Briefings in Bioinformatics, 22(5), 2021, 1–10

https://doi.org/10.1093/bib/bbab048
Problem Solving Protocol

DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration

Yongcui Wang, Yingxi Yang, Shilong Chen and Jiguang Wang

Corresponding authors: Yongcui Wang, Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, Qinghai, 810008, China. Tel: +86 09716143065; E-mail: ycwang@nwipb.cas.cn; Jiguang Wang, Division of Life Science and Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China. Tel: +852 34692672; E-mail: jgwang@ust.hk.

Abstract

Recent pharmacogenomic studies that generate sequencing data coupled with pharmacological characteristics for patient-derived cancer cell lines led to large amounts of multi-omics data for precision cancer medicine. Among various obstacles hindering clinical translation, lacking effective methods for multimodal and multisource data integration is becoming a bottleneck. Here we proposed DeepDRK, a machine learning framework for deciphering drug response through kernel-based data integration. To transfer information among different drugs and cancer types, we trained deep neural networks on more than 20 000 pan-cancer cell line-anticancer drug pairs. These pairs were characterized by kernel-based similarity matrices integrating multisource and multi-omics data including genomics, transcriptomics, epigenomics, chemical properties of compounds and known drug-target interactions. Applied to benchmark cancer cell line datasets, our model surpassed previous approaches with higher accuracy and better robustness. Then we applied our model on newly established patient-derived cancer cell lines and achieved satisfactory performance with AUC of 0.84 and AUPRC of 0.77. Moreover, DeepDRK was used to predict clinical response of cancer patients. Notably, the prediction of DeepDRK correlated well with clinical outcome of patients and revealed multiple drug repurposing candidates. In sum, DeepDRK provided a computational method to predict drug response of cancer cells from integrating pharmacogenomic datasets, offering an alternative way to prioritize repurposing drugs in precision cancer treatment. The DeepDRK is freely available via https://github.com/wangyc82/DeepDRK.

Key words: drug repurposing; multi-omics data sources; kernel-based data integration; machine learning; cancer precision medicine

INTRODUCTION

High-throughput screening of small molecules against a large number of cancer cell lines (CCLs) provided an unprecedented opportunity to characterize the genetic contexts for distinct cancer vulnerabilities [1–5]. Significant efforts have been devoted to decoding the relationship between genetic signatures and drug response based on emerging multi-omics data, including transcriptomics, proteomics, genomics and epigenomics [6–12].

Yongcui Wang is an associate professor in Key Laboratory of Adaptation and Evolution of Plateau Biota at Northwest Institute of Plateau Biology, Chinese Academy of Sciences, China. Her research activity is focused on the cancer genomics and cancer precision treatment.
Yingxi Yang is a PhD student in Department of Chemical and Biological Engineering at The Hong Kong University of Science and Technology, China.
Shilong Chen is a professor in Key Laboratory of Adaptation and Evolution of Plateau Biota at Institute of Sanjiangyuan National Park, Chinese Academy of Sciences, China.
Jiguang Wang is an assistant professor in Division of Life Science, Department of Chemical and Biological Engineering, and State Key Laboratory of Molecular Neuroscience at The Hong Kong University of Science and Technology, China. His research activity includes cancer evolution, cancer precision medicine and noncoding genome.
Submitted: 28 August 2020; **Received (in revised form):** 16 January 2021

In addition, properties of chemical compounds and their protein targets have been used to further improve the prediction of drug sensitivity [13]. Yet, substantial complexity and heterogeneity of multi-omics and multisource data limited the predictive power of the computational models challenging potential clinical translation. Therefore, novel integrative models/frameworks that efficiently incorporate heterogeneous multi-omics data and simultaneously integrate *in vitro* and/or clinical response of multiple drugs from various sources are urgently needed.

Recently, deep learning framework (DLF) has been widely used to solve complex problems using multilayer feature extraction and transformation via a cascade of nonlinear processing units, forming a hierarchical representation of either labeled or unlabeled subjects. Unlike previous methods that manually select features, DLF adopts efficient algorithms for feature extraction, leading to scalable, flexible and stable models for either unsupervised or supervised problems [14–18]. DLF outperformed the state-of-the-art methods in various fields including but not limited to imaging analysis, video/audio classification and recognition [19–24]. Moreover, DLF has been applied to the field of biomedicine, advancing a variety of areas such as genome-guided precision cancer medicine [25–38]. However, the DLF methods typically require a big number of data points, and few studies integrated heterogeneous data types and multiple data sources [39–40].

Here we proposed DeepDRK to integrate heterogeneous multi-omics data, for predicting drug response of cancer cells. To simultaneously consider different available drugs across CCLs, we applied a multitasking strategy under the assumption that drugs with similar chemical properties should have similar treatment outcomes. To overcome the difficulty of heterogeneous data integration, we adapted a kernel method to construct similarity matrix based on different types of feature. First of all, DeepDRK was trained on two well-established pharmacogenomic resources, *i.e.* Cancer Therapeutics Response Portal (CTRP) [41] and the Genomics of Drug Sensitivity in Cancer (GDSC) [42]. We compared DeepDRK with previous prediction methods on both datasets and found that DeepDRK outperformed the previous methods in terms of accuracy and robustness. Secondly, we applied DeepDRK to the early-passage patient-derived tumor cells and found that the predictions were compatible with the experimental outcome. Thirdly, we predicted clinical response of TCGA (The Cancer Genome Atlas) cancer patients and showed that DeepDRK prediction was significantly correlated with real clinical outcome in various cancer types. Lastly, DeepDRK has been developed into an open-access software for research usage.

METHODS

The DeepDRK model

To predict drug response, we developed DeepDRK, a deep learning model to integrate data from different sources, diverse cancer types and various chemical compounds. As illustrated in Figure 1, a kernel-based approach was employed to generate integrative representation of interacting partners of a CCL and an anticancer drug, which was subsequently used to train the deep neural networks (DNNs) for drug response prediction. In particular, we firstly collected various types of multi-omics data to, respectively, construct multiple kernel-based similarity matrices of CCLs (Figure 1B1, Methods Multi-omics data integration via kernel methods), and then used chemical feature and drug-target interaction of compounds to, respectively, calculate two similarity matrices of anticancer drugs (Figure 1B2). In addition, the drug screening data from different sources were binarized to represent the treatment response of CCLs by anticancer drugs (Methods Discretization of drug response), based on which a bipartite graph of CCL and drug was constructed with edges labeling the digitalized sensitivity value (Figure 1B3). Furthermore, CCL-drug pairs were represented by concatenating multiple similarity vectors (Figure S1), followed by the training of a classification model to predict drug efficacy. More details of the DeepDRK framework will be explained in the following subsections.

Discretization of drug response

Instead of estimating the continuous response value, we categorized response value into three classes, *i.e.* sensitive, resistant and unclear. To achieve this goal, we visualized the overall distribution of drug response data and manually selected cutoffs for digitalization. For example, when using CTRP data, we generated a histogram to represent the overall distribution of the area under the dose–response curve (AUCDR) of all drugs, and labeled CCL-drug pairs with AUCDR<6 as sensitive (red in Figure S2) and those with AUCDR>16 as resistant (blue in Figure S2). Subsequently, a bipartite graph was constructed to represent the above defined associations (lower panel of Figure S2). In this graph, two types of nodes in the bipartite graph were used to represent drugs and cancer cells, respectively. The edges were used to represent either sensitive (S) or resistant (R) response of the corresponding cancer cells upon drug treatment.

Multi-omics data integration via kernel methods

To integrate heterogeneous data, we generated multiple similarity matrices of both CCLs and anticancer drugs using kernel methods. For CCLs, different types of features were individually used to calculate the similarity matrices between all available CCLs (upper panel of Figure S1A and B). In this process, the radial basis function kernel was adopted to map different data types into the same kernel Hilbert space. Particularly, for copy number alteration, DNA methylation and gene expression data, the following formula was used:

$$Sc\left(c,c'\right)=\exp\left(-\gamma\left\|x_c-x_{c'}\right\|^2\right)$$

where c and c' were two CCLs; x_c and $x_{c'}$ were values of the corresponding feature type; and γ was a predefined kernel parameter. Here we used different γ in different data types, with 0.001 for copy number alteration, 0.1 for DNA methylation and 0.01 for gene expression in order to normalize all data into a comparable scale. Regarding to mutation data, a Hamming distance [43] of CCL's mutation profile (x_c and $x_{c'}$) was integrated into the kernel function:

$$Sc\left(c,c'\right)=\exp\left(-HD\left(x_c,x_{c'}\right)\right)$$

where $HD(x_c, x_{c'})$ represented the Hamming distance.

In addition, two similarity matrices of anticancer drugs were also calculated (lower panel of Figure S1A and B). The similarity matrix based on drug targets was calculated using the above Hamming distance-based kernel function, while the Tanimoto coefficients [44] obtained from chemical molecular descriptors (QuaSAR-Descriptor in the Molecular Operating Environment (MOE)) were used to represent the drug similarities based on chemical property.

Next, for a given CCL and an anticancer drug, we extracted a corresponding vector from each feature similarity matrix and

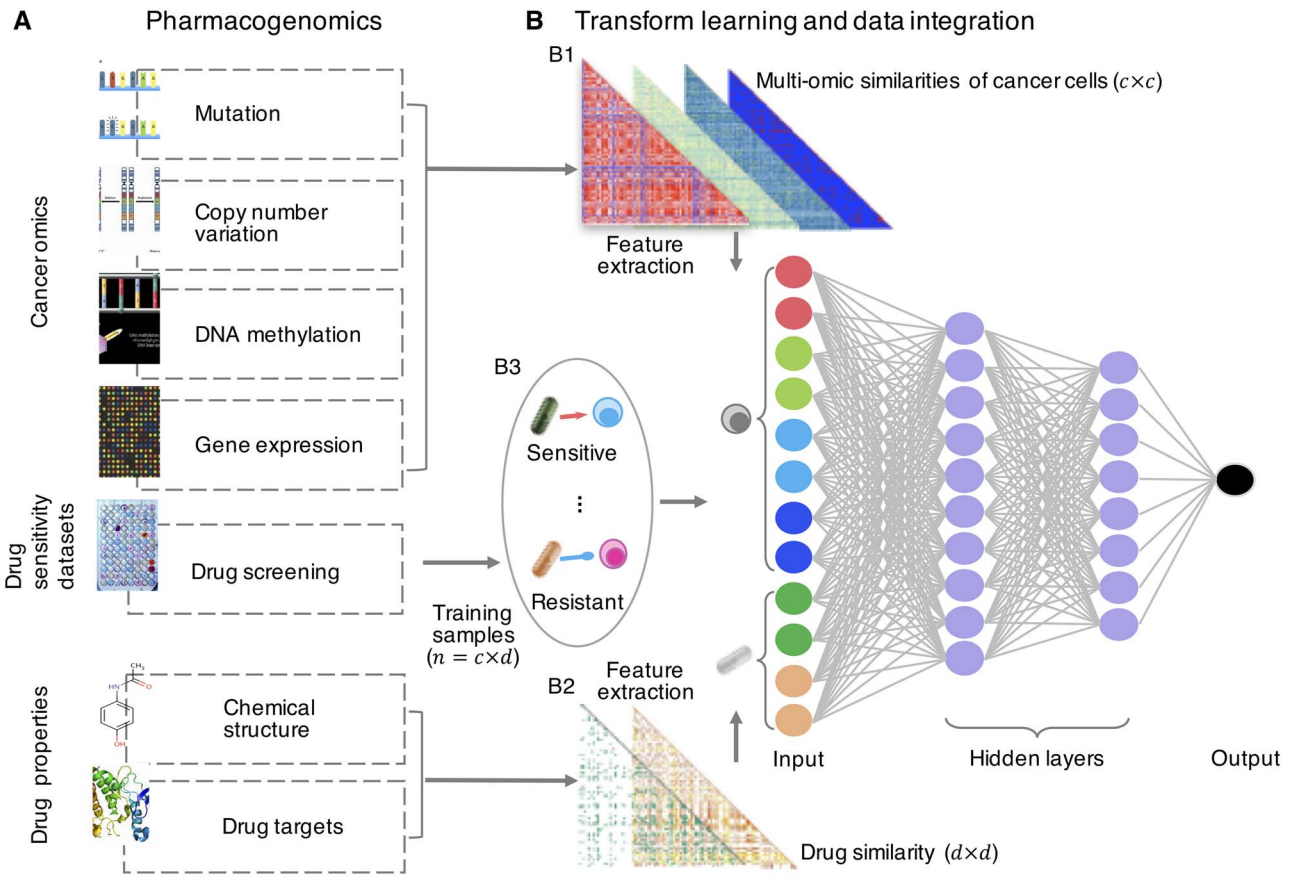

Figure 1. The framework of DeepDRK. (A) Heterogenous high-throughput pharmacogenomic data sources were integrated. Multi-omics data characterizing cancer cell lines (CCLs) and drug properties characterizing anticancer compounds were used to represent CCL-drug pairs, while drug screening data from multiple sources labeled these pairs. (B) A kernel-adapted deep neural network was proposed to learn the drug response in cancer cells. (B1) The cancer omics data were converted into similarity matrices to form the multi-omics representation of CCLs. (B2) Drug chemical structures and previously reported target proteins were converted into similarity matrices to form an integrated representation of anticancer drugs. (B3) A bipartite graph was constructed to label the relationship between CCL and anticancer drug based on digitalized drug sensitivity.

concatenated all relevant vectors of this CCL and the anticancer drug to represent the CCL-drug pair for further investigation (Figure S1C). All available CCL-drug pairs with the proper labels (defined as sensitive or resistant in section 2.1.1) were used as the training samples. Subsequently, we used this strategy to integrate multiple datasets to train fully connected feedforward neural networks with multiple hidden layers. In these DNNs, the input layer was the above concatenated vectors, and the output layer was a single node indicating the predicted drug sensitivity. The hyperbolic tangent function was selected as the activation function.

Model implementation and evaluation

To implement the DeepDRK model, the 'h2o' R package [45] was applied to train the feedforward DNN. Particularly, the hyperbolic tangent function (Tanh), a continuous function producing an output for all input values in scale of [−1, 1], was used as the activation function. In the training process, the cross-entropy function was used as the loss function, and we allowed each hidden layer to contain at most 200 nodes and the maximal iteration times to be 10. To optimize the model parameters, the Nesterov accelerated gradient method was carried out, and the area under the precision-recall curve (AUPRC) was used as the stopping metric. Moreover, the input dropout ratio was set as 0, while the dropout ratio for hidden layer was set as 0.5. Model performance was then evaluated through the 5-fold cross-validation based on different metrics including AUC (the area under receiver operating characteristic (ROC) curve) [46], AUPRC [47], Accuracy, Sensitivity, Specificity, Precision and F1 score (harmonic mean of precision and recall). The output of the well-trained DeepDRK classifiers was named as the DeepDRK score to represent the level of sensitivity of a CCL-drug pair.

Pharmacogenomic datasets

Two well-established cancer genomic resources, *i.e.* CTRP and GDSC, were used to train the DeepDRK model. The CTRP dataset contained the AUCDR value for 545 drugs across 887 CCLs. These cell lines were characterized by genomic and transcriptomic sequencing, leading to the multi-omics profiles that include genomic mutation, copy number alteration and gene expression (https://portals.broadinstitute.org/ccle). The GDSC dataset included the dose–response curves for 265 anticancer drugs across 1074 CCLs. We acquired genomic mutation, copy number alteration, DNA methylation and gene expression data of these GDSC CCLs from https://www.cancerrxgene.org/. Moreover, the chemical properties of compound for anticancer drugs were extracted from a collection of 2D molecular descriptors and calculated by QuaSAR-Descriptor in the Molecular Operating Environment (MOE v. 2011.10, Chemical Computing Group Inc., Montreal, Canada) based on chemical structures from PubChem

SDF files. The 2D MOE descriptors included physical properties, atom counts and bond counts. The target proteins were collected from DrugBank [48] and KEGG [49]. More details of these two datasets were listed in Supplementary Table S1.

TCGA patients

The molecular properties for cancer patients used in this study, including somatic mutation, copy number variation, DNA methylation and gene expression were acquired from TCGA data portal (https://gdac.broadinstitute.org/). The clinical drug response of TCGA cancer patients was extracted from Ding *et al.* [50] including 2182 patient-drug pairs between 1029 pan-cancer patients and 130 clinical drugs.

RESULTS

Evaluation of DeepDRK in benchmark datasets

We first assessed the performance of DeepDRK on two benchmark datasets: CTRP and GDSC (Table S1). Based on drug screening data in each dataset, all CCL-drug pairs were grouped into sensitive (S), resistant (R) and unclear (Method, Figure S2). Pairs in categories S and R were used as the training samples of DeepDRK. To evaluate the performance within a benchmark dataset, 5-fold cross-validation procedure was carried out. Considering that the number of samples labeled as positive and that labeled as negative was unbalanced, the AUPRC was applied as the evaluation criterion. As a result, DeepDRK achieved the AUPRC of 0.97 on CTRP dataset, and 0.96 on GDSC dataset. Compared to conventional methods such as Support Vector Machine (SVM) and Random Forest (RF), our model showed better accuracy (high AUPRC) and robustness (low standard deviation) (Figure 2C). Yet, we realized that the deep learning model might be overfitted due to the potential information linking during the calculation of similarity matrix in cross-validation. We therefore performed the hold-out validation on both benchmark datasets (CTRP and GDSC). Particularly, 80% of the drug-cell pairs of a given dataset was used to construct the similarity matrix and train the DeepDRK model, which was then tested in the other 20% of unseen pairs. We found that although the performance of the hold-out validation was slightly worse than that of the initial cross-validation, AUC and AUPRC of DeepDRK were over 0.85 on both benchmark datasets (Figure 2D). To demonstrate the effectiveness of multi-omics data integration, more experiments were carried out to evaluate the model performance of DeepDRK with a single data type. As expected, the model integrating all data types achieved the best performance (Tables S2 and S3).

Generalization ability of DeepDRK

To test whether DeepDRK works for drugs not seen by the classifier, we trained the DeepDRK model using both CTRP and GDSC (including 10,754 sensitive and 10,607 resistant CCL-drug pairs) and tested it in 16 anticancer drugs that were not included in the combined dataset. These 16 drugs were experimentally screened across 49 CCLs in another study [3], based on which we extracted 196 sensitive and 195 resistant CCL-drug pairs as an independent test set (Figure S3). Interestingly, we found that DeepDRK scores of the sensitive pairs were significantly higher than that of the resistant ones (fold change = 3.35, $P < 2e^{-16}$, Figure 3A). To compare the prediction performance of different training datasets, another two DeepDRK models were trained based on a single dataset (either CTRP or GDSC). We demonstrated that the model integrating both datasets achieved AUC of 0.93 (Figure 3B) and AUPRC of 0.92 (Figures 3C), which outperformed both CTRP-based model and GDSC-based model. More comparisons using various metrics including accuracy, sensitivity, specificity, precision and F1 score were summarized in Figure S4.

Although the 16 tested drugs were not seen in the training set, DeepDRK was able to infer their profile of treatment outcome, mainly due to the rule of 'guilty by association', which implied that if two drugs had the same drug targets or shared chemical properties, they might have a higher chance to behave similarly in inhibiting CCLs (Figure 3D). For instance, Lapatinib, a dual tyrosine kinase inhibitor targeting *HER2* and *EGFR* pathways in breast cancer and other solid tumors, was not studied in CTRP or GDSC, while the CTRP project investigated Neratinib across 850 CCLs. As these two compounds have similar chemical properties (similarity = 0.913) and they share the same targets (Figure 3E), the deep learning model automatically used Neratinib as a 'template' to infer Lapatinib's treatment outcome. Accordingly, DeepDRK predicted potential response of 27 CCLs upon Lapatinib treatment. Overall, this prediction achieved 0.85 (23/27) accuracy in comparing with experimental data (Figure 3F).

Collectively, our deep learning model successfully integrated multisource and multimodal data, outperformed traditional machine learning methods, and was able to predict potential efficacy of new drugs with proper characterization.

Application of DeepDRK to patient-derived CCLs

We then applied DeepDRK to a more complex dataset containing drug screening data of a cohort of patient-derived cancer cell lines (PDCs). These PDCs were early passages of the three-dimensional tumor culture derived from newly diagnosed cancer patients, better representing heterogeneous genetic and molecular backgrounds of cancer patients in real world [51]. We acquired the drug sensitivity data of 60 drugs on 74 glioma PDCs, and separated the PDC-drug pairs into sensitive and resistant groups (Figure S5). The DeepDRK model integrating CTRP and GDSC was applied to predict sensitivity of these PDC-drug pairs. We found that the DeepDRK model achieved AUC of 0.84 and AUPRC of 0.77 (Figure 4A and B). Moreover, we found that the DeepDRK score of the sensitive group was significantly higher than that of the resistant group, further demonstrating the efficacy of our method in segregating PDCs into groups with different treatment outcome (Figure 4C).

Notably, DeepDRK provided an *in silicon* way to infer drug response of PDCs whose experimental results were missing. Figure 4D displayed the DeepDRK prediction on the anticancer drugs with partially known and partially unknown experimental outcomes. DeepDRK correlated tremendously well with the known response measurements with accuracy 0.91 (70 out of 77), and importantly, it revealed novel PDC-drug interactions when biological experiments were not available.

Application of DeepDRK to cancer patients

We then applied the DeepDRK model to predict drug response in cancer patients. According to a recent study, some patients in the cancer genome atlas (TCGA) had drug response information and therefore the patient-drug pairs can be divided into two classes: the responders and the nonresponders [50]. To apply the DeepDRK model to these patient-drug pairs, we first calculated the similarity between the TCGA patients and CCLs in our training datasets via multi-omics data and then calculated the similarity between the corresponding drugs that were applied to TCGA patients and the anticancer drugs that were

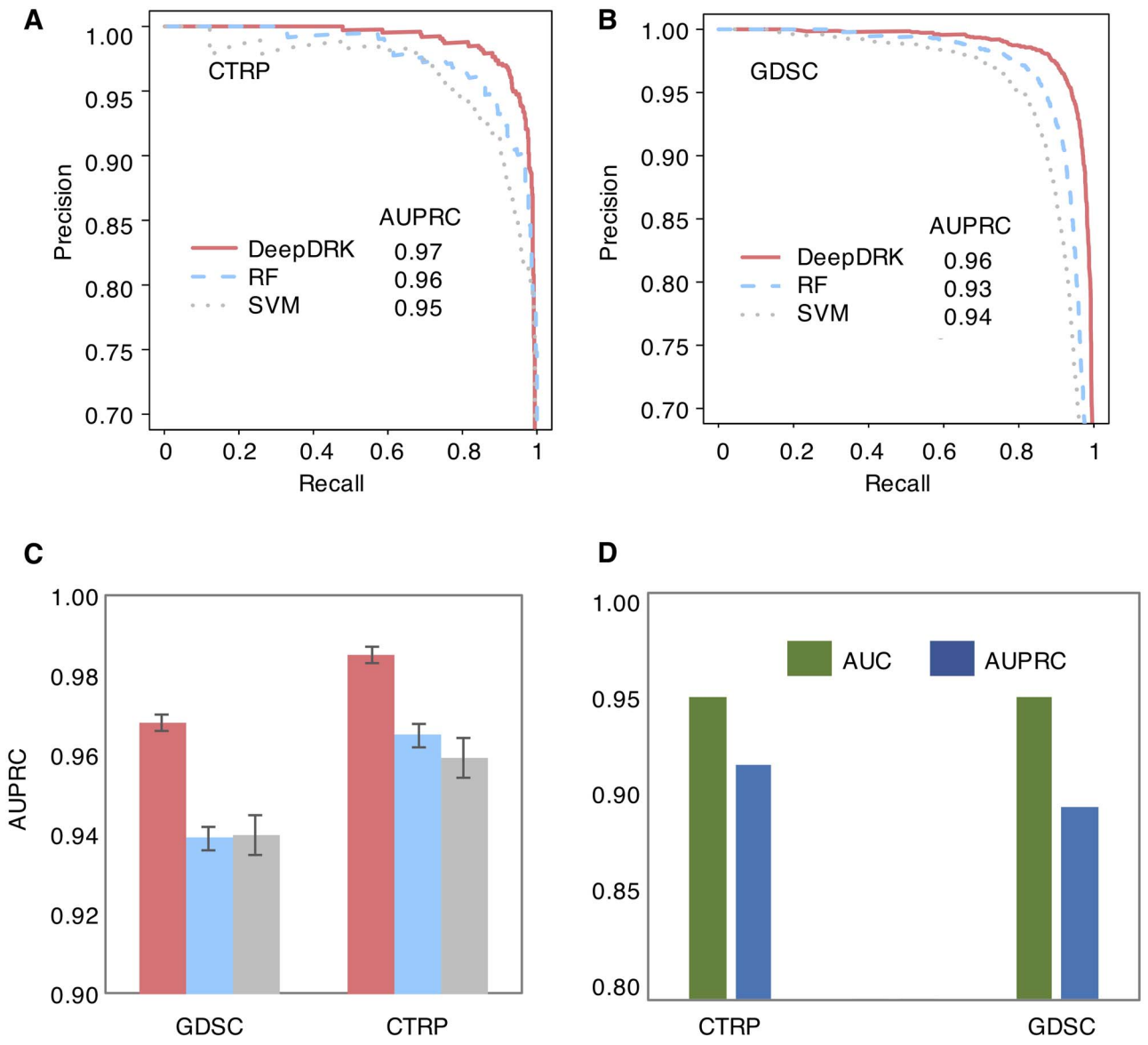

Figure 2. Comparison of DeepDRK with the baseline machine learning approaches on two benchmark datasets. (A-B) Precision-recall curves of different methods on CTRP and GDSC. (C) Comparison of the area under the precision-recall curve (AUPRC) of different methods. Error bars indicated the standard deviation of AUPRC in 10 repeated procedures of 5-fold cross-validation. RF: random forest; SVM: support vector machine. (D) Hold-out validation performance on these two benchmark datasets.

included in the training datasets. Concatenated similarity vectors were then used to represent the patient-drug pairs, which were subsequently adopted as the input to calculate DeepDRK score. Interestingly, we found that the DeepDRK score in the responder group was significantly higher than that in the nonresponder group (P-value < 0.001, Figure 5A). This observation was consistent in various cancer types, especially Breast Invasive Carcinoma (BRCA), Head–Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), Colon adenocarcinoma (COAD), Sarcomas (SARC) and Testicular Germ Cell Tumors (TGCT) (Figure 5B).

Note that some cancer types such as SARC and TGCT had no CCLs available in the combined training dataset, but DeepDRK can transfer knowledge from available tumor types to patients with novel cancer types. For instance, the SARC patient TCGA-LI-A67I was similar to a glioma CCL (LN405) in not only the profile of genomic mutation (Figure 5C) but also the overall gene expression (Figure 5D). Similarly, the TGCT patient TCGA-2G-AAHA was mirrored by a lung CCL SW900 (Figure 5C and D). Taking advantage of the similarities between cancer cells, DeepDRK was able to provide a meaningful prediction for pan-cancer patients based on multi-omics data integration.

To elaborate on the prediction of DeepDRK, we focused on the cancer types with relatively good performance where the responders had elevated DeepDRK scores compared to that of the nonresponders (Figure 5B). For each of these cancer types, we evaluated the impact of a driver mutation on the DeepDRK score of an anticancer drug. Significant interactions between driver genes and the selected FDA approved drugs (that were not only applied to TCGA patients but also investigated in at least one of the training datasets) were extracted and visualized via the gene-drug networks (Figure 6), providing a list of

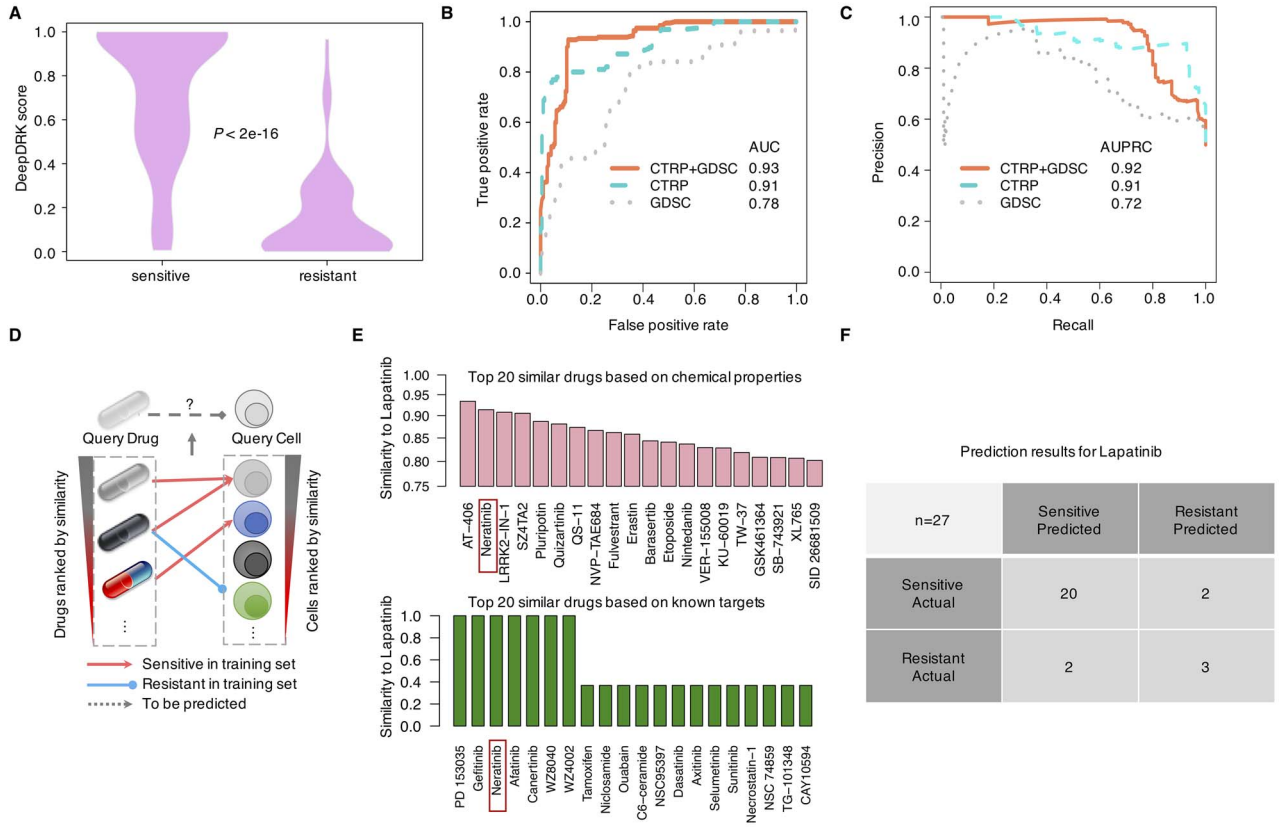

Figure 3. DeepDRK revealed treatment efficacy of unseen drugs. (A) Comparison of DeepDRK score between the sensitive and resistant groups. (B) The receiver operating characteristic (ROC) curves for DeepDRK in predicting novel CCL-drug interactions (using CTRP, GDSC or both). (C) The precision-recall curves for DeepDRK in predicting novel CCL-drug interactions (using CTRP, GDSC or both). (D) The 'guilty by association' assumption for inference of CCL-drug interactions. (E) The profiles of Lapatinib represented by its top 20 similar compounds, respectively, based on chemical properties (upper panel) and known targets (lower panel). (F) The confusion matrix of prediction for Lapatinib.

potential targeting strategies via drug repurposing. Notably, the gene-drug networks pinpointed novel candidate drugs (highlighted in red squares) for genome-characterized cancer cells, potentially benefiting patients whose conventional therapeutic interventions had failed. For example, we reported 23 sensitive anticancer drugs against *KIT* mutations in the TGCT patients, and the top prioritized candidates included Tamoxifen (inhibiting testicular spermatogenesis and steroidogenesis), Cetuximab (*EGFR* inhibitor), Pazopanib (*VEGFR* Inhibitor) and Bicalutamide (an antiandrogen drug).

DeepDRK with incomplete feature types

In addition, in case that some feature types were not available, we further constructed alternative versions of DeepDRK using different list of features via the kernel-based methods. As shown in Figure S6A, all possible feature combinations have been used to train DeepDRK. Not surprisingly, models with more feature types had higher accuracy in both cross-validation and independent testing (Figure S6B). Importantly, although the training accuracies were scarified when losing important feature types, a number of updated models could achieve acceptable accuracies (AUPR > 0.7 for independent testing), providing applicable models for various application scenarios. Furthermore, we also trained DeepDRK models in case that some features might be partially lost. To infer the missing values, we introduced the hot deck imputation method which uses similar samples' value to fill the missing ones. To test the model performance, we manipulated the training dataset by randomly removing certain percentage of values and retrained DeepDRK models. Figure S7 showed the change of model performance according to the proportion of missing values in the training datasets. We found that DeepDRK achieved reasonable accuracy with less than 10% missing values.

Lastly, DeepDRK model was developed into an open-access software for academic usage, which provided a translational tool for predicting the potential drug effects in cancer patients based on the integration and deep learning of existing knowledge and large-scale pharmacogenomic data on a remarkable number of CCLs.

DISCUSSION

In this work, we developed a kernel adapted DLF, namely DeepDRK, to computationally predict anticancer drug response via large-scale heterogeneous data integration from multiple datasets. We showed that the deep learning model outperformed conventional machine learning approaches in both accuracy (high average AUC and AUPRC) and robustness (low variation of AUC and AUPRC). Note that there were other deep learning methods previously developed on the same problem such as DeepDSC, tCNNS and CDRscan [9–12, 36–37, 52]. Most of these studies were regression models based on single-source data

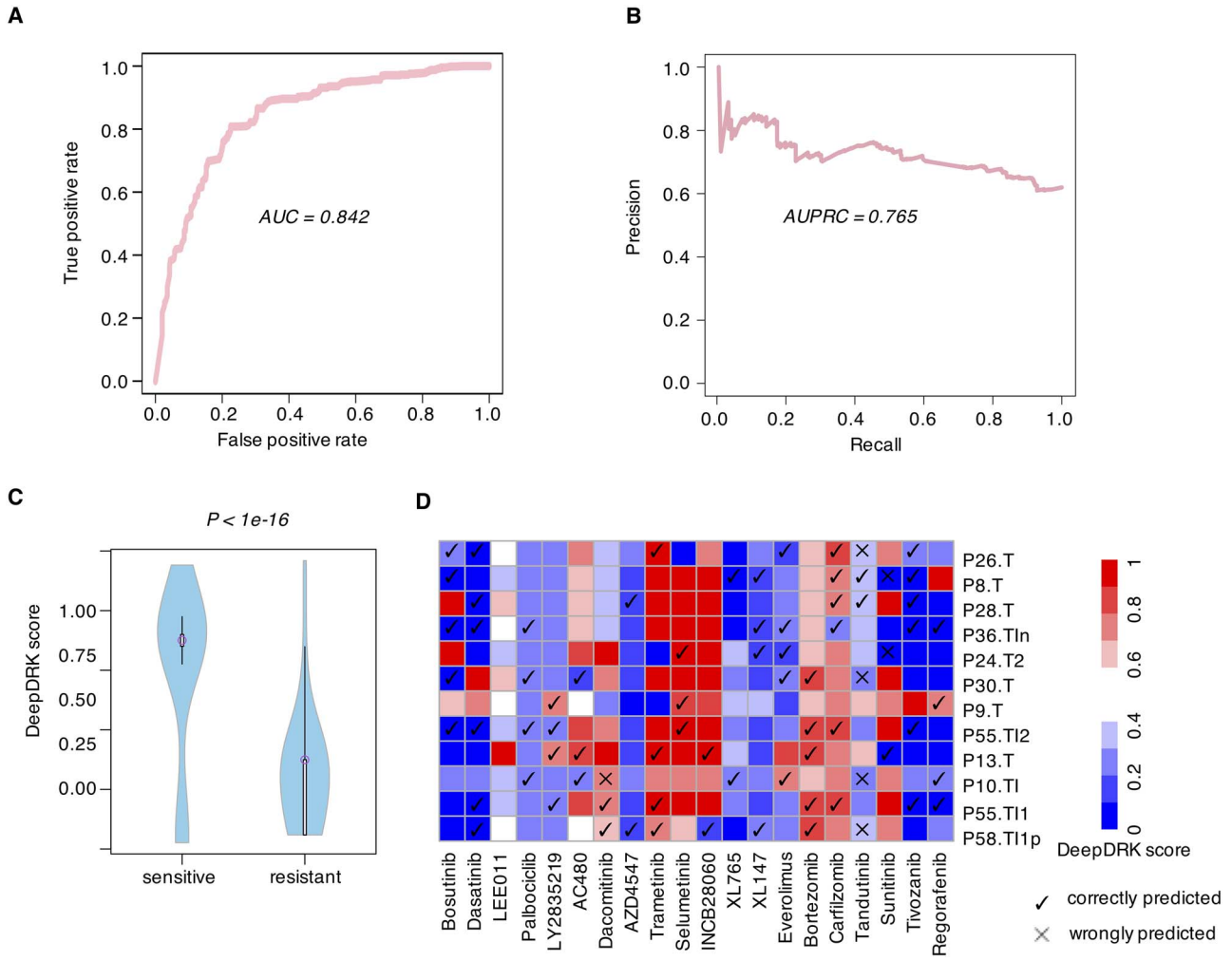

Figure 4. The application of DeepDRK in glioma patient-derived cancer cells (PDCs). (A-B) The receiver operating characteristic (ROC) curve and precision-recall curve of DeepDRK predicting drug response of glioma PDCs. (C) The comparison of DeepDRK score in sensitive and resistant PDC-drug groups. *P*-value was calculated by Wilcoxon signed-rank test. (D) Heatmap comparing DeepDRK score and the experimental results. DeepDRK score of selected pairs of PDCs and anticancer drugs were visualized in heatmap and compared to the experimental data in [51]. We used the 'tick marks' to indicate consistent results between prediction and experimental data, and 'wrong crosses' indicating inconsistent results. Notably, the PDC-drug pairs without any marks either had no available experimental data or had experimental values falling in unclear group but our computational model provided predictions for further biological and clinical investigation.

or single-omics sample characterization, while DeepDRK is a classification model integrating multimodal and multi-omics data from different sources. In our model, we transformed the continuous responses into three classes: sensitive, resistant and unclear, which denoised the raw experimental measurement of drug responses, and provided explicit prediction to guide clinical decision. Moreover, DeepDRK utilized the kernel methods to construct CCL-CCL and drug–drug similarity matrices for sample representation, which naturally reduced the feature space and model parameters. In addition, the model carried out a multitasking strategy by training models using the edges of a bipartite graph connecting CCLs to anticancer drugs. Lastly, the multimodal integration framework of DeepDRK enables the potential extension of the model by including additional features. Future work on incorporating protein and/or noncoding RNA expression to further characterize CCLs [53] or incorporating drug therapeutic annotations (ATC-code), drug side-effects and targeted proteins to characterize anticancer drugs [54–60] might further increase the model performance.

From the biological perspective, we have listed several findings with biological significance, and some of them were already supported by the literatures. The nice performance of DeepDRK in PDCs indicated an advantage of DeepDRK in predicting clinical drug responses in data with complex genetic and molecular backgrounds, which enabled the potential application in patients. When applied to TCGA patients, DeepDRK performed well on BRCA and HNSC, but not on cancer types like CESC and UCEC. This is mainly due to small cohort size. As shown in Figure S8, the prediction performance denoted by the fold change of DeepDRK score between responders and nonresponders was significantly related to total number of responders and nonresponders in each TCGA cancer type.

To eventually achieve clinical application of DeepDRK, another main challenge in drug response prediction *i.e.* intratumoral heterogeneity (ITH) shall be considered. ITH, a phenomenon observed in a number of aggressive cancers, is precluding efficacy of targeted therapies. Actually, many tumors consist of heterogeneous cell populations with a wide range of

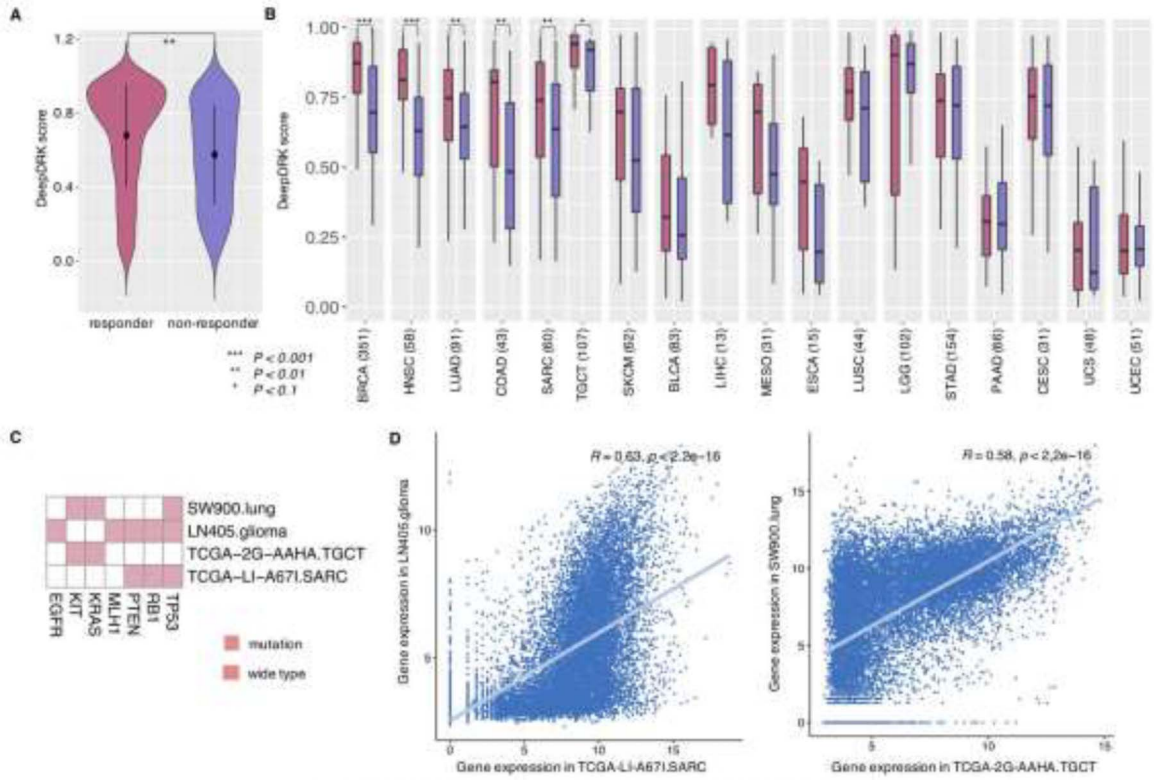

Figure 5. The application of DeepDRK in TCGA cancer patients. (A) The violin plot showing DeepDRK score of responders and nonresponders of TCGA patients. (B) DeepDRK score comparison of responders and nonresponders in individual TCGA cancer types. The red bars represent responders and the blue bars represent nonresponders. All cancer types with patient-drug pairs larger than 10 were considered. * indicated *P*-value < 0.1; ** indicated *P*-value < 0.01; *** indicated *P*-value < 0.001. (C) Heatmap showing the mutation profile of training CCLs and SARC and TGCA patient, implying that the DeepDRK model might use these CCLs as template to infer treatment response of SARC and TGCA patients. (D) The scatter plot showing the correlation between expression of training CCL and test patient.

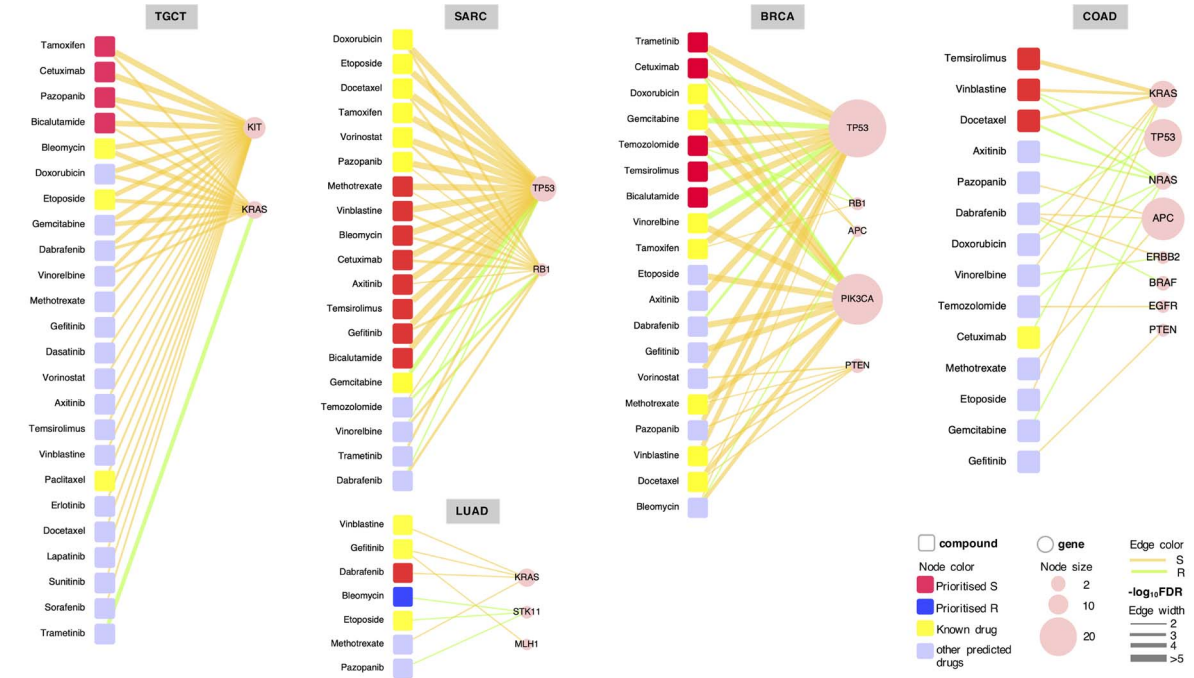

Figure 6. DeepDRK repurposed anticancer drugs based on cancer type-specific molecular alterations in TCGA. Nodes with different shape types represent the genomic alterations (circle) and inhibitors (square), respectively, and size of a circle represents number of patients harboring mutations of the gene. The known response anticancer drugs and prioritized compounds identified by DeepDRK are visualized by different colors. The width of each edge is proportional to -log10 (FDR) in which False Discovery Rate (FDR) was calculated by two-sided ranksum test with Benjamini–Hochberg correction.

morphologies, genotypes and phenotypes. Increasing evidence suggested that ITH plays an important role in drug resistance and treatment failure [61–63]. To address this challenge, more efforts will be needed to further develop DeepDRK. Moreover, recent studies suggested that the identification of synergistic drug combination might provide an efficient way for drug repositioning [64]. More efforts will be needed to collect enough amount of drug combination data to further extend the prediction models to advance computational precision medicine.

Key Points

- Deep learning provided an accurate model to predict drug response of cancer cells.
- Kernel-based methods integrated multi-omics and multisource data.
- DeepDRK was developed into an open-access software in R.
- DeepDRK offered a computational framework for precision cancer medicine.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work was supported by the National Natural Science Foundation of China (No. 11671396, No. 11371365, No. 31270270, No. 31922088), a grant from Qinghai Sciences and Technology Department for Basic Research Program (No. 2020-ZJ-719) and grants from Department of Science and Technology of Guangdong Province (GDST20EG61), Hong Kong RGC (N_HKUST606/17, 26102719, C7065-18GF, C4039-19GF, R4017-18), Hong Kong ITC (ITCPD/17-9) and the Hong Kong Epigenomics Project (LKCCFL18SC01-E).

References

1. Bussey KJ, Chin K, Lababidi S, *et al*. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol Cancer Ther* 2006;**5**:853–67.
2. Lamb J, Crawford ED, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.
3. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.
4. Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer* 2010;**10**:241–53.
5. Venkatesan K, Stransky N, Margolin A, *et al*. Prediction of drug response using genomic signatures from the cancer cell line Encyclopedia. *Clin Cancer Res* 2010;**6**(19 Supplement):PR2.
6. Caponigro G, Sellers WR. Advances in the preclinical testing of cancer therapeutic hypotheses. *Nat Rev Drug Discov* 2011;**10**:179–87.
7. Garnett MJ, Edelman EJ, Heidorn SJ, *et al*. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**:570–5.
8. 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**:e61318.
9. Costello JC, Heiser LM, Georgii E, *et al*. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**(12):1202–12.
10. 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.
11. Zhang F, Wang M, Xi J, *et al*. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci Rep* 2018;**8**(1):3355.
12. 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):8857.
13. 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**:32679.
14. Chin C, Brown DE. Learning in science: a comparison of deep and surface approaches. *J Res Sci Teach* 2000;**37**:109–38.
15. Bengio Y. Learning deep architectures for AI. *Found Trends Mach Learn* 2009;**2**:1–127.
16. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013;**35**:1798–828.
17. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**(7553):436–44.
18. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2014;**61**:85–117.
19. Lee H, Pham P, Largman Y, *et al*. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Adv Neural Inf Process Syst* 2009;**22**:1096–104.
20. Le QV, Zou WY, Yeung SY. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *IEEE Xplore* 2011;3361–8.
21. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Proc Mag* 2012;**29**:82–97.
22. Graves, A., Mohamed, A.R. and Hinton, G. (2013) Speech recognition with deep recurrent neural networks. In: *2013 IEEE international conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, IEEE, 2013, 6645–9.
23. Wu, Z., Jiang, Y.G., Wang, J., Pu, J. and Xue, X. (2014) Exploring Inter-feature and Inter-class Relationships with Deep Neural Networks for Video Classification. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014, 167–76.
24. Xiong, C., Merity, S. and Socher, R. (2016) Dynamic memory networks for visual and textual question answering. In: *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, 2016, 2397-406.
25. Lena DP, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics* 2012;**28**(19): 2449–57.
26. Leung, M.K.K., Xiong, H.Y., Lee, L.J. and Frey, B.J. (2014) Deep learning of the tissue-regulated splicing code. *Bioinformatics*, **30**(12):i121–i129.
27. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2014;**31**(5):761–3.

28. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**(8):831–8.
29. Jo T, Hou J, Eickholt J, *et al*. Improving protein fold recognition by deep learning networks. *Sci Rep* 2015;**5**:17573.
30. Xu YG, Wang YC, Luo JS, *et al*. Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res* 2017;**45**(21):12100–12.
31. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**(5):851–69.
32. Aliper A, Plis S, Artemov A, *et al*. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol Pharm* 2016;**13**(7):2524–30.
33. Vougas K, Krochmal M, Jackson T, *et al*. Deep learning and association rule Mining for Predicting Drug Response in cancer. *bioRxiv* 2017. doi: 10.1101/070490.
34. Chiu YC, Chen HH, Zhang TH, *et al*. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics* 2019;**12**(Suppl 1):18.
35. 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):513.
36. Li M, Wang Y, Zheng R, *et al*. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Trans Comput Biol Bioinform* 2019;1. doi: 10.1109/TCBB.2019.2919581.
37. 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):408.
38. Sakellaropoulos T, Vougas K, Narang S, *et al*. A deep learning framework for predicting response to therapy in cancer. *Cell Rep* 2019;**29**(11):3367–73.
39. Miotto R, Wang F, Wang S, *et al*. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018;**19**(6):1236–46.
40. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* 2019;**20**(5):1878–912.
41. Rees MG, Seashore-Ludlow B, Cheah JH, *et al*. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016; **12**:109–16.
42. 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**:D955–61.
43. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J* 1950;**29**(2):147–60.
44. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Chem* 2015;**7**(1):20.
45. The H2O.ai team. *h2o: R Interface for H2O*. R package version 3.10.5.3. https://CRAN.R-project.org/package=h2o.
46. Gribskov M, Robinson NL. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Comput Chem* 2015;**20**:25–33.
47. Powers DM. Evaluation: from precision, recall and F- measure to ROC, Informedness, Markedness and correlation. *J Mach Learn Tech* 2011;**2**(1):37–63.
48. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2017;**34**(Database issue):D668–72.
49. Kanehisa M, Furumichi, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.
50. Ding Z, Zu S, Gu J. Evaluating the molecule- based prediction of clinical drug responses in cancer. *Bioinformatics* 2016;**32**(19):2891–5.
51. Lee JK, Liu Z, Sa JK, *et al*. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nat Genet* 2018;**50**(10):1399–411.
52. Basu A, Mitra R, Liu H, *et al*. RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics* 2018;**34**(19):3332–9.
53. Chen X, Guan NN, Sun YZ, *et al*. MicroRNA-small molecule association identification: from experimental results to computational models. *Brief Bioinform* 2020;**21**(1): 47–61.
54. Chen X, Yan CC, Zhang X, *et al*. Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016;**17**(4):696–712.
55. Campillos M, Kuhn M, Gavin AC, *et al*. Drug target identification using side-effect similarity. *Science* 2008;**321**:263–6.
56. Wang YC, Zhang CH, Deng NY, *et al*. Kernel- based data fusion improves the drug-protein interaction prediction. *Comput Biol Chem* 2011;**35**(6):353–62.
57. Duran-Frigola M, Aloy P. Recycling side-effects into clinical markers for drug repositioning. *Genome Med* 2012; **4**(1):3.
58. Wang YC, Chen SL, Deng NY, *et al*. Computational probing protein-protein interactions targeting small molecules. *Bioinformatics* 2016;**32**(2):226–34.
59. Wang YC, Chen SL, Deng NY, *et al*. Drug repositioning by kernel integration molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;**8**(11):e78518.
60. Wang YC, Chen SL, Deng NY, *et al*. Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics* 2013;**29**(10):1317–24.
61. Gerlinger M, Swanton C. How darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br J Cancer* 2010;**103**:1139–43.
62. Yap TA, Gerlinger M, Futreal PA, *et al*. Intratumor heterogeneity: seeing the wood for the trees. *Sci Transl Med* 2012;**4**(127):127ps10.
63. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* 2013;**108**:479–85.
64. Chen X, Ren B, Chen M, *et al*. NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Comput Biol* 2016;**12**(7):e1004975.