Alexandria Engineering Journal 105 (2024) 88–97

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

DRPO: A deep learning technique for drug response prediction in oncology cell lines

Muhammad Shahzad [a,b], Adila Zain Ul Abedin Kadani [b], Muhammad Atif Tahir [b], Rauf Ahmed Shams Malick [b], Richard Jiang [a,*]

[a] School of Computing and Communication, Lancaster University, Lancaster, Lancashire, United Kingdom
[b] School of Computing, National University of Computer and Emerging Sciences (NUCES-FAST), Karachi, Sindh, Pakistan

ARTICLE INFO

Keywords:
Deep learning
Cancer cell lines
Personalized medicine
Matrix factorization
Sensitivity score
Molecular features
Predictive modeling
Drug response mechanism

ABSTRACT

With the invention of high-throughput screening technologies, innumerable drug sensitivity data for thousands of cancer cell lines and hundreds of compounds have been produced. Computational analysis of these data has opened a new horizon in the development of novel anti-cancer drugs. Previous deep-learning approaches to predict drug sensitivity showed drawbacks due to the casual integration of genomic features of cell lines and compound chemical features. The challenges addressed include the intricate interplay of diverse molecular features, interpretability of complex deep learning models, and the optimization of drug combinations for synergistic effects. Through the utilization of normalized discounted cumulative gain (NDCG) and root mean squared error (RMSE) as evaluation metrics, the models aim to concurrently assess the ranking quality of recommended drugs and the accuracy of predicted drug responses. The integration of the DRPO model into cancer drug response prediction not only tackles these challenges but also holds promise in facilitating more effective, personalized, and targeted cancer therapies. This paper proposes a new deep learning model, **DRPO**, for efficient integration of genomic and compound features in predicting the half maximal inhibitory concentrations (IC50). First, matrix factorization is used to map the drug and cell line into latent 'pharmacogenomic' space with cell line-specific predicted drug responses. Using these drug responses, we next obtained the essential drugs using a Normalized Discounted Cumulative Gain (NDCG) score. Finally, the essential drugs and genomic features are integrated to predict drug sensitivity using a deep model. Experimental results with **RMSE 0.39** and **NDCG 0.98** scores on Genomics of drug sensitivity in cancer (GDSC1) datasets show that our proposed approach has outperformed the previous approaches, including DeepDSC, CaDRRes, and KMBF. These good results show great potential to use our new model to discover novel anti-cancer drugs for precision medicine.

1. Introduction

Cancer is one of the leading causes of death, as it is a genetic disease that impacts gene expression and causes uncontrolled cell proliferation and metastasis. Compared to commonly used cancer treatment methods like chemotherapy or radiotherapy, targeted drugs would be a better way to destroy the cancerous cells with a minimized toxicity effect on non-cancerous tissues [1]. However, based on individual molecular and genomic variants, the patient responds to the drug in a different way, i.e., drugs do not have a uniform response for everyone [2]. These varieties of drug responses in cancer treatment draw attention, to the use of genomics features to predict the drug response (sensitivity), which is one of the aims of precision medicine [3,4].

Recently, the genomic information of cancer cell lines and drug sensitivity profiles from hundreds of compounds have been cataloged in public datasets such as cancer cell line encyclopedia (CCLE) [5], GDSC [6] and human tumor cell line screen NCI60 [7]. These datasets originate from numerous human cancer tissues like the lungs, kidneys, nervous, breast, and kidney. Moreover, the different types of genomic information including gene expression, mutation, methylation, and copy number variant against each tissue are also provided in these public datasets. The drug sensitivity can be measured in terms of IC50, maximum effect attributable to the drug (Emax), or area above the dose–response curve AUC, but mostly IC50 has been used in the literature to represent the drug sensitivity prediction. This study utilized

* Corresponding author.
E-mail address: r.jiang2@lancaster.ac.uk (R. Jiang).

https://doi.org/10.1016/j.aej.2024.06.052
Received 26 September 2023; Received in revised form 17 May 2024; Accepted 20 June 2024
Available online 3 July 2024
1110-0168/© 2024 The Authors. Published by Elsevier B.V. on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

gene expression data to represent cell lines, as it is considered more informative than other omics data [8]. Additionally, the drug response was measured using the IC50 metric.

The drug sensitivity prediction challenge organized by the DREAM Project (http://www.the-dream-project.org/) has motivated data scientists and researchers around the world to solve this problem. Numerous predictive models, like the random forest, support vector machine (SVM), and neural network, have been used to solve drug sensitivity prediction [9–11]. These kinds of challenges on public datasets, particularly CCLE and GDSC1, enable researchers and scholars to use computational and analytical approaches for drug response prediction, such as [12–14], and [15]. All these studies help drug discovery societies to come up with more effective drugs. [16,17] also presented a comprehensive review on drug–target interaction prediction that contributed to the novel drug discovery challenge. But none of these models incorporated drug information that limits our study includes drug information.

To cope with these limitations, more advanced machine learning techniques like a dual-layer cell line drug network (DLN) model [18] and similarity-regularized matrix factorization (SRMF) [19] model were proposed that used drug and cell line similarity information together to predict the drug response. CaDRReS [20] the matrix factorization-based recommender system used essential genes and produced better results than DLN and SMRF models in terms of RMSE. The authors in [21] has also been widely used to integrate multi-omics data along with the chemical features of compounds to boost performance. Some of the variations in matrix factorization approaches also produced better drug response predictions like [22,23] used weighted graph regularized matrix factorization technique and kernelized similarity-based matrix factorization (KSRMF) respectively. Another work [24] is a hybrid recommender system that uses neighbor-based collaborative filtering with global effect removal (NCFGER) method to compute the drug response. A dataset is preprocessed by applying a global effect removal technique on drug similarity in addition to a cell line similarity network to reduce biases of the unknown responses. The result of NCFGER on the CCLE and GDSC datasets in terms of averaged root-mean-square error (RMSE) and Pearson correlation coefficient (PCC) are better than DLN, SMRF, CaDRReS, KSRM overall drugs. However, all these works suffered from the curse of dimensionality problems. As there are limited numbers of cell lines i.e. rows and thousands of compounds' chemical and structural features and cell lines genomics features i.e. columns, therefore the chance of underfitting increases and prediction quality decreases.

The recent achievement of the Deep learning (DL) model has been revealing the outstanding capability to contribute to many challenging applications with high-dimensional data [25]. The problems related to drug–target interaction, drug repositioning, and visual screening, DL models have outperformed the traditional machine learning models [26–28]. Deep learning is also serving in drug response problems like [29] proposed a deep neural network model based on the expression and mutation profiles of cancer cell lines. The model comprised three networks (expression and mutation encoders and a feed-forward network), where both the expression and mutation encoders were used to reduce the dimension of the cancer genome atlas (TCGA) dataset and then fed into the feed-forward network to make predictions based on the IC50 values. [30] proposed a multi-model attention-based neural network that integrated compounds' molecular structure, cells' genetic profile, prior knowledge of protein interaction. which increased the overall model's performance), and SMILES encoding of compounds. [31] used the DL model to quantify survival and drug responses using multi-omics datasets of breast cancer cell lines with a 1.154 mean squared error (MSE). DeepDSC [32], CDRScan [33] and tCNN [34] were some recent models that predict IC50 values, but amongst them, DeepDSC produced the best coefficient of determination (R2) and RMSE on both the CCLE and GDSC datasets.

With the invention of ensemble-based deep learning models, the aggregated result of the different predictions would give deterministic estimates [35]. Considering these characteristics, [36] used an ensemble of transfer learning to predict drug responses in applications like drug repurposing, precision oncology, and new drug development. The approach discussed in this paper was limited to that particular application but did not contribute specifically to drug sensitivity prediction. In [37], authors proposed new machine learning techniques based on a transfer learning approach to predict drug sensitivity. They worked on breast cancer, triple-negative breast cancer, and multiple myeloma datasets. The authors claimed that their work was the first attempt at comparing different transfer learning approaches in the clinical informatics domain. The results of this work were remarkable in terms of classification problems. The drug sensitivity problem refers to the regression problem with respect to predicting the half-maximal inhibitory concentration, i.e., the IC50 value. Recently, article 37 proposed a k-means ensemble support vector regression model (kESVR) for predicting drug responses of patients using gene expression data sets. The proposed model was a combination of supervised and unsupervised algorithms and used principal component analysis and regression methods for predicting drug responses. All of these simple to complex models, integrated genomics, and drug features have exhaustive combinations.

Recently [38] gathered gene expression data of 49 different breast cancer cell lines, as well as data on how these cells responded to 220 different drugs, from the GDSC dataset. With this data, they created a complex network, called a multiple-layer cell line-drug response network (ML-CDN2), with two different networks: one learns the similarities between the cell lines, and another network learns similarities between the drugs. Using this network, the authors were able to make predictions about how new breast cancer cell lines or samples from patients would respond to different drugs. DrugCell [39] is another deep-learning model that was trained on how 1,235 tumor cell lines reacted to 684 drugs and mapped how human cancer cells respond to therapy. They also attempted drug combinations, that also enhanced results. However, the downside of this approach is that it is limited to using mutation data only as drug features, where other molecular information such as gene expression, epigenetic states, and so on, may be more useful features than mutations. Therefore, there is a need to come up with a solution using targeted drugs and genomic features, which should have biological significance rather than causal integration.

In conclusion, we have identified two potential challenges: firstly, offering a substitute for informal integration between genes and drugs, and secondly, attaining results meeting the threshold deemed acceptable by the medical community. This is crucial for adopting our approach in hospital practices for cancer prognosis and treatment planning. To make these challenges our objectives, in this work, we propose DRPO: a deep learning technique for drug response prediction in oncology cell lines and a novel neural network framework for drug response prediction on cell lines. DRPO represents a computational model designed to provide personalized drug recommendation based on genetic profile and molecular information of the patients. The aim is to determine which drug would be most effective on cancer cell lines using oncogenes expression data and the molecular structure of essential drugs. DRPO model relies mainly on the genetic and molecular data, extracting meaningful patterns to obtain insights into patient specific drug responses. The computational requirements associated with intricate and extensive dataset requires a HPC (High Performance Computing), large memory, GPU (General Processing Unit) and storage capacity which is crucial for the DRPO model training. Utilizing a GPU accelerated the training time due to it's parallel processing capability. Without a GPU, reliance on a CPU alone may result in limited model complexity, slow training time and increased computational cost. Hence causing it not being feasible enough in a typical laboratory setting. The work is carried out in multiple stages, where initially

the cell lines and drugs are mapped into latent space using a matrix factorization approach, and then a list of essential drugs is determined using the NDCG algorithm. Later, a deep learning model is applied to the resultant list of essential drugs for both CCLE and GDSC to determine the final predicted sensitivity score. In comparison to all the existing models, the performance of our proposed framework on the GDSC1 dataset shows that it has achieved better results concerning RMSE and NDCG.

In summary, the novelty of our work mainly resides in the following aspects:

- We move towards identifying the optimal drug-gene pairing by employing a unique methodology. Our novel approach involves extracting essential drugs through the utilization of the NDCG score and matrix factorization. Subsequently, we merge these identified drugs with oncogenes, drawing from the CRISPR experiment [40], to construct the final input matrix. To the best of our knowledge, no previous studies have suggested mapping targeted drugs with essential genes using this particular approach.
- We apply and test our DRPO approach on the CCLE and GDSC1 datasets based on RMSE and NDCG scores. In comparison with prominent models in existing literature, the lowest RMSE and highest NDCG score of DRPO on the GDSC1 dataset and the cumulative RMSE score on both the CCLE and GDSC1 datasets show the feasibility of the proposed technique.

As discussed in Sections 5.1 and 5.2, we have compared our proposed **DRPO** model with some of the similar previous studies like [41–44] [9,20,45,46] [32]. Tables 5 and 6 show that with the above novel proposed approach, we have achieved better results in terms of NDCG and RMSE scores in our experiments. This would hopefully be another potential step toward drug discovery and precision medicine.

The rest of the paper is organized as follows: in Section 2 material and methods are presented. While Section 3 presents the performance measure, Section 4 presents the result and discussion and finally conclusion and future work are presented in Section 5.

2. Materials and method

This work presents the deep learning regression model, (DRPO) that predicts the drug sensitivity score in terms of IC50 value on cancer cell lines. Drug responses and cell line expression data are obtained from CCLE and GDSC1 data sources. The drugs that are available in the GDSC1 and CCLE datasets, their molecular 2D structures, are obtained from PubChem [47] in the form of Standard Delay Format (SDF) files. These 2D structures are then converted into Morgan fingerprints using Camb to make a compound feature vector of 256 bits. The input for our model is cell line expression data for oncogenes and compound fingerprints of essential drugs.

In this section, the most commonly used datasets and our proposed **DRPO** model are described, so that one can replicate our experiments easily.

2.1. Materials

CCLE. The CCLE dataset contains gene expression data of 1037 cell lines. There are approximately 20,042 genes against each cell line thus making a vector of the same length. Each gene value against each cell line expresses the transcription level of genes. For drugs, we have also used drug response data of 24 drugs provided by CCLE. The metric for the drug's sensitivity on cancers is IC_{50} value which is then converted into -$log_{10}IC_{50}(\mu M)$. The lower IC_{50} value, the more effective drug is and the higher IC_{50} value means the opposite.

GDSC1. The gene expression data from the GDSC1 dataset contains 17419 genes for 1074 cell-line. Where each genes' value is a transcription level of genes. The drug responses to a cell line data are also downloaded from GDSC1. There are 367 drug responses in terms of IC_{50} value to each cell line taken from the GDSC1 dataset. The summary of these two datasets are given in Table 1.

Table 1
Total cell-lines, gene expression profile and drugs matrix.

Source	Cell lines	Genes expression profile	Drugs
CCLE	1037	20 042 Genes	24
GDSC1	1074	17 419 Genes	367

Table 2
Cell lines with essential genes and drug responses.

Source	Cell-lines	Essential genes	Final data matrices shape
CCLE	491	1718	491 × 24 = 11 784
GDSC1	987	1610	987 × 226 = 223 062

Table 3
Merged drug response and gene expression data for both CCLE and GDSC1.

Source	Drug response data shape	Gene expression data shape	Resultant matrix
CCLE	491 Cell-lines × 11 Essential Drugs = 5401 rows × 3 columns (Cell-line, Drug ID, IC50)	491 Cell-lines × 1718 Genes	5401 Cell-lines × 1720 Columns (1718 Genes, Drug ID and IC50)
GDSC1	985 Cell-lines × 23 Essential Drugs = 22,655 rows × 3 columns (Cell-line, Drug ID, IC50)	985 Cell-lines × 1610 Genes	22655 Cell-lines × 1612 Columns (1610 Genes, Drug ID and IC50)

2.2. Method

Our novel approach is split into two stages i.e. at first stage, matrix factorization and drug ranking approaches are used to build a final input matrix from both CCLE and GDSC1 datasets. Whereas at the second stage, deep learning models are used for drug response prediction.

As not all cell lines can be tested against each drug, therefore we have extracted those cell lines that have drug responses. After this extraction, the final data matrix contains 987 cell lines versus 226 drugs from the GDSC1 dataset and 491 cell lines versus 24 drugs CCLE dataset. The resultant shape of these data matrices can be seen in Table 2.

Both data matrices represent gene expressions as features of cell lines. To make our model simple, we have reduced the dimension, i.e. gene expression features from both matrices and have only considered the known oncogenes obtained from the CRISPR experiment [40]. The reduced feature dimension contains 1610 essential genes from GDSC1 and 1718 essential genes from CCLE. These essential genes are also known as oncogenes. The resultant shape of our input data matrices can be seen in Table 2.

Pre-processing steps are applied to the resultant input matrix, where null values are replaced by the mean of IC_{50} values. To create a combined dataset, the gene expression data from two sources (CCLE and GDSC1) were used. The CCLE dataset included 491 cell lines and 1718 genes, while the GDSC1 dataset included 985 cell lines and 1610 genes. The drug response data was also included, resulting in a matrix that contained the IC_{50} value and gene expression value for each cell line in relation to its corresponding Drug ID. The summary of resultant matrices can be seen in Table 3.

Various methods are applied in our work for determining the sensitivity score of $cell-lines \times drugs$ relation and ranking the drugs based on their sensitivity scores using various mathematical techniques and machine learning models.

2.2.1. DRPO - matrix factorization technique

Matrix factorization (MF) has gained popularity primarily due to its effectiveness in predicting missing values. For the past couple of

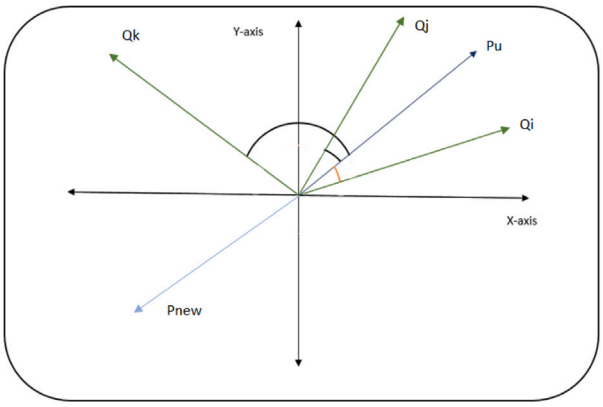


Fig. 1. Latent space representation of cell-line and drugs.

years, it has extended its application into the realm of personalized medicine, showing promise in modern drug discovery analyses. Its potential lies in its ability to integrate various heterogeneous datasets, making it particularly valuable in this context. The reason for the use of this collaborative filtering technique lies in constructing a model that emphasizes information from similar drugs, thereby not assigning equal importance to all drugs in predicting responses. Furthermore, it is a method to come up with latent features through the product of two matrices. The contribution of matrix factorization to predicting drug responses lies in its ability to capture complex relationships and dependencies between drugs and cell lines. By mapping drugs and cell lines into a latent pharmacogenomic space, the model gains a more compact and meaningful representation of the underlying biological interactions. This facilitates improved generalization to new drug-cell line pairs do not present in the training data and enhances the model's ability to predict responses for targeted cancer drugs. In recommender systems, collaborative filtering is one of the well-known applications of matrix factorization, which maps two different entities into a latent space. This latent space gives the relationship amongst entities. MF has become a famous method in the personalized medicine domain because of its capability to solve linear problems and predict unknown drug responses [48]. MF is an unsupervised algorithm that factorizes a high dimensional matrix into two low-rank matrices as shown in Eq. (1).

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^{\mathrm{T}}, \quad \text{where} \quad \mathbf{P} \in \mathbb{R}^{N\times K}, \mathbf{Q} \in \mathbb{R}^{M\times K}. \tag{1}$$

Here R is a drug response matrix that represents the association between cell lines and drugs, P is a matrix that gives the association between cell lines and their genomics features, whereas Q is another matrix i.e. association between the drug and its features.

We have used the same approach that factorized the known drug response matrix into cell lines and drug matrices with their latent features in latent space.

A latent space is a vector representation showing the relationship between the cell line vector and drug vector as depicted in Fig. 1. Here P represents the cell line vector and Q represents the drug vector. The cell lines P_u are sensitive to the drug Q_i and Q_j but not to the Q_k in the latent space. As the cell line P_u and drug Q_i have a smaller angle between them compared to cell line P_u and drug Q_k that has a larger angle. Smaller angles between the vectors indicate a higher degree of similarity than the larger angles. This is because two vectors having smaller angles between them tend to point in the same direction which suggests that they share some common characteristics.

The model is trained by defining a drug sensitivity score, where a higher value of the drug sensitivity score determines that the cell lines are more sensitive to the drug. The model was trained against CCLE and GDSCI datasets independently. For training the DRPO model, the MF model was used. Our MF model learned the latent features of cell lines and drugs using a done product between cell line vector and drug vector resulting in a cell line-specific drug response as shown in the Eq. (2).

$$\hat{S}_{ui} = \mu + b_i^Q + b_u^P + q_i.p_u = u + b_i^Q + b_u^P + q_i(x_u W_P)^T \tag{2}$$

Here $\hat{S}_{ui}$ is the predicted drug sensitivity score, P represents the cell-lines vector and Q represents the drug vector. While the b_i^Q and b_i^P are the biases hyper-parameter for cell-line u and drug i. μ represent the mean for all drug responses and W_p represents the transformation matrix for projecting cell-line features x_u onto the pharmacogenomic space (latent space).

For calculating the sensitivity score, P and Q values are determined, i.e., P cell line vector and Q drug vector, corresponding to the CCLE and GDSC1 datasets. At the start of training, random weights (biases) are considered for cell lines and drugs, represented as W_p and W_q respectively.

During the training process, the objective is to minimize the loss using the sum of squared error as shown in Eq. (3)

$$L(\theta) = \frac{1}{2|K|} \sum_u \sum_i R_{ui} \tag{3}$$

$$R_{ui} = \left(s_{ui} - \hat{s}_{ui}\right)^2 \tag{4}$$

Here $L(\theta)$ represents the loss function for the mean sum of squared error. u and i represent cell lines and drugs, S_{ui} and $\hat{S}_{ui}$ represent the observed and predicted sensitivity scores for cell lines and drugs, and $|K|$ represents the number of drug response experiments in the training dataset. R_{ui} represents the squared difference between the observed and predicted drug sensitivity score.

Fig. 2 depicts step-by-step processes of applying MF. Initially, cell line features are calculated based on gene expression data. Gene expression data is one of the genomic features often used as input features for the DRPO model. These gene expression features significantly influence the prediction of IC50 values as they provide insights into the molecular mechanisms governing drug response. Specific gene expression patterns may be indicative of sensitivity or resistance to particular drugs, thereby influencing the predicted IC50 values. Each gene has multiple values corresponding to each cell line, it is then normalized by computing fold changes in comparison to the mean values across the cell lines. Once gene expression data is normalized, the Pearson correlation coefficient value for each cell-line pair is computed to determine the similarity between each cell-line pair. Once GDSC1 and CCLE datasets are pre-processed, the matrix factorization technique is applied to the resultant cell-line similarity matrix and drug response data to compute the final predicted sensitivity score. During training, the model learns to recognize patterns in the gene expression profiles that are associated with different IC50 values. This training allows the model to generalize to new data and predict IC50 values for drug-cell line pairs not present in the training set.

2.2.2. DRPO - drug ranking

One of the reasons for using the Matrix Factorization technique in our work is to rank the drugs based on their predicted sensitivity score. The drugs are ranked using the normalized discounted cumulative gain (NDCG) [49], which gives a score that indicates the ranking. Our top-ranked drugs are calculated using Eq. (5) [49] and Eq. (6) [49], which are as follows:

$$N\,DCG(\hat{r}, s) = \frac{DCG(\hat{r}, s)}{DCG(r, s)} \tag{5}$$

$$DCG(\hat{r}, s) = \sum_i \frac{2^{s_i} - 1}{log_2 \hat{r}_i + 1} \tag{6}$$

The $\hat{r}_i$ is the predicted rank obtained from MF, s_i is the drug-sensitive score and r is the actual or known rank of the drug of the

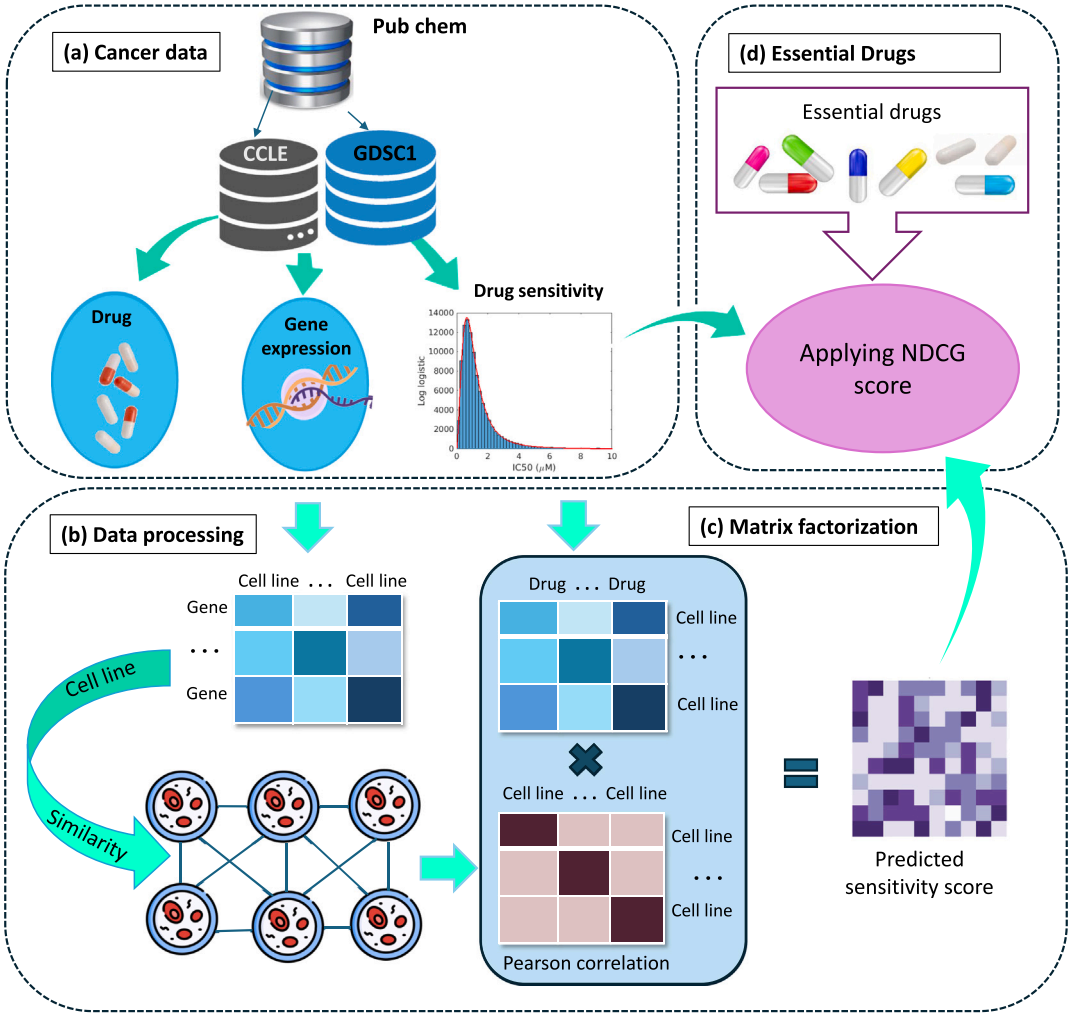


Fig. 2. The proposed framework shows step by step process of applying MF technique and obtaining essential drugs. (a) The drug data, gene data and sensitivity score were obtained from two main datasets: GDSC1 and CCLE. (b) The gene expression data is normalized by computing fold changes across the cell lines. A cell line similarity matrix is created where Pearson correlation is evaluated for each cell line pair. (c) The predicted sensitivity score is evaluated using the matrix factorization technique. (d) The final NDCG score is determined from actual and predicted sensitivity scores of both CCLE and GDSC1.

Table 4
List of essential drugs in CCLE and GDSC1 datasets.

Data source	Essential drugs
GDSC	Masitinib, BMS-509744, Parthenolide, XMD8-85, Cytarabine, CGP-082996, PD0325901, JNK Inhibitor VIII, NSC-87877, OSI-930, FTI-277, Olaparib, LFM-A13, CP724714, NVP-TAE684, PIK-93, GNF-2, BMS-536924
CCLE	Nilotinib, RAF265, PHA-665752, PD-0325901, ZD-6474, AZD0530, Lapatinib, Sorafenib, Erlotinib, PLX4720, TKI258

Table 5
NDCG scores of CCLE and GDSC over various approaches.

Model	CCLE	GDSC
Elastic net regression model using cross-validation [41]	0.89	0.61
Optimal drug prediction algorithm [42]	0.79	0.38
CwKBMF (Kernelized matrix factorization with component wise kernel learning) [43]	0.79	0.4
SRMF (Matrix factorization with similarity regularization) [44]	0.75	0.42
CaDRRes [20]	0.89	0.6
Our DRPO	**0.99**	**0.98**

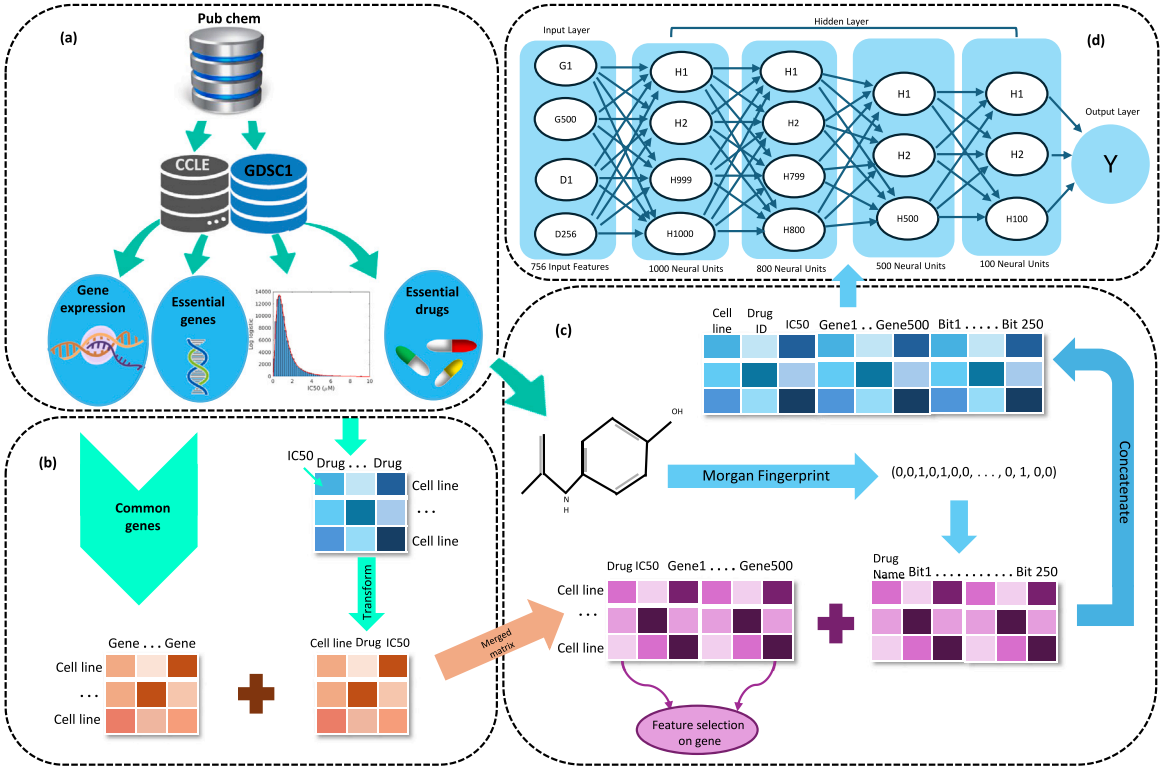


Fig. 3. The proposed framework shows DRPO model architecture. (a) The essential gene, essential drug, gene expression and sensitivity score were obtained from two main datasets: GDSC1 and CCLE (b) The data is pre-processed through common genes and drugs (c) Feature selection is applied to the gene expression dataset and Morgan fingerprint technique is applied to the drug data (d) The processed data is passed to the DRPO model for training and prediction.

Table 6
RMSE scores of CCLE and GDSC dataset over various ML algorithms.

Model	CCLE	GDSC
Feed forward multilayer perceptron neural network using chemical and genomic features [45]	–	0.83
KMBF (Kernelized Bayesian Matrix Factorization) [46]	–	0.83
Random forest [9]	0.44	0.75
DeepDSC [32]	0.23	0.52
Our DRPO with all Drugs	**0.30**	**0.67**
Our DRPO with Essential Drugs	**0.26**	**0.39**

ith cell line which is calculated on the basis of drug response values. NDCG score closer to 1 indicates that the model has correctly ranked the drugs based on their sensitivity score. This technique is used to determine essential drugs for both CCLE and GDSC1 datasets as shown in Table 4 which is then used in our deep learning model to predict the drug responses.

2.2.3. DRPO - deep learning technique

At this stage, we have used a deep learning regression model, in which gene expressions and compounds' fingerprints are integrated and used as input to predict the sensitivity score. As our input data matrix has high dimensions i.e. more features than the samples, therefore we used feature selection techniques to reduce the model over-fitting.

Fig. 3 is a model architecture diagram, depicting the entire process of the proposed approach, initially, GDSC1 and CCLE data is collected and processed based on common genes, cell lines, and drugs. After the pre-processing of the dataset, the feature selection technique is applied to the gene expression dataset, and Morgan's fingerprint technique is applied to the drug data to obtain a 256-bit vector. The processed dataset is passed to the DRPO model for training where drug response (IC50) is the label and 500 genes and 256 vector bits of drug data are the input features.

Feature Selection is a technique that gives us those features from our data that contributes the most to the target variable. During our work feature selection technique is applied to the gene expression dataset for various reasons:

1. To reduce the feature dimensions and choose the features that contribute the most to the target data.
2. To reduce the training time and improve the model's performance.
3. To avoid overfitting.

In our proposed work, SelectKBest [50] feature selection technique is applied to an essential gene expression dataset. This is done in 2 steps:

1. By finding the cross-correlation between the regressor i.e. each gene from gene expression data and the target variable i.e. IC_{50} value using Eq. (7).
2. By converting this to F score which will then be used to find the top rank gene expression features that are positively correlated with IC_{50} value.

$$F_k = \frac{\sum_{i=1}^{n}(g_i - \overline{g})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n}(g_i - \overline{g})^2 (s_i - \overline{s})^2}} \tag{7}$$

Where g_i represents the gene expression value of the ith cell line, and s_i is the corresponding drug sensitivity score in terms of IC_{50} value. $\overline{g}$ and $\overline{s}$ are the means of gene expression values and drug sensitivity scores respectively.

Finally, top-ranked genes are selected that are indicated by F_k given in Eq. (7), where k is a hyper-parameter for selecting the genes which we explicitly defined to 500. The reason for selecting the top-ranked 500 gene expression data features is to make them more meaningful genes in terms of their potential contribution toward predicting drug sensitivity.

For our drug (compound) features, we used the Morgan fingerprint technique using camb [51] to make a compound feature vector of 256 bits. The molecular fingerprints convey information about the presence of a molecular sub-structure. Drug molecular structures are obtained in the form of an SDF file from PubChem [52]. In the case of GDSC1, we have obtained 18 essential drugs sdf files from PubChem, and in the case of CCLE, we have obtained 11 essential drugs sdf files from PubChem. To determine the essential drugs, we selected the highest-scoring drugs from the GDSC dataset by considering the top 10% based on the NDCG score. Meanwhile, when creating a list of essential drugs from the CCLE dataset, we chose the top 50% of drugs based on their NDCG scores. The resultant matrix is a binary vector of 256 bits corresponding to each drug in the dataset.

Both the compound's fingerprints and the selected features of the gene expression datasets are concatenated to make the final input matrix and to feed into a deep feedforward network for predicting drug sensitivity scores.

The algorithm 1 outlines a methodology for predicting drug sensitivity scores based on gene expression data and drug molecular structures. Overall, this algorithm provides a structured approach to leveraging gene expression data and drug molecular structures for predicting drug sensitivity, incorporating feature selection techniques and deep learning methodologies.

3. Experimental setup

We have used the same architecture to train our DRPO model on CCLE and GDSC1 datasets. After performing the pre-processing steps, at the first step, we performed feature selection on gene expression data using the f_regression() function, a Python scikit-learn library function that uses correlation statistics to learn the feature's relationship. The parameter K is set to 500, resulting in 500 gene expression features that contributed the most to the IC50 value.

The deep forward network consists of five stacked layers where each neural unit is connected to all the units of the next layer and the obtained output is the sensitivity score of each drug-cell-line pair. Implementation of this model is done on Keras [53]. The dimension for the input layer is 756 i.e. 500 features are gene expression features and 256 bits vectors are drug features obtained by using Morgan's fingerprints. The hidden layers consisted of 4 layers having 1000, 800, 500, and 100 neural units respectively. The activation function for all these hidden layers is an exponential linear unit (elu) [54]. The reason for selecting elu is due to its ability to handle the vanishing gradient problem, improved accuracy, faster convergence, and lack of saturation, especially in deep learning models. There is a single neural unit at the output layer with no activation function. For good estimations of the model's performance, a 10-fold cross-validation approach is used on the input data with 1000 epochs during model training. Where the RMSE is used as the loss function for the feed-forward network and for the dropout [55] rate value we have adopted the same value as used in [34] to avoid overfitting after hidden layers. Furthermore learning rate in the AdaMax optimization algorithm and patience values for early stopping is also the same as employed in [34] experiments.

Algorithm 1 DRPO (Drug Response Prediction and Optimization)

Input: GDSC1 and CCLE datasets, Gene expression data, Drug molecular structures in SDF format
Output: Predicted drug sensitivity scores
Algorithm Steps:
Collect and process GDSC1 and CCLE datasets based on common genes, cell lines, and drugs.
Pre-process the dataset:
• Apply feature selection technique to gene expression dataset.
• Apply Morgan's fingerprint technique to drug data.
Pass the processed dataset to the DRPO model for training.
• Use drug response (IC50) as the label and 500 genes and 256 vector bits of drug data as input features.
Feature Selection:
• Apply SelectKBest feature selection technique to gene expression dataset.
• Find cross-correlation between each gene g_i and the IC50 value s_i using the formula:

$$F_k = \frac{\sum_{i=1}^{n}(g_i-\bar{g})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(g_i-\bar{g})^2 \sum_{i=1}^{n}(s_i-\bar{s})^2}}$$

where n is the number of genes, $\bar{g}$ is the mean of gene expression values, and $\bar{s}$ is the mean of IC50 values.
• Convert cross-correlation to F score to find top rank gene expression features positively correlated with IC50 value.
• Select top-ranked 500 gene expression data features.
For drug (compound) features:
• Use Morgan fingerprint technique to create compound feature vector of 256 bits.
• Obtain drug molecular structures in SDF format from PubChem.
• Select essential drugs based on NDCG scores from GDSC1 and CCLE datasets.
• Concatenate compound fingerprints and selected gene expression features.
Use deep feedforward network for predicting drug sensitivity scores.
Calculate Predicted Drug Sensitivity Score:
• Use the equation:

$$\hat{S}_{ui} = \mu + b_{Qi} + b_{Pu} + q_i \cdot (x_u W_P)^T$$

where:
$-\hat{S}_{ui}$ is the predicted drug sensitivity score.
$-P_u$ represents the cell-lines vector.
$-Q_i$ represents the drug vector.
$-b_{Qi}$ and b_{Pu} are the biases hyper-parameters for cell-line u and drug i.
$-\mu$ represents the mean for all drug responses.
$-W_P$ represents the transformation matrix for projecting cell-line features x_u onto the pharmacogenomic space.

4. Performance metrics

To assess the effectiveness of our proposed model, we employ multiple evaluation metrics. Prior research on predicting cancer cell line drug responses has often been limited in its evaluation methods, relying primarily on RMSE metrics. Our study addresses this gap by introducing two evaluation metrics, RMSE and NDCG, to measure the performance of the model. NDCG score focuses mainly on the ranked list of recommendations, emphasizing the importance of placing more relevant items higher in the list, whereas RMSE focuses on the accuracy of the predicted things without considering their order. Together, both metrics provide a comprehensive evaluation of the proposed system. Using these two evaluation metrics in the proposed model not only gives accurate predictions but also presents those predictions in a way that reflects users' likely preferences in the recommended list. Using both metrics helps to address different aspects of the recommendation

quality: accuracy of predictions and quality of the recommended ranking. Both the CCLE and GDSC1 datasets are evaluated using these two evaluation metrics. The following subsections discuss the details of both RMSE NDCG metrics.

4.1. Root Mean Squared Error (RMSE)

RMSE metrics are used to determine the loss achieved from the model by comparing the actual sensitivity score with the predicted sensitivity score. A smaller RMSE value indicates better model performance in terms of prediction accuracy. The Eq. (8) is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}} \tag{8}$$

Where y_i is the observed drug sensitivity score and $\hat{y}_i$ is the predicted drug sensitivity score by the deep forward network for the ith input data and N is the size of the test data.

4.2. Normalized discounted cumulative gain (NDCG)

Initially in Section 2.2.2, we used the NDCG score to determine the list of essential drugs from both GDSC and CCLE datasets. Here the NDCG score is again used as an evaluation metric for our proposed **DRPO** model. To evaluate our DRPO model for each cell, we calculate the NDCG score for each cell line based on the predicted drug response values. We then average NDCG scores over the entire test dataset to obtain a final NDCG score for the model. Calculating scores based on NDCG is already defined in Eqs. (5) and (6).

5. Result and discussion

Numerous models for predicting drug response have been suggested in existing literature, primarily concentrating on forecasting how various cell lines will respond to a specific drug. Consequently, the assessment of these models has been conducted for each drug separately, relying on the correlation between predicted and observed drug responses. Nevertheless, although predicting the response of cell lines to individual drugs can offer insights into distinct drug response mechanisms, the clinical utility is likely to be higher when ranking drugs for unseen cell lines or patients. Therefore, to obtain a more accurate measurement of the predictive performance of our deep learning models trained on the CCLE and GDSC datasets, we utilized a 10-fold cross-validation technique. This was done to reduce bias in the results. We have randomly divided our experimental data into ten equally sized folds. One fold is selected as the validation set, while the remaining nine folds are used for training purposes, and this process is repeated iteratively. The final results are determined by calculating the average of the root mean square error (RMSE), and NDCG score values obtained from the ten folds. In both evaluation metrics, DRPO consistently generated effective models and demonstrated strong performance on datasets not previously encountered.

5.1. Comparison with other models on CCLE and GDSC1 based on NDCG scores

Based on the NDCG performance metric, we have compared our DRPO model with four previous studies: [20,41–44]. All of these models use the NDCG metric to evaluate their model strength on drug response prediction problems on the CCLE and GDSC1 datasets. The result shown in Table 5 has been obtained from their papers for comparison purposes. To avoid any biased performance measurement, we have performed 10-fold cross-validation for our DRPO evaluation. The result shows that our DRPO has outperformed the previous studies by achieving the highest NDCG score. Fig. 6 shows the visual representation of comparisons between our proposed model and other existing models in literature.

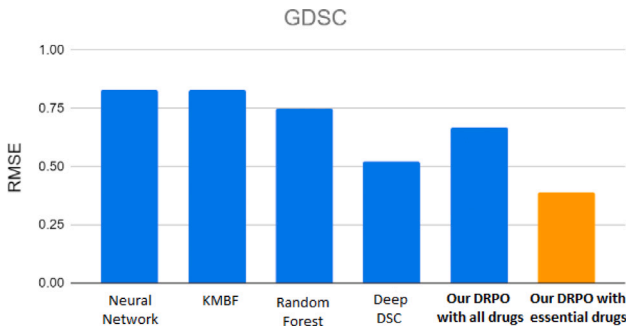


Fig. 4. RMSE score comparison of different machine learning models over GDSC datasets.

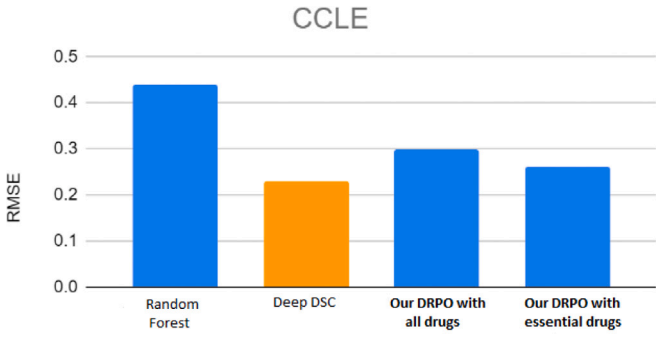


Fig. 5. RMSE score comparison of different machine learning models over CCLE datasets.

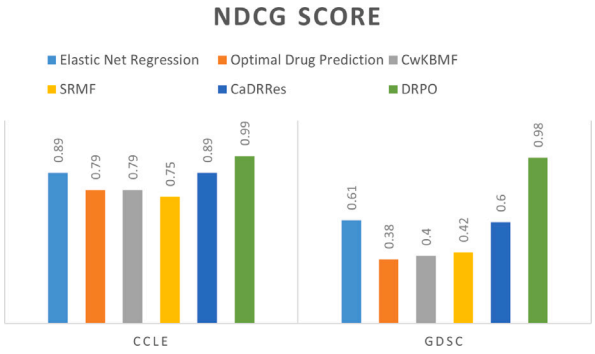


Fig. 6. NDCG score comparison of different machine learning models over CCLE and GDSC datasets.

5.2. Comparison with other models on CCLE and GDSC1 based on RMSE scores

We have compared our DRPO model with the previous studies [45,46] based on the RMSE performance metric [9,32], namely multi-layer NN, KBMF, RF, and DeepDSC. It is observed that our proposed DRPO model, when trained on essential drugs, outperformed these prominent models in the existing literature based on RMSE scores. We have also performed 10-fold cross-validation to obtain less biased performance measurements. The average RMSE with essential drugs on CCLE is 0.26, and on GDSC1, it is found to be 0.39, as shown in Table 6 and Figs. 4 & 5. These missing values at CCLE columns in Table 6 are because the [45,46] approaches were not applied to the CCLE dataset. The result indicates that our proposed approach works well on the CCLE and GDSC datasets using the gene expression profile and Drug Response datasets when trained on essential drugs.

6. Conclusion

In conclusion, we successfully propose a new method, DRPO, in which we first find the essential anticancer drugs using matrix factorization and then get the optimized cell line's oncogene expression features using the SelectKBest technique. The final input matrix is built by integrating the essential drugs' chemical features with optimized cell lines' gene expression features. This final input matrix is then fed into a deep feed-forward network to train our DRPO model. Our experimental results show that our new method has outperformed the previous approaches discussed in the literature with respect to NDCG and RMSE scores on both the CCLE and GDSC datasets. To avoid biased performance measurement, we evaluated our experimental results based on 10-fold cross-validation using sub-sampling.

Our research presents a novel approach for predicting drug response to cancer cell lines using gene expression data and the chemical structure of drugs. The limitation of the present study is the consideration of gene expression data only, without considering the multi-omics perspective of the target gene. Along with the multi-omic data, the network biology-based dimensions will strengthen the biological significance. This work can be extended in future work by incorporating other molecular features like mutation, copy number variation, and methylation along with gene expression data. Moreover, explainability and interpretability will also be added to this framework so that biological significance can be explicitly drawn from the results.

CRediT authorship contribution statement

Muhammad Shahzad: Data curation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Adila Zain Ul Abedin Kadani:** Writing – review & editing. **Muhammad Atif Tahir:** Conceptualization, Writing – review & editing. **Rauf Ahmed Shams Malick:** Writing – review & editing. **Richard Jiang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

We have no conflicts of interest to disclose.

Acknowledgments

This work is supported in part by the Engineering and Physical Sciences Research Council (EPSRC) Grant EP/P009727/2 and in part by the Leverhulme Trust Grant RF-2019-492.

References

[1] A.C. Begg, F.A. Stewart, C. Vens, Strategies to improve radiotherapy with targeted drugs, Nat. Rev. Cancer 11 (4) (2011) 239–253.
[2] M. Bachtiar, C.G. Lee, Genetics of population differences in drug response, Curr. Genet. Med. Rep. 1 (3) (2013) 162–170.
[3] N.B. La Thangue, D.J. Kerr, Predictive biomarkers: a paradigm shift towards personalized cancer medicine, Nat. Rev. Clin. Oncol. 8 (10) (2011) 587–596.
[4] L.J. Van't Veer, R. Bernards, Enabling personalized cancer medicine through analysis of gene-expression patterns, Nature 452 (7187) (2008) 564–570.
[5] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C.J. Wilson, J. Lehár, G.V. Kryukov, D. Sonkin, et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (7391) (2012) 603–607.
[6] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (D1) (2012) D955–D961.
[7] R.H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen, Nat. Rev. Cancer 6 (10) (2006) 813–823.
[8] J.C. Costello, L.M. Heiser, E. Georgii, M. Gönen, M.P. Menden, N.J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S.A. Khan, et al., A community effort to assess and improve drug sensitivity prediction algorithms, Nat. Biotechnol. 32 (12) (2014) 1202–1212.
[9] I. Cortés-Ciriano, G.J. van Westen, G. Bouvier, M. Nilges, J.P. Overington, A. Bender, T.E. Malliavin, Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel, Bioinformatics 32 (1) (2016) 85–95.
[10] T. Turki, Z. Wei, A link prediction approach to cancer drug sensitivity prediction, BMC Syst. Biol. 11 (5) (2017) 1–14.
[11] C. Huang, R. Mezencev, J.F. McDonald, F. Vannberg, Open source machine-learning algorithms for the prediction of optimal cancer drug therapies, PLoS One 12 (10) (2017) e0186906.
[12] F. Azuaje, Computational models for predicting drug responses in cancer research, Brief. Bioinform. 18 (5) (2017) 820–829.
[13] I.S. Jang, E.C. Neto, J. Guinney, S.H. Friend, A.A. Margolin, Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data, in: Biocomputing 2014, World Scientific, 2014, pp. 63–74.
[14] M.J. Garnett, E.J. Edelman, S.J. Heidorn, C.D. Greenman, A. Dastur, K.W. Lau, P. Greninger, I.R. Thompson, X. Luo, J. Soares, et al., Systematic identification of genomic markers of drug sensitivity in cancer cells, Nature 483 (7391) (2012) 570–575.
[15] J. Chen, L. Zhang, A survey and systematic assessment of computational methods for drug response prediction, Brief. Bioinform. 22 (1) (2021) 232–246.
[16] K. Sachdev, M.K. Gupta, A comprehensive review of feature based methods for drug target interaction prediction, J. Biomed. Inform. 93 (2019) 103159, http://dx.doi.org/10.1016/j.jbi.2019.103159, URL https://www.sciencedirect.com/science/article/pii/S1532046419300772.
[17] B. Agyemang, W.-P. Wu, M.Y. Kpiebaareh, Z. Lei, E. Nanor, L. Chen, Multi-view self-attention for interpretable drug–target interaction prediction, J. Biomed. Inform. 110 (2020) 103547, http://dx.doi.org/10.1016/j.jbi.2020.103547, URL https://www.sciencedirect.com/science/article/pii/S1532046420301751.
[18] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, X.S. Liu, Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model, PLoS Comput. Biol. 11 (9) (2015) e1004498.
[19] L. Wang, X. Li, L. Zhang, Q. Gao, Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization, BMC Cancer 17 (1) (2017) 1–12.
[20] C. Suphavilai, D. Bertrand, N. Nagarajan, Predicting cancer drug response using a recommender system, Bioinformatics 34 (2018) 3907–3914.
[21] A. Emdadi, C. Eslahchi, Dsplmf: a method for cancer drug sensitivity prediction using a novel regularization approach in logistic matrix factorization, Front. Genet. 11 (2020) 75.
[22] N.-N. Guan, Y. Zhao, C.-C. Wang, J. Li, X. Chen, X. Piao, Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization, Mol. Ther. Nucleic Acids 17 (2019) 164–174.
[23] R. Rani, A. Sharma, KSRMF: Kernelized similarity based regularized matrix factorization framework for predicting anti-cancer drug responses, J. Intell. Fuzzy Systems 35 (2018) http://dx.doi.org/10.3233/JIFS-169713.
[24] H. Liu, Y. Zhao, L. Zhang, X. Chen, Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal, Mol. Therapy - Nucleic Acids 13 (2018) 303–311, http://dx.doi.org/10.1016/j.omtn.2018.09.011, URL https://www.sciencedirect.com/science/article/pii/S2162253118302555.
[25] R. Jiang, D. Crookes, Shallow unorganized neural networks using smart neuron model for visual perception, IEEE Access 7 (2019) 152701–152714.
[26] L. Wang, Z.-H. You, X. Chen, S.-X. Xia, F. Liu, X. Yan, Y. Zhou, K.-J. Song, A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network, J. Comput. Biol. 25 (3) (2018) 361–373.
[27] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, F. Cheng, deepDR: a network-based deep learning approach to in silico drug repositioning, Bioinformatics 35 (24) (2019) 5191–5198.
[28] M. Karimi, D. Wu, Z. Wang, Y. Shen, DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, Bioinformatics 35 (18) (2019) 3329–3338.
[29] Y.-C. Chiu, H.-I.H. Chen, T. Zhang, S. Zhang, A. Gorthi, L.-J. Wang, Y. Huang, Y. Chen, Predicting drug response of tumors from integrated genomic profiles by deep neural networks, 2018, arXiv:1805.07702.
[30] A. Oskooei, J. Born, M. Manica, V. Subramanian, J. Sáez-Rodríguez, M. Rodriguez Martinez, PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks, 2018.
[31] V. Malik, Y. Kalakoti, D. Sundar, Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer, BMC Genom. 22 (2021) http://dx.doi.org/10.1186/s12864-021-07524-2.
[32] M. Li, y. Wang, R. Zheng, X. Shi, Y. Li, F. Wu, J. Wang, DeepDSC: A deep learning method to predict drug sensitivity of cancer cell lines, IEEE/ACM Trans. Comput. Biol. Bioinform. PP (2019) 1, http://dx.doi.org/10.1109/TCBB.2019.2919581.
[33] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T.S. Kim, J. Jung, J.-M. Shin, Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature, Sci. Rep. 8 (1) (2018) 1–11.

[34] P. Liu, H. Li, S. Li, K.-S. Leung, Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network, BMC Bioinform. 20 (1) (2019) 1–14.
[35] C.-Y. Chiang, C. Barnes, P. Angelov, R. Jiang, Deep learning-based automated forest health diagnosis from aerial images, IEEE Access 8 (2020) 144064–144076.
[36] Y. Zhu, T. Brettin, Y.A. Evrard, A. Partin, F. Xia, M. Shukla, H. Yoo, J.H. Doroshow, R.L. Stevens, Ensemble transfer learning for the prediction of anti-cancer drug response, Sci. Rep. 10 (1) (2020) 1–11.
[37] T. Turki, J.T. Wang, Clinical intelligence: New machine learning techniques for predicting clinical drug response, Comput. Biol. Med. 107 (2019) 302–322.
[38] S. Huang, P. Hu, T.M. Lakowski, Predicting breast cancer drug response using a multiple-layer cell line drug response network model, BMC Cancer 21 (1) (2021) 648.
[39] B.M. Kuenzi, J. Park, S.H. Fong, K.S. Sanchez, J. Lee, J.F. Kreisberg, J. Ma, T. Ideker, Predicting drug response and synergy using a deep learning model of human cancer cells, Cancer Cell 38 (5) (2020) 672–684.
[40] T. Wang, K. Birsoy, N. Hughes, K. Krupczak, Y. Post, J. Wei, E. Lander, D. Sabatini, Identification and characterization of essential genes in the human genome, Science 350 (2015) http://dx.doi.org/10.1126/science.aac7041.
[41] F. Iorio, T. Knijnenburg, D. Vis, G.R. Bignell, M. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R.K. Egan, Q. Liu, M. Garnett, A landscape of pharmacogenomic interactions in cancer, Cell 166 (2016) http://dx.doi.org/10.1016/j.cell.2016.06.017.
[42] J. Sheng, F. Li, S.T.C. Wong, Optimal drug prediction from personal genomics profiles, IEEE J. Biomed. Health Inf. 19 (2015) 1264–1270.
[43] M. Ammad-ud din, S. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio, S. Kaski, Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization, Bioinformatics 32 (2016) i455–i463, http://dx.doi.org/10.1093/bioinformatics/btw433.
[44] L. Wang, X. Li, L. Zhang, Q. Gao, Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization, BMC Cancer 17 (2017) http://dx.doi.org/10.1186/s12885-017-3500-5.
[45] M.P. Menden, F. Iorio, M.J. Garnett, U. McDermott, C.H. Benes, P.J. Ballester, J. Sáez-Rodríguez, Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties, PLoS ONE 8 (2013).
[46] M. Ammad-ud din, E. Georgii, M. Gönen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, S. Kaski, Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization, J. Chem. Inf. Model. 54 (2014) http://dx.doi.org/10.1021/ci500152b.
[47] E. Bolton, Y. Wang, P. Thiessen, S. Bryant, Chapter 12 PubChem: Integrated platform of small molecules and biological activities, Annu. Rep. Comput. Chem. 4 (2008) 217–241, http://dx.doi.org/10.1016/S1574-1400(08)00012-1.
[48] C. Suphavilai, D. Bertrand, N. Nagarajan, Predicting cancer drug response using a recommender system, Bioinformatics 34 (22) (2018) 3907–3914.
[49] C. Distinguishability, A theoretical analysis of normalized discounted cumulative gain (NDCG) ranking measures, 2013.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[51] C. Brooksbank, M.T. Bergman, R. Apweiler, E. Birney, J. Thornton, The European bioinformatics institute's data resources 2014, Nucleic Acids Res. 42 (D1) (2014) D18–D25.
[52] S. Kim, P. Thiessen, E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. Shoemaker, J. Wang, B. Yu, J. Zhang, S. Bryant, PubChem substance and compound databases, Nucleic Acids Res. 44 (2015) http://dx.doi.org/10.1093/nar/gkv951.
[53] F. Chollet, "Keras: Deep learning library for theano and tensorflow, 2015, GitHub repository, GitHub https://github.com/fchollet/keras.
[54] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), 2015, arXiv preprint arXiv:1511.07289.
[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.