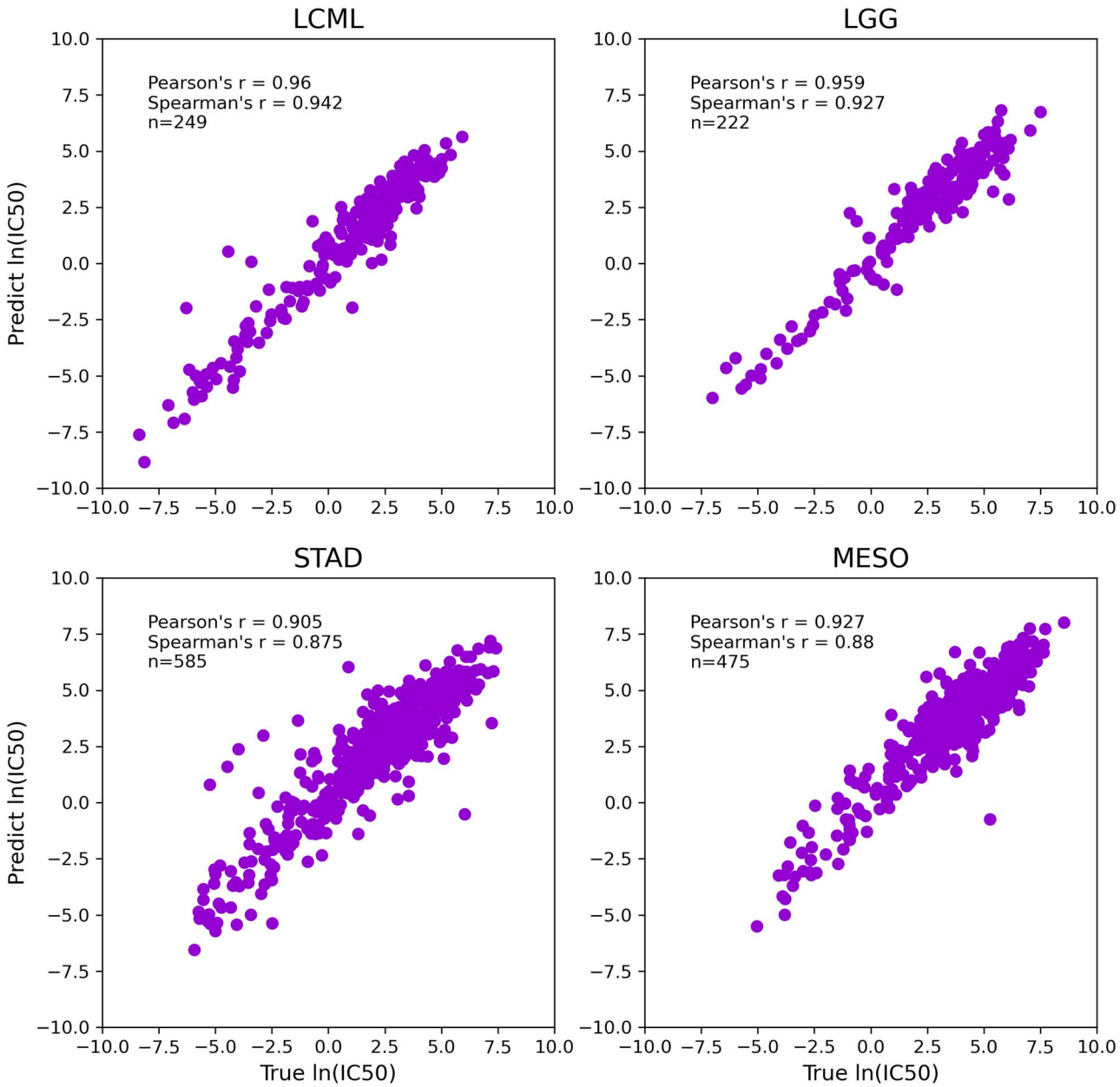

Figure 2. Performance of DeepTTA in different TCGA cancer types. The horizontal axis represents the DeepTTA predicted IC50 value, and the vertical axis represents the database record IC50 value. The best performance cancer types are chronic myelogenous leukemia (LCML) and brain lower grade glioma (LGG). The worst cancer types are stomach adenocarcinoma (STAD) and mesothelioma (MESO).

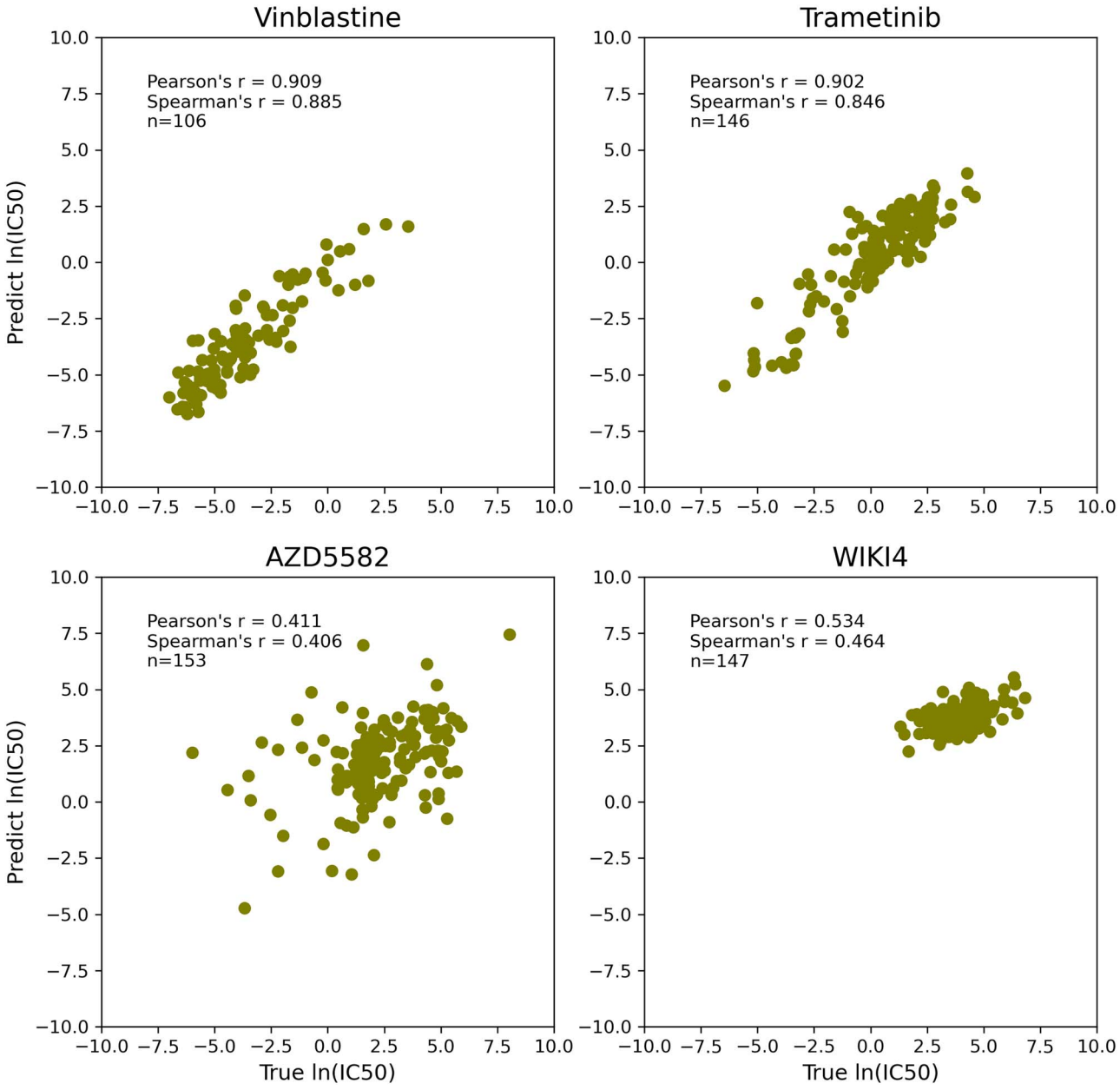

Figure 3. Performance of DeepTTA in different drugs. The horizontal axis represents the DeepTTA predicted IC50 value, and the vertical axis represents the database record IC50 value. The two most effective drugs are vinblastine and trametinib. The worst drugs are AZD5582 and WIKI4.

Results

Performance comparison of our method and existing methods

To evaluate the effectiveness of our proposed model, we compared the most advanced models on the GDSC datasets.

- CDRscan [7] employs a two-step CNN architecture, in which the genomic mutational fingerprints of cell lines and the molecular fingerprints of drugs are processed individually, then combined by 'virtual docking', an *in silico* modeling of drug treatment.
- tCNNS [8] adopts a CNN to learn representations for drugs from SMILES format and uses another CNN to extract features for cancer cell lines from the genetic feature vectors, respectively, to predict the interaction between the drugs and the cancer cell lines.
- MOLI [6] uses type-specific encoding sub-networks to learn features for each omics type (i.e. somatic mutation, copy number aberration and gene expression data) and concatenates them into one representation for drug response prediction.
- DeepCDR [5] is a hybrid graph convolutional network, which consists of a uniform graph convolutional network and multiple subnetworks. It integrates multi-omics profiles of cancer cells and explores the intrinsic chemical structures of drugs for predicting CDR. DeepCDR is by far the most reliable model in CDR prediction.

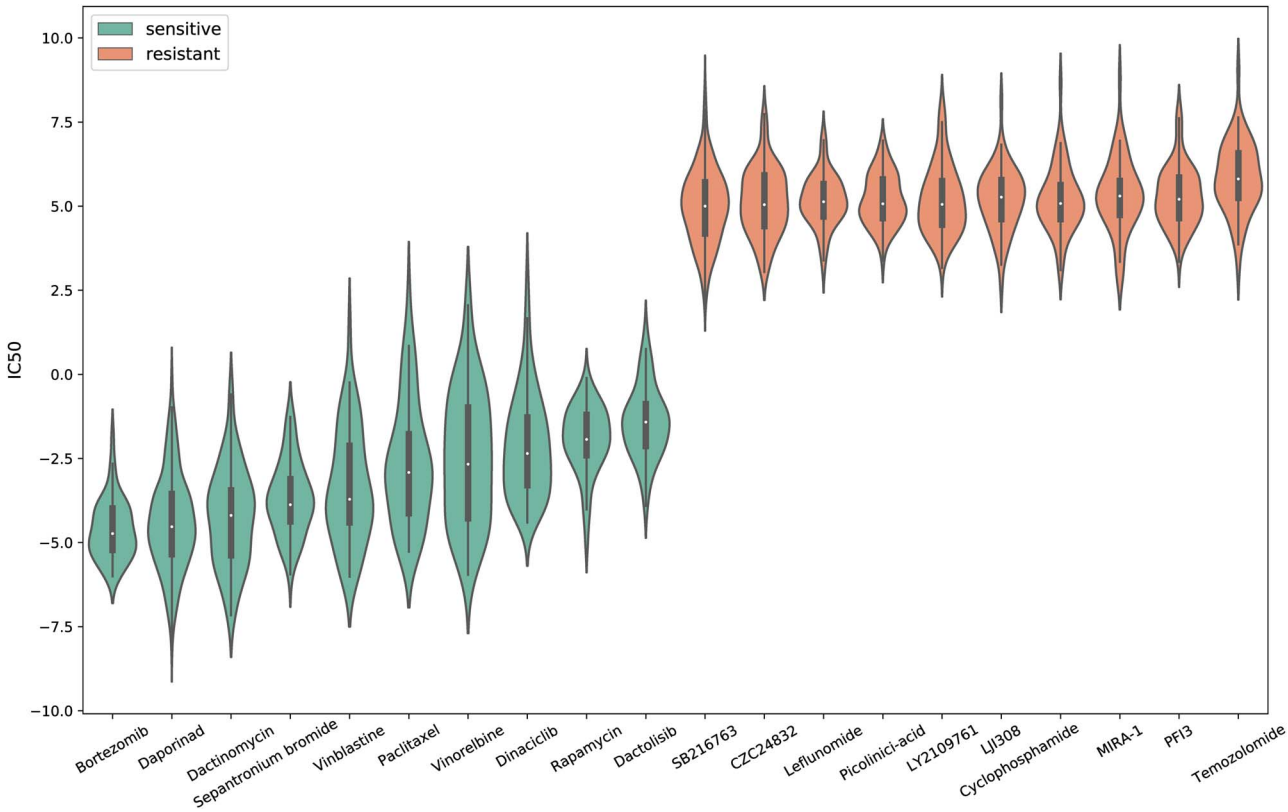

Figure 4. Predicting missing CDRs and sorting the average Ic50 values for each drug. This figure shows the top 10 'sensitivity' and last 10 'resistance' drugs, respectively. IC50 is a measure of the potency of a substance in inhibiting a specific biochemical function. We consider that the drug is more effective when IC50 is lower, so the drugs were called 'sensitivity'. Conversely, ineffective drugs were called 'resistance'.

Among the total of 103 492 instances across 31 cancer types, 95% of instances 98 211 for each cancer were selected to train on DeepTTA. The corresponding to 5% of total instances 5184 as a test on DeepTTA. For the compared model, we reproduced their work. We used three common metrics RMSE, PCC and SRCC as evaluation indicators to evaluate the performance of our model and existing models. The experimental results are shown in Table 1, indicating that our model DeepTTA achieved state-of-the-art performance among the existing models. Specifically, the lowest RMSE 0.952 indicates the most accurate prediction, while the highest PCC 0.941 and SRCC 0.914 proved a strong agreement between recorded and the predicted IC50 values.

Predicting cell line and drug

In order to assess the prediction accuracy of DeepTTA, we follow 'According Cancer' segmentation strategy and randomly selected 80% (82 703) of the total 103 492 instances spanning 30 cancer types to train the model, the left 20% (20 692) for case prediction across multiple TCGA cancer or different drugs. Then, we examined the Pearson's correlation between the observed IC50 in GDSC and the values predicted by DeepTTA using the test set. In all cell lines, the PCC score ranges from 0.905 to 0.959, which illustrates that DeepTTA achieved outstanding performance. Especially, in the best performance cancer types, chronic myelogenous leukemia and brain lower grade glioma achieved 0.960 and 0.959 (Figure 2). In the worst cancer types, stomach adenocarcinoma and mesothelioma got 0.905 and 0.927, respectively (Figure 2). All cancer types were recorded in Supplementary Table S2 available online at http://bib.oxfordjournals.org/.

From the point of view of drugs, the prediction effect of the model is uneven; the PCC ranges from 0.411 to 0.909. The PCC values of the two most effective regression prediction are 0.902 and 0.909 for camptothecin and oxaliplatin. The two worst drugs are AZD5582 (PCC = 0.411) and AIKI4 (PCC = 0.514), which are shown in Figure 3 and recorded in Supplementary Table S3 available online at http://bib.oxfordjournals.org/.

Predicting missing CDRs

As described in Data section, we follow 'Missing' segmentation strategy and applied 103 492 instances of all recorded data for model training and then predicted the 20 478 missing CDRs in the GDSC database (~17% of all pairs). By sorting the average IC50 value for each drug, Figure 4 shows the top 10 'sensitive' and last 10 'resistant' drugs, respectively. No coincidence, we found the same most effective drug as DeepCDR, bortezomib, which was proved to be effective in a variety of cancer cell line experiments or tumor treatment [43]. As explained in a review, bortezomib (Velcade, formerly

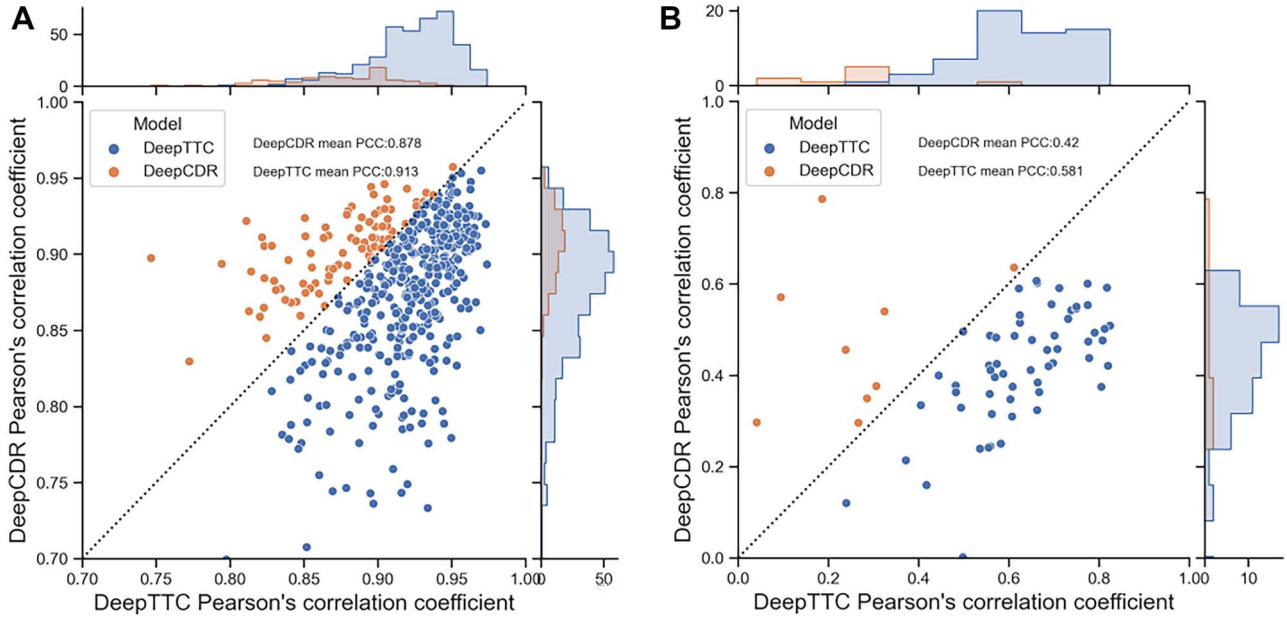

Figure 5. Comparison of PPC between DeepTTA and DeepCDR. The x-axis and y-axis of each dot represent the PCC of DeepTTA and DeepCDR, respectively. If DeepTTA's PCC is higher than DeepCDR, the dot is blue. Otherwise, the dot is orange. (**A**) Among the total share of 497 cell lines in both models, 393 (79%) was more accurately predicted by DeepTTA. (**B**) There are 69 drugs shared by DeepCDR and DeepTTA, of which DeepCDR achieved average PCC value of 0.42, and DeepTTA got a higher value of 0.58.

PS-341) represents the first proteasome inhibitor to show anti-tumor activity in both solid and hematological malignancies. It blocks activation of nuclear factor-kappa B (NF-κB), resulting in increased apoptosis, decreased production of angiogenic cytokines and inhibition of tumor cell adhesion to the stroma. Bortezomib has shown significantly *in vitro* activity, different clinical trials are currently ongoing in multiple myeloma as well as in many other hematologic and solid tumors (mantle cell and follicular non-Hodgkin's lymphoma; peripheral T-cell lymphoma; Waldenström's macroglobulinemia, chronic lymphocytic leukemia; head and neck; gastroesophageal junction; stomach; colorectal; prostate; non-small cell lung cancer) [44].

Specifically, the natural logarithm IC50 predicted by DeepTTA is −6.007 for bortezomib in chronic myeloid leukemia (CML) cell line JurL-MK1. The result shows that bortezomib provides a potential therapeutic option in CML [45].

Dactinomycin is the third score chemotherapy drug used to treat several types of cancer (https://pubchem.ncbi.nlm.nih.gov/compound/Dactinomycin), including Wilms tumor, rhabdomyosarcoma, Ewing's sarcoma, trophoblastic neoplasm, testicular cancer and certain types of ovarian cancer. The predicted IC50 value of its effect on the neuroblastoma cell line CHP-134 is −7.164, indicating its value in the treatment of neuroblastoma.

Independent test for both drugs and cell lines

In this part, by using the 'Independent' data segmentation strategy, we applied DeepTTA to predict drugs and cell lines separately. To verify the capability of DeepTTA, we compared it with the best baseline DeepCDR in the previous study. As shown in Figure 5A and B, the PCC values of both cell lines and drugs mostly exceed DeepCDR.

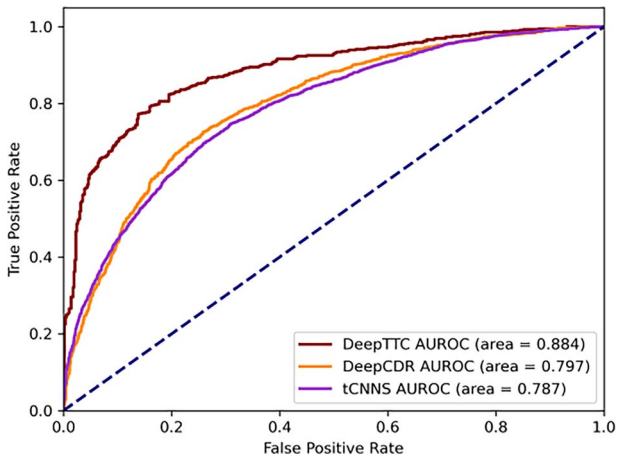

Figure 6. Comparison of AUC between DeepTTA and DeepCDR. This figure shows the comparison between DeepTTA and DeepCDR based on the binarization datasets.

For cell lines, DeepTTA was again significantly higher than DeepCDR with an average PCC of 0.913 versus 0.878 for DeepCDR.

Among the total share of 497 cell lines in both models, 393 (79%) was more accurately predicted by DeepTTA. In the independent test for the drug, there are 69 drugs shared by DeepCDR and our model, of which DeepCDR achieved an average PCC value of 0.420, and DeepTTA got a higher value of 0.580.

Drug sensitivity prediction

According to the study of DeepCDR, simply using a uniform threshold is too reckless, we binarized the IC50 of each drug through the standard offered by Iorio *et al.* [30]. Due to missing thresholds for some drugs, we

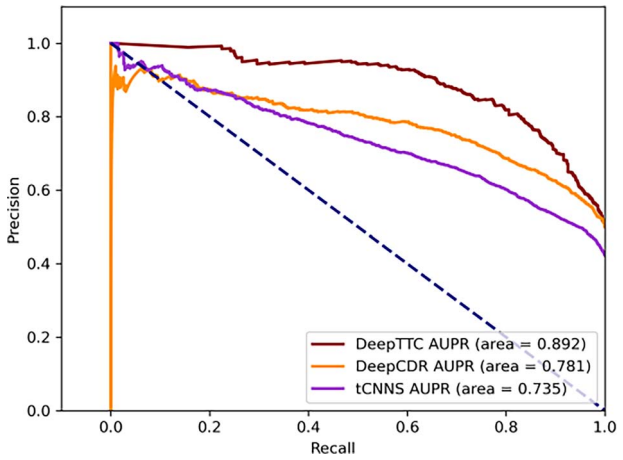

Figure 7. Comparison of AUPR between DeepTTA and DeepCDR. This figure shows the comparison between DeepTTA and DeepCDR based on the binarization datasets.

obtain 2811 sensitive and 31 871 resistance instances. For DeepCDR and tCNNS, we get the 'sensitive' and 'resistant' pairs of 7488 versus 52 210 and 18 481 versus 127 134. Since the data after binarization are unbalanced, we sample the resistance instances down to 1:1. Then, we use AUROC and AUPR to compare different models. As shown in Figures 6 and 7, DeepTTA demonstrated consistently high performance by achieving the highest AUROC (0.884) and AUPR (0.892), reaffirming the superiority of DeepTTA in extracting relationships between drug and cell lines.

Conclusion

In this paper, we proposed a novel end-to-end deep learning model to predict the anti-cancer drug response based on transformer architecture. To the best of our knowledge, DeepTTA is the first work to apply the transformer architecture in the CDR problem. Benchmark comparison results show that compared with existing prediction models, our proposed model has improved performance in terms of RMSE, PCC and SRCC on multiple test sets. Based on the results of missing data, we identify two potential multi-clinical indication drugs, bortezomib and dactinomycin. For the independent tests, especially for cell lines, DeepTTA also gains a significantly higher PCC of 0.913 than the existing model. In addition, DeepTTA achieves high performance in cancer drug sensitive prediction with the highest AUROC and AUPR. In conclusion, experiment results suggest that DeepTTA has potential ability in precision oncology where currently only ~5% of all patients benefit from precision oncology. With further improvement, we envision that DeepTTA will contribute to the growing field of oncology by facilitating the use of omics data for precision cancer medicine.

Obviously, our model still has some limitations. As a deep learning model, DeepTTA has a certain degree of inexplicability. Meanwhile, although gene expression data are useful enough for model construction, we considered that only learning the transcriptomic gene expression features is limitation. Integrating more multi-omics data into our model is our future research work. In addition, despite DeepTTA having strong predictive power, the model was built on *in vitro* data. There are still challenges in its application.

Key Points

- We proposed a model that combined gene expression data and drug information for anti-cancer drug response prediction.
- We introduced transformer architecture to extract drug useful information to improve model prediction performance.
- Compared with existing models, our model achieved the state-of-the-art performance.

Supplementary data

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

Funding

National Key R&D Program of China (2017YFE0130600); National Natural Science Foundation of China (grant nos. 61772441, 61872309, 62072384 and 62072385); Basic Research Program of Science and Technology of Shenzhen (JCYJ20180306172637807).

References

1. Zhu J, Wang J, Wang X, *et al.* Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat Biotechnol* 2021;**39**(11):1444–52.
2. Kohane IS. Ten things we have to do to achieve precision medicine. *Science* 2015;**349**(6243):37–8.
3. Rubin MA. Health: make precision medicine work for cancer care. *Nature* 2015;**520**(7547):290–1.
4. Chiu Y-C, Chen HIH, Gorthi A, *et al.* Deep learning of pharmacogenomics resources: moving towards precision oncology. *Brief Bioinform* 2020;**21**(6):2066–83.
5. Liu Q, Hu Z, Jiang R, *et al.* DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;**36**(Supplement_2):i911–8.
6. Sharifi-Noghabi H, Zolotareva O, Collins CC, *et al.* MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 2019;**35**(14):i501–9.
7. Chang Y, Park H, Yang HJ, *et al.* Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep* 2018;**8**(1):1–11.
8. Liu P, Li H, Li S, *et al.* Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics* 2019;**20**(1):1–14.

9. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet* 2018;**19**(2): 93–109.
10. Chen J-J, Knudsen S, Mazin W, *et al.* A 71-gene signature of TRAIL sensitivity in cancer cells. *Mol Cancer Ther* 2012;**11**(1): 34–44.
11. Rosenwald A, Wright G, Wiestner A, *et al.* The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* 2003;**3**(2):185–97.
12. Carter SL, Eklund AC, Kohane IS, *et al.* A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet* 2006;**38**(9):1043–8.
13. Ramaswamy S, Ross KN, Lander ES, *et al.* A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003;**33**(1): 49–54.
14. Bild AH, Yao G, Chang JT, *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;**439**(7074):353–7.
15. Manica M, Oskooei A, Born J, *et al.* Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol Pharm* 2019;**16**(12): 4797–806.
16. Born J, Manica M, Oskooei A, *et al.* PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *Iscience* 2021;**24**(4):102269.
17. Yang W, Soares J, Greninger P, *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2012;**41**(D1): D955–61.
18. Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 2015;**528**(7580): 84–7.
19. Wang L, Li X, Zhang L, *et al.* Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 2017;**17**(1):1–12.
20. Yuan H, Paskov I, Paskov H, *et al.* Multitask learning improves prediction of cancer drug sensitivity. *Sci Rep* 2016;**6**(1): 1–11.
21. Stanfield Z, Coşkun M, Koyutürk M. Drug response prediction as a link prediction problem. *Sci Rep* 2017;**7**(1):1–13.
22. Liu H, Zhao Y, Zhang L, *et al.* Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol Ther-Nucleic Acids* 2018;**13**:303–11.
23. Zhang L, Chen X, Guan NN, *et al.* A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Front Pharmacol* 2018;**9**:1017.
24. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 2014;**15**(3):1–12.
25. Menden MP, Iorio F, Garnett M, *et al.* Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 2013;**8**(4):e61318.
26. Ammad-ud-din M, Georgii E, Gönen M, *et al.* Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J Chem Inf Model* 2014;**54**(8):2347–59.
27. Zhang N, Wang H, Fang Y, *et al.* Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol* 2015;**11**(9):e1004498.
28. Wang Y, Fang J, Chen S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. *Sci Rep* 2016;**6**(1):1–11.
29. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* 2016;**32**(19): 2891–5.
30. Iorio F, Knijnenburg TA, Vis DJ, *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**(3):740–54.
31. Ding MQ, Chen L, Cooper GF, *et al.* Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol Cancer Res* 2018;**16**(2):269–78.
32. Irizarry RA, Hobbs B, Collin F, *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**(2):249–64.
33. Devlin J, Chang M-W, Lee K, *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* 2019:4171–4186.
34. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Mind Mach* 2020;**30**(4):681–94.
35. Liu Y, Ott M, Goyal N, *et al.* Roberta: a robustly optimized bert pretraining approach. *arXiv Preprint*. arXiv:1907.11692 2019.
36. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. 2017, NIPS: Long Beach, CA, USA.
37. Huang K, Xiao C, Glass L, *et al.* Explainable substructure partition fingerprint for protein, drug, and more. In: *Learning Meaningful Representation of Life Workshop at NeurIPS*, 2019.
38. Willett P. The calculation of molecular structural similarity: principles and practice. *Mol Inform* 2014;**33**(6–7):403–13.
39. Kubinyi H. Similarity and dissimilarity: a medicinal chemist's view. *Persp Drug Discov Design* 1998;**9/11**:225–52.
40. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. *arXiv Preprint*. arXiv:1508.07909 2015.
41. Gage P. A new algorithm for data compression. *C Users J* 1994;**12**(2):23–38.
42. Huang K, Xiao C, Glass LM, *et al.* MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* 2021;**37**(6):830–6.
43. Richardson PG, Barlogie B, Berenson J, *et al.* A phase 2 study of bortezomib in relapsed, refractory myeloma. *N Engl J Med* 2003;**348**(26):2609–17.
44. Roccaro AM, Vacca A, Ribatti D. Bortezomib in the treatment of cancer. *Front Anti-Cancer Drug Discov* 2006;**1**(3):397–403.
45. Heaney NB, Pellicano F, Zhang B, *et al.* Bortezomib induces apoptosis in primitive chronic myeloid leukemia cells including LTC-IC and NOD/SCID repopulating cells. *Blood* 2010;**115**(11): 2241–50.