BMC Biology

RESEARCH ARTICLE

Open Access

TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion

Xiaoqiong Xia[1], Chaoyu Zhu[2], Fan Zhong[2*] and Lei Liu[2,3*]

Abstract

Background Accurate and robust drug response prediction is of utmost importance in precision medicine. Although many models have been developed to utilize the representations of drugs and cancer cell lines for predicting cancer drug responses (CDR), their performances can be improved by addressing issues such as insufficient data modality, suboptimal fusion algorithms, and poor generalizability for novel drugs or cell lines.

Results We introduce TransCDR, which uses transfer learning to learn drug representations and fuses multi-modality features of drugs and cell lines by a self-attention mechanism, to predict the IC_{50} values or sensitive states of drugs on cell lines. We are the first to systematically evaluate the generalization of the CDR prediction model to novel (i.e., never-before-seen) compound scaffolds and cell line clusters. TransCDR shows better generalizability than 8 state-of-the-art models. TransCDR outperforms its 5 variants that train drug encoders (i.e., RNN and AttentiveFP) from scratch under various scenarios. The most critical contributors among multiple drug notations and omics profiles are Extended Connectivity Fingerprint and genetic mutation. Additionally, the attention-based fusion module further enhances the predictive performance of TransCDR. TransCDR, trained on the GDSC dataset, demonstrates strong predictive performance on the external testing set CCLE. It is also utilized to predict missing CDRs on GDSC. Moreover, we investigate the biological mechanisms underlying drug response by classifying 7675 patients from TCGA into drug-sensitive or drug-resistant groups, followed by a Gene Set Enrichment Analysis.

Conclusions TransCDR emerges as a potent tool with significant potential in drug response prediction.

Keywords Drug response prediction, Multimodal learning, Cancer cell line, Deep learning, Drug representation learning, Transfer learning

*Correspondence:
Fan Zhong
zonefan@163.com
Lei Liu
liulei_sibs@163.com
[1] Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
[2] Intelligent Medicine Institute, Fudan University, Shanghai 200032, China
[3] Shanghai Institute of Stem Cell Research and Clinical Translation, Shanghai 200120, China

Background

Tumors exhibit intra- and inter-tumoral heterogeneity [1], contributing to the variable efficacy of anticancer drugs among different tumor subtypes and patients. In order to enhance clinical outcomes and patient survival rates, precision/personalized medicine [2] seeks to individualize treatments based on the specific molecular characteristics of each patient [3]. Genomics, epigenomics, and transcriptomics have emerged as invaluable tools for providing unprecedented insights into the underlying

molecular mechanisms of cancer [4]. Precision medicine and drug repurposing can be considerably facilitated by performing a systematic analysis of drug properties and multi-omics features of cancer cell lines and accurately predicting cancer cell drug responses.

The advent of large-scale drug sensitivity data and the genomic data for over 1000 cultured cancer cell lines, such as Genomics of Drug Sensitivity in Cancer (GDSC) [5], NCI-60 [6], and Cancer Cell Line Encyclopedia (CCLE) [7], has enabled the development of computational models to predict cancer drug responses (CDR). Several novel models, including DeepCDR [8], DeepTTA [9], and GraphDRP [10], have been reported using standard datasets extracted from the GDSC. These end-to-end models share a similar architecture, with drug and cell line encoders learning representations for drugs and cell lines. Stacked fully connected layers then utilize these representations to predict drug sensitivities. Consequently, generating an accurate and robust prediction model requires appropriate representation learning for drugs and cell lines. The emergence of novel deep learning modules, such as convolutional neural networks (CNN), graph neural networks (GNN) [11, 12], and Transformer [13], has motivated their application to CDR models [14]. For example, GraphDRP [10] and GraOmicDRP [12] utilized GNNs (e.g., GIN and GAT) to learn drug features from their graph representation. DeepTTA employed a Transformer module for drug representation learning from extended-connectivity fingerprints (ECFP) [9]. CNN blocks were utilized to extract features from multi-omics data for cell lines [10, 12, 15]. Previous studies have demonstrated the superiority of GraphDRP and DeepCDR over traditional machine learning methods (e.g., ENet and random forest) and three deep learning methods (CDRscan, tCNN, and MOLI) [16]. In this regard, we will present a comprehensive comparison of our proposed method with GraphDRP and DeepCDR in the following section, highlighting their performance differences. Additionally, models such as TGSA [17] and DRPreter [18] were proposed to make better use of prior domain knowledge (e.g., protein–protein interaction). They applied GNN to extract cell line features from gene networks.

Despite the considerable progress achieved in CDR models, several limitations still exist. Firstly, labeled drugs for CDR tasks are often scarce, leading to deficient representation learning of drugs. Secondly, while these CDR models aim to learn more appropriate drug representations from 1D Simplified Molecular Input Line Entry System (SMILES) strings [19] or 2D molecular graphs, or ECFPs and achieve high accuracy, the potential interplay among multiple drug representations has yet to be fully explored [20]. Thirdly, the fusion representation of CDR is obtained by concatenating representations of drugs and cell lines, thereby limiting CDR models' performance. Finally, the accuracy of prior methods significantly drops when predicting the response of an unrepresented drug in the training set, and their inability to accurately predict CDRs in cold start scenarios has not been thoroughly evaluated and discussed. These substantial limitations will hinder the effectiveness of CDR models in precision medicine and drug repurposing.

Transfer learning is a technique that aims to enhance models' performance on small-volume datasets by transferring knowledge extracted from related large-scale datasets [21]. Although this technique is wildly applied in natural language processing [22] and computer vision, its development in computational chemistry is yet to be effectively realized. Recently, several pre-trained drug encoders have been made available. For example, ChemBERTa is a BERT-like transformer model pre-trained on a vast corpus of SMILES strings through masking language modeling of chemical SMILES strings [23]. Gin_supervised_masking is a graph isomorphism network (GIN) model pre-trained with supervised learning and attribute masking [24]. These pre-trained drug encoders can be implemented to learn global and expressive drug representation and transferred to various downstream tasks, such as drug response prediction, drug-target prediction, drug design, and property prediction.

We proposed an end-to-end regression/classification model, TransCDR (Fig. 1), to overcome the abovementioned limitations. TransCDR captured high-dimensional features from the drug's SMILES strings (S), molecular graphs (G), and ECFPs (F), as well as the associations between drug and cell line representations, to predict the half maximal inhibitory concentration (IC_{50}) value when presented with a drug-cell line pair. TransCDR significantly outperformed SOTA models for predicting IC_{50} values or sensitive states under warm and cold start. Several innovative aspects of the model's architecture contributed to the success of TransCDR. First, we introduced transfer learning to extract the chemical features of drugs automatically. Second, we integrated 3 drug structural representations (i.e., S, G, F). Third, we leveraged a multi-head attention mechanism to fuse the representations of drugs and cell lines. Finally, we evaluated the prediction ability of TransCDR on external verification sets: CCLE and applied the trained TransCDR to screening drugs for clinical patients. Furthermore, we elucidated the biological mechanisms of candidate CDRs via Gene Set Enrichment Analysis (GSEA). Thus, TransCDR contributed to cancer drug prediction and drug repurposing/discovery.

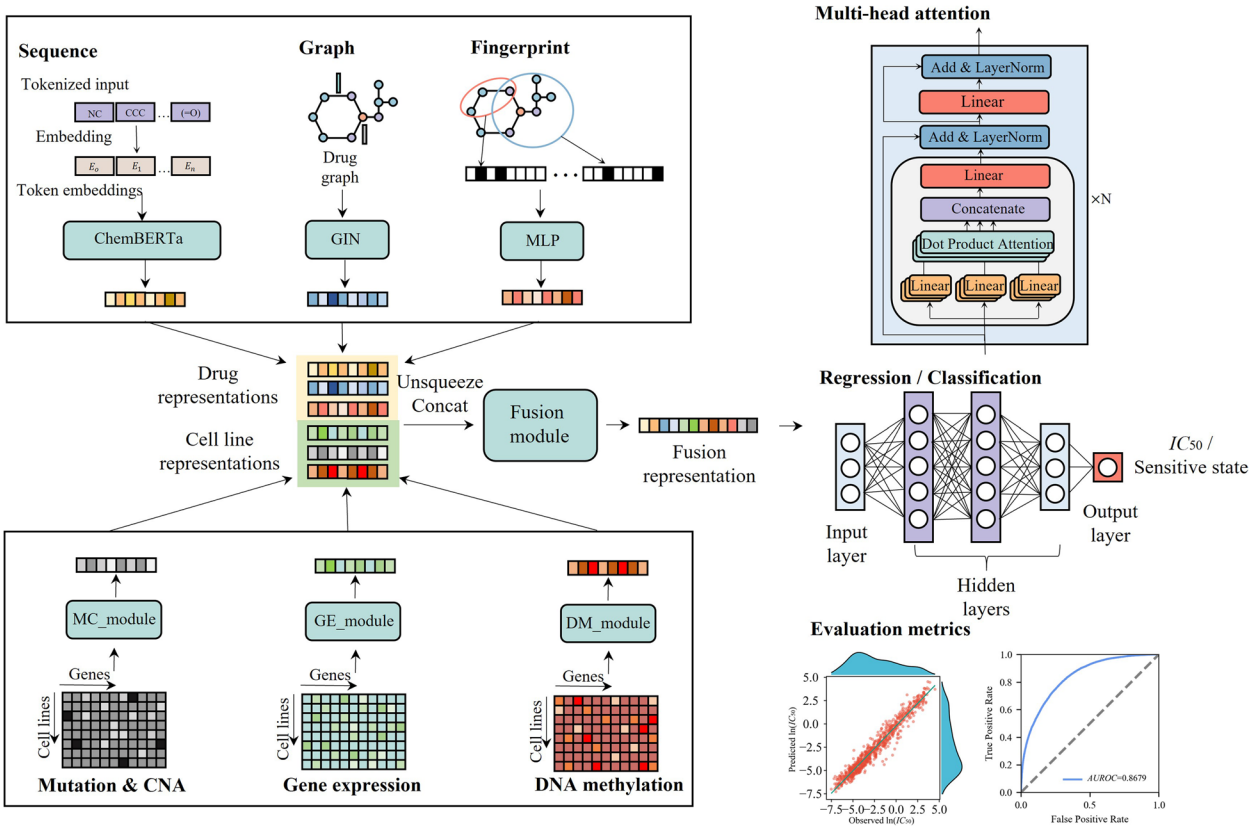

Fig. 1 The framework of TransCDR includes three drug modules (ChemBERTa, GIN, MLP) for extracting drug features from SMILES strings, molecular graphs, and fingerprints, respectively. Similarly, there are three cell line modules (MC_module, GE_module, and DM_module) for extracting cell line features from genomic mutation, gene expression, and DNA methylation data, respectively. The drug and cell line representations are fused using a self-attention-based fusion module. Finally, the fusion representation is fed to a regression/classification network consisting of four fully connected layers to predict the $\ln(IC_{50})$ or sensitive state

Results

Sensitivity analysis and performance evaluation of TransCDR

We have analyzed the effects of learning rate, batch size, and epochs on the model's predictive performance. Our results demonstrated that TransCDR exhibits robustness to variations in the learning rate and batch size, indicating a reduced likelihood of overfitting or underfitting (Additional file 1: Fig. S1). Notably, TransCDR achieved optimal predictive performance on *PC*, *SC*, and *C-index* [25] when the batch size was set to 64 and the learning rate was 1E−5. Therefore, we adopted these hyperparameters in our subsequent experiments. Furthermore, our preliminary results suggested that the model can converge within 100 epochs. Consequently, we set the maximum number of epochs to 100 and employed early stopping to determine the optimal epoch for model training. For activation functions, optimizer, and dropout rate, we followed the previous research [9].

The evaluation performance of TransCDR exhibited significant variations across 5 distinct sample scenarios (i.e., warm start, cold cell (10 clusters), cold drug, cold scaffold, cold cell and scaffold), underscoring the diverse efficacy of TransCDR and its applicability in real-world contexts. For the warm start scenario, TransCDR exhibited relatively high prediction performance with an *RMSE* of 0.9703 ± 0.0102 and *PC* of 0.9362 ± 0.0014 in regression tasks, indicating its precise application in predicting missing IC_{50} of drugs on cell lines in GDSC. However, the cold start scenario was more challenging due to the inclusion of scaffold/cell lines that were unseen during the training process. TransCDR performed worse with more strict data segmentation strategies (Additional file 2: Fig. S2). As demonstrated in Table 1, the regression *PC* of TransCDR was 0.8639 ± 0.0103 under the strictest cold cell scenario, highlighting its generalizability in predicting drug responses of unseen omics profiles, particularly for patients with known anticancer drugs, which can greatly aid precision medicine. The *PC* values were found to be 0.5467 ± 0.1586, 0.4816 ± 0.1433, and 0.4146 ± 0.1825 for cold drug, cold scaffold, and cold cell and scaffold scenarios, respectively, suggesting its potential in predicting

Table 1 Evaluation performance of TransCDR under the 5 scenarios

Sample scenarios	*RMSE*	*PC*	*SC*	*C-index*
Warm start	**0.9703 ± 0.0102**	**0.9362 ± 0.0014**	**0.9146 ± 0.0020**	**0.8797 ± 0.0013**
Cold cell (10 clusters)	1.3949 ± 0.0897	0.8639 ± 0.0103	0.8243 ± 0.0085	0.8213 ± 0.0051
Cold drug	2.2756 ± 0.3785	0.5467 ± 0.1586	0.4678 ± 0.1367	0.6651 ± 0.0523
Cold scaffold	2.3722 ± 0.3794	0.4816 ± 0.1433	0.4470 ± 0.1423	0.6571 ± 0.0522
Cold cell and scaffold	2.4518 ± 0.4201	0.4146 ± 0.1825	0.3681 ± 0.1918	0.6283 ± 0.0693

The performance of the TransCDR regression model is assessed using metrics such as *RMSE*, *PC*, *SC*, and *C-index*. All results are obtained by 10-CV

Bold indicates the best predictive performance under the 5 scenarios

massive unseen drug/compound responses on seen/unseen cell lines, hence, offering a powerful tool for drug repurposing and discovery.

Performance comparison of TransCDR and other models

To verify the effectiveness of our proposed TransCDR, we compared TransCDR with DeepCDR [8], GraphDRP [10], DeepTTA [9], TGSA [17], and DRPreter [18] on the GDSC dataset. TransCDR achieved the best performance with the highest *PC*, *SC*, and *C-index* compared to DeepCDR, GraphDRP_GAT_GCN, GraphDRP_GINConvNet, GraphDRP_GATNet, and GraphDRP_GCNNet under all scenarios (Fig. 2). TransCDR demonstrated significant superiority over DeepCDR (9.134E−5), GraphDRP_GAT_GCN (1.649E−4), GraphDRP_GATNet (9.134E−5), GraphDRP_GCNNet (9.134E−5), and DeepTTA (9.134E−5) on *PC* under the warm start scenario (Fig. 2A–C). These findings indicated that leveraging knowledge learned from large chemical datasets using ChemBERTa and GIN modules can lead to substantial improvements in predictive performance when applied to CDR prediction tasks under the warm start scenario. Notably, while DRPreter and TGSA demonstrated comparable performance to TransCDR on a warm start, with no significant difference in their *PC* ($p > 0.05$), they faltered under the cold scaffold and drug scenarios, highlighting the robustness of TransCDR across different settings. The results indicated that DRPreter and TGSA were overfitting to training sets and thus cannot generalize to the novel drugs and scaffolds. TransCDR displayed superior generalization capabilities, particularly in the challenging cold scaffold task. TransCDR had comparable performance with DRPreter, TGSA, and DeepTTA under cold cell cluster, even though TransCDR was trained without prior knowledge: protein–protein interactions. These findings suggested that the transfer learning strategy could effectively transfer the knowledge learned from a large-scale chemical dataset, thereby improving the prediction performance of TransCDR on novel drugs and scaffolds. From the perspective of real application scenarios, TransCDR was the best model to efficiently integrate information and extract features from the structures of drugs and multi-omics data of cell lines for drug response predictions.

Transfer learning exhibits superior performance in comparison to training a model from scratch

We investigated the effectiveness of transfer learning by converting the pre-trained drug representation modules into drug encoders trained from scratch (Section TransCDR variants without pre-training). As depicted in Fig. 3, TransCDR with pre-trained drug encoders demonstrated superior performance compared to its variants, including sequence-based (i.e., TransCDR_CNN and TransCDR_RNN), graph-based (i.e., TransCDR_AttentiveFP, and TransCDR_NeuralFP), and FP-based (i.e., TransCDR_ECFP) models. Specifically, the *RMSE* of TransCDR variants increased to over 0.9845, while the *PC*, *SC*, and *C-index* of TransCDR variants dropped below 0.9342, 0.9124, and 0.8780, respectively under the warm start scenarios (Wilcoxon test, $p < 0.05$). Moreover, our comprehensive evaluation revealed that TransCDR consistently outperforms other model variants across diverse scenarios, including warm start, cold drug, and cold cell line and scaffold settings (Wilcoxon test, $p < 0.05$). Specifically, under warm start conditions, TransCDR's *PC* surpassed those of TransCDR_ECFP ($p = 0.0245$), TransCDR_NeuralFP ($p = 0.001$), TransCDR_AttentiveFP ($p = 8.981\text{E}-5$), TransCDR_CNN ($p = 8.981\text{E}-5$), and TransCDR_RNN ($p = 9.032\text{E}-5$). In the cold cell line and scaffold scenarios, the *PC* values of TransCDR_ECFP, TransCDR_NeuralFP, TransCDR_CNN, TransCDR_AttentiveFP, and TransCDR_RNN decreased by 45.43%, 46.69%, 45.98%, 28.52%, and 39.96%, respectively, compared to the pre-trained TransCDR model. The results suggested that transfer learning was reliable for learning drug representations by leveraging the chemical knowledge extracted from large-scale datasets like ZINC and PubChem. Notably, TransCDR variants inherently learned drug representations by training an end-to-end model on the training set. However, their performance on the test set with unseen drugs

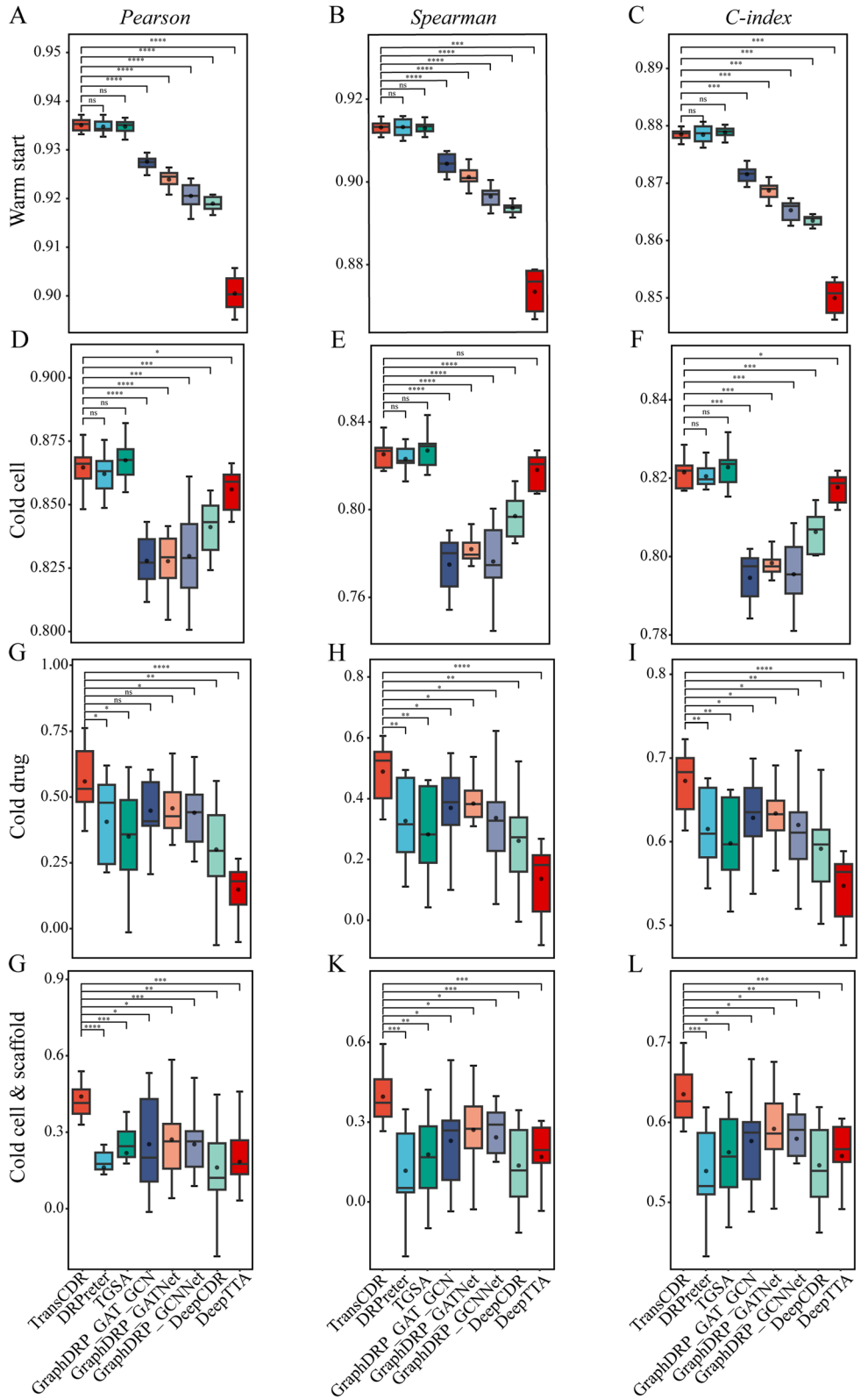

Fig. 2 Performance comparisons are conducted between TransCDR and 7 other models, namely, DRPreter, TGSA, GraphDRP_GAT_GCN, GraphDRP_GINConvNet, GraphDRP_GATNet, GraphDRP_GCNNet, DeepCDR, and DeepTTA on 5 scenarios

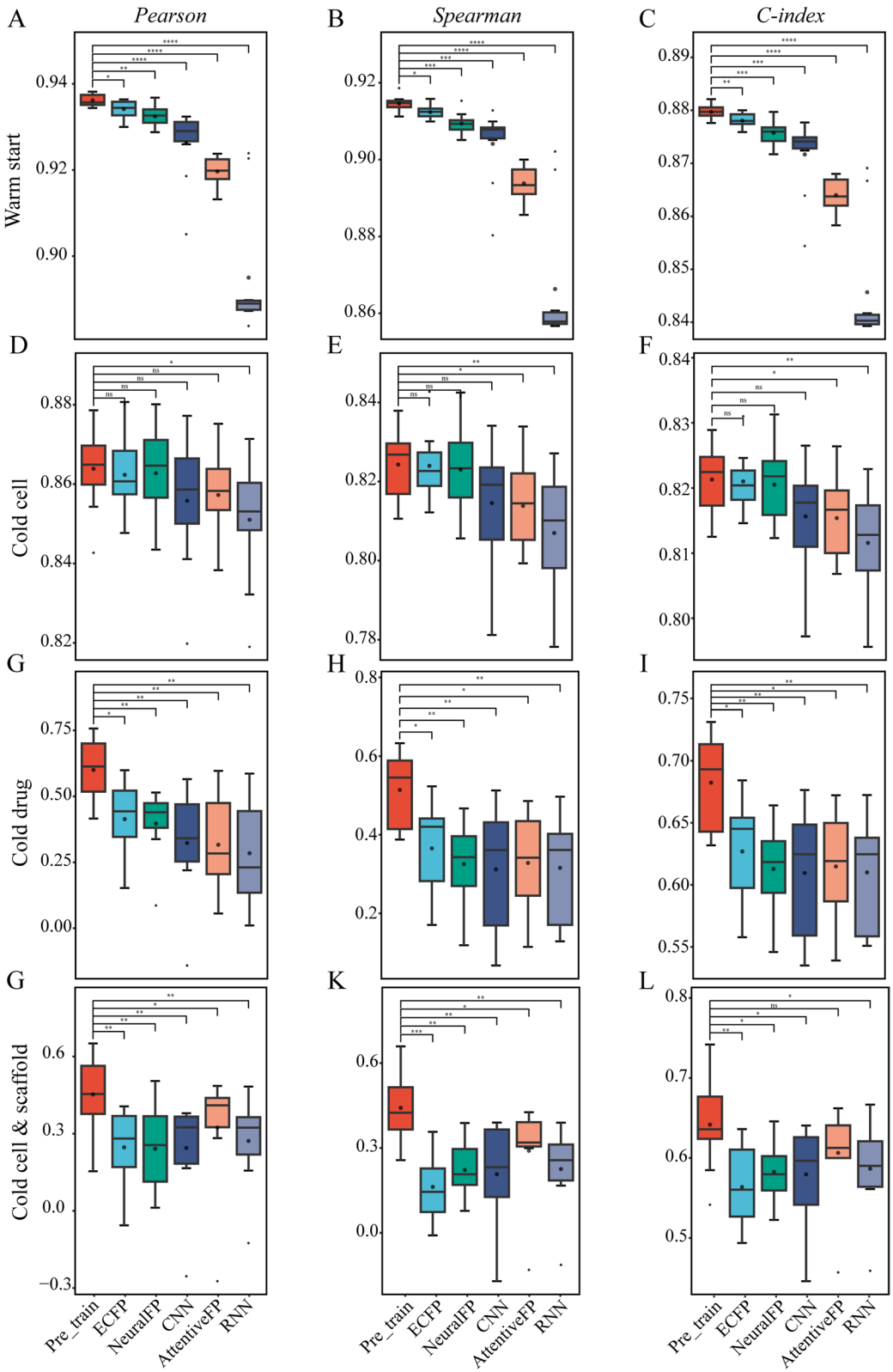

Fig. 3 The performances of TransCDR and its variants with various drug representation modules

could be better. TransCDR_ECFP attained better performance than other variants thanks to the generation of informative FP representations through Morgan's algorithm. Interestingly, in the cold cell line scenario, while TransCDR's mean *PC* was higher than those of other models, the differences were not statistically significant (Wilcoxon test, $p > 0.05$), with the exception of TransCDR_RNN ($p = 0.0225$). This finding implied that TransCDR's strength lies not in predicting the responses of unknown cell lines on drugs, but rather in predicting the responses of cell lines to unknown drugs, highlighting its potential in facilitating drug discovery and development.

Impact of each modality in TransCDR

The present study provided insights into the effectiveness of the proposed framework for drug response prediction. Ablation studies were conducted by removing each feature (i.e., S, G, and F of drugs, and MC, GE, and DM of cell lines) from TransCDR, and the resulting decrease in predictive performance was analyzed. Figure 4 demonstrated that removing these features affected the performance of TransCDR. Specifically, the performance dropped by 0.14% (S), 0.17% (G), 0.22% (F), 0.18% (GE), 0.19% (DM), and 0.27% (MC) when each modality was removed. MC was found to be the most critical among cell line features, followed by GE and DM. F was identified as the most significant for drug features, followed by G and S. These findings corroborated the comparison results presented in Fig. 4 and highlighted how multi-modality fusion could enhance model performance by complementing the limitations of individual modalities. In summary, MC and F contributed the most among different omics profiles for cell lines and drug notations for drugs, respectively.

Furthermore, we invested relative contributions of the ChemBERTa and GIN module or the transfer learning strategy in TransCDR to the generalization capability of TranCDR on cold cell and scaffold scenarios. Our analysis suggested that the ChemBERTa module contributes significantly to the generalization capability of TransCDR on cold cell and scaffold scenarios, with an ablation study showing a 4.19% drop in performance when ChemBERTa was removed and a 35.91% drop in performance when GIN was removed.

The effectiveness of self-attention

Two variants of cross-attention, DCA and CDA, were implemented along with concatenation operation to assess the efficacy of self-attention in the fusion module. Notably, self-attention outperformed other fusion methods in all regression evaluation metrics, including *RMSE*, *PC*, *SC*, and *C-index* (Table 2). For instance, the *RMSE* achieved by self-attention was recorded as 0.9703 ± 0.0102, surpassing the second-best option of concatenation, 0.9845 ± 0.0147 ($p = 0.001$), with an improvement of 0.25% in *PC*. The attention map (Additional file 3: Fig. S3) clearly showed that F is the most important modality for drugs, while MC is the most important modality for cell lines, consistent with the results of the ablation experiments. Furthermore, the attention weights for F-S, MC-DM, and F-DM pairs are high, indicating that self-attention effectively integrates multi-modal features of drugs, cell lines, and their

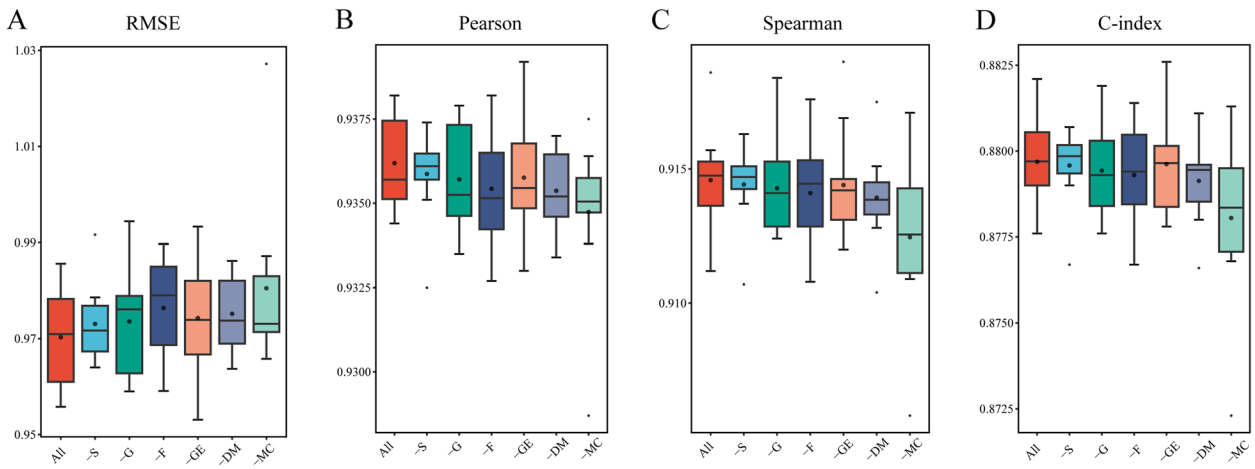

Fig. 4 Model ablation experiments results. The x-axis denotes the removal of a specific modality. MC is the most critical characteristic of cell lines, as the exclusion of this feature significantly ($p < 0.05$) increases the *RMSE* values of TransCDR compared to models without GE or DM. Similarly, F is the most significant feature for drugs. The TransCDR model without F demonstrates significantly ($p < 0.05$) increased *RMSE* values compared to that without G or S

Table 2 The performance of TransCDR with distinct fusion methods

Fusion module	*RMSE*	*PC*	*SC*	*C-index*
Self-attention	**0.9703 ± 0.0102**	**0.9362 ± 0.0014**	**0.9146 ± 0.0020**	**0.8797 ± 0.0013**
DCA	0.9962 ± 0.0208	0.9326 ± 0.0029	0.9099 ± 0.0025	0.8761 ± 0.0021
CDA	1.0303 ± 0.0104	0.9275 ± 0.0015	0.9048 ± 0.0024	0.8720 ± 0.0017
Concatenate	*0.9845 ± 0.0147*	*0.9339 ± 0.0020*	*0.9117 ± 0.0022*	*0.8773 ± 0.0017*

The best results are emphasized using bold font, and the second-best results are italicized

interactions, which contributes to the improvement of model performance. The performance of DCA and CDA was inferior as they only focused on the cross-effect between drugs and cell lines while disregarding the internal feature interaction of either.

TransCDR predicts binary drug response

We subsequently assessed the predictive power of TransCDR in cell line responses to drugs. TransCDR demonstrated high performance across varying ratios of positive and negative samples in the warm start scenario. Specifically, when the dataset was balanced, TransCDR yielded superior performance with an *AUROC* of 0.8213 ± 0.0067 and an *AUPR* of 0.8138 ± 0.0085. When the dataset was unbalanced of 1:2, 1:5, and 1:8, TransCDR displayed a slight increase in *AUROC* and a decline in *AUPR*, with reductions over 8.76%, 20.99%, and 26.93% for *AUPR*, respectively. These findings highlighted the impact of dataset imbalance on the predictive power of TransCDR, with *AUPR* exhibiting sensitivity to sample ratio variations. Therefore, we utilized *AUPR* as the primary evaluation metric. In the cold test setting, the *AUPR* of TransCDR reduced more compared with a warm start when the dataset was imbalanced. Specifically, when the sample ratio was 1:1, TransCDR in cold cell achieved an *AUPR* of 0.7492 ± 0.0227, which was 39.52% higher than that of the 1:8 sample ratio of (*AUPR* = 0.3540 ± 0.0381). Similarly, in the cold drug setting, a sample ratio 1:1 yielded optimal performances (Fig. 5). Consequently, subsequent experiments were conducted using the sample ratio of 1:1.

Application of TransCDR on GDSC

The pre-trained TransCDR exhibited excellent performance across a diverse range of cancer types (Fig. 6A–C), cell lines (Fig. 6D–F), and drugs (Fig. 6G–I). In all tested cancer types, the *PC* and *SC* values ranged from 0.9624 to 0.9763 and 0.9345 to 0.9591, respectively (Additional file 4: Table S1). The *PC* and *SC* values for cell lines ranged from 0.9192 to 0.9886 and 0.8723 to 0.9676, respectively (Additional file 5: Table S2). The performance of TransCDR on drugs varied considerably, with the *PC* and *SC* ranging from 0.3949 to 0.9838 and 0.3983 to 0.9814, respectively (Additional file 6: Table S3). Employing the trained TransCDR, we predicted 34,662 missing IC_{50} values for drug-cell line

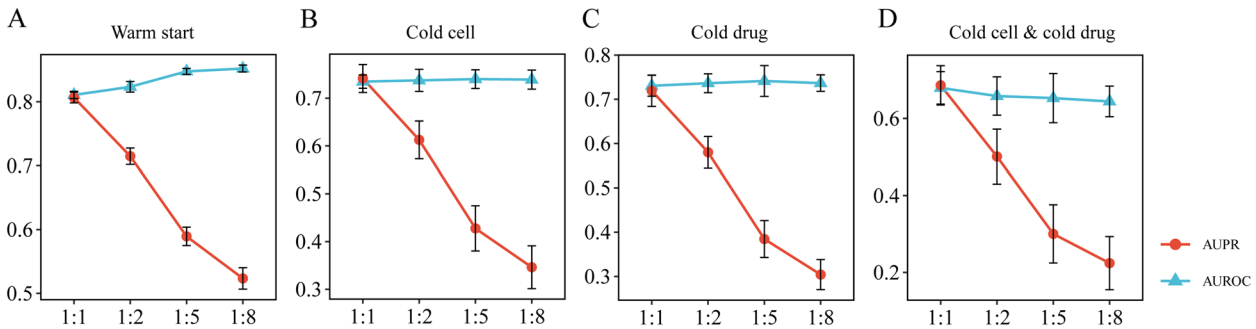

Fig. 5 TransCDR's performance evaluation is assessed across 4 sampling scenarios utilizing 4 sampling ratios between positive and negative samples (1:1, 1:2, 1:5, and 1:8)

(See figure on next page.)

Fig. 6 **A**–**C** The scatter plots of CDRs of specific cancer types, specifically adrenocortical carcinoma (ACC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and multiple myeloma (MM), with the top 3 prediction performances. **D**–**F** The scatter plots of CDRs of specific cell line types, specifically T-lymphoid cell line (CML-T1), NCI-H1105, and BALL-1, with the top 3 prediction performances. **G**–**I** The scatter plots of CDRs of specific drugs, including FK866, Gemcitabine, and GSK1070916, with the top 3 prediction performances. **J** The study categorizes drugs based on their average predicted IC_{50} values in ascending order, with the top 10 drugs being sensitive and the bottom 10 being resistant

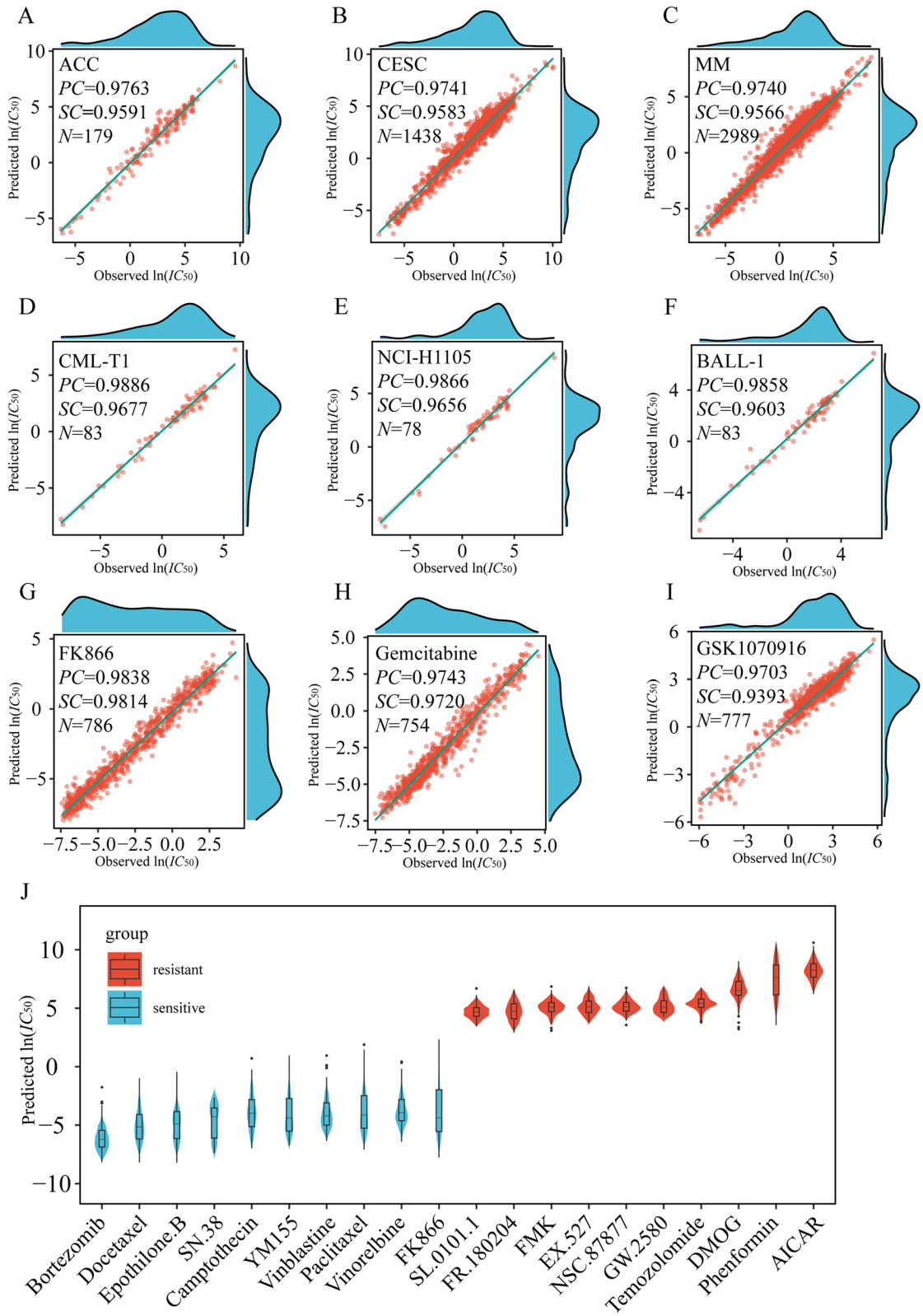

Fig. 6 (See legend on previous page.)

pairs in the GDSC database, corresponding to approximately 18.10% of all 191,475 pairs involving 851 cancer cell lines and 225 drugs. We ranked the IC_{50} values predicted by the regression model in ascending order and selected the top 10% (3,466) drug-cell line pairs inclusive of 610 cancer cell lines and 75 drugs (Additional file 7: Table S4). Our work confirmed previous research findings on the top 15 (the lowest IC_{50}) drug-cell line pairs that were molecularly effective in cancer treatment, involving 15 cell types, 4 drugs, 10 tissues, and 8 cancers (see additional file 8: Table S5 [26–35]). Notably, bortezomib, one of the approved proteasome inhibitors for treating various malignancies (e.g., SKCM, OV, and BRCA) [36], was predicted to be sensitive to different cell lines and cancer types. The top 10 "sensitive" and the last 10 "resistant" drugs are depicted in Fig. 6J. As anticipated, several sensitive/resistant drugs were also identified by DeepCDR [8]. For instance, bortezomib, docetaxel, epothilone B, vinblastine, vinorelbine, and SN-38 [37] were predicted as sensitive drugs, and FR-180204, NSC-87877, GW-2580, DMOG, phenformin, and AICAR were predicted as resistant drugs by DeepCDR. Additionally, the effectiveness of the most potent drugs, bortezomib [38], docetaxel [39], and vinblastine [40], has been established in multiple cancer types.

External validation results

The present study assessed the efficacy of TransCDR trained on the GDSC dataset by evaluating the external in vitro dataset CCLE. The results demonstrated TransCDR's outstanding performance with a *PC* range varying from 0.6736 to 0.8931 when tested across diverse cancer types. Bile duct cancers exhibited the highest performance (*PC* of 0.8931), while kidney cancer demonstrated the lowest (*PC* of 0.6736) (Additional file 9: Table S6). These findings suggested that TransCDR could effectively predict drug response in new cell lines specific to certain cancer types. Furthermore, our comparative analysis of the predictive performance of TransCDR and other models on the CCLE dataset revealed that TransCDR exhibits superior performance in multiple cancer types, including bile duct cancer, sarcoma, and skin cancer (Fig. 7), thereby highlighting its robust generalizability across a diverse range of cancer types (Additional file 9: Table S6).

We applied TransCDR and other models to real-world drug screening scenarios. Notably, TransCDR successfully identified numerous CDRs, whereas other methods failed to do so (Additional file 10: Table S7). For example, our model correctly predicted the sensitivity of 17-AAG to A172 cells, which is particularly significant given the promising anti-tumor activity of 17-AAG in glioblastoma and the widespread use of A172 as a model for this cancer type [41, 42]. Furthermore, TransCDR identified the

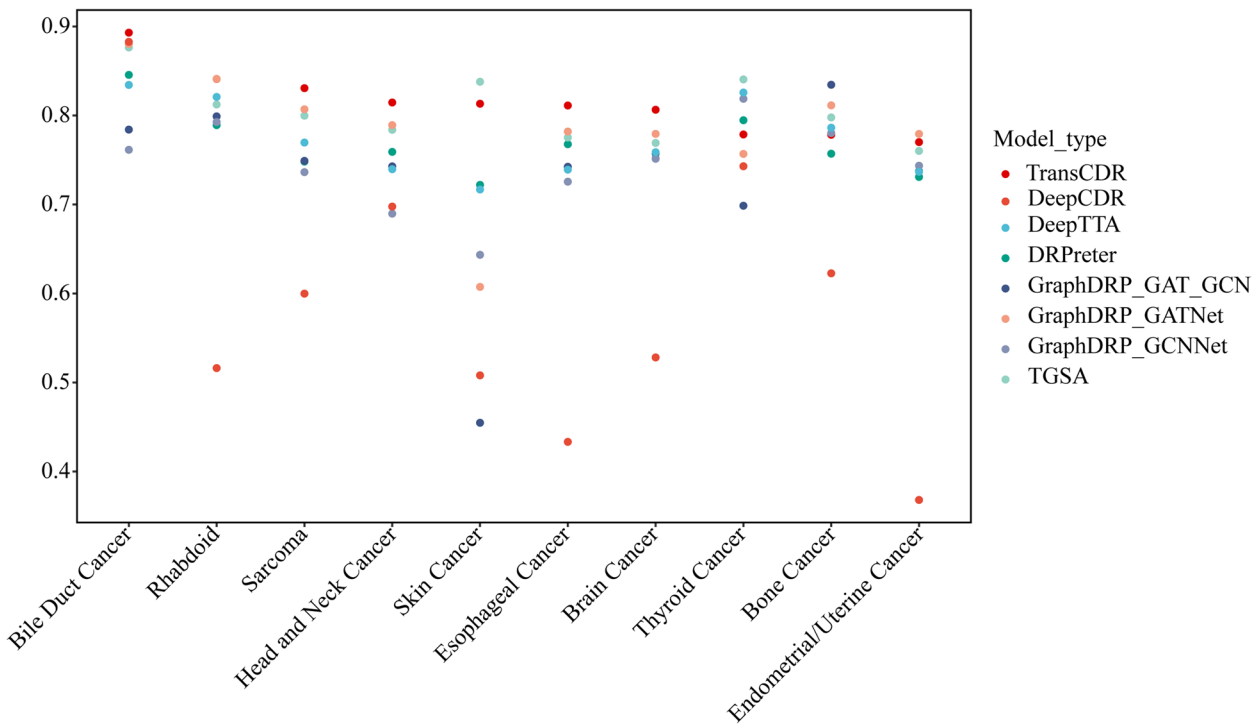

Fig. 7 Predictive performance (*PC*) of TransCDR and other models on the CCLE dataset across various cancer types

PD-0325901-OV90 pair, which is consistent with previous reports that PD-0325901 inhibits OV90 cell growth by blocking the PI3K/mTOR and RAS/ERK signaling pathways, commonly hyperactivated in ovarian cancer [43, 44]. These results demonstrated the effectiveness of TransCDR in predicting CDRs and its potential to facilitate the discovery of novel therapeutic strategies.

TransCDR recognizes biological mechanisms under drug response

To recognize biological mechanisms under drug response, we utilized the pre-trained TransCDR to screen 225 drugs for 7675 patients from TCGA. The predicted drug sensitivities of these patients were presented in Additional file 11: Table S8. We selected the top 10 drugs, namely CX-5461, lapatinib, dasatinib, erlotinib, afatinib, trametinib, utlin-3a, A-770041, CHIR-99021, and AZD-0530 for GSEA. By performing GSEA, we were able to elucidate the biological mechanisms underlying the predicted drug sensitivities of patients and explore possible underlying mechanisms (Additional file 12: Table S9). Our observation showed that differential expression genes caused by afatinib medication demonstrated a significant enrichment within gene sets associated with breast and lung cancer. This observation aligned with evidence supporting afatinib's efficacy in treating breast and lung cancer [45, 46]. Furthermore, the enriched gene sets offered insight into afatinib's therapeutic mechanisms. For instance, up-regulated genes observed in afatinib-sensitive patients exhibited a significant enrichment in COLDREN_GEFITINIB_RESISTANCE_DN ($NES = 1.983$, $p = 0.0005$), which pertained to genes that down-regulated in non-small cell lung carcinoma cell lines resistant to Gefitinib in comparison to those that were sensitive [47]. This finding indicated that gefitinib and afatinib operated through similar mechanisms [48]. In contrast, the up-regulated genes observed in afatinib-sensitive patients showcased a significant enrichment in HOLLERN_EMT_BREAST_TUMOR_DN ($NES = 2.211$, $p = 4.53\text{E}-6$), which consisted of genes with low expression levels in mammary tumors marked by epithelial-mesenchymal transition histology and could result in resistance to afatinib [49].

Discussion

In comparison with existing SOTA models, TransCDR exhibited several improvements. Firstly, it outperformed other models across diverse prediction tasks under different sample scenarios (warm and cold start). Secondly, TransCDR fused the most extensive data modalities, incorporating 3 drug representations and 3 omics profiles, whereas DeepTTA only considered SMILE strings and gene expression profiles. Thirdly, TransCDR learned the fusion representations by a self-attention-based module which was more effective than a simple concatenation operation. Thirdly, we comprehensively assessed the generalizability of TransCDR across diverse scenarios. Our proposed model enhanced the performance in cold drug/scaffold and cold cell and scaffold scenarios, essential for predicting cancer drug response and screening novel candidates from a vast drug/compound space.

We demonstrated that generalizing TransCDR to novel scaffolds posed a greater challenge than cell line clusters. Several factors contributed to this phenomenon. Cell lines were characterized by gene expression profiles obtained via omics measurements, providing a comprehensive representation of cellular biology features. Conversely, compounds were encoded using SMILES strings, which may lead to loss of structural information. Furthermore, TransCDR learned drug embedding from SMILES strings or molecular graphs using end-to-end training, requiring substantial drug structures. Lastly, minor structural differences between similar compounds may result in significant disparities in SMILES strings, yielding distinct embeddings. TransCDR can serve as an effective tool for the cancer-drug response prediction. Additionally, TransCDR have promising applications in drug discovery. Specifically, we can initially assess the scaffold similarity of a new compound/drug against known drugs; if a similarity scaffold is identified, our predicted CDRs will hold greater credibility. If not, TransCDR stands as the optimal model to predict CDRs in cold scaffold and cold cell and scaffold scenarios.

However, several limitations and potential directions for further improving TransCDR have been identified. The study requires large-scale, highly qualified datasets, including multiple drugs and cell lines. Although drug response data have increased dramatically over the past decades, cell lines with multi-omics profiles are limited. The performance of TransCDR on the cold scaffold is significantly better than other SOTA models through transfer learning but still has much room for improvement. The current TransCDR cannot capture the drugs' three-dimensional structural information, which inevitably affects drug representation learning. A better drug representation model that can extract discriminating features from drug notations will be designed, such as GeoGNN, which encodes molecules' topology and geometry information by a geometry-based GNN architecture [50]. Therefore, to further improve the prediction performance and interpretability of TransCDR, we will propose the next version of TransCDR, trained on the larger and more reliable CDR dataset, considering the multimodal features of drugs and cell lines and making full use of prior domain knowledge. We also evaluated the performance of TransCDR on the TCGA dataset,

where patient-drug response data were collected from previous research [51]. However, we regret to report that TransCDR's predictive performance on this dataset fell short of our expectations. We speculated that the model, trained on cellular data, may not be directly applicable to TCGA samples. We acknowledged the limitations of our current approach and proposed fine-tuning the TransCDR model on patient-related datasets in future studies to enhance its predictive power for patient responses to drugs.

We leveraged the self-attention mechanism for feature fusion, which is invented for long sequence data such as languages, but has been extensively applied to feature fusion tasks [52–54]. In contrast to the concatenate module, which represents the most straightforward and conventional data fusion approach, we adopted the self-attention-based fusion strategy to augment the innovativeness of our model. Although the attention-based method does not exhibit a pronounced advantage over concatenation in the present study, the improvement is statistically significant ($p < 0.05$), we hypothesize that its superiority will become more pronounced as the number of modalities increases in future research, thereby underscoring the potential of self-attention-based fusion in fusing complex multimodal data.

Recently, the emergence of large-scale foundation models on single-cell transcriptomics, such as Geneformer [50], scGPT [55], and scFoundation [56], has brought new ideas and insights to the field of cell line-drug response prediction. Similar to TransCDR, which utilized the pre-trained chemical language model, ChemBERTa to learn embeddings from SMILES strings, these foundation models can also serve as feature extraction modules to learn embeddings from cell lines' transcriptomics. For instance, Minsheng Hao et al. replaced the MLP-based cell line transcriptional feature extraction module in DeepCDR with scFoundation and demonstrated that the scFoundation-based model significantly outperformed the original model [56]. This result further suggested that foundation models trained on single-cell transcriptomics can also be applied to learn transcriptomics generated by bulk sequencing. However, it is noteworthy that these foundation models have a large number of parameters, requiring substantial computational resources for fine-tuning. Fortunately, as a foundation model, scFoundation can be directly used as a cell line transcriptional feature extraction module. In future work, we can also explore the impact of DNA methylation foundation models and gene mutation foundation models on CDR prediction models.

More drug response prediction methods were developed with different training objectives and application scenarios. For example, GraphCDR was a binary classification model used for predicting the drug-cell line response, rather than predicting IC_{50} value [57]. To evaluate the model performance, GraphCDR used *AUROC* and *F1-score*. In contrast to GraphCDR, TransCDR and some other compared models were regression models that predict IC_{50}. Consequently, they typically used the *PC* as the evaluation metric. The pioneering work of DeepCoVDR [58] has demonstrated the effectiveness of pre-training a CDR prediction model on GDSC and fine-tuning it on the SARS-CoV-2 dataset, resulting in outstanding predictive performance in COVID-19 drug response prediction tasks. By contrast, TransCDR was designed for generic cell-line drug response prediction. In our future research, we plan to investigate the potential of fine-tuning TransCDR on specialized datasets aligned with specific objectives, with the goal of further improving its predictive accuracy on targeted tasks.

Conclusions

In this study, we presented an end-to-end CDR prediction model called TransCDR, which fused multi-modality representations of drugs, including SMILES string, molecular graph, ECFP, and omics profiles, including genetic mutation, gene expression, and DNA methylation to learn the $\ln(IC_{50})$ values or sensitive states of drugs on cell lines. TransCDR outperformed the 8 SOTA models and showed high performance under different sample scenarios. In addition, TransCDR outperformed multiple variants with drug encoders that were trained from scratch. We confirmed that F and genetic mutation contributed the most among multiple drug notations and omics profiles, respectively. Furthermore, TransCDR showed high prediction performance on the external test sets CCLE. Finally, we predicted $\ln(IC_{50})$ values of missing CDRs in GDSC and screened the drug response of cancer patients to drugs. These candidate CDRs were verified by existing literature and GSEA. In summary, our deep learning model, TransCDR, offers a powerful tool for drug response prediction.

Methods

Data preparation

This study utilized GDSC, CCLE, and TCGA datasets. Specifically, GDSC was employed to assess the effectiveness of TransCDR across various application scenarios, including predicting missing CDRs for known cell lines and drugs, unseen cell lines, unseen drugs, and unseen cell-drug combinations. Additionally, all CDRs from the GDSC database were utilized for training the final TransCDR model, which was subsequently evaluated on external datasets: CCLE.

GDSC v2 constitutes a vital asset in the endeavor of discovering therapeutic biomarker for cancer cells [5].

We have gathered a total of 156,813 CDRs that satisfied three specific criteria, including 851 cancer cell lines and 225 drugs. (1) CDRs encompassed drug sensitivity profiles ascertained through the measurement of the half maximal inhibitory concentration (IC_{50}) or sensitive state, which indicated the capacity of a drug to impede the growth of specific cell lines. (2) The selected CDRs exhibited the presence of three omics data sets: genetic mutation, gene expression, and DNA methylation for the corresponding cell lines. (3) The included drugs possessed SMILES strings.

This study obtained mutation and copy number aberration (MC), gene expression (GE), and DNA methylation (DM) profiles from GDSC. Specifically, gene expression profiles were downloaded for 1000 human cancer cell lines using transcriptional profiling arrays E-MTAB-3610, and were pre-processed using the R package *affy*. The Affymetrix GeneChip system, along with the robust multiarray average method, was employed for measuring gene expression [59], resulting in 18,451 gene expression values for each cell line. Subsequently, the gene expression matrix was then normalized using *z*-score. The MC data consisted of a binary matrix with 735 features, where 1 indicated a mutation or copy number aberration in the gene, and 0 indicated the absence of such aberrations. The DM matrix was obtained by downloading the processed matrix of GSE68379 from GEO, where continuous values represented the methylation score of each CpG. The methylation scores of CpG sites were then averaged to obtain methylation scores for genes, resulting in 20,617 methylation values for each cell line. The DM matrix was also normalized by *z*-score. Drug SMILES strings were retrieved from PubChem [60] and converted to canonical SMILES using open-source cheminformatics software RDKit [61].

For each combination of cell line and drug CDP_{ij}, cell line i was characterized using 3 types of omics data (i.e., MC, GE, DM); drug j was represented by SMILES strings; and the label of CDP_{ij} was the natural logarithm-transformed IC_{50}. A total of 156,813 CDPs were utilized in the development of the regression model. In classification experiments, IC_{50} values were binarized based on the provided threshold for each drug [62]. Consequently, a total of 154,603 CDPs were obtained, with $CDP_{ij} \in \{0,1\}$, consisting of 18,143 sensitive CDPs and 136,460 resistant CDPs. We treated all sensitive CDPs as positive samples. Then, we randomly sampled CDPs from all resistant CDPs as negative samples, with positive-to-negative sample ratios of 1:1, 1:2, 1:5, and 1:8. Finally, we combined the positive and negative samples to construct datasets with different positive-to-negative ratios.

For the CCLE dataset, this study accessed MC, GE, and DM profiles as well as pharmacological profiling files from the Broad DepMap Portal. The processing steps for CDPs in GDSC were followed to extract 9242 CDPs, which consisted of 401 cancer cell lines and 24 drugs, with IC_{50} values transformed via natural logarithm. For the TCGA dataset, a total of 7675 patients with multi-omics profiles, including MC (MC3 gene-level non-silent mutation), GE (Illumina HiSeq), and DM (Methylation 450 k) were obtained from UCSC Cancer Genome Browser Xena [63] using TCGA patient ID. Due to differences in the feature dimensions of MC, GE, and DM between CCLE, TCGA, and GDSC, the features of CCLE and TCGA were aligned with those of GDSC. Standardization of GE and DM profiled across different platforms was ensured through *z*-score normalization.

Data segmentation strategies

We employed tenfold cross-validation (10-CV) to evaluate TransCDR's generalizability comprehensively. Datasets were divided based on 5 strategies: warm start, cold drug, cold scaffold, cold cell, and cold cell and scaffold.

1. Warm start: A warm start approach was adopted to assign a random selection of 80%, 10%, and 10% of the CDRs to the training, validation, and testing sets, respectively. Notably, it was possible for a drug/cell line from the test or validation set to also be present in the training set. The models trained using the warm start strategy were then employed to predict the missing IC_{50} values in the GDSC dataset.
2. Cold drug: Drugs present in the test/validation set were carefully excluded from the training set. Among the drug-associated CDRs, a random selection of 80% (180) drugs were assigned to the training set, 10% (22) to the validation set, and the remaining CDRs with 10% (23) drugs were designated for the test set. This experimental design aimed to assess the model's performance on unforeseen drugs. It was important to note that despite these efforts, there may be instances where different drugs share similar scaffolds, resulting in scaffold overlap between train, validation, and test data. Consequently, this overlap may potentially overestimate the generalization ability of the CDR model to novel drugs.
3. Cold scaffold: First, we utilized the MurckoScaffoldSmiles function from the RDKit library to extract the Murcko scaffold of each drug. Then, we grouped the drugs based on their shared Murcko scaffolds. Next, we partitioned these Murcko scaffolds into training, test, and validation sets at a ratio of 8:1:1. This ensures that the Murcko scaffolds present in the training set are not observed in the test and validation sets. We then selected all the drugs associated with the scaffolds assigned to the training set and further

extracted the corresponding drug-cell line pairs to constitute the final training set. The same approach was applied to construct the validation and test sets, where the drug-cell line pairs were selected based on the Murcko scaffolds assigned to the respective sets. This cold scaffold-based data splitting method can better evaluate the model's generalization capability when faced with novel compound structures and reliably assess the model's performance in predicting the activity of new, unseen compounds.

4. Cold cell: First, we utilized the *K*-means clustering function from the sklearn Python package to group the cell lines based on the similarity of their omics features (concatenation of three different omics data in the feature dimension). We experimented with 10, 50, 100, and 200 clusters. Taking the 10-cluster scenario as an example, we split the 10 cell line clusters into training, validation, and test sets at a ratio of 8:1:1. We then selected all the cell lines belonging to the clusters assigned to the training set from the entire dataset, and extracted the corresponding cell line-drug pairs to constitute the final training set. The same approach was applied to construct the validation and test sets. We named the resulting datasets as dataset_10C, dataset_50C, dataset_100C, and dataset_200C, respectively, each containing the corresponding training, validation, and test sets. The purpose of this experiment is to investigate the model's generalization ability to predict the drug response of unknown cell lines. Furthermore, the fewer the number of clusters, the greater the differences between the cell line groups. This allows us to explore the model's predictive performance on cell lines with varying degrees of dissimilarity, which is crucial for assessing the model's robustness.
5. Cold cell and scaffold: We adopted a data partitioning approach that simultaneously satisfies both cold cell and cold scaffold conditions. we partitioned drug scaffolds and 10 cell line clusters separately into training, validation, and test sets at an 8:1:1 ratio. From the entire dataset, we selected the cell line-drug pairs where the cell line belongs to the training set clusters and the scaffold belongs to the training set, to form the final training set. For the validation and test sets, we used the same approach, selecting the cell line-drug pairs where the cell line and scaffold belong to the respective set. The strictest data partitioning scenario can effectively evaluate the model's generalization on unseen cell lines and drug scaffold.

Overall architecture of TransCDR

We proposed TransCDR, an end-to-end deep learning model that employed drugs' chemical structures and cell lines' multi-omics data to predict drug responses. TransCDR consisted of two prediction modes: regression for predicting IC_{50} values and classification for predicting drug sensitivity or resistance on cell lines. The model was composed of four main components (Fig. 1): (1) We employed ChemBERTa, a pre-trained model, to learn drugs' representations from SMILES strings, gin_supervised_masking, another pre-trained model, to learn drugs' molecular graph representations, and a stacked full connected (FC) layers module to acquire high-dimensional features from ECFPs. We opted not to fine-tune the pre-trained models, instead utilizing them as feature extractors to obtain embeddings of drugs' structures. This approach leveraged transfer learning, where the pre-trained models were employed to extract generalizable features that can be adapted to our specific task. (2) We used three FCs to learn numerical representations of MC, GE, and DM data. (3) These drug and cell line representations were fused in a fusion module, a stacked multi-head attention layer module with 6 layers and 8 heads. The fusion module integrated multi-modality features of drugs and cell lines. (4) A regression/classification network with four FCs used fusion representations to predict drug responses. The components above of TransCDR were further elaborated in the subsequent paragraphs.

Drug representations

We employed three drug encoders to acquire numerical representations from the three basic molecular notations (S, G, and F), followed by applying three notation-specific networks that extracted 256-dimension features from the numerical representations. Given the relatively small number of drugs in our dataset, we determined that using the pre-trained models as feature extractors was more suitable, as fine-tuning would likely result in overfitting and require additional training time and computational resources.

Sequence representation

We utilized the SMILES format to represent drugs, which involved a series of characters indicating atom and bond symbols and a few grammar rules resembling natural language. To this end, we proposed employing a pre-trained BERT-like model called ChemBERTa to acquire the numerical representations from SMILES strings. ChemBERTa [64] is pre-trained on 10 M SMILES strings from PubChem using the masked language modeling approach. Figure 1 displays the specifications of

the sequence representation module used in our experiments. Subsequently, the SMILE string was tokenized into sub-word token strings using the Byte Pair Encoding tokenizer and then converted into token IDs with a maximum sequence length of 512. Next, the token IDs were inputted into the pre-trained ChemBERTa to obtain the sequence representation. The numeric representation of a SMILES string is computed as follows:

$$T = \text{tokenizer}(\text{SMILES}) \tag{1}$$

$$h_s = \text{ChemBERTa}(T) \tag{2}$$

where T indicates the token IDs of a SMILES string $T = \{t_1, t_2, \ldots, t_{512}\}$, and h_s represents the learned numeric representation of the SMILES, with a dimension of 768. The ChemBERTa and tokenizer were downloaded from HuggingFace [65]. Furthermore, we extracted features from the numeric representations utilizing a neural network, with two hidden layers comprising 1024 and 256 neural units, respectively. Every layer is formulated according to the following equation:

$$h_s = \text{ReLU}(W_i h_s + b_i) \tag{3}$$

where W_i and b_i represent learnable matrices. The output size of the network is set to 256 to facilitate fusion operation.

Graph representation

The recent emergence and success of GNNs have inspired their application to drug representations. Specifically, we represented drugs as molecular graphs as $G = (V, E)$, where V denotes the atoms, and E denotes the chemical bonds node. Each node $v \in V$ is associated with node features h_v and each edge $(u, v) \in E$ is associated with edge features e_{uv}

$$h_v^{(l+1)} = \text{MLP}^{l+1}\left(\left(1 + \epsilon^{l+1}\right) * h_v^l + \sum_{u \in N(v)} e_{uv} * h_u^l\right) \tag{4}$$

$$h'_v = \text{CONCAT}(h_v^0, h_v^1, \ldots, h_v^l) \tag{5}$$

$$h_g = \frac{1}{N} \sum_N h'_v \tag{6}$$

where v represents the target node, u represents the neighboring node of v, and e_{uv} denotes the weight assigned to the edge from u to v. The model includes a learnable parameter ϵ and employs h_v^l, the node representation of layer l, and h_G, the graph representation. The pre-trained GIN model, gin_supervised_masking, performed well in learning local and global representations at the individual node and whole graph level. To learn the appropriate representations with a dimension of 300 from molecular graphs, we applied a neural network with 2 hidden layers to extract features from these representations.

Fingerprint representation

The topological fingerprints of drugs were captured using ECFP representations [66] based on Morgan's algorithm with RDKit. Specifically, each atom was assigned a unique integer identifier and updated to represent larger circular substructures with a radius of 2. The final substructures were hashed into a binary vector with a length of 1,024, defined as $FP = \{fp_1, fp_2, \ldots, fp_{1024}\}$, where $fp_i \in \{0,1\}$. ECFP features were also extracted using a neural network with 2 hidden layers.

TransCDR variants without pre-training

To examine the effectiveness of transfer learning, we replaced the pre-trained drug representation modules, namely ChemBERTa and gin_supervised_masking, with non-pre-trained modules, including CNN, RNN, AttentiveFP, NeuralFP, and ECFP. All parameters were initialized randomly and subsequently learned from scratch through a back-propagation algorithm. One-hot encoding was used to represent drugs in CNN and RNN, whereas AttentiveFP and NeuralFP represented drugs as pre-defined graph structures with atomic and bond features. The drug module architecture was identical to that of DeepPurpose [67].

Cell line representations

We employed a late fusion strategy to process high-dimensional and heterogeneous omics data and capture complex relationships from mutation and copy number aberration, gene expression, and DNA methylation profiles. We used omics-specific networks to extract features of cell lines from each omics and fuse these 3 types of features using multi-head attention. The fully-connected networks had 2 hidden layers with 1024 and 256 neural units. We mapped the 3 types of omics data into a latent space with an embedded dimension fixed at 256.

$$h_{\text{MC}} = \text{Network}_{\text{MC}}(X_{\text{MC}}) \tag{7}$$

$$h_{\text{GE}} = \text{Network}_{\text{GE}}(X_{\text{GE}}) \tag{8}$$

$$h_{\text{DM}} = \text{Network}_{\text{DM}}(X_{\text{DM}}) \tag{9}$$

where h_{GE}, h_{MC}, and $h_{\text{DM}} \in \mathbb{R}^{n*d}$, d=256, and n is the batch size.

Multi-head attention for feature fusion

We proposed utilizing the multi-head attention mechanism to model the relationships between drug features (i.e., sequences, graphs, and ECFPs) and cell line features (i.e., MC, GE, and DM). Initially introduced in Transformer [68], the multi-head attention method has been widely adopted for multi-modality fusion [69, 70]. Specifically, the attention module mapped a query and a set of key-value pairs to an output generated as a weighted sum of the values. The attention is formulated as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}})V \tag{10}$$

where $Q,K,V \in \mathbb{R}^{n*6*d_k}$, derived from the concatenation of 3 drug and 3 cell line features, n is the batch size, d_k represents the feature dimension, and T is a transpose operation. To learn the features from distinct representation subspaces, we projected the Q, K, and V h times, and calculated the multi-head attention function as follows:

$$Q_i = QW_i^Q + b_i^Q, i \in \{1,2,\ldots,h\} \tag{11}$$

$$K_i = KW_i^K + b_i^K, i \in \{1,2,\ldots,h\} \tag{12}$$

$$V_i = VW_i^V + b_i^V, i \in \{1,2,\ldots,h\} \tag{13}$$

$$\text{MultiHeadAtt}(Q,K,V) = \text{CONCAT}(head_1, head_2, \ldots, head_n)W^O \tag{14}$$

$$head_i = \text{Attention}(Q_i, K_i, V_i) \tag{15}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d*d_k}$, $W^O \in \mathbb{R}^{h*d_k*d}$ b_i^Q, b_i^K, and b_i^V are learnable matrices, h is set at 8, $d_k = \frac{d}{h} = \frac{256}{8} = 32$.

The multi-head attention mechanism was the primary constituent in constructing the fusion module of TransCDR. More specifically, the fusion module consisted of 6 identical multi-head attention layers. The output of this module was then flattened and incorporated into a regression module. Our study delved into the inquiry of 3 attention modules, namely self-attention, drug-cell line attention (DCA), and cell line-drug attention (CDA) (Fig. 8). First, we created a single input matrix that contains information from both domains by concatenating the drug and cell line features. This allowed the multi-head attention mechanism to compute attention weights for each feature, which indicated the importance of that feature relative to others. For example, in the self-attention case, the attention weights were computed for each of the 6 features (3 drug and 3 cell line features). This enabled the model to learn to weigh the importance of each feature based on its relevance to the prediction task, ultimately improving model performance.

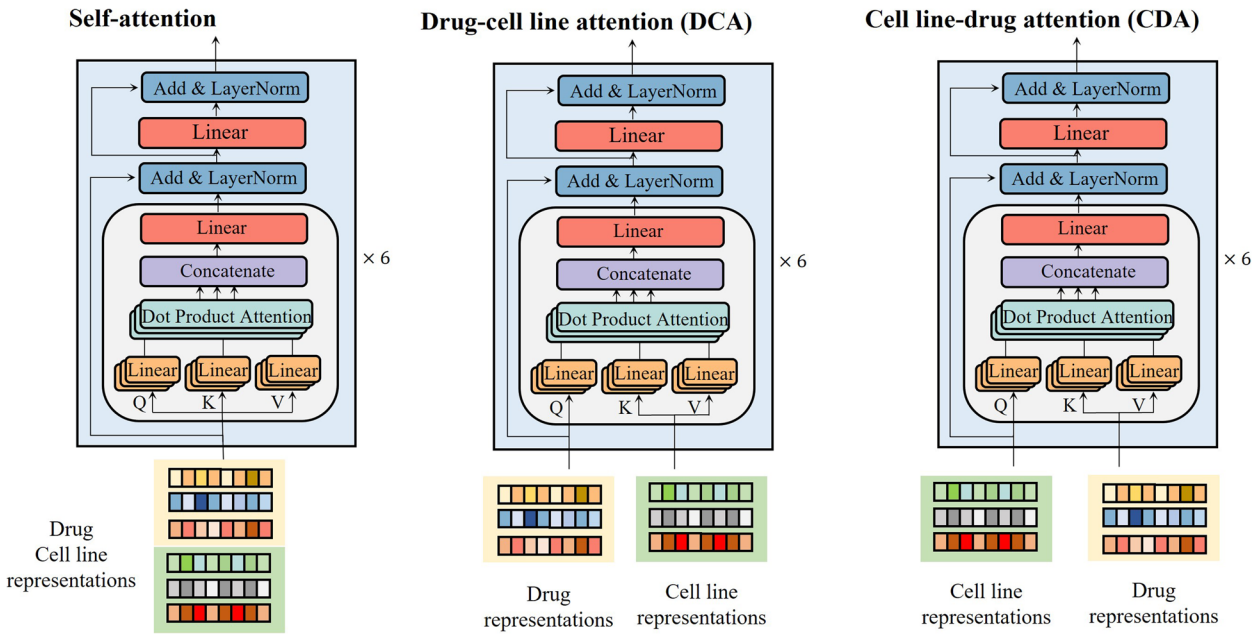

Fig. 8 Illustration of the three attention modules employed in this study. **A** Self-attention: a module concatenates drug and cell line representations to serve as the Q, K, and V parameters. **B** Drug-cell line attention: Q denotes the drug representations, and K and V correspond to the cell line representations. **C** Cell line-drug attention: Q denotes the cell line representations, and K and V correspond to the drug representations

Prediction module

The prediction module comprised a four-layer neural network incorporating rectified linear unit (ReLU) activation functions and dropout layers characterized by a dropout rate of 0.1. The output value pertained to a cell line's predicted drug response/sensitive state. To train regression/classification models, we adopted either the mean squared error (MSE) or binary cross entropy (BCE) loss function, which we subsequently back-propagated to the network and then updated all parameters end-to-end.

Performance metrics

For the regression experiments predicting $\ln(IC_{50})$ values of drugs and cell lines, we assessed TransCDR's performance using 4 evaluation measures: root mean square error (*RMSE*), Pearson correlation coefficient (*PC*), Spearman's rank correlation coefficient (*SC*), and concordance index (*C-index*) [25]. *RMSE* was used to calculate the difference between predicted and ground truth IC_{50} values:

$$RMSE = \sqrt{\frac{1}{N}\sum (y_i - \widetilde{y}_i)^2} \tag{16}$$

where N denotes the size of the test set. y_i and $\widetilde{y}_i$ represent the ground truth and predicted IC_{50} values, respectively. *PC* and *SC* measured the linear and rank-based correlations between ground truth and predicted IC_{50} values across all test samples. Additionally, we evaluated TransCDR's predictions using the *C-index* across all test samples. Upon deploying the trained model to predict CDRs in the GDSC and CCLE datasets, we evaluated its predictive performance across diverse cell types and drugs by calculating the *PC* and *SC* and the *C-index* for individual cell lines or drugs.

For the classification experiments, we evaluated the performance of each method using the area under the receiver operating characteristics (*AUROC*) and the Area Under the Precision-Recall (*AUPR*) curves across all test samples. *AUPR* was used as the primary metric, especially when negative samples were much more extensive than positive ones [71].

Lastly, we employed the two-sided Wilcoxon rank sum test with a significance threshold 0.05 to demonstrate the significant performance difference between TransCDR and other compared models. We reported the mean and standard deviation of metrics obtained by executing 10-CV for each method.

GSEA

Performing GSEA on the omics data of patients from TCGA can offer valuable biological insight into TransCDR. We employed the trained TransCDR classification model to evaluate 225 drugs on 7675 patients with the available 3 omics profiles from TCGA. For each drug, patients were ranked based on their prediction score, and the top and bottom 5% (384 of each) were classified as drug-sensitive and drug-resistant, respectively (Fig. 9). The difference in predicted score between drug-sensitive and drug-resistant patients was calculated using the formula:

$$Diff_{\text{drug}} = \overline{S}_{\text{sen}} - \overline{S}_{\text{res}} \tag{17}$$

Furthermore, we sorted $Diff_{\text{drug}}$ in descending order to further identify the top 10 drugs to analyze the biological mechanisms underlying drug sensitivity/resistance. We calculated the $\log_2$ fold change ($\log_2 FC$) of genes for each drug between drug-sensitive and resistant patients:

$$\log_2 FC = \log_2(\overline{X}_{\text{sen}}/\overline{X}_{\text{res}}) \tag{18}$$

where $\overline{X}_{\text{sen}}$ and $\overline{X}_{\text{res}}$ represent the gene-wise mean expression levels for the drug-sensitive and drug-resistant patient cohorts, respectively. We conducted GSEA on the differentially expressed genes with $\log_2 FC$ using the *clusterProfiler* R package and Molecular Signature Database v2023, which contains 33,591 gene sets across

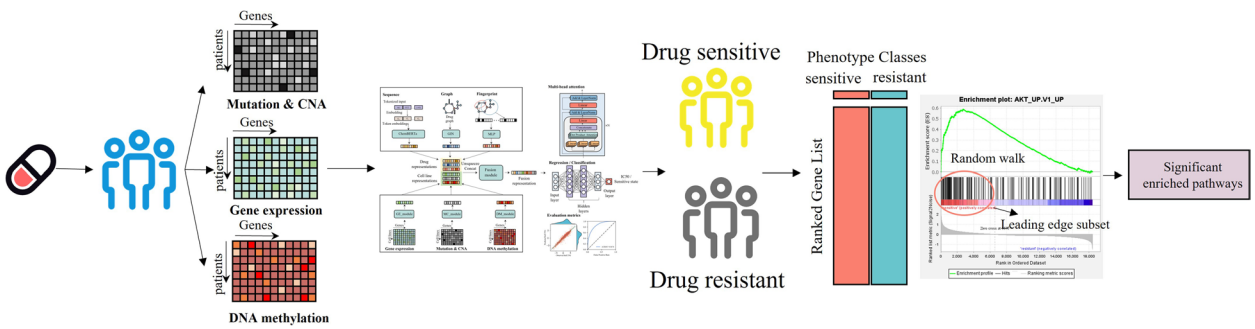

Fig. 9 The workflow of GSEA. The patients are classified into drug-sensitive or drug-resistant groups using the TransCDR classification model. The GSEA method is employed to identify pathways that are significantly enriched

9 major collections. Gene sets were significantly enriched if they had both Benjamini–Hochberg corrected p-value and FDR q-value < 0.01 and $|NES| \geq 1.9$.

Implementation details

All models based on neural networks, including TransCDR, DeepTTA, GraphDRP, TGSA, DRPreter, and TransCDR varieties, were developed using Pytorch. The training process was limited to 100 epochs for all training sets and models. The Adam optimizer with a learning rate of 10^{-5} was used to update the model parameters during the back-propagation process. The batch size was 64, MSELoss was employed as the loss function for regression models, and BCELoss for classification models. A dropout rate of 0.1 was specified, and the validation set was used to fine-tune hyperparameters and stop the training process. All models employed early stopping to prevent overfitting. The specific procedure is as follows: after each epoch of training, the model's *MSE* on the validation set is calculated. If the current *MSE* is less than the best *MSE*, the current model is retained as the best model, and the best *MSE* is updated to the current *MSE*. If the current *MSE* is greater than the best *MSE*, the training process is terminated, and the best model is the one retained from the previous iteration. The TransCDR further improves upon early stopping by introducing a more stringent criterion, where training is terminated if the best *MSE* does not update for 5 consecutive epochs, thereby enabling the model to converge more effectively to the optimal solution. For specific implementation details, please refer to the original code repositories for these models.

TransCDR: https://github.com/XiaoqiongXia/TransCDR;

DeepTTA: https://github.com/jianglikun/DeepTTC;

GraphDRP: https://github.com/hauldhut/GraphDRP;

TGSA: https://github.com/violet-sto/TGSA;

DRPreter: https://github.com/babaling/DRPreter; and.

DeepCDR: https://github.com/kimmo1019/DeepCDR.

All experiments were conducted on Tesla A100 GPUs with 40 GB of memory. GSEA was conducted on RStudio. For further details, please refer to the respective GitHub repository: https://github.com/XiaoqiongXia/TransCDR and https://doi.org/10.5281/zenodo.7912777.

Abbreviations

CDR	Cancer drug responses
GDSC	Genomics of Drug Sensitivity in Cancer
CCLE	Cancer Cell Line Encyclopedia
CNN	Convolutional neural networks
GNN	Graph neural networks
ECFP	Extended-connectivity fingerprints
SMILES	Simplified Molecular Input Line Entry System
GIN	Graph isomorphism network
IC_{50}	Half maximal inhibitory concentration
GSEA	Gene Set Enrichment Analysis
MC	Copy number aberration
GE	Gene expression
DM	DNA methylation
10-CV	10-Fold cross-validation
FC	Full connected layers
DCA	Drug-cell line attention
CDA	Cell line-drug attention
ReLU	Rectified linear unit
MSE	Mean squared error
BCE	Binary cross entropy loss
RMSE	Root mean square error
PC	Pearson correlation coefficient
SC	Spearman's rank correlation coefficient
C-index	Concordance index
AUROC	Area under the Receiver Operating Characteristics
AUPR	Precision-Recall curves
log_2FC	Log_2 fold change

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-02023-8.

Additional file 1: Fig. S1. Sensitivity analysis of TransCDR's parameters, revealing robustness to learning rate and batch size.

Additional file 2: Fig. S2. Comparative performance analysis of TransCDR with different data segmentation strategies. Model evaluation results of TransCDR using more strict strategies (A) drug split and (B) cell line strategies. In (B), the cell lines are clustered into 10, 50, 100, and 200 clusters using the K-means algorithm.

Additional file 3: Fig. S3. The visualization of attention map, where the color intensity corresponds to the attention weight value.

Additional file 4: Table S1. The regression performance of TransCDR on GDSC across cancer types.

Additional file 5: Table S2. The regression performance of TransCDR on GDSC across cell lines.

Additional file 6: Table S3. The regression performance of TransCDR on GDSC across drugs.

Additional file 7: Table S4. The top 10% predicted IC50 values of missing CDRs in GDSC.

Additional file 8: Table S5. The top 15 drug-cell line pairs predicted by TransCDR.

Additional file 9: Table S6. Test results of TransCDR and other compared models on CCLE dataset with specific cancer types.

Additional file 10: Table S7. The CDRs identified by TransCDR.

Additional file 11: Table S8. Patient screening results for drugs.

Additional file 12: Table S9. Summary of GSEA results.

Acknowledgements

We acknowledge the support provided by the project of Shanghai's Double First-Class University Construction, the Development of High-Level Local Universities: Intelligent Medicine Emerging Interdisciplinary Cultivation Project. The research is supported by the Medical Science Data Center of Fudan University. The authors express their gratitude to Professor YQZ for his insightful suggestions on the paper.

Authors' contributions

XXQ collected and analyzed the data, and drafted the manuscript. CYZ assisted with data analysis and reviewed the manuscript. FZ and LL provided guidance on the project and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai.

Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. The Python and Torch implementation of the TransCDR model is accessible at https://doi.org/10.5281/zenodo.7912777 or https://github.com/XiaoqiongXia/TransCDR.

• GDSC: https://www.cancerrxgene.org. Cancer cell line and drug response data, as well as cell lines' mutation, gene expression, and DNA methylation profiles, and drug SMILES data, are available for download.

• CCLE: https://sites.broadinstitute.org/ccle/. Cancer cell line and drug response data can be downloaded.

• TCGA: https://www.cancer.gov/ccg/research/genome-sequencing/tcga. Cancer patients' mutation, gene expression, and DNA methylation profiles are available for collection.

• MSigDB: https://www.gsea-msigdb.org/gsea/msigdb/. Annotated gene sets for use with GSEA can be downloaded.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 20 February 2024 Accepted: 30 September 2024
Published online: 09 October 2024

References

1. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21(8):846–53.
2. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526(7573):336–42.
3. Vargas AJ, Harris CC. Biomarker development in the precision medicine era: lung cancer as a case study. Nat Rev Cancer. 2016;16(8):525–37.
4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. 2017;18:83.
5. Yang WJ, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(D1):D955–61.
6. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6(10):813–23.
7. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
8. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics. 2020;36(Supplement_2):i911–8.
9. Jiang L, Jiang C, Yu X, Fu R, Jin S, Liu X. DeepTTA: a transformer-based model for predicting cancer drug response. Briefings in Bioinformatics. 2022;23(3):bbac100.
10. Nguyen T, Nguyen GTT, Nguyen T, Le DH. Graph convolutional networks for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf. 2022;19(1):146–54.
11. Sun MY, Zhao SD, Gilvary C, Elemento O, Zhou JY, Wang F. Graph convolutional networks for computational drug development and discovery. Brief Bioinform. 2020;21(3):919–35.
12. Nguyen GTT, Vu HD, Le DH. Integrating molecular graph data of drugs and multiple-omic data of Cell Lines for drug response prediction. IEEE/ACM Trans Comput Biol Bioinf. 2022;19(2):710–7.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv preprint. 2017. arXiv:1706.03762.
14. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Brief Bioinform. 2021;22(1):360–79.
15. Cheng X, Dai C, Wen Y, Wang X, Bo X, He S, et al. NeRD: a multichannel neural network to predict cellular response of drugs by integrating multi-dimensional data. BMC Med. 2022;20(1):368.
16. Chen YR, Zhang LX. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Brief Bioinform. 2022;23(1):bbab378.
17. Zhu Y, Ouyang Z, Chen W, Feng R, Chen DZ, Cao J, et al. TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. Bioinformatics. 2022;38(2):461–8.
18. Shin J, Piao Y, Bang D, Kim S, Jo K. DRPreter: interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. Int J Mol Sci. 2022;23(22):13919.
19. Weininger D. Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules. J Chem Inform Comput Sci. 1988;28(1):31–6.
20. An X, Chen X, Yi DY, Li HY, Guan YF. Representation of molecules for drug response prediction. Brief Bioinform. 2022;23(1):bbab393.
21. Zhuang FZ, Qi ZY, Duan KY, Xi DB, Zhu YC, Zhu HS, et al. A comprehensive survey on transfer learning. Proc IEEE. 2021;109(1):43–76.
22. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res. 2011;12:2493–537.
23. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das PJae-p. Large-scale chemical language representations capture molecular structure and properties2021 June 01, 2021:arXiv:2106.09553 p. Available from: https://ui.adsabs.harvard.edu/abs/2021arXiv210609553R.
24. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks. arXiv preprint. 2019. arXiv:1905.12265.
25. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. EVALUATING THE YIELD OF MEDICAL TESTS. JAMA. 1982;247(18):2543–6.
26. Moxley KM, McMeekin DS. Endometrial carcinoma: a review of chemotherapy, drug resistance, and the search for new agents. Oncologist. 2010;15(10):1026–33.
27. Takemura K, Noguchi M, Ogi K, Tokino T, Kubota H, Miyazaki A, et al. Enhanced Bax in oral SCC in relation to antitumor effects of chemotherapy. J Oral Pathol Med. 2005;34(2):93–9.
28. Punzo F, Tortora C, Di Pinto D, Pota E, Argenziano M, Di Paola A, et al. Bortezomib and endocannabinoid/endovanilloid system: a synergism in osteosarcoma. Pharmacol Res. 2018;137:25–33.
29. Zhang LS, He M, Zhang YQ, Nilubol N, Shen M, Kebebew E. Quantitative high-throughput drug screening identifies novel classes of drugs with anticancer activity in thyroid cancer cells: opportunities for repurposing. J Clin Endocrinol Metab. 2012;97(3):E319–28.
30. Durkin A, Vu HY, Lee H. The VR23 antitumor compound also shows strong anti-inflammatory effects in a human rheumatoid arthritis cell model and acute lung inflammation in mice. J Immunol. 2020;204(4):788–95.
31. Matsuda R, Sakagami H, Amano S, Iijima Y, Sano M, Uesawa Y, et al. Inhibition of neurotoxicity/anticancer activity of bortezomib by caffeic acid and chlorogenic acid. Anticancer Res. 2022;42(2):781–90.
32. Rao RJR, Rao AKSB, Swapna K, Rani BB, Kumar SP, Awantika S, et al. Design, synthesis and biological evaluation of novel analogs of bortezomib. J Korean Chem Soc. 2011;55(5):765–75.
33. Lesinski GB, Raig ET, Guenterberg K, Brown L, Go MR, Shah NN, et al. IFN-alpha and bortezomib overcome Bcl-2 and Mcl-1 overexpression in melanoma cells by stimulating the extrinsic pathway of apoptosis. Can Res. 2008;68(20):8351–60.
34. Amiri KI, Horton LW, LaFleur BJ, Sosman JA, Richmond A. Augmenting chemosensitivity of malignant melanoma tumors via proteasome inhibition: implication for bortezomib (VELCADE, PS-341) as a therapeutic agent for malignant melanoma. Can Res. 2004;64(14):4912–8.
35. Calastretti A, Rancati F, Ceriani MC, Asnaghi L, Canti G, Nicolin A. Rapamycin increases the cellular concentration of the BCL-2 protein and exerts an anti-apoptotic effect. Eur J Cancer. 2001;37(16):2121–8.
36. Manasanch EE, Orlowski RZ. Proteasome inhibitors in cancer therapy. Nat Rev Clin Oncol. 2017;14(7):417–33.
37. Syed YY. Sacituzumab govitecan: first approval. Drugs. 2020;80(10):1019–25.

38. Moreau P, Richardson PG, Cavo M, Orlowski RZ, San Miguel JF, Palumbo A, et al. Proteasome inhibitors in multiple myeloma: 10 years later. Blood. 2012;120(5):947–59.
39. Das T, Anand U, Pandey SK, Ashby CR, Assaraf YG, Chen ZS, et al. Therapeutic strategies to overcome taxane resistance in cancer. Drug Resist Updates. 2021;55:100754.
40. Caputi L, Franke J, Farrow SC, Chung K, Payne RME, Nguyen TD, et al. Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. Science. 2018;360(6394):1235–8.
41. García-Morales P, Carrasco-García E, Ruiz-Rico P, Martínez-Mira R, Menéndez-Gutiérrez MP, Ferragut JA, et al. Inhibition of Hsp90 function by ansamycins causes downregulation of cdc2 and cdc25c and G(2)/M arrest in glioblastoma cell lines. Oncogene. 2007;26(51):7185–93.
42. Jane EP, Pollack IF. The heat shock protein antagonist 17-AAG potentiates the activity of enzastaurin against malignant human glioma cells. Cancer Lett. 2008;268(1):46–55.
43. Sheppard KE, Cullinane C, Hannan KM, Wall M, Chan J, Barber F, et al. Synergistic inhibition of ovarian cancer cell growth by combining selective PI3K/mTOR and RAS/ERK pathway inhibitors. Eur J Cancer. 2013;49(18):3936–44.
44. Wainberg ZA, Alsina M, Soares HP, Braña I, Britten CD, Del Conte G, et al. A multi-arm phase I study of the PI3K/mTOR inhibitors PF-04691502 and gedatolisib (PF-05212384) plus irinotecan or the MEK inhibitor PD-0325901 in advanced cancer. Target Oncol. 2017;12(6):775–85.
45. Hurvitz SA, Shatsky R, Harbeck N. Afatinib in the treatment of breast cancer. Expert Opin Investig Drugs. 2014;23(7):1039–47.
46. Jain P, Khanal R, Sharma A, Yan F, Sharma N. Afatinib and lung cancer. Expert Rev Anticancer Ther. 2014;14(12):1391–406.
47. Coldren CD, Helfrich BA, Witta SE, Sugita M, Lapadat R, Zeng C, et al. Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. Mol Cancer Res. 2006;4(8):521–8.
48. Park K, Tan EH, O'Byrne K, Zhang L, Boyer M, Mok T, et al. Afatinib versus gefitinib as first-line treatment of patients with EGFR mutation-positive non-small-cell lung cancer (LUX-Lung 7): a phase 2B, open-label, randomised controlled trial. Lancet Oncol. 2016;17(5):577–89.
49. Hollern DP, Swiatnicki MR, Andrechek ER. Histological subtypes of mouse mammary tumors reveal conserved relationships to human cancers. PLoS Genet. 2018;14(1):e1007135.
50. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24.
51. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics. 2016;32(19):2891–5.
52. Cao R, Fang LY, Lu T, He NJ. Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification. IEEE Geosci Remote Sens Lett. 2021;18(1):43–7.
53. Yan CG, Meng LX, Li L, Zhang JH, Wang Z, Yin J, et al. Age-Invariant Face Recognition by Multi-Feature Fusion and Decomposition with Self-attention. Acm Transactions on Multimedia Computing Communications and Applications. 2022;18(1).
54. Jia S, Min ZC, Fu XY. Multiscale spatial-spectral transformer network for hyperspectral and multispectral image fusion. Information Fusion. 2023;96:117–29.
55. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods. 2024.
56. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nature Methods. 2024.
57. Liu X, Song C, Huang F, Fu H, Xiao W, Zhang W. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. Brief Bioinform. 2022;23(1):bbab457.
58. Huang Z, Zhang P, Deng L. DeepCoVDR: deep transfer learning with graph transformer and cross-attention for predicting COVID-19 drug response. Bioinformatics. 2023;39(39 Suppl 1):i475–83.
59. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
60. Kim S, Chen J, Cheng TJ, Gindulyte A, He J, He SQ, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021;49(D1):D1388–95.
61. RDKit: Open-source cheminformatics. https://www.rdkit.org.
62. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54.
63. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena platform for cancer genomics data visualization and interpretation. bioRxiv preprint. 2018. bioRxiv:326470.
64. Chithrananda S, Grand G, Ramsundar BJae-p. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction2020 October 01, 2020:[arXiv:2010.09885 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv201009885C.
65. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al., editors. Transformers: State-of-the-Art Natural Language Processing. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. Electr Network2020.
66. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
67. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics. 2020;36(22–23):5545–7.
68. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need2017 June 01, 2017:[arXiv:1706.03762 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V.
69. Bao H, Wang W, Dong L, Liu Q, Mohammed OK, Aggarwal K, et al. VLMo: unified vision-language pre-training with mixture-of-modality-experts. arXiv preprint. 2021. arXiv:2111.02358.
70. Kim W, Son B, Kim I. ViLT: vision-and-language transformer without convolution or region supervision. arXiv preprint. 2021. arXiv:2102.03334.
71. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

Publisher's Note