DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines

Min Li, Yake Wang, Ruiqing Zheng, Xinghua Shi, Yaohang Li, Fang-Xiang Wu, and Jianxin Wang

Abstract—High-throughput screening technologies have provided a large amount of drug sensitivity data for a panel of cancer cell lines and hundreds of compounds. Computational approaches to analyzing these data can benefit anticancer therapeutics by identifying molecular genomic determinants of drug sensitivity and developing new anticancer drugs. In this study, we have developed a deep learning architecture to improve the performance of drug sensitivity prediction based on these data. We integrated both genomic features of cell lines and chemical information of compounds to predict the half maximal inhibitory concentrations (IC_{50}) on the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC) datasets using a deep neural network, which we called DeepDSC. Specifically, we first applied a stacked deep autoencoder to extract genomic features of cell lines from gene expression data, and then combined the compounds' chemical features to these genomic features to produce final response data. We conducted 10-fold cross-validation to demonstrate the performance of our deep model in terms of root-mean-square error (RMSE) and coefficient of determination R^2. We show that our model outperforms the previous approaches with RMSE of 0.23 and R^2 of 0.78 on CCLE dataset, and RMSE of 0.52 and R^2 of 0.78 on GDSC dataset, respectively. Moreover, to demonstrate the prediction ability of our models on novel cell lines or novel compounds, we left cell lines originating from the same tissue and each compound out as the test sets, respectively, and the rest as training sets. The performance was comparable to other methods.

Index Terms—Deep learning, cancer cell lines, drug sensitivity, autoencoder, predictive models

1 Introduction

Cultured Cancer cell lines with heterogeneous genomic backgrounds and gene expressions are fundamental materials to study the molecular basis of drug activity [1] and to discover novel anticancer drugs in cancer biology, despite their genomic differences from original tissue or tumor samples [2], [3]. Several large-scale high-throughput screening efforts have catalogued genomic information of a panel of in vitro cell lines as well as their drug sensitivity profiles against hundreds of compounds. The US National Cancer Institute (NCI) made the first effort to screen over 60 human tumor cell lines with thousands small molecules aiming to discover potential anticancer compounds in the late 1980s [4]. Since then, the NCI60 has been serving to study the mechanisms of growth inhibition and killing of tumor cell lines. The Cancer Cell Line Encyclopedia (CCLE) project compiled genomic profiles of 947 human cancer cell lines, and pharmacologic profiles of 24 anticancer drugs across 479 cancer cell lines to benefit personalized medicine [5]. The Genomics of Drug Sensitivity in Cancer (GDSC) is another effort to catalogue genomic profiles of 639 human cancer cell lines and their drug response data to 130 drugs aiming to identify genomic biomarkers of drug sensitivity in cancer cells [6]. Cell lines in CCLE and GDSC datasets originate from several human cancer tissues, such as lung, breast, and kidney. Both CCLE and GDSC datasets have abundant genomic information data including gene expression, DNA copy number, oncomap mutations, and so on.

These high-throughput screening datasets have made a great contribution to cancer biology research and cancer treatment. Previous studies on these datasets are wide-ranging, such as novel anticancer drug discovery [1], [7], human cancer pathogenesis analysis [8], [9], anticancer drug repositioning [10], [11], and so on. In the context of drug discovery and repositioning, computational approaches have been widely adopted to predict drug sensitivity data over human cancer cell lines. Initial methods known as Quantitative Structure-Activity Relation-ship (QSAR) generally build models on a set of fixed cell lines or tissues using drug properties under the assumption that chemically and structurally similar compounds may have similar biological effect on known cell lines [12], [13], [14]. Though widely adopted to explore on chemical space to discover novel anticancer compounds, QSAR models could not generalize across cancer cell lines since they treat the drug response prediction as a single-task learning problem on given cell lines whose properties were not considered.

- *M. Li, Y. Wang, R. Zheng and J. Wang are with the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P. R. China. E-mail: {limin, jxwang}@mail.csu.edu.cn, {wangyk, rqzheng}@csu.edu.cn.*
- *X. Shi is with the Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC 28223 USA. E-mail: x.shi@uncc.edu.*
- *Y. Li is with the Department of Computer Science, Old Dominion University, Norfolk, VA 23529 USA. E-mail: yaohang@cs.odu.edu.*
- *F. Wu is with the Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. E-mail: faw341@mail.usask.ca.*

Manuscript received 8 Oct. 2018; revised 28 Apr. 2019; accepted 23 May 2019. Date of publication 28 May 2019; date of current version 1 Apr. 2021. (Corresponding author: Min Li.)
Digital Object Identifier no. 10.1109/TCBB.2019.2919581

However, modeling chemical features and cell line features simultaneously may benefit drug sensitivity predictions, drugs side effects' analysis and the models' ability to extrapolate to novel compounds and cell lines [14]. Menden et al. [15] made the first effort to integrate cell line genomic features, including microsatellites, sequence variation and copy number variation, and compounds' one dimensional (1D) and two dimensional (2D) chemical properties to model the half maximal inhibitory concentration (IC_{50}) values of 111 drugs across 608 cell lines using a three-layer neural network and random forest (RF). As a result, a coefficient of determination R^2 of 0.64 and an RMSE of 0.97 was obtained on a hold-out test set. Amid-ud-din et al. [16] proposed an extended QSAR model via integrating cell line genomic characteristics, such as gene expression data, copy number variationas, and cancer gene mutations, as side information to predict drug response data of 650 cell lines to 116 drugs using Kernelized Bayesian Matrix Factorization (KBMF). The R^2 for novel drug prediction is 0.32. Another study [17] built a dual-layer integrated cell line-drug network containing a cell line similarity network (CSN) and a drug similarity network (DSN) based on the Pearson correlation coefficients of cell lines' gene expression profiles and compounds' 1D and 2D information on CCLE and Cancer Genome Project (CGP) datasets, respectively. This work was under the assumption that similar drugs may have similar responses on given cell lines and vice versa. The drug responses were deduced by each network, then a weighted linear formula was used to get the final responses, with the weights customized for every drug. This model got a Pearson correlation coefficient of 0.6 between the predicted and the observed responses for most drugs. Cortes-Ciriano et al. [18] compared seven genomic profiles and their combinations of cell lines, and found that protein, gene transcript level and miRNA abundance have the highest predictive ability when modeling the 50 percent growth inhibition bioassay end point (GI_{50}) values on NCI60 dataset. After that, they integrated the transcript profiles of top 1000 genes that display the highest variance across 59 cell lines and Morgan fingerprints of 17142 compounds to predict the drug responses using RF and support vector machine (SVM). This work grouped more than seventeen thousand compounds in NCI60 dataset into sets at random and not based on chemical similarity, which likely made the extrapolation easier [18]. A very recent study [19] proposed a similarity regularized matrix factorization (SRMF) method to predict drug responses on CCLE and GDSC datasets using drug chemical profiles and gene expression profiles. This work adopted the same assumption as Zhang et al. [17] and obtained better results on GDSC than KBMF [16] and Zhang et al. [17].

Although many studies have been devoted to predicting drug response based on computational approaches, it is still a challenge task to develop accurate drug sensitivity predictive models. One of the major solutions to this challenge is to collect as much high quality bioactivity data as possible, while it can be costly [14]. Another way is to develop more accurate and robust computational models based on current available datasets to improve the predictive performance.

Inspired by the rapid development of deep learning technology, in this paper, we propose a deep learning architecture to predict drug sensitivity of cancer cell lines by integrating genomic profiles of cell lines and chemical profiles of compounds. Our model consists of a stacked deep autoencoder at first to extract cell lines feature from gene expression data in an unsupervised way. The stacked deep autoencoder also serves as a dimension reduction instrument in our model, since the dimension of gene expression data is extremely huge. After that, we integrate the chemical features of compounds into this model to bring drug sensitivity data of the given cell line-compound pairs. We first perform 10-fold cross-validation on CCLE and GDSC datasets to show its interpolation ability on known cell lines and drugs. Then, we perform leave-one-out validation both on tissues and compounds to show its extrapolation ability on novel cell lines and compounds, respectively. The lowest prediction errors show that our deep learning model outperforms the state-of-the-art approaches on drug sensitivity prediction. To our knowledge, this study is the first attempt to leverage deep learning to model drug sensitivity. The outstanding performance of our model indicates that deep learning can promote the study of drug sensitivity prediction considerably.

2 Materials and Methods

2.1 Materials

In this study, we propose a deep learning regression model, namely DeepDSC, to predict drug sensitivity on cancer cell lines. Both cell lines expression data and compounds fingerprints were integrated as the input of our model.

The expression profiles of cell lines and drug sensitivity data were collected from two public datasets, CCLE and GDSC. The compound chemical structure files were downloaded from PubChem [20].

CCLE. The Robust Multi-Array Average-normalized (RMA) Affymetrix $\mathrm{U133}+2$ arrays gene expression data of 1037 cell lines in CCLE datasets were downloaded from CCLE website. The gene expression data of a sort of cell lines contain the transcript level of about 20,000 genes, thus correspond to a vector of the same length. We further extracted 504 cell lines with response data against 24 drugs for this study. The drug responses are the IC_{50} values, with the unit of uM. A low IC_{50} value indicates that the given cell line is sensitive to the given drug while a high IC_{50} value means the opposite. We converted the response data IC_{50} to $-\log_{10}IC_{50}(\mu M)$, the negative logarithm of IC_{50} values, to compare with the previous study [18]. The 1D and 2D compound structures of 23 drugs were downloaded from PubChem in Standard Delay Format (sdf), except for LBW242, whose chemical structure file is not available. We standardized the compound structures and computed their hashed count Morgan fingerprints using camb [21] by setting the size of 256 bits, thus obtained compounds' feature vectors of a length of 256. The final data matrix contains 491 cell lines and 23 drugs with 96.25 percent data completeness.

GDSC. The gene expression data of around 789 cell lines were downloaded from The European Bioinformatics Institute (EMBL-EBI) [22] in Affymetrix CEL file format. Every original cel file also contains the transcript level of around 20,000 genes, but they are not normalized. Most cel files and cell lines are one-to-one correspondence, with a few cel files correspond to the same cell lines. We perform RMA normalization using R package oligo [23], and average the normalized

TABLE 1
The Two Datasets Used in Experiments

dataset	cell lines	drugs	data points	completeness
CCLE	491	23	10,870	96.25%
GDSC	655	139	73,075	80.26%

Cell lines and drugs denote the number of cancer cell lines and compounds included in this study, respectively. Data points denote the number of cell line-compound pairs in corresponding dataset. Completeness denotes the ratio of data points to the product of number of cell lines and drugs.

data corresponding to the same cell lines. The cell line drug response data IC_{50} are downloaded from GDSC website (release 5.0) and converted to $\log_{10}IC_{50}(\mu M)$ for comparison. Cell lines with more than one sensitivity data against the same drugs are filtered out. The compounds' 1D and 2D structure files are also downloaded from PubChem and processed in the same way as those in CCLE. The final data matrix contains 655 cell lines and 139 drugs with a completeness of 80.26 percent. Table 1 summarizes these two datasets.

2.2 Deep Learning and Autoencoder

Deep learning are a set of algorithms to learn hierarchical concepts built on top of each other through deep architectures which can extract increasingly abstract features of original inputs through a series of hidden layers with non-linear transformations [24]. Particularly, deep learning has achieved remarkable success in computer vision [25], [26], speech recognition [27], [28], natural language processing [29], [30] and bioinformatics [31], [32], [33] in recent years. A classic deep network contains three kinds of layers, input layer, hidden layers and output layer. Neural units at the same layer share the same non-linear transformation, also known as activation function. The forward propagation procedure begins at the original inputs and the output of each neural unit is the value of the activation (typically nonlinear) function of the linear combination of its inputs, which are the outputs of neural units at the previous layer. This procedure is defined as:

$$z^l = f(W^l x + b^l), \tag{1}$$

where $x^l \in \mathbb{R}^m$ and $z^l \in \mathbb{R}^n$ are the input vector and output vector of the lth layer with a length of m and n, respectively. $f(\cdot)$ is the activation function. $W \in \mathbb{R}^{n \times m}$ is the weight matrix, and $b^l \in \mathbb{R}^n$ is the bias vector of the l-th layer. The outputs of the deep network can either be discrete or continuous depending on the task it performs (classification or regression). Autoencoder is a typical unsupervised deep learning model, and it tries to reconstruct the input under some constraints, such as low-dimension, sparse, and noise-free [34], [24]. An autoencoder consists of an encoder and a decoder. Both can be considered as a function described by (2) and (3). An encoder takes the original input $x \in \mathbb{R}^d$ and perform some transformations $f_\theta(\cdot)$ on it to get the hidden representation $h \in \mathbb{R}^t$ of the input, where d and t are the dimension of the input and hidden representation, and in general, t is much smaller than d, since the encoder is often hoped to generate low dimension and abstract features useful for downstream tasks. In contrast, a decoder takes the hidden representation h and performs some transformations $g_{\theta'}(\cdot)$ to produce the reconstruction vector x'.

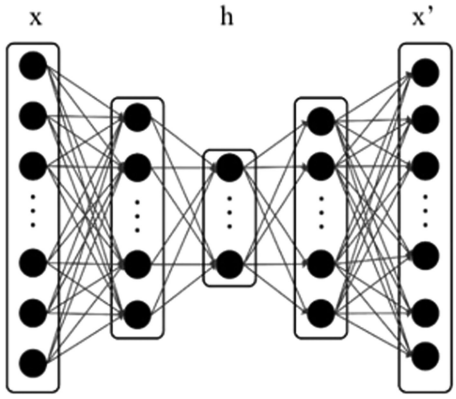


Fig. 1. Architecture of a stacked deep autoencoder, with a three-layer encoder and a three-layer decoder. Where x and x' denote the input and the reconstruct output, and h denotes the encoded feature representation. An autoencoder first encodes the input x to hidden representation h and then use a decoder to reconstruct the input.

$$h = f_\theta(x) \tag{2}$$

$$x' = g_{\theta'}(h) = g_{\theta'}(f(x)) \tag{3}$$

$$L(x, x') = L(x, g_{\theta'}(f_\theta(x))). \tag{4}$$

The reconstruction loss $L(x, x')$ between the original inputs and the reconstruction vectors are used in back propagation (BP) to guide the training procedure, which means to forward and backward propagate through the network many times to update the parameters θ and θ' and reduce the reconstruction loss. Fig. 1 shows the architecture of a stacked deep autoencoder.

2.3 DeepDSC

Given enough neural units or deep layers, a feedforward deep network can become a universal approximator [29]. Powerful modeling ability makes it extensively used in many fields [35], [36], [37], [38], [39]. This ability can be used to promote drug sensitivity prediction considering the existence of large amount of drug pharmacological profiling data, either in a regression task or in a classification task. The former tries to construct a computational model to predict the continuous drug sensitivity values, such as IC_{50} values and the area under the dose activity curve (AUC), while the latter tries to model drug sensitivity categories by discretizing drug sensitivity data. However, Jang et al. [40] pointed out that discrete drug sensitivity data lose some information when comparing to continuous data and regression models in general outperform classification in drug sensitivity prediction. Following their suggestion, we have trained a deep neural network to model the continuous drug response values (IC_{50}s), which outperforms start-of-the-art studies in this field. Fig. 2 shows the flowchart of our deep learning model. Our mode can be summarized in two parts. Since autoencoder is in general used to learn abstract features and reduce feature dimensions, we first adopt a stacked deep autoencoder (inside the yellow box) to extract features from gene transcript profiles (blue circles) in an unsupervised way and produce cell lines' features (green circles). Before fed into the autoencoder, the gene expression data in both datasets is normalized to the range of $[0, 1]$ by subtracting the corresponding minimal

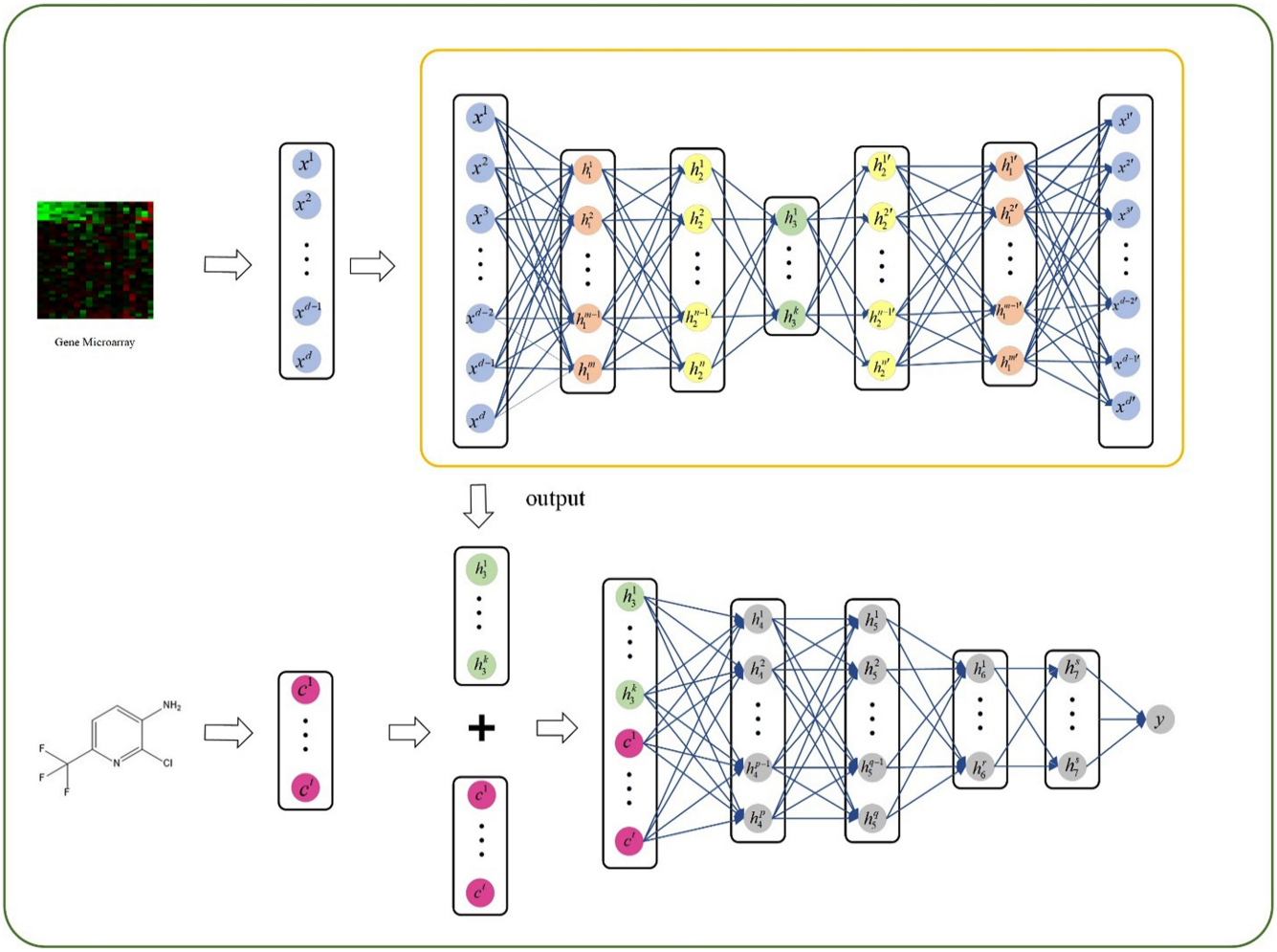


Fig. 2. Flowchart of a DeepDSC, with a three-layer encoder and a three-layer decoder. Where x and x' denote the input and the reconstruction output, and h denotes the encoded feature representation. An autoencoder first encodes the input x to hidden representation h and then use a decoder to reconstruct the input.

value and divided by the maximum value. The dimension of the extracted features is much smaller than the original data. More specifically, the dimension of the learned features is reduced to the fortieth of the primary data, of which the dimension is extrmently large, about 20,000. The decoder only takes effect in the stage of training and is discarded after that, since we only need the learned features for downstream drug sensitivity prediction. The drug chemical features, Morgan fingerprints (dark red circles) of the 1D and 2D compound structure, are integrated into the model at the second parts, combined with the learned cell line features. The reason why we don't combine compound features and cell line gene expression data to get the combined features at the very beginning of this procedure and take them as the inputs of autoencoder directly is that comparing with the length of gene expression data, compound fingerprints, with a length of 256, only holds 1.2 percent of the total length or so. This can be a severe problem for the subsequent autoencoder since the loss of reconstructing gene expression data will dominate the reconstruction loss described in (4) and training an autoencoder on such data will ultimately lead to bad hidden features in which the compound information is almost covered by cell line information. After feature dimension reduction using the autoencoder, the length of cell line features is comparable to that of compound features. The second part of DeepDSC is a feedforward neural network (gray circle). It contains five stacked layers with each neural unit connecting to all the units at next layer and the outputs are drug sensitivity data (denoted as y) of given cell line-drug pairs. Both the learned hidden features and compound fingerprints are concatenated together simply as the inputs of this part. Errors between the predicted drug responses and the ground truth are used to train this part, to adjust the network parameters and reduce the predictive error.

Generally, our deep learning model contains two steps. At the first step, the stacked autoencoder is adopted to extract genetic features of cell lines from gene expression data, which also takes an effect of dimension reduction. At the second step, the learned cell line features are combined with the compound Morgan fingerprints as the inputs of a deep feedforward neural network to predict drug sensitivity data.

2.4 Experimental Setup

Owing to the data inconsistency between GDSC and CCLE datasets, models trained on one dataset using the original pharmacologic profiles could not generalize to the other directly [41]. We first trained models on CCLE dataset and applied the same architecture to train models on GDSC

TABLE 2
Results on GDSC Dataset

	method	NN	KBMF	RF	DeepDSC
CV	RMSE	0.83	0.83+/−1.00	0.75+/−0.01	0.52+/−0.01
	R^2	0.72	$0.32 + / - 0.37$	$0.74 + / - 0.01$	$0.78 + / - 0.01$
LOTO	RMSE	0.99	NA	$0.81 + / - 0.16$	$0.64 + / - 0.05$
	R^2	0.61	NA	$0.72 + / - 0.08$	$0.66 + / - 0.07$
LOCO	RMSE	NA	$0.85 + / - 0.41$	$1.40 + / - 0.80$	$1.24 + / - 0.74$
	R^2	NA	$0.52 + / - 0.37$	$0.13 + / - 0.11$	$0.04 + / - 0.06$

dataset, which means these models share the same numbers of layers and corresponding neural units. At the first step, except for the input layer, the encoder has three hidden layers: the first layer contains 2,000 neural units, the second layer contains 1,000, and the third layer contains 500. The third layer is the encoded layer, corresponding to the extracted features with a length of 500. The decoder has two hidden layers with numbers of units of 1,000 and 2000 respectively, and an output layer with the same number of units as the input layer in the encoder. The activation function of these layers, except for the input layer and the output layer, is selu (Scaled Exponential Linear Unit) [42]. The activation function of the output layer is sigmoid function [43], which maps the outputs to the range of (0, 1), in order to reconstruct the normalized gene expression data. The loss function of the autoencoder is cross entropy [44].At the stage of training the deep autoencoder, we initialize the parameters using Xavier uniform initializer and choose AdaMax as the optimization algorithm with a learning rate of 0.0001 [45]. To avoid the problem of gradient explosion, we clip the parameter gradients to make sure that their absolute values are no more than 1.

The dimension of the deep feedforward network's input layer is 756, as a result of an addition of 500 (length of the extracted cell lines' features) and 256 (length of the compounds' fingerprints). Except for the input layer and output layer, the rest hidden layers have 1,000, 800, 500, and 100 neural units, and all these four layers share the same activation function elu [46]. The output layer only has one neural unit and there is no activation function to restrict the output range. The loss function of the deep feed forward network is RMSE. These parameters are initialized via He normal initializer [47]. To avoid overfitting, the dropout strategy [48] is adopted after these layers, and the drop rate is set to 0.1. We also use early stopping with a patience of 30. The optimization algorithm for training is also AdaMax, and the learning rate is set to 0.0004. We clip the gradient absolute values to no more than 5 to overcome the gradient explosion problem. The model is implemented via Keras [49].

2.5 Performance Metrics

There are many metrics to measure the error of a regression model, such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE) and so on. Among them, the RMSE is more acute to unusual lager error and is more suitable to measure the drug sensitivity model, since the model is always encouraged to try to avoid big mistakes. Thus, we chose root mean squared error (RMSE) described in (5) as the error metric of our deep learning model. We also choose coefficient of determination R^2 [50] described in (6) to measure the predictive performance of our deep learning models.

$$RMSE = \sqrt{\sum (y_i - \tilde{y}_i)^2 / N} \tag{5}$$

$$R^2 = 1 - \sum (y_i - \tilde{y}_i^{r0})^2 / \sum (y_i - \bar{y}_i)^2, \tag{6}$$

where N is the size of the test data, y_i and $\tilde{y}_i$ are the target drug sensitivity data and the predicted counterpart of ith input data, and $\bar{y}$ is the average value of the target drug data. $\tilde{y}_i^{r0} = k\tilde{y}_i$, where k is the slope defined in (7).

$$k = \sum y_i \tilde{y}_i / \sum \tilde{y}_i^2. \tag{7}$$

3 Results and Discussion

We have performed 10-fold cross-validation to obtain a less biased predictive performance measurement of our deep learning models trained on CCLE and GDSC datasets. The experimental data are randomly divided into ten equal parts. Each part is chosen as the validation set and the rest are used as the training set, iteratively. We average 10 folds RMSE and R^2 values to obtain the final results, respectively. The cross-validation results can be considered as the measurement of missing drug sensitivity data imputation. We also left cell lines from each tissue out (LOTO, leave-one-tissue-out) and each compound out (LOCO, leave-one-compound-out) to test the ability of the models trained on the rest data to predict drug sensitivity data on novel cell lines and compounds. The final RMSE and R^2 values are averaged across the tissues and compounds.

3.1 Comparison on GDSC Dataset

On GDSC dataset, we compared with three previous studies [15], [16], [18], which we termed as NN, KBMF and RF, using the same performance metrics, RMSE and coefficient of determination R^2. All of them modeled cell line features and compound chemical features simultaneously in different ways. We adopted the original results described in their papers and the same output unit, $\log_{10} IC_{50}(\mu M)$, for comparison. All of three studies performed cross-validation experiments. As a result, DeepDSC achieved the best prediction performance with the lowest RMSE of 0.52 and the highest coefficient of determination R^2 of 0.78, which indicated that DeepDSC has the best ability to impute missing drug sensitivity data among the four approaches. To test the denovo performance, NN and RF performed LOTO experiments, while KBMF and RF performed LOCO experiments. Table 2 shows the results of DeepDSC, NN, KBMF and RF on GDSC dataset.

From Table 2 we can see that the prediction error of DeepDSC is lower than all other three approaches. However, the R^2 values of LOTO and LOCO experiments of DeepDSC are slightly lower. This is because the ranges of the response data left are quite small compared to the whole data. When the data range considered is small, R^2 cannot reflect the goodness of the prediction performance of the model, and R^2 may get worse even the model accuracy is improved [18]. This can be illustrated by Fig. 3. We first randomly sampled 100 points in [0, 3] and [0, 100] as x-axis, and then added noises sampled from a standardized

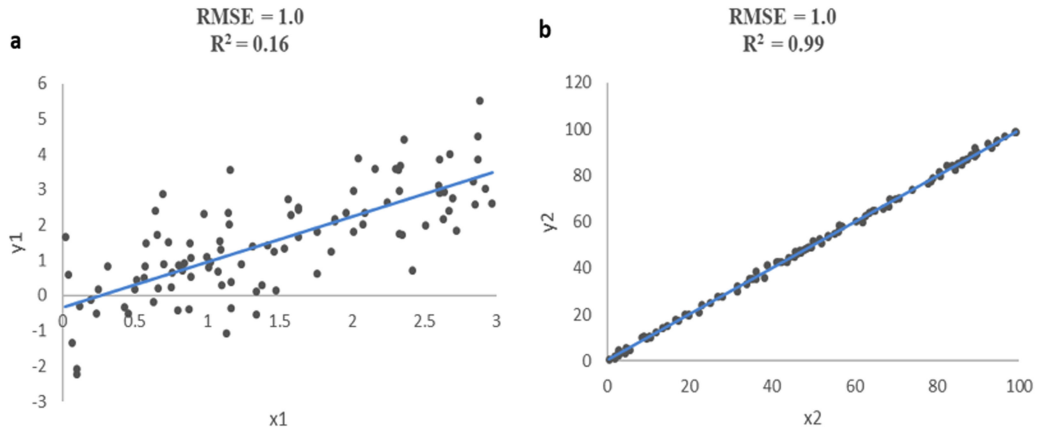


Fig. 3. Significant change of R^2. We randomly sampled 100 values in a range of 0-3 and 0-100 for a and b as x-axis, respectively. Noises sampled from a standardized normal distribution were added to obtain y-axis. Even a and b share the same RMSE value, the R^2 value of a is smaller than b.

TABLE 3
Results on CCLE Dataset

	CV		LOTO		LOCO	
Metrics	RMSE	R^2	RMSE	R^2	RMSE	R^2
RF	$0.44 + / - 0.02$	$0.74 + /0.03$	$0.42 + / - 0.11$	$0.75 + / - 0.12$	$0.71 + / - 0.57$	$0.14 + / - 0.18$
DeepDSC	$0.23 + /0.02$	$0.78 + / - 0.04$	$0.28 + / - 0.08$	$0.73 + / - 0.12$	$0.61 + / - 0.64$	$0.05 + / - 0.06$

normal distribution to generate y-axis. In both cases, the RMSE values keep unchanged, while R^2 value are different dramatically.

3.2 Comparison on CCLE Dataset

We compared DeepDSC with the state-of-the-art study [18] on CCLE dataset, which was termed as RF. For comparison, the response data were already converted to the same unit, i.e., $-\log_{10} IC_{50}(\mu M)$. Cross-validation results show that DeepDSC outperforms RF with lower RMSE value of 0.23 and higher R^2 value of 0.78. LOTO and LOCO experiments were also performed. Table 3 shows the results of DeepDSC and RF on CCLE datasets.

As we illustrated above, the R^2 value are less important than the RMSE value when the data range considered is small. The LOTO and LOCO experiment results show that DeepDSC outperforms RF in predicting drug sensitivity data involving novel cell lines or compounds.

Although the prediction errors of DeepDSC are smaller than the previous studies in the area, it does have the limitation of training on a merged dataset for the data inconsistency among GDSC, CCLE and high-throughput screening datasets [41]. Future studies may achieve better results if this problem is solved. More experimental data obtained would speed up the drug sensitivity prediction.

4 Conclusion

In this study, we proposed DeepDSC, a deep learning model, to predict drug sensitivity of cancer cell lines. The model was trained on two available pharmacologic datasets, GDSC and CCLE. We chose gene expression data as cell line genomic features and Morgan fingerprints as chemical features of compounds. Both cell line features and compound features were combined as the inputs of DeepDSC. As a result, DeepDSC outperformed state of the art methods with the lowest prediction errors and high coefficient determination. The 10-fold cross-validation results showed that DeepDSC had the best interpolation ability to fill in missing drug sensitivity values, and the LOTO and LOCO results also showed that DeepDSC had very low extrapolation errors. These results indicated DeepDSC can benefit clinical cancer therapy and future study on drug sensitivity prediction.

Acknowledgments

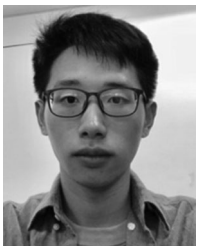
The authors would like to thank Isidro Cortés Ciriano at the Department of Biomedical Informatics, Harvard Medical School for providing the research data and discussing with us during research. This work is supported by the National Natural Science Foundation of China under Grant No. 61832019, No. 61622213, No.61772552, the 111 Project (No. B18059), the Hunan Provincial Science and Technology Program (2018WK4001), and the Fundamental Research Funds for the Central Universites of Central South University under grant No. 2018zzts560 and No. 2018zzts028. Parts of this paper appeared *in the Proceedings of the 2018 International Conference on Intelligent Computing* (ICIC2018) [51].

References

[1] J. N. Weinstein, "Drug discovery: Cell lines battle cancer," *Nature*, vol. 483, no. 7391, pp. 544–545, Mar. 28, 2012.
[2] J. P. Gillet, A. M. Calcagno, S. Varma, et al., "Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance," *Proc. Nat. Acad. Sci. United States America*, vol. 108, no. 46, pp. 18708–18713, Nov. 15, 2011.
[3] W. D. Stein, T. Litman, T. Fojo, et al., "A serial analysis of gene expression (SAGE) database analysis of chemosensitivity: Comparing solid tumors with cell lines and comparing solid tumors from different tissue origins," *Cancer Res.*, vol. 64, no. 8, pp. 2805–2816, 2004.

[4] R. H. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Rev. Cancer*, vol. 6, no. 10, pp. 813–823, Oct. 2006.
[5] J. Barretina, G. Caponigro, N. Stransky, et al., "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 28, 2012.
[6] M. J. Garnett, E. J. Edelman, S. J. Heidorn, et al., "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, Mar. 28, 2012.
[7] J. L. Wilding and W. F. Bodmer, "Cancer cell lines for drug discovery and development," *Cancer Res.*, vol. 74, no. 9, pp. 2377–2384, May 1, 2014.
[8] M. Imielinski, A. H. Berger, P. S. Hammerman, et al., "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing," *Cell*, vol. 150, no. 6, pp. 1107–1120, 2012.
[9] F. W. Huang, E. Hodis, M. J. Xu, et al., "Highly recurrent TERT promoter mutations in human melanoma," *Sci.*, vol. 339, pp. 957–959, 2013.
[10] G. Jin and S. T. Wong, "Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines," *Drug Discovery Today*, vol. 19, no. 5, pp. 637–644, May 2014.
[11] F. Cheng, H. Hong, S. Yang, et al., "Individualized network-based drug repositioning infrastructure for precision oncology in the panomics era," *Brief Bioinf.*, vol. 18, no. 4, pp. 682–697, Jul. 1, 2017.
[12] A. Bender and R. C. Glen, "Molecular similarity: A key technique in molecular informatics," *Organic Biomolecular Chemistry*, vol. 2, no. 22, pp. 3204–3218, 2004.
[13] A. Cherkasov, E. N. Muratov, D. Fourches, et al., "QSAR modeling: Where have you been? Where are you going to?," *J. Med. Chemistry*, vol. 57, no. 12, pp. 4977–5010, Jun. 26, 2014.
[14] I. Cortes-Ciriano, L. H. Mervin, and A. Bender, "Current trends in drug sensitivity prediction," *Current Pharmaceutical Des.*, vol. 22, no. 46, pp. 6918–6927, 2016.
[15] M. P. Menden, F. Iorio, M. Garnett, et al., "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *Plos One*, vol. 8, no. 4, 2013, Art. no. e61318.
[16] M. Ammad-ud-din, E. Georgii, M. Gonen, et al., "Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization," *J. Chemical Inf. Model.*, vol. 54, no. 8, pp. 2347–2359, Aug. 25, 2014.
[17] N. Zhang, H. Wang, Y. Fang, et al., "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *Plos Comput. Biol.*, vol. 11, no. 9, 2015, Art. no. e1004498.
[18] I. Cortes-Ciriano, G. J. van Westen, G. Bouvier, et al., "Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel," *Bioinf.*, vol. 32, no. 1, pp. 85–95, Jan. 1, 2016.
[19] L. Wang, X. Li, L. Zhang, et al., "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization," *BMC Cancer*, vol. 17, no. 1, Aug. 2, 2017, Art. no. 513.
[20] E. E. Bolton, Y. Wang, P. A. Thiessen, et al., "PubChem: Integrated platform of small molecules and biological activities," *Annual Reports In Computational Chemistry*, Amsterdam, The Netherlands: Elsevier, 2008pp. 217–241.
[21] D. S. Murrell, I. Cortes-Ciriano, G. J. P. van Westen, et al., "Chemically aware model builder (camb): An R package for property and bioactivity modelling of small molecules," *J. Cheminform.*, vol. 7, 2015, Art. no. 45.
[22] C. Brooksbank, M. T. Bergman, R. Apweiler, et al., "The european bioinformatics institute's data resources 2014," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D18–D25, 2013.
[23] B. S. Carvalho and R. A. Irizarry, "A framework for oligonucleotide microarray preprocessing," *Bioinf.*, vol. 26, no. 19, pp. 2363–2367, 2010.
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015, Art. no. 436.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[26] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[27] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 6645–6649.
[28] G. Hinton, L. Deng, D. Yu, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[29] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.
[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[31] M. Li, Z. Fei, M. Zeng, et al., "Automated ICD-9 coding via a deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2018.2817488.
[32] M. Zeng, M. Li, Z. Fei, et al., "A deep learning framework for identifying essential proteins by integrating multiple types of biological information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2897679.
[33] F. Zhang, H. Song, M. Zeng, et al., "DeepFunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions," Proteomics., to be published, doi: 10.1002/PMIC.201900019.
[34] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transfer Learn.*, 2012, pp. 37–49.
[35] N. An, W. Zhao, J. Wang, et al., "Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting," *Energy*, vol. 49, pp. 279–288, 2013.
[36] K. Bhaskar and S. Singh, "AWNN-assisted wind power forecasting using feed-forward neural network," *IEEE Trans. Sustain. Energy*, vol. 3, no. 2, pp. 306–315, Apr. 2012.
[37] D. Tran and Y. K. Tan, "Sensorless illumination control of a networked LED-lighting system using feedforward neural network," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2113–2121, Apr. 2014.
[38] H. Luo, J. Wang, M. Li, et al., "Computational drug repositioning with random walk on a heterogeneous network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2018.2832078.
[39] H. Luo, M. Li, S. Wang, et al., "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinf.*, vol. 34, no. 11, pp. 1904–1912, 2018.
[40] I. S. Jang, E. C. Neto, J. Guinney, et al., "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," in *Proc. Pac Symp Biocomput.*, 2014, pp. 63–74.
[41] B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, et al., "Inconsistency in large pharmacogenomic studies," *Nature*, vol. 504, no. 7480, 2013, Art. no. 389.
[42] G. Klambauer, T. Unterthiner, A. Mayr, et al., "Self-normalizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst*, 2017, pp. 972–981.
[43] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989.
[44] P.-T. De Boer, D. P. Kroese, S. Mannor, et al., "A tutorial on the cross-entropy method," *Ann. Operations Res.*, vol. 134, no. 1, pp. 19–67, 2005.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv.org>cs>arXiv:1412.6980, 2014.
[46] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," arXiv preprint arXiv:1511.07289, 2015.
[47] K. He, X. Zhang, S. Ren, et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[48] N. Srivastava, G. Hinton, A. Krizhevsky, et al., "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[49] F. Chollet, "Keras: Deep learning library for theano and tensorflow," *Io/K*, vol. 7, 2015, Art. no. 8, https://github.com/fchollet/keras
[50] A. Golbraikh and A. Tropsha, "Beware of q2!," *J. Mol. Graph. Modell.*, vol. 20, no. 4, pp. 269–276, 2002.
[51] Y. Wang, M. Li, R. Zheng, et al., "Using deep neural network to predict drug sensitivity of cancer cell lines," in *Proc. Int. Conf. Intell. Comput.*, 2018, pp. 223–226.

Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently the vice dean and a professor at the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her research interests include computational biology, systems biology, and bioinformatics. She has published more than 80 technical papers in refereed journals such as *Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Proteomics*, and conference proceedings such as BIBM, GIW, and ISBRA.



Yake Wang is working toward the master's degree in computer science from Central South University. His main research interests include bioinformatics and systems biology.

Ruiqing Zheng received the BSc and MSc degrees from Central South University, China, in 2013 and 2016, respectively. He is currently working toward the PhD degree in bioinformatics at Central South University. His main research interests include bioinformatics and systems biology.

Xinghua Shi received the BEng and MEng degrees in computer science from the Beijing Institute of Technology, China, and the MS and PhD degrees in computer science from the University of Chicago. She is an assistant professor in the Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte. Before joining UNC Charlotte, she was a postdoctoral research fellow with Brigham and Women's Hospital and Harvard Medical School, an NIH T32 medical genetics training fellow at Harvard Medical School, a visiting research fellow in the Medical and Population Genetics program at the Broad Institute, and an associate in the Quantitative Genetics Program at Harvard School of Public Health. Her research interest include bioinformatics and computational systems biology. Particularly, she works on the design and development of tools and algorithms to solve large-scale computational problems in biology and biomedical research. She is currently focused on integrating genetic and epigenetic datasets to study how genetic architecture affects biological processes and complex phenotypes at the systems level. She is also interested in genetic privacy, complex network analysis, and big data analytics in biomedical research. Her work is supported by multiple agencies and foundations including Wells Fargo Foundation Fund, NSF, NIH, and DARPA.

Yaohang Li received the BS degree in computer science and engineering from the South China University of Technology, in 1997, and the MS and PhD degrees in computer science from Florida State University, Tallahassee, FL, USA, in 2000 and 2003, respectively. He is an associate professor of computer science at Old Dominion University, Norfolk, VA, USA. His research interests are in protein structure modeling, computational biology, bioinformatics, Monte Carlo methods, big data algorithms, and parallel and distributive computing. After graduation, he worked at the Oak Ridge National Laboratory as a postdoctoral researcher for a short period in 2003. He was a summer research fellow at the National Center of Supercomputing Applications (NCSA), in 2007 and a Summer Faculty Research Participation member at the Oak Ridge National Laboratory in 2006 and 2008, respectively. He is the author of more than 70 papers in international journals and refereed conference proceedings. He is the program committee co-chair of the 2015 International Symposium on Bioinformatics research and Applications (ISBRA2015). He also serves on the editorial boards of the *International Journal of Computational Mathematics* and *Computational Biology Journal*. He received the best poster awards at ISBRA2015 and the best paper awards at Modeling, Simulation, and Visualization Capstone Conferences in 2014 and 2015, respectively. He is the recipient of the Ralph E. Powe Award in 2005 and an NSF CAREER Award in 2009.

Fang-Xiang Wu (M'06-SM'11) received the BSc and MSc degrees in applied mathematics, both from the Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from North-western Polytechnical University, Xi'an, China, in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004–2005, he worked as a postdoctoral fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a professor in the Division of Biomedical Engineering and the Department of Mechanical Engineering at the U of S. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, applications of control theory to biological systems. He has published more than 260 technical papers in refereed journals and conference proceedings. He is serving as an editorial board member of three international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals. He is a senior member of the IEEE.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the dean and a professor with the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics, and computer network. He has published more than 150 papers in various international journals and refereed conferences. He is a senior member of the IEEE.