Dual-layer Strengthened Collaborative Topic Regression Modeling for Predicting Drug Sensitivity

Journal:	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
Manuscript ID	TCBB-2017-12-0469
Manuscript Type:	Regular Paper
Keywords:	drug response, collaborative regression, matrix decomposition, E.5.b Optimization < E.5 Files < E Data

Dual-layer Strengthened Collaborative Topic Regression Modeling for Predicting Drug Sensitivity

Hang Wang, Jianing Xi, *Student Member, IEEE,* Minghui Wang, *Member, IEEE,* and Ao Li, *Member, IEEE*

Abstract—An effective way to facilitate the development of modern oncology precision medicine is the systematical analysis of the known drug sensitivities that have emerged in recent years. Meanwhile, the screening of drug response in cancer cell lines provides an estimable genomic and pharmacological data towards high accuracy prediction. To classify or regress the drug response, existing works primarily utilize genomic or functional genomic features to classify or regress the drug response. Here in this work, by the migration and extension of the conventional merchandise recommendation method, we introduce an innovation model on accurate drug sensitivity prediction by using dual-layer strengthened collaborative topic regression (DS-CTR), which incorporates not only the graphic model to jointly learn drugs and cell lines feature from pharmacogenomics data but also drug and cell line similarity network model to strengthen the correlation of the prediction results. Using Genomics of Drug Sensitivity in Cancer project (GDSC) as benchmark datasets, our 5-fold cross-validation experiment demonstrates that DS-CTR model significantly improves drug response prediction performance compared with four categories of state-of-the-art algorithms as for both Receiver Operator Curve (ROC) and the Area Under Receiver Operator Curve (AUC $>$ 0.9). By uncovering the unknown cell-drug associations with advanced literature evidences, our novel model DS-CTR is validated and supported. The model also provides the possibility to make the discovery of new anti-cancer therapeutics in the preclinical trials cheaper and faster.

Index Terms—drug response, collaborative regression, matrix decomposition,optimization.

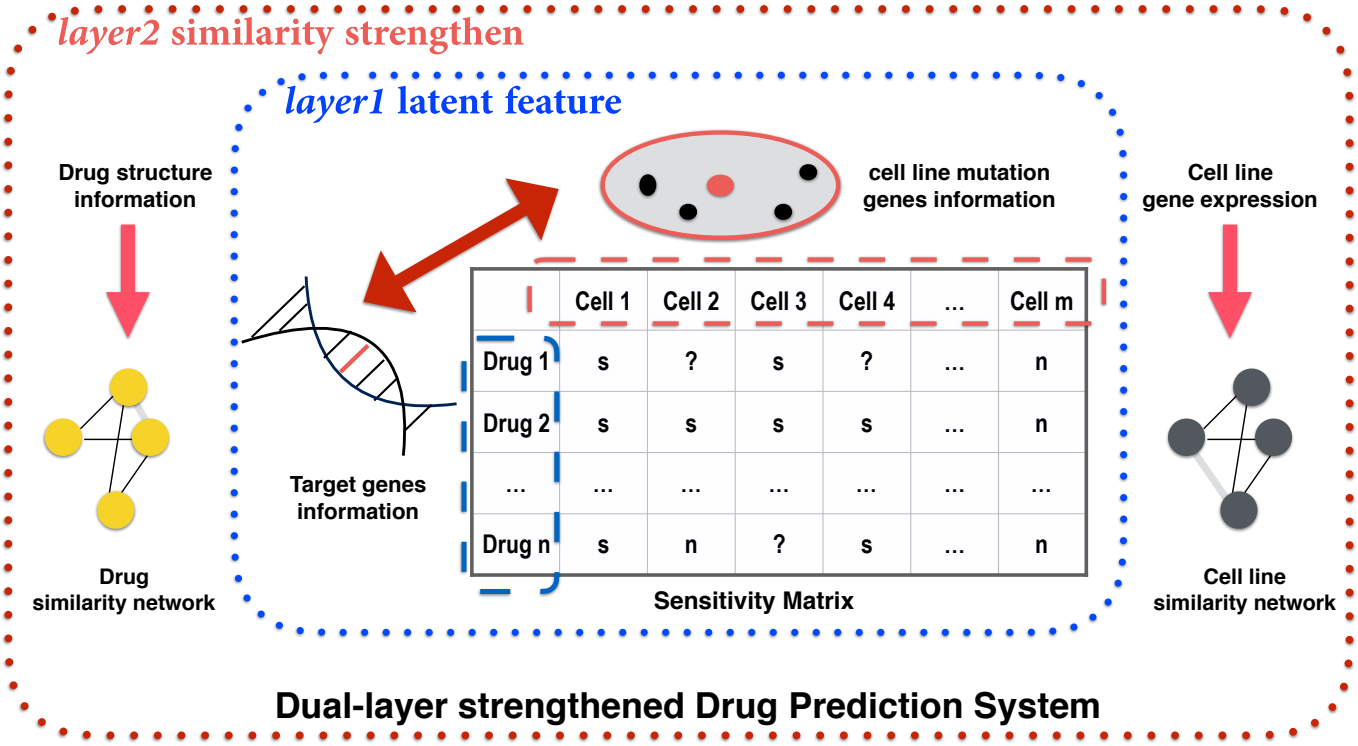
1 Introduction

Development or discovery of the new molecular entities with specified pharmacological properties is an expensive and high attrition rate process. For example, considering about the entities' toxicity or lack of efficacy which consequently brings the failure of many phase II or III clinical trials on the candidates, the cost greatly increases [1]. It's clear to us that even the single drug may act on multiple targets, and also on the other side, lots of targets are involved in multi-pathway. This fact motivates the practitioners' interest in drug repositioning which will largely reduce the cost and effort of developing new drugs [2][3]. Moreover, the goal of precision medicine is to achieve effective individualize treatment by selecting proper therapeutics referring to the given molecular profiles of a patient's tumor [4][5][6]. It eventually leads to the research on systemic computation techniques which are able to facilitates the drug development process by predicting novel drug-cell interactions based on the existing drugs and cell lines information for further experimental confirmation. Howbeit, initiation and progression of cancer involve multiple molecular mechanisms and also the diversity across cancer cells from the same patient or even across tumors from different patients makes the picture very complex [7]. Recent years, the molecular nature of cancer is widely studied by researchers which makes significant advances in this field, such as the successes in large scale high-throughput screening efforts. This kind of studies profile large amount of panels of human's cancer cell lines and drugs [8][9][3] and thus, open the door for researchers. The problem of predictive biomarker identification has been systematically addressed by analyzing the pharmacological profiles of clinical-relevant human cell lines and corresponding drugs in the work done by some consortiums, such as Cancer Cell Line Encyclopedia (CCLE) [9] and Cancer Genome Project (CGP) [10]. Meanwhile, for each cell line, the gene mutation status and expression profiles were included in their work [11]. Consequently, based on the advances, the establishment of high accuracy prediction system of the cancer cell response to medication which can systematic translate cancer genomic information to the tumor biology and therapeutic knowledge[12] will be easier and more reliable.

In order to provide the clue on deregulated mechanisms that would guide to specific treatment, various drug response prediction approaches have been exerted [13]. Recent advances demonstrate that when predicting drug response in certain type of cell lines, genomic and molecular features are necessary and useful [14]. Therefore, many

- *H. Wang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei AH 230027, China. E-mail: wanghang@mail.ustc.edu.cn.*
- *J. Xi is with the School of Information Science and Technology, University of Science and Technology of China, Hefei AH 230027, China. E-mail: xjn@mail.ustc.edu.cn.*
- *M. Wang is with the School of Information Science and Technology, and Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH 230037, China. E-mail: mhwang@ustc.edu.cn.*
- *A. Li is with the School of Information Science and Technology, and Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH 230037, China. E-mail: aoli@ustc.edu.cn.*

Manuscript received XXXX; revised XXXX.



	Cell 1	Cell 2	Cell 3	Cell 4	...	Cell m
Drug 1	s	?	s	?	...	n
Drug 2	s	s	s	s	...	n
...	...	...	...	...	...	...
Drug n	s	n	?	s	...	n

Fig. 1. The framework of drug response prediction model DS-CTR. In the first layer, we integrate the drugs' target genes and cell lines' mutation genes information to capture the semantic relationship of each drug-cell line pair. In the second layer, we utilize the gene's structure similarity network and cell lines' gene expression similarity information to strengthen the correlation of the prediction result.

works are on the basis of in vitro drug sensitivity and before treatment baseline gene expression levels in cells to predict the anti-cancer drug response [15][16][17][10]. Based on the dataset we mentioned before, Geeleher et al. focused on the breast cancer cell lines and in order to predict the drug response, applied a sparse regression model using baseline gene expression data [13]. The chemical properties of drugs were also being used by Menden et al. together with the genomic features of cell lines to represent the feature of each cell line-drug pair and then do the prediction [18]. In Ammad-ud-din et al.'s work, except for the genomic properties of cell lines, by systematically utilizing the known primary targets of the drugs (drug properties matrix) as prior knowledge which is another kind of pathway information of drugs, the prediction results were finally received through kernelized Bayesian matrix factorization [9][14]. Further on, based on the assumption that similarity network of drugs or cell lines could potentially improve the quality of the drug response prediction, a lot of work have been done by researchers. For example, in Zhang et al.'s work, they used cell similarity network and drug similarity network to predict drug response separately which indicates that the prediction was done twice. After that, the final prediction was obtained as the weighted average of the two network's result we got before [11]. Meanwhile, in the work done by Berlow et al., researchers present a novel approach for drug response prediction based on integrated function and genomic characterizations [19]. The inhibition profile of drugs is counted to calculate the drug similarity.

As shown in the previous paragraph, most of the approaches rely on genetic measurements alone, such as mutation genes and target genes [15][13][18]. Thus, two quite important features of the cancer cell-drug response screens are overlooked. One is that the cell lines or samples with similar genetic features might respond to the given drug quite similarly, and the other clue is that the similar therapeutic effects might appear if we use the structurally similar drugs when considering about the shared molecular structure or their targeting patterns [11]. Related studies demonstrate that the prediction based on genomic information alone usually result in low accuracy and limitations [19][20][21]. Similarly, in Zhang et al. work, even though the cell similarity network and drug similarity network are stressed to enforce the prediction result, the basic molecular and genomic features are missed which have already been proved useful for the prediction [14]. To sum up, a limited statistical strength should be stressed here because of the small sample size in most of drug response studies which makes the prediction task challenging and possibly make the prediction result less reliable. [2].

So as to overcome the aforementioned problems and limitations, firstly, it's necessary for us to build up an effictive and sufficient computational prediction framework to combine the available multi-view data sources, and at the same time, to improve the prediction performance and provide maximum reasonable information fundamentally. The underlying assumption of this statement is that, when predicting the response variable, a possible useful signal/feature for it will be contributed by all or some of the data sources [2]. For example, the similarity relationship between two

drugs will show up or down regulation of their sensitivity with a certain cell line. The statistical relationships may be revealed through the integration of the shared features from different sources, which may not be observed from the data itself but are strongly relevant to the high accuracy drug response prediction. Aim to effectively infer these shared features from different data sources, those kinds of computational methods are commonly referred to as multi-view learning [2][22].

In this paper, we proposed a novel collaborative filtering (CF) based approach to predict interactions with reliance on both genomic information and similarity network of drugs and cells. In this work, the novel proposed dual-layer strengthened topic regression model (DS-CTR) contains two layers to integrate two kinds of necessary information: ***Layer1.*** topic layer, ***Layer2.*** relationship layer. The interpretable feature structure for drugs and cells is given in the first layer by using traditional collaborative filtering and probabilistic topic modeling [23][24]. While the second layer reflects so called similarity relationship of drugs or cells and intensifies the correlation in recommendation results to strengthen the prediction performance. Note that different from Zhang et al.'s work (see before), both the two layers will update simultaneously by setting one graphic model and optimizing a cost function. By this way, our model takes both the genomic information and similarity information into account because it has been proved that both the information mentioned above comes into play and influence partly in the whole procedure of drug-cell interaction [14][15][16][17][10]. To indicate the model structure and data resources clearly, the framework of DS-CTR can be seen in **Fig. 1**, and the layer function of DS-CTR model will be detailed in the third section. Finally, in this study, we choose a real world dataset GDSC [25] to evaluate the superiority of the novel proposed DS-CTR model. And the result shows that our proposed model outperforms another four state-of-the-art drug response prediction systems in terms of both Receiver Operator Curve (ROC) and Area Under Receiver Operator Curve (AUC). With advanced literature evidences, by uncovering the unknown cell-drug associations, the performance of DS-CTR was also validated.

The reminder of the paper is structured as follows. **Section 2** details the rationale and learning process of the new proposed DS-CTR model. We present the real-world dataset based experiments and results in **Section 3** to evaluate the method. And finally, the conclusion and future work are given in **Section 4**.

2 Methods

In this section, we first give a brief review of the related techniques in our work, including matrix factorization, probabilistic topic models and collaborative filter method. After that, we discuss the modeling and learning process of DS-CTR in detail. Finally, the evaluation metric is given.

2.1 Notations

Suppose there are n drugs $U = \{u_1, u_2, \cdots, u_n\}$, m cells $V = \{v_1, v_2, \cdots, v_m\}$, p drugs target genes $P = \{p_1, p_2, \cdots, p_n\}$ and q cells mutation genes $Q = \{q_1, q_2, \cdots, q_n\}$, Let $U \in \mathbb{R}^{K\times n}$ and $V \in \mathbb{R}^{K\times m}$be the latent drug and cell feature matrices, with column vectors U_i and V_j representing the K-dimensional drug-specific and cell-specific latent feature vector of drug i and cell j, respectively. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ be the similarity network matrix of drugs and cells. The sensitivity of the given drugs to certain cells are given in a sensitivity matrix $R \in \mathbb{R}^{n\times m}$ in which R_{ij} denotes the sensitivity of drugs i with cell j. And here we define confidence parameter c_{ij} which is the precision parameter for R_{ij}:

$$c_{ij} = \begin{cases} a & r_{ij} = 1 \\ b & r_{ij} = 0 \end{cases} \quad (1)$$

Note that where a and b are turning parameter satisfying $a > b > 0$. If c_{ij} is larger, we trust R_{ij} more. The situation when $r_{ij} = 0$ can be interpreted into two ways: the drug i is either negative on cell line j or no existed work has been done on exploring the sensitivity of the pair. All in all, our work focuses on predicting novel drug-cell sensitivity pair based on the genomic and similarity information of cells and drugs (see **Fig.** 1).

2.2 Probabilistic topic models

Probabilistic graphic model originally aims to utilize thematic information to annotate and discovery large archives of documents [26]. Here in our research, we explore to utilize the principle of this model to serve our research purpose. Based on the fact that if two drugs have similar target genes, we probably classify them with the same type and have similar response with a given cell lines. And also, the same for the cell lines with similar mutation genes, they will interact with a certain drug simultaneously. To be more specific, by using the topic modeling algorithms, we are able to analyze the genomic information (include the mutation genes and target genes) of the cell lines and drugs to discover the so called 'themes' that run through them, how those themes are connected to each other, and how they change over. To achieve that, we divide the whole procedure of this model into two steps: first, to define a new drugs or cells, one chooses a distribution over so called 'themes', a probability distribution over genes is provided by the individually interpretable theme which partly represents the type of the cells or drugs, and the distribution picks out the coherent cluster of terms with certain correlation. And after that, for each gene in that target genes group or mutation genes group, one chooses a 'theme' at random according to the provided distribution, and draws a gene from that theme [27]. One of the most commonly used models to solve this problem is latent Dirichlet allocation(LDA). By using LDA, we mine the hidden theme or type of the cells or drugs and in this way, to correlate the drug-cell response pairs together. This method has been used for tasks like corpus exploration, document classification and information retrieval. The main generative process is as follows. We assume ω_j is the j-th cell or drug in the research library and meanwhile there are totally K types $\beta = \beta_{1:K}$ in the dataset:

1) Generate the cell or drug's type:

$$\theta_j \sim Dirichlet(\alpha).$$

2) Generate the n-th mutation or target gene in this cell or drug:

- Draw gene types assignment

$$z_{jn} \sim Mult(\theta_j).$$

- Draw gene $\omega_{jn} \sim Mult(\beta_{z_{jn}})$.

2.3 Matrix factorization

The basic idea of matrix factorization (MF) methods [28] is to factorize a matrix into two specific matrix and, for example, drug-specific matrix and cell-specific matrix as (2), where k denotes the feature dimension. Then the original matrix can be approximated by multiplying the two factorized matrix. Researchers always use regularization terms to avoid over-fitting as (3).

$$\hat{\boldsymbol{R}}_{\boldsymbol{n}\times\boldsymbol{m}} = \boldsymbol{U}^T_{n\times k} \times \boldsymbol{V}_{k\times m} \tag{2}$$

$$\arg\min \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij} - \boldsymbol{U}_i^{\boldsymbol{T}} V_j\right)^2 + \lambda_u \|\boldsymbol{U}\|_F^2 + \lambda_v \|\boldsymbol{V}\|_F^2 \tag{3}$$

where $\| \ \|_F^2$ denotes the Frobenius norm. The constant $\lambda_u$and $\lambda_v$controls the extent of regularization. Based on basic MF, similarity network matrix factorization(SMF) is exerted to improve prediction performance by incorporating relationship matrix [29][30]. While Chen et al.'s [31] mentioned that the conventional SMF assigns a single prior to all objects which neglect the different between them. In that case, in Chen et al.'s proposed CTR-SMF2 model [31], it considers similarity network information by assigning a different prior to each objects based on the social network [31]. As the similarity regulation term in (4) shows, these features allow the models to take full advantage of the network information:

$$min_{i=1}^{m}\sum_{j=1}^{m} T_{ij}\|U_i - U_j\|^2 \tag{4}$$

where T_{ij} represents the relationship value between user$_i$ and user$_j$ which is based on the users' influential contexts such as the users average rating [17]. While in medical field, luckily, the relationship among drugs or cell lines can be easily calculated by analyzing cell's gene expression profiles and drug's chemical structures [29] and therefore, multi-source information should also be stressed here.

2.4 Collaborative topic regression

Collaborative topic regression(CTR) is put forward by Wang et al. [23] to solve scientific articles recommendation problem. This method combines the merits of probabilistic topic modeling and the traditional collaborative filtering and this model has already been validated in real world multi-view information dataset [23][31][24]. In this model, we first draw each drug latent vector, then generate the cell lines latent vector by using LDA and finally we draw the sensitivity as the prediction for each given drug-cell pair. In CTR model, Wang et al. propose to use a variational expectation-maximization(EM) algorithm to learn the maximum a posterior(MAP) estimates. For more details about this algorithm

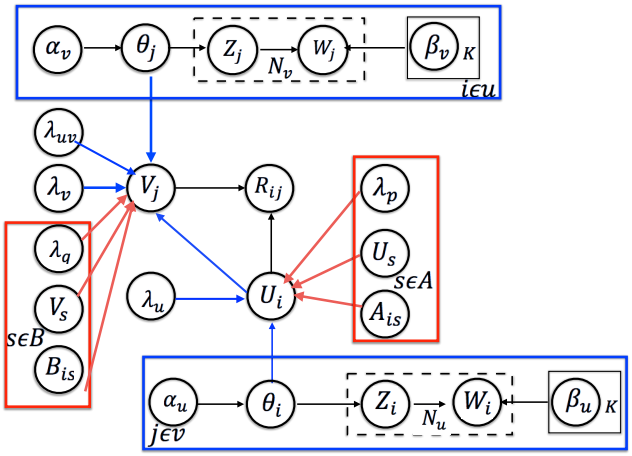


Fig. 2. Graphic model of DS-CTR, where the first layer is marked in blue and the second layer is marked in red.

please refer to [23]. To further improve the performance of CTR, similarity network need to be included, e.g. CTR-SMF [32] and CTR-SMF2 [31]. Wang et al. also proposed a model to consider about item similarity relationship into CTR. Based on the idea given by Chen et al.'s work [24], for our research purpose, in order to mine the semantic information between both drugs and cells embedded in genomic information and bridge feature of genes and sensitivity between drugs and cells [24]. In our model, we especially mine semantic information of genes for drugs and after that, we mine semantic information of genes for cells. Finally, we capture the implicit preference between genes and cells by using the following term (5):

$$\prod_i I_{ij}^R N(U_i,\ \lambda_{uv}^{-1} I_K) \tag{5}$$

2.5 DS-CTR: Dual-layer collaborative topic regression

Based on the analysis we talked above, our proposed model DS-CTR is a hierarchical Bayesian model that jointly learns the user and item latent spaces. Basically, the model is divided into two layers: *Layer1.* **topic layer**, *Layer2.* **relationship layer**. The first layer we group the genomic inormation for each drugs and cells respectively, hence, use LDA to mine the semantic information of genes for each drug and cell respectively [24] (plotted in red). While in the second layer we use two-side matrix factorization to handle sensitivity matrix and similarity network information to strengthen the prediction result (plotted in blue). And finally we incorporate the dual-layer information into matrix factorization to factorize sensitivity information. The two layer update simultaneously to optimize the set loss function by designed EM algorithm. The graph model can be seen in (2). In this research, two different kinds of similarity are used to strengthen the recommendation result. In this first layer(**Layer 1**), when a user and an item are linked by genomic information and sensitivity matrix,

their latent features are similar to each other to some extent [24]. Based on the fact that functional related drugs may have similar therapeutic effect [12] and the possibility to identify the patients through molecular biomarkers which can be shown from genomic changes [25], we incorporate relationship/similarity of drugs or cells into our system as the extra strengthened layer of the whole model by matrix factorization. Hence in the second layer(*Layer 2*), when a drug has close relationship (the similarity value is high) with another drug, their recommender result should be similar to each other, too. The same with the cells. Thus, DS-CTR can capture the semantic correlation more accurately than conventional drug response prediction model, which only bridges the drugs and cells by genomic information. Assuming there are *K* topics for both users and items, the generative process of DS-CTR works as follows:

1. Mining semantic information for drugs.

For each drug u_i

(a) Draw drug's type proportions

$$\theta_i \sim Dirichlet(\alpha_u);$$

(b) Draw drug's latent vector set, $U \sim p(U)$, where
Layer 1.

$$p(U) \propto N(\theta_i, \lambda_u^{-1} I_K)$$

Layer 2.

$$p(U) \propto \prod_{i=1}^{n} \prod_{s \in A_i} N(U_s, \lambda_p^{-1} A_{is}^{-1})$$

To sum up:

$$p(U) \propto N(\theta_i, \lambda_u^{-1} I_K) \prod_{i=1}^{n} \prod_{s \in A_i} N(U_s, \lambda_p^{-1} A_{is}^{-1}) \quad (6)$$

(c) For each target gene ω_{in_u} of drug u_i

i. Draw target gene's type assignment

$$z_{in_u} \sim Mult(\theta_i)$$

ii. Draw target gene as

$$\omega_{in_u} \sim Mult(\beta_{z_{in_u}});$$

2. Mining semantic information for cells, capture the implicit preference between drugs and cells.

For each cell v_j

(a) Draw cell's type proportions

$$\theta_j \sim Dirichlet(\alpha_v);$$

(b) Draw cell's latent vector set, $V \sim p(V)$, where
Layer 1.

$$p(V) \propto N(\theta_i, \lambda_v^{-1} I_K)$$

$$p(V) \propto \prod_i I_{ij}^R N(U_i, \lambda_{uv}^{-1} I_K)$$

Layer 2.

$$p(V) \propto \prod_{i=1}^{n} \prod_{s \in B_i} N(V_s, \lambda_q^{-1} B_{is}^{-1})$$

To sum up:

$$p(V) \propto N(\theta_i, \lambda_v^{-1} I_K) \prod_i I_{ij}^R N(U_i, \lambda_{uv}^{-1} I_K) \prod_{i=1}^{n} \prod_{s \in B_i} N(V_s, \lambda_q^{-1} B_{is}^{-1}) \quad (7)$$

(c) For each mutation gene ω_{jn_v} of the cell line v_j,

i. Draw mutation gene's type assignment

$$z_{jn_v} \sim Mult(\theta_j);$$

ii. Draw genes

$$\omega_{jn_v} \sim Mult(\beta_{z_{jn_v}});$$

3. Draw the rating.
For each drugs and cell line pair $(i,\ j)$, the sensitivity prediction

$$R_{ij} \sim N(U_i^T V_j,\ c_{ij}^{-1}). \quad (8)$$

In this above process procedure, we define I_K is the $K \times K$ identity matrix; I_{ij}^R denotes an indicator matrix with the value equal to 1 if drug u_i is sensitive to cell v_j, 0 otherwise; $N(x|\mu, \sigma^2)$ represents a Gassian distribution with a mean μ and a variance σ^2. And as we mentioned in the above section, matrix A and B represent the drugs and cell similarity network separately, e.g. B_{ij} denotes the similarity value between drug u_i and drug u_j.

In DS-CTR, we set parameter λ_u, λ_v to balance the contribution of drug and cell's semantic information provided by genomic and sensitivity information to the model performance [18]. Here we set λ_p, λ_q to balance the contribution of drug and cell information provided by similarity network (***L2***) to the model performance. The parameter λ_{uv} balances the contribution of implicit preference on model performance.

In order to use Bayesian inference method, we firstly obtain the probabilistic distribution of drugs and cells based on the above procedure:

$$p(U|\lambda_u,\ \lambda_p, U_s, A) \propto p(U|\lambda_u) \times \prod_{i=1}^{n} \prod_{s \in A_i} p(U|U_s,\ A_{is}^{-1}\lambda_p\ )$$

$$p(V|\lambda_v,\ \lambda_{uv}, \lambda_q, U, V_s, B\ ) \propto p(V|\lambda_v) \times \prod_i I_{\mathrm{ij}}^R p(U|U_i, \lambda_{uv}^{-1}) \times \prod_{i=1}^{n} \prod_{s \in B_i} p(V|V_s, \lambda_q^{-1} B_{is}^{-1})$$

Meantime, the conditional distribution of observed sensitivity can be formalized as

$$p\left(R|U,\ V, C\right) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\mathrm{N}\left(R_{ij} \middle| U_i^T V_j, \sigma_R^2\right)\right]^{I_{ij}^R}$$

Finally, we use Bayesian inference, we are able to get the following equation for the posterior probability of latent feature vectors given the sensitivity matrix, two-side similarity network and two-side genomic information:

$$\begin{aligned} &p\left(U,V|R,\ A,\ B,C,\ \lambda_u,\ \lambda_v,\ \lambda_p,\ \lambda_q,\lambda_{\text{uv}}\right) \\ &\propto p(R|U,V,C)\times p(U|\lambda_u,\ \lambda_p,U_s,A) \\ &\times\ p(V|\lambda_v,\ \lambda_{\text{uv}},\lambda_q,U,V_s,B\ ) \end{aligned} \tag{9}$$

2.6 Parameter learning of DS-CTR

It's intractable to compute the full posterior of drug and cell latent vector U_i and V_j based on the given parameter and information matrix. In that case, we develop a coordinate ascent algorithm to learn the MAP estimates [23][31][24]. We firstly fix the hyper-parameter, and then maximizing the posterior over the two latent vectors, which is equivalent to maximizing the following complete log likelihood of U, V, A, B, $\theta_{1:n}$, $\theta_{1:m}$ and R, given λ_u, λ_v, λ_p, $\lambda_q$and λ_{uv}:

$$\begin{aligned} L=&-\frac{\lambda_u}{2}\sum_i\left(U_i-\theta_i\right)^T\left(U_i-\theta_i\right)-\frac{\lambda_v}{2}\sum_j\left(V_j-\theta_j\right)^T\left(V_j-\theta_j\right) \\ &-\frac{\lambda_{uv}}{2}\sum_{ij}\left(U_i-V_j\right)^T\left(U_i-V_j\right)-\frac{\lambda_p}{2}\sum_{s\in A}A_{is}\left\|U_i-U_s\right\|_F^2 \\ &-\frac{\lambda_q}{2}\sum_{r\in B}B_{ir}\left\|V_i-V_r\right\|_F^2-\sum_{ij}\frac{c_{ij}}{2}\left(R_{ij}-U_i^TV_j\right)^2 \\ &+\sum_i\sum_{n_u}\log\left(\sum_k\theta_{ik}\beta_{k,\ \omega_{in_u}}\right) \\ &+\ \sum_j\sum_{n_v}\log\left(\sum_k\theta_{jk}\beta_{k,\ \omega_{jn_v}}\right) \end{aligned} \tag{10}$$

The first three items of the function represents the first layer of the model while the next two items represents the second layer. Here the Dirichlet prior is set to 1. We optimize the MF variable using coordinate descent which is iteratively the MF variables U_i, V_j and the topic proportions θ_i, θ_j with the fixed hyper-parameter. In order to do that, firstly we update U_i, V_j, given the current estimate of θ_i, θ_j. We calculate the gradient of the loss function with respect U_i, V_j and set it to zero. Finally, the equation leads to:

$$\begin{aligned} U_i\leftarrow&\left(\lambda_uI_K+VC_iV^T+\lambda_p\sum_jA_{ij}I_K+\lambda_{uv}\sum_jI_{ij}^RI_K\right)^T \\ &\left(\lambda_u\theta_i+VC_iR_i+\lambda_p\sum_jA_{ij}U_j+\lambda_{uv}\sum_jI_{ij}^RV_j\right) \end{aligned} \tag{11}$$

$$\begin{aligned} V_j\leftarrow&\left(\lambda_vI_K+UC_iU^T+\lambda_q\sum_iB_{ji}I_K+\lambda_{uv}\sum_jI_{ij}^RI_K\right)^T \\ &\left(\lambda_v\theta_j+UC_jR_j+\lambda_q\sum_iB_{ji}V_i+\lambda_{uv}\sum_iI_{ij}^RU_i\right) \end{aligned} \tag{12}$$

where C_i is a diagonal matrix with $c_{ij}(j=1,\ \cdots,m)$ as its diagonal elements and $R_i=(r_{ij})_{j=1}^m$ for drug i. For user j, C_j, R_j are similar defined. The above equation shows how parameter λ_u, λ_v, λ_p, $\lambda_q$and λ_{uv} influence both the drug and cell latent feature. Based on the two-side information, we are able to get the aforementioned symmetric update equation of U_i and V_j. In our model, we use five parameters to balance the proportion of different information. For example, when parameter λ_p is bigger, the proportion of drug's similarity network information becomes larger. Especially, when $\lambda_p=\ \lambda_q=0$, DS-CTR model collapses to TRCF [18]. Hence, if $\lambda_{uv}=0$, the model collapses to CTR-SMF2 [31]. When set all the parameters to 0, the model will collapse to matrix fatorization.

After getting feature matrix U and V for drugs and cell lines simultaneously, we now learn the topic proportions θ_i, θ_j. The procedure is as the same as TRCF model [18]. For θ_i, we first define $q\left(z_{in_u}=k\right)=\Phi_{in_uk}$, and then separate the drugs that contain θ_i and apply Jensen's inequality [23],

$$\begin{aligned} L\left(\theta_i\right)\geq&-\frac{\lambda_u}{2}\left(U_i-\theta_i\right)^T\left(U_i-\theta_i\right) \\ &+\sum_{n_u}\sum_k\Phi_{in_uk}\left(\log\theta_{ik}\beta_{k,\omega_{in_u}}-\log\Phi_{in_uk}\right) \\ =&\ L(\theta_i,\phi_i) \end{aligned} \tag{13}$$

Note that $\Phi_i=(\Phi_{in_uk})_{n_u=1,k=1}^{N_u\times K}$, N_u is the target gene's number in the content of drug i, and the Φ_{in_uk} satisfies $\Phi_{in_uk}\propto\theta_{ik}\beta_{k,\omega_{in_u}}$. Then projection gradient [33] is utilized to optimize θ_i. For θ_j, we are able to use the same method to update.

In other hand, we update *beta* by using the method mentioned in [33]:

$$\beta_{kw_i}\propto\sum_i\sum_{n_u}\Phi_{in_u}k1[\omega_{in_u}=\omega]$$

After the optimization of parameters include U^*, V^*, $\theta_{1:n}{}^*$, $\theta_{1:m}^*$, β_u^* and β_v^*, we are able to draw the predict sensitivities as:

$$R_{ij}^*\approx\left(U_i^*\right)^TV_j^*.$$

The whole learning procedure is summarized in **Algorithm. 1** as a simple example.

2.7 Evaluation metric

Since our model aims to predict the novel association between drug and cell line, in our work, we follow related study [12] and choose to use ROC and AUC as metrics to measure the performance of the models. In medical decision making, ROC graphs are commonly used. And in recent years it has been widely used in other fields like data mining and machine learning research [34]. The Receiver Operating Characteristics graphs aims to organize classifier and realize the visualization of their performance. When compared with Precision-Recall curve (PRC), in class distribution, ROC are insensitive to changes, e.g. the ROC won't change even if the proportion of positive to negative instances changes in the test set, while on the contrary, for PRC, will change a lot [34]. In that case, we set the two metrics to evaluate the model

Algorithm 1: Learning algorithm of DS-CTR

Input: Sensitivity matrix **R**, drugs' target gene group T_1, cell lines' mutation genes group T_2, drug similarity network matrix A, cell lines similarity network matrix B, regularization parameter λ_u, λ_v, λ_p, $\lambda_q$and λ_{uv}
Output: drug and cell feature matrix U and V
Generate tag topic proportions θ_u and θ_v, vocabulary distributions β_u and β_v for drugs and cells separately by using LDA;
repeat
  Initialize U, V
  Update drug latent matrix U
  Update cell latent matrix V
  Update parameter θ_u and θ_v
  Update parameter β_u and β_v
  Calculate the likelihood
until *Convergence or the number of iterations more than preset threshold;*

performance. The *x*-axis and *y*-axis of ROC are specificity (FP) and sensitivity (TP) with the definition as follows:

$$\begin{aligned} \text{specificity} &= \frac{\text{FP}}{FP+TN} \\ \text{sensitivity} &= \frac{\text{TP}}{FN+TP} \end{aligned} \tag{14}$$

3 Experiment and Evaluation

In this section, we describe the real-life medical dataset we used in our work to evaluate the performance of our proposed model. The experiments are conducted by using Matlab (MathWorks Inc., Natick, MA), and tested on machines with Linux OS, Intel(R) Core(R) CPU 1.4 GHz.

3.1 Experimental setup

In our experiment, we chose to use Genomics of Drug Sensitivity in Cancer (GDSC, July, 2016 version) project dataset. For the purpose of therapeutic biomarker discovery and further preclinical validation in the cancer cells, the GDSC database provides a large amount of drug response information and related genomic datasets to identify putative therapeutic biomarkers [25]. 1001 cancer cell lines and 265 anticancer drugs were included in this version and more detail of the dataset is available at [35]. Drug response results are recorded as log IC_{50} values [15]. Tissue type can also be obsearved as the annotation of the cell lines. And drugs were annotated with their primary therapeutic targets [14]. Also, the cell lines' gene expression data is also provided in Iorio et al.'s work. In GDSC dataset, researchers generate the sensitivity matrix based on the following principle [25][35]:
for drug ***i*** and cell line ***j***, we define sensitivity r_{ij}

$$r_{ij} = \begin{cases} 1 & if\ drug_i\ is\ sensitive\ to\ cell_j \\ 0 & else \end{cases} \tag{15}$$

Then, the chemical structure files were downloaded from PubChem [2]. Finally, we are able to calculate the drug similarity based on the drugs structure and the cell line similarity based on the gene expression follow Zhang et al.'s work in [10]. Those two kinds of information are analogized as strengthened layer information of drugs and cells. We generate the heatmap of the aforementioned information in **Fig.** (3). And the data detailed information is described in **Table** 1.

TABLE 1
Overview of the data source

Data source	Drug	Target gene	Drug similarity
Drug	265	125571	65536
Data source	Cell line	Mutation gene	Cell similarity
Cell	1001	127035	1002001

In our experiments, to evaluate the methods' performance, we choose to use 5-fold cross-validation. Hence, we split each dataset into three parts: 80% for training dataset, 10% for a held-out validation dataset and 10% for the test dataset. We train our model on the training dataset and then use validation dataset to find optimal parameters in the model. Finally, we use test data to evaluate the model.

3.2 Comparison with collapsed DS-CTR model

Since the core structure of our proposed model is collaborative filtering based regression, in this subsection, depends on the different types of information from drugs or cells, the proposed model DS-CTR is compared with the following three collapsed recommendation methods. When considering about the cell line genomic information only, we apply conventional collaborative topic regression [23] model on our dataset (mark as R+G). Hence, we add the similarity information of one side (drug similarity) and assign a different prior to each drugs by utilizing an improvement version of collaborative topic regression[31], which incorporates both the collaborative topic regression and social matrix factorization(mark as R+G+D). Moreover, we are able to draw the influence of two-side (cell lines and drugs) genomic information and semantic relationship between them through tag and rating based collaborative filtering model[24], which is marked as R+(GD) in **Fig. 4**. Finally our proposed model is marked as R+(GD)+C to emphasize the extra cell line similarity network we added.

On the other hand, during the comparison, we use grid search to find the best value for all the hyper parameters for all the four models. For the topic based model (CTR, CTR-SMF2, TRCF), we arbitrarily set the topic number *K* as 50 to generate the topic distribution by using LDA. Separately, as for CTR, we set the confidence parameters *a*=1 and *b*=0.01, also we set $\lambda_u = 0.01$, $\lambda_v = 10$. As for CTR-SMF2, we set the confidence parameters *a*=1 and *b*=0.1, also we set $\lambda_u = 1$, $\lambda_v = 1$, $\lambda_f = 1$. As for TRCF, we set the confidence parameters *a*=1 and *b*=0.1, also we set $\lambda_u = 10$, $\lambda_v = 1$, $\lambda_{uv} = 0.01$. As for DS-CTR, we set the confidence parameters *a*=1 and *b*=0.1, and we set $\lambda_u = 10$, $\lambda_v = 1$, $\lambda_{uv} = 0.0001$, $\lambda_p = \lambda_q = 0.001$.

The results of five-fold cross-validation ROC and AUC comparison are given in **Fig. 4,** through which we have the following observations: 1) Recommender model that

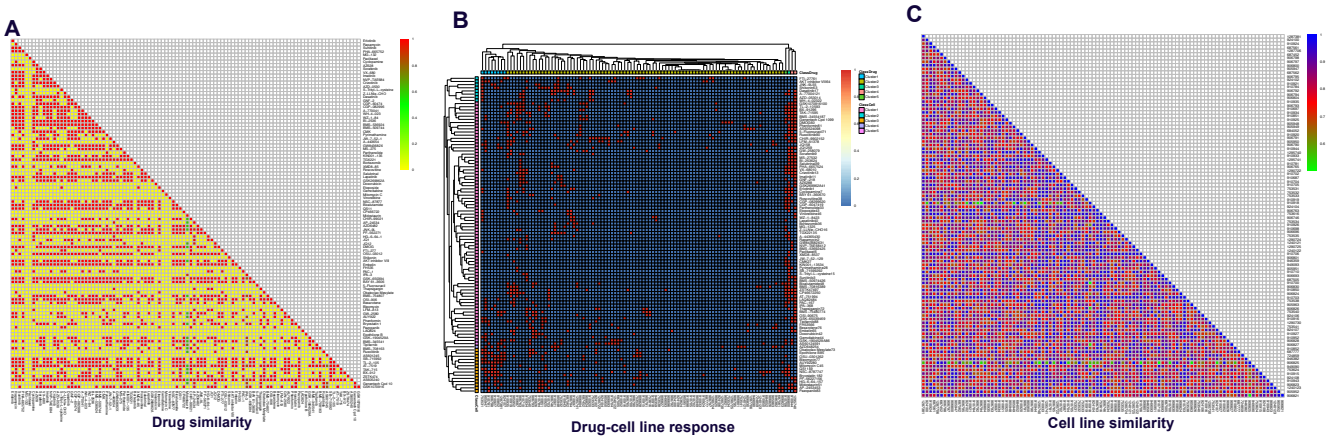


Fig. 3. Heterogeneous source information we used in DS-CTR. (A)The heatmap of drug's similarity based on the structure information. (B)The heatmap of sensitivity matrix based on IC50 value provided by GDSC project. (C)The heatmap of cell line's similarity based on the gene expression information. Note that only the first 100 drugs and cells information are chosen here.

consider extra information (similarity, genomic) significantly improve the prediction performance in terms of AUC and ROC. Especially the similarity network based CF (e.g. R+D+G) and two-side genomic information based CF (e.g. R+(DG)) have the better performance than conventional collaborative topic regression model, with a 7% and 15% higher evaluation metric value. 2) The proposed method, DS-CTR, which consider the both-side similarity network and genomic information from cells and drugs have the best performance. When compared with R+(DG) and R+G, our proposed method achieves 24% and 33% improvement on GDSC dataset. 3) According to ROC, with the same True Positive rate, our proposed model approaches the lowest False Positive rate compared with another three methods which denotes our model provides the most accurate prediction result among all. It can be seen more clearly from **Fig. 4** where we set the fixed specificity value as 0.1 follow related work and then compare the correspondent sensitivity value. The results show that our proposed method provide the better efficiency on realizing a chosen positive example ranks higher than a chosen negative example [34].

3.3 Performance comparison with existed drug prediction methods

A variant of drug response prediction methods have been put forward by researchers [14][15][16][17][10][13][18] in recent years. In this subsection, the comparison were done between our proposed model and the four existing methods, all the four state-of-the-art analysis methods considers more than only genomic information. In Muhammad Ammad-uddin et al.'s work [15], in order to accomplish the prediction task, researchers present a kernelized Bayesian Matrix Factorization method (cwKBMF). This method integrates genomic information from both drugs and cells for prediction of drug responses and encode the similarity network between samples in the side-data views. The similarity network including Drug Similarity Network(DSN) and Cell line Similarity Network(CSN) is emphasized first in Zhang et al.'s work [11]. As we mentioned in the above section, after integrating the two networks(C+D), the prediction is made. The comparison is done among the above four

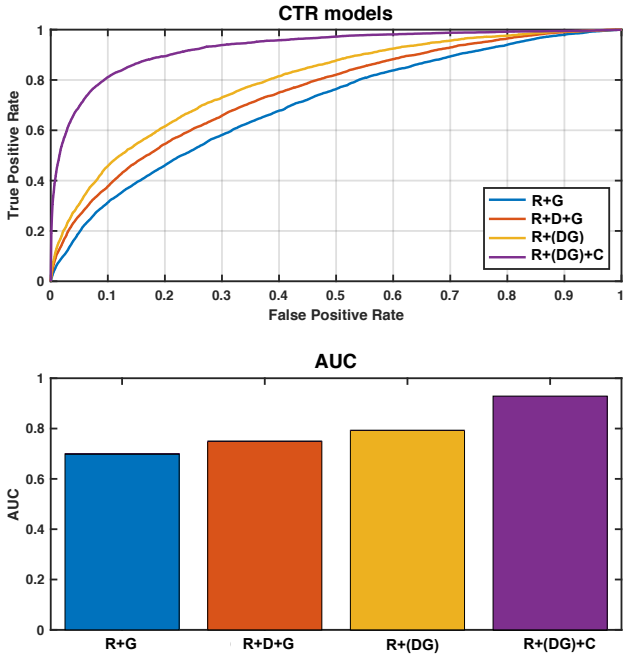


Fig. 4. AUC and ROC comparison result among CF-based topic models. Method abbreviation: R+G, collaborative topic regression with the drug-cell response information and the genomic information from cell lines; R+D+G, collaborative topic regression with extra drug structure similarity network; R+(DG), an improvement model of R+D+G, where the drug-cell response pair is connected with each other by the two-side genomic information and the drug similarity is counted in this model; R+(DG)+C, our proposed dual-layer strengthened collaborative topic regression which is an extension of R+(DG), the extra cell line similarity information is counted. The upper figure is the ROC comparison and the lower figure is the AUC value in bar format.

methods(CSN, DSN, C+D, cwKBMF and proposed DS-CTR) in **Fig. 5**. Note that when we utilize the CSN, DSN, C+D and cwKBMF model, we apply 90% for training and 10% for testing, while for DS-CTR, we split the dataset as the same as we mentioned before. All the aforementioned methods are configured as the default settings.

We first analyze the results of five-fold cross-validation in terms of ROC. Obviously, higher sensitivity and speci-

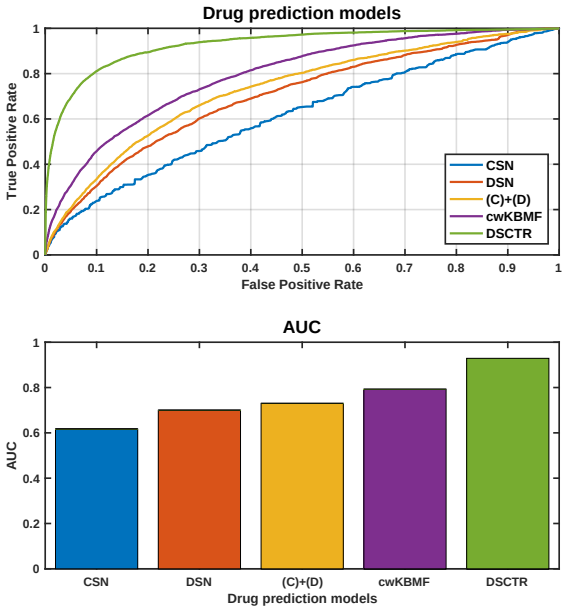


Fig. 5. AUC and ROC comparison result among state-of-the-art prediction models. Method abbreviation: CSN, cell line similarity network; DSN, drug similarity network; (C)+(D), combine both side similarity information to predict the response; cwKBMF, kernelized Bayesian matrix factorization method; DSCTR, our proposed dual-layer strengthened collaborative topic regression. The upper figure is the ROC comparison and the lower figure is the AUC value in bar format.

ficity appears when the curve is nearer to the left and top borders. Especially from **Fig. 5** when we fix the specificity as 0.1, our proposed method owns the highest sensitivity among all the methods. Overview, from **Fig. 5**, our proposed method notably outperforms (AUC=0.9289) the other methods. To be more specific, when compared among the CSN, DSN and C+D method, we found that DSN-based predictions were better (17%) than that CSN-based alone. In addition, more information integrated gave superior performance to either CSN (20%) or DSN (5%) which indicates again that multi-view sources are helpful in the prediction [51]. At the same time, the cwKBMF which considers the genomic similarity of drugs and cells, performs better than Zhang et al.'s method (5%). Our proposed method, in terms of AUC, is 49%, 31% and 26% higher than the above four methods separately which denotes the excellent priority of the proposed CF-based model when combining the multi-view information.

3.4 Case study

By utilizing DS-CTR model, we firstly generate the prediction result and compare it with the known response relationship provided by GDSC data set. The top-20 predicted result is given in **Table 2**. The sensitivity column in this table demonstrates the experiment result provided by the dataset. And the index value in the third column denotes a specific cell lines which is defined by GDSC project. For more information about the index, please refer to [25] and [35]. We are able to see it clearly that all the top-20 prediction result is agreed with the orignal dataset. Meanwhile, the top-200 prediction result is also provided in the supplementary materia (**Table S1**). Both of them are indicative of the superior performance of our model.

The proposed method was also able to be supported by its effectiveness in discovering novel drug response associations. In order to validate our model's functional performance, we also generate the top-10 novel predicted drug-cell line pairs which is not included in the GDSC dataset and then search for related advanced research evidence or database. The result is summarized in **Table 3**, where if the sensitivity of the pair has not been given in GDSC dataset but has been observed by advanced research, we mark the pair in bold. For example, in our experiments, the sensitivity between drug AC220 and cell line EoL-1-cell is not given in GDSC dataset, while based on our recommendation result, it's 'likely' to obtain some relationship between them. In Reiter et al.'s work published on Nature in September, 2017 [36], researchers noticed that FLT3-D835Y or FLT3-ITD protein was still retained in the perinuclear ER and after addition of AC220, a cell membrane localization similar to FLT3-WT (Eol-1 is included) or FLT3-N676K was observed. For more details, please refer to [36]. This evidence partly supports that the association between AC220 and EOL-1-cell from the side view. Hence, the top-50 novel prediction pairs are included in our supplementary materia (**Table S2**). To sum up, the prediction results of our system are proved to be able to produce necessary biological predication for practical medical test, which is also the promising purpose of our work. The proposed method DS-CTR which combines two layers to strengthen the accuracy of the results is capable to make the development of novel drug-cell therapy pairs faster and cheaper when compared with pre-clinical trials.

TABLE 2
Top-20 predicted response pairs

Rank	Drug	Cell line	Sensitivity
1	CX-5461	908158	1
2	QL-XI-92	1330960	1
3	Tubastatin A	909715	1
4	Tubastatin A	906856	1
5	QL-XI-92	908156	1
6	CX-5461	909715	1
7	Y-39983	949155	1
8	PIK-93	909715	1
9	QL-XI-92	1331033	1
10	KIN001-260	909715	1
11	KIN001-260	909702	1
12	Nutlin-3a	908156	1
13	Tubastatin A	1330982	1
14	KIN001-260	908158	1
15	Tubastatin A	906800	1
16	Tubastatin A	1331034	1
17	Y-39983	909715	1
18	CX-5461	1331033	1
19	QL-XI-92	1331037	1
20	CX-5461	909702	1

4 Discussion and Conclusion

A number of critical computation studies are performed to predict response sensitivity based on the existing drugs and cell lines genomic information. The extraordinary progress

TABLE 3
Top-10 novel predicted response pairs

Rank	Drug	Cell line
1	CX-5461	1327774
2	CH5424802	908156
3	Tubastatin A	907275
4	TG101348	1331037
5	GSK1070916	910706
6	**AC220**	**906856**
7	CX-5461	688026
8	CX-5461	908457
9	KIN001-260	907275
10	CX-5461	906862

in screening anti-cancer drugs and oncology cell lines opens the opportunities to build an integrative prediction and holistic methods with heterogeneous-source data which also promotes the research on drug repurposing [11] and personalized medicine exploration. However, two key challenges underlying prediction models tremendously affect the performance of the system. From the modeling perspective, it's necessary to combine the multi-view information together in one model [2]. Simultaneously, the efficacy of the system should also be stressed [37].In this paper, we proposed a novel CF based recommendation framework DS-CTR to solve this problem and deal with the shortcoming in the former methods. To the best of our knowledge, prediction of drug sensitivity using both the two-side genomic information and similarity network has not been explored before.

The main potential reasons considered for the outstanding performance in our accuracy prediction might be summarized into two factors. First, the highly information fusion degree. The first layer mainly emerges the genomic information including drug's target genes and cell line's mutation genes to highlight the heterogeneity relationships and substantial complexity between drug responses and gnomic alterations [12]. While the second layer emphasize both the drugs' structure similarity and cell lines' gene expression similarity. For one point, the fact mentioned by Wang et al. [12] that functional or structural similar drugs may have similar therapeutic effect and influence on the whole prediction system. For the other point, similar gene expression profiles are believed to exhibit similar response [11]. Based on the aforementioned evidence, it's more reasonable for us to draw the multi-source information in our proposed DS-CTR model and provide more reliable prediction results for researches. Consequently, by this way, we are able to overcome the limitation of statistical strength [2]. A promising direction for simplify the prediction procedure is to optimize the learning process through the way of formulating the problem with integrated multi-view sources as known prior knowledge in one time, which in other word, so called holistic methods [22], such as collaborative filtering (CF) methods mentioned in Cobanoglu et al.'s work. Thus, our collaborative filtering based method is shown to group drugs and cell lines according to their feature similarity learned by the multi-view information and after that, by using EM algorithms, to optimal the setting loss function. Different from Zhang et al. work, the prediction need to be done in two steps which probably makes the practical application difficult. In this work, we are able to provide more efficient and simple learning procedure compared with conventional CSN and DSN based models [11], which can also possibly make the proposed model easily be to applied in other multi-source information problems.

Despite the successful results achieved, attention should also be paid on the limitation of the proposed methods. One of the manifest feature of the model is the learning process. The newly propose gradient descent algorithm suffered from a heavy and tough computation load [37] which leads to the difficulty on applying to large-scale dataset. Thus, at the cost of the computation complexity, the model provides us the higher accuracy prediction result. With more and more different types of information accumulated and vast amount of heterogeneous in the biomedical community, faster algorithms are quite an urgently must for practical application. In our model, we choose the structure of drugs and the gene expression of the cell lines as extra strengthen information. We also have another information choices such as the copy number of DNA and oncogene mutation, which can provide the molecular alterations information, or ATC-code, which provides the therapeutic effect at molecular level. Though we have already got high accuracy prediction result by utilizing DS-CTR, the possibility of using another useful information is still worth to be further investigated by researchers.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61571414, 61471331 and 31100955). Correspondence should be addressed to Dr. Ao Li. The authors would like to thank M. I. Mohamed Refai (Biomedical Signals and Systems group, University of Twente, Netherlands) for his review and advice on this study.

References

[1] P. H. van der Graaf and N. Benson, "Systems pharmacology: bridging systems biology and pharmacokinetics-pharmacodynamics (pkpd) in drug discovery and development," *Pharmaceutical research*, vol. 28, no. 7, pp. 1460–1464, 2011.
[2] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, "Predicting drug–target interactions using probabilistic matrix factorization," *Journal of chemical information and modeling*, vol. 53, no. 12, p. 3399, 2013.
[3] 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*, vol. 483, no. 7391, pp. 603–607, 2012.
[4] L. A. Garraway, "Genomics-driven oncology: framework for an emerging paradigm," *Journal of Clinical Oncology*, vol. 31, no. 15, pp. 1806–1814, 2013.
[5] H. Yuan, I. Paskov, H. Paskov, A. J. González, and C. S. Leslie, "Multitask learning improves prediction of cancer drug sensitivity," *Scientific reports*, vol. 6, 2016.
[6] A. Ezzat, P. Zhao, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 3, pp. 646–656, 2017.
[7] K. H. Allison and G. W. Sledge, "Heterogeneity and cancer," *Oncology*, vol. 28, no. 9, pp. 772–772, 2014.
[8] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang *et al.*, "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules," *Cell*, vol. 154, no. 5, pp. 1151–1161, 2013.

[9] 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*, vol. 483, no. 7391, p. 570, 2012.

[10] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS computational biology*, vol. 11, no. 9, p. e1004498, 2015.

[11] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome biology*, vol. 15, no. 3, p. R47, 2014.

[12] Y. Wang, J. Fang, and S. Chen, "Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties," *Scientific reports*, vol. 6, 2016.

[13] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS one*, vol. 8, no. 4, p. e61318, 2013.

[14] M. Ammad-ud din, S. A. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio, and S. Kaski, "Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization," *Bioinformatics*, vol. 32, no. 17, pp. i455–i463, 2016.

[15] M. Ammad-Ud-Din, E. Georgii, M. Gonen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski, "Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization," *Journal of chemical information and modeling*, vol. 54, no. 8, pp. 2347–2359, 2014.

[16] A. Cichonska, J. Rousu, and T. Aittokallio, "Identification of drug candidates and repurposing opportunities through compound–target interaction networks," *Expert opinion on drug discovery*, vol. 10, no. 12, pp. 1333–1345, 2015.

[17] I. Cortés-Ciriano, G. J. van Westen, G. Bouvier, M. Nilges, J. P. Overington, A. Bender, and T. E. Malliavin, "Improved large-scale prediction of growth inhibition patterns using the nci60 cancer cell line panel," *Bioinformatics*, vol. 32, no. 1, pp. 85–95, 2015.

[18] M. Ammad-ud din, S. A. Khan, K. Wennerberg, and T. Aittokallio, "Systematic identification of feature combinations for predicting drug response with bayesian multi-view multi-task linear regression," *Bioinformatics*, vol. 33, no. 14, pp. i359–i368, 2017.

[19] N. Berlow, S. Haider, Q. Wan, M. Geltzeiler, L. E. Davis, C. Keller, and R. Pal, "An integrated approach to anti-cancer drug sensitivity prediction," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 11, no. 6, pp. 995–1008, 2014.

[20] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehr, G. V. Kryukov, and D. Sonkin, "The cancer cell line encyclopedia enables predictive modeling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, p. 603, 2012.

[21] M. L. Sos, K. Michel, T. Zander, J. Weiss, P. Frommolt, M. Peifer, D. Li, R. Ullrich, M. Koker, F. Fischer *et al.*, "Predicting drug susceptibility of non–small cell lung cancers based on genetic lesions," *The Journal of clinical investigation*, vol. 119, no. 6, p. 1727, 2009.

[22] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot *et al.*, "A landscape of pharmacogenomic interactions in cancer," *Cell*, vol. 166, no. 3, pp. 740–754, 2016.

[23] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 448–456.

[24] C. Chen, X. Zheng, Y. Wang, F. Hong, D. Chen *et al.*, "Capturing semantic correlation for item recommendation in tagging systems." in *AAAI*, 2016, pp. 108–114.

[25] 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 research*, vol. 41, no. D1, pp. D955–D961, 2012.

[26] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[27] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[28] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[29] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 931–940.

[30] C. Chen, J. Zeng, X. Zheng, and D. Chen, "Recommender system based on social trust relationships," in *e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference on*. IEEE, 2013, pp. 32–37.

[31] C. Chen, X. Zheng, Y. Wang, F. Hong, Z. Lin *et al.*, "Context-aware collaborative topic regression with social matrix factorization for recommender systems." in *AAAI*, vol. 14, 2014, pp. 9–15.

[32] S. Purushotham, Y. Liu, and C.-C. J. Kuo, "Collaborative topic regression with social matrix factorization for recommendation systems," *arXiv preprint arXiv:1206.4684*, 2012.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[34] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

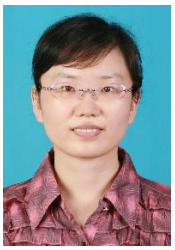
[35] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "Pubchem: a public information system for analyzing bioactivities of small molecules," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W623–W633, 2009.

[36] K. Reiter, H. Polzer, C. Krupka, A. Maiser, B. Vick, M. Rothenberg-Thurley, K. Metzeler, D. Dörfel, H. Salih, G. Jung *et al.*, "Tyrosine kinase inhibition increases the cell surface localization of flt3-itd and enhances flt3-directed immunotherapy of acute myeloid leukemia." *Leukemia*, 2017.

[37] L. Zhang, H. Liu, Y. Huang, X. Wang, Y. Chen, and J. Meng, "Cancer progression prediction using gene interaction regularized elastic net," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 14, no. 1, pp. 145–154, 2017.

Hang Wang is the bachelor student who will graduate in automation from the School of Information Science and Technology, University of Science and Technology of China (USTC) in 2018. His current research interest is data mining and processing on computational biology and bioinformatics.

Jianing Xi received the BS degree in electronic science and technology from the School of Information Science and Technology, University of Science and Technology of China (USTC) in 2013, where he is currently working toward the PhD degree. His current research interest is matrix factorization algorithms for computational biology and bioinformatics. He is a student member of the IEEE.



Minghui Wang received the BS degree from the School of Gifted Youth, University of Science and Technology of China (USTC), and the PhD degree in Biomedical Engineering from School of Information Science and Technology, USTC in 2006. She is an associated professor in the School of Information Science and Technology and Centers for Biomedical Engineering, USTC. Her research interests include bioinformatics, biostatistics and machine learning. She is a member of the IEEE.

Ao Li received the BS degree in biophysics from the School of Life Science, University of Science and Technology of China (USTC), in 2000 and the PhD degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2005. Currently, he is an associated professor in the School of Information Science and Technology and Centers for Biomedical Engineering, USTC. His research contributions are in computational cancer genomics and bioinformatics with a focus on issues concerning systematic identification and evaluation of genome-wide variants in cancer. He is a member of the IEEE.

Additional file 1 — Supplementary Tables

Dual-layer Strengthened Collaborative Topic Regression Modeling for Predicting Drug Sensitivity

Hang Wang[1], Jianing Xi[1], Minghui Wang[1,2], Ao Li[1,2*]

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, PR China
[2]Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, PR China
*Corresponding authors: mhwang@ustc.edu.cn; aoli@ustc.edu.cn

List of Tables

Table S1. Top 200 drug-cell line pair detected by DS-CTR model on GDSC dataset(July, 2016 version)[1][2]. The ranks in this table demonstrates the prediction value of the given drugs on the cell lines. Note that the index number in the cell line column represents the specific cell line in the dataset. The detailed information of it is available at GDSC project website. The sensitivity column denotes the response result based on GDSC project, where 1 means the drug is sensitive to the cell line and 0 means the sensitivity is unknown or negative.

Rank	Drug	Cell line	Sensitivity	Rank	Drug	Cell line	Sensitivity
1	CX-5461	908158	1	101	QL-XI-92	1331034	1
2	QL-XI-92	1330960	1	102	CX-5461	713899	1
3	Tubastatin A	909715	1	103	Tubastatin A	909252	1
4	Tubastatin A	906856	1	104	PIK-93	1323913	1
5	QL-XI-92	908156	1	105	Y-39983	1330947	1
6	CX-5461	909715	1	106	XMD15-27	908156	1
7	Y-39983	949155	1	107	QL-XI-92	753548	1
8	PIK-93	909715	1	108	CX-5461	910944	1
9	QL-XI-92	1331033	1	109	QL-XI-92	910947	1
10	KIN001-260	909715	1	110	Tubastatin A	684059	1
11	KIN001-260	909702	1	111	TPCA-1	909252	1
12	Nutlin-3a	908156	1	112	Trametinib	1240149	1
13	Tubastatin A	1330982	1	113	Tubastatin A	1330947	1
14	KIN001-260	908158	1	114	STF-62247	908156	1
15	Tubastatin A	906800	1	115	JW-7-24-1	1327774	1
16	Tubastatin A	1331034	1	116	Y-39983	908156	1
17	Y-39983	909715	1	117	GSK690693	949163	1
18	CX-5461	1331033	1	118	QL-XI-92	1331035	1
19	QL-XI-92	1331037	1	119	CX-5461	908156	1
20	CX-5461	909702	1	120	ABT-263	906856	1
21	CX-5461	1331036	1	121	QL-XI-92	684059	1
22	KIN001-260	1330984	1	122	CX-5461	949158	1
23	KIN001-260	1295740	1	123	XMD14-99	910706	1
24	GSK690693	1331036	1	124	CX-5461	1331037	1
25	Y-39983	905965	1	125	GSK690693	909701	1
26	CX-5461	1240145	1	126	Tubastatin A	908156	1
27	Y-39983	905952	1	127	KIN001-260	910944	1
28	KIN001-260	910906	1	128	Y-39983	905958	1
29	CX-5461	949155	1	129	GSK690693	909703	1
30	Tubastatin A	1323913	1	130	BX-912	910944	1
31	Y-39983	1295740	1	131	Y-39983	1327769	1
32	CX-5461	905958	1	132	GSK690693	683665	1
33	GSK690693	1331033	1	133	GSK690693	908448	1
34	Tubastatin A	910947	1	134	QL-XI-92	910706	1
35	GSK690693	684059	1	135	GSK690693	906870	1
36	CX-5461	1330984	1	136	XMD14-99	909715	1
37	PIK-93	908134	1	137	Y-39983	684059	1
38	CX-5461	684059	1	138	Tubastatin A	1330984	1
39	CX-5461	1331034	1	139	GSK1070916	908156	1
40	ABT-263	908156	1	140	CX-5461	1247871	1
41	KIN001-260	1330947	1	141	GSK690693	1331034	1
42	Tubastatin A	910944	1	142	KIN001-260	910947	1
43	CX-5461	753610	1	143	KIN001-260	908156	1
44	GSK690693	910944	1	144	Tubastatin A	907272	1
45	Nutlin-3a	1247871	1	145	KIN001-260	1297446	1
46	Y-39983	1331036	1	146	KIN001-260	906856	1
47	GSK1070916	1295740	1	147	GSK690693	906846	1
48	CX-5461	908146	1	148	TPCA-1	906800	1
49	QL-XI-92	1330947	1	149	Y-39983	909252	1
50	QL-XI-92	908146	1	150	Trametinib	905974	1
51	STF-62247	1331037	1	151	QL-XI-92	906846	1
52	GSK690693	684057	1	152	CX-5461	1330960	1
53	TG101348	908156	1	153	PIK-93	909252	1
54	Tubastatin A	908158	1	154	A-770041	906800	1
55	KIN001-260	907272	1	155	KIN001-260	907277	1
56	CX-5461	907073	1	156	Ruxolitinib	910706	1
57	Trametinib	909713	1	157	CX-5461	909251	1
58	A-770041	1331037	1	158	GSK1070916	1247871	1
59	Trametinib	924238	1	159	PIK-93	908158	1
60	TG101348	1330947	1	160	Tubastatin A	1297446	1
61	QL-XI-92	908158	1	161	Tubastatin A	1295740	1
62	KIN001-260	1323913	1	162	Tubastatin A	1331037	1
63	CX-5461	1295740	1	163	CX-5461	910546	1
64	CX-5461	1299080	1	164	GSK1070916	909715	1
65	Y-39983	906800	1	165	Tubastatin A	909702	1
66	Trametinib	906855	1	166	AP-24534	1330960	1
67	QL-XI-92	1331036	1	167	Tubastatin A	683665	1
68	GSK1070916	684059	1	168	BIX02189	906870	1
69	A-770041	1330960	1	169	KIN001-260	908134	1
70	PIK-93	1330984	1	170	BIX02189	909715	1
71	QL-XI-92	1295740	1	171	QL-XI-92	909251	1
72	JW-7-24-1	908158	1	172	QL-XI-92	909260	1
73	Y-39983	1330984	1	173	GSK690693	909715	1
74	GSK690693	907275	1	174	PIK-93	909703	1
75	QL-XI-92	1330933	1	175	XMD14-99	1323913	1
76	KIN001-236	909715	1	176	QL-XI-92	906836	1
77	Y-39983	906856	1	177	QL-XI-92	1330985	1
78	Y-39983	908158	1	178	Trametinib	753545	1
79	TPCA-1	909715	1	179	GSK690693	907789	1
80	QL-XI-92	907799	1	180	KIN001-260	909252	1
81	Trametinib	909756	1	181	CX-5461	907320	1
82	GSK690693	1327774	1	182	VNLG/124	909715	1
83	KIN001-260	906800	1	183	Ruxolitinib	1331037	1
84	QL-XI-92	1330950	1	184	BX-912	908156	1
85	CX-5461	909703	1	185	GSK690693	1331040	1
86	T0901317	908156	1	186	QL-XI-92	908448	1
87	GSK690693	906824	1	187	QL-XI-92	907275	1
88	CX-5461	909701	1	188	QL-XI-92	1323913	1
89	KIN001-260	906870	1	189	Tubastatin A	905958	1
90	QL-XI-92	909715	1	190	QL-XI-92	909255	1
91	GSK1070916	909252	1	191	TPCA-1	1330984	1
92	Trametinib	907061	1	192	GSK690693	1330947	1
93	GSK690693	908158	1	193	GSK690693	910688	1
94	CX-5461	906800	1	194	XMD14-99	1331037	1
95	CX-5461	753548	1	195	Nutlin-3a	905952	1
96	TPCA-1	1330947	1	196	QL-XI-92	906800	1
97	QL-XI-92	906870	1	197	CX-5461	949165	1
98	CX-5461	924247	1	198	Y-39983	1331037	1

99	QL-XI-92	910906	1	199	CX-5461	909256	1
100	XMD14-99	908158	1	200	Ruxolitinib	908156	1

Table S2. Top 100 novel drug-cell line pair detected by DS-CTR model on GDSC dataset(July, 2016 version)[1][2]. The ranks in this table demonstrates the prediction value of the given drugs on the cell lines. Note that the index number in the cell line column represents the specific cell line in the dataset. The detailed information of it is available at GDSC project website. All the pairs shown in the table are either negative or unclear in the GDSC the dataset.

Rank	Drug	Cell line	Sensitivity	Rank	Drug	Cell line	Sensitivity
1	CX-5461	1327774	0	26	GSK1070916	1327774	0
2	CH5424802	908156	0	27	Tubastatin A	907783	0
3	Tubastatin A	907275	0	28	Trametinib	908440	0
4	TG101348	1331037	0	29	OSI-930	906856	0
5	GSK1070916	910706	0	30	BX-912	949155	0
6	AC220	906856	0	31	VNLG/124	909703	0
7	CX-5461	688026	0	32	Bleomycin (50 uM)	1298215	0
8	CX-5461	908457	0	33	FMK	1330983	0
9	KIN001-260	907275	0	34	KIN001-102	910906	0
10	CX-5461	906862	0	35	T0901317	909252	0
11	CX-5461	1524419	0	36	VNLG/124	1295740	0
12	GSK1070916	1330960	0	37	Trametinib	905956	0
13	STF-62247	907275	0	38	CP466722	909715	0
14	Belinostat	908156	0	39	AV-951	906856	0
15	CAL-101	1327774	0	40	Zibotentan	910944	0
16	NPK76-II-72-1	910906	0	41	Y-39983	1327774	0
17	Zibotentan	1327774	0	42	GDC0941	907275	0
18	CX-5461	949178	0	43	KIN001-102	1330984	0
19	Tubastatin A	1327774	0	44	XMD13-2	907789	0
20	Y-39983	907275	0	45	AV-951	908158	0
21	AZD6244	906820	0	46	Trametinib	753614	0
22	GSK1070916	949158	0	47	Ruxolitinib	910906	0
23	CX-5461	1240130	0	48	Belinostat	1323913	0
24	KIN001-260	1330960	0	49	VNLG/124	909255	0
25	CX-5461	908133	0	50	AV-951	907272	0

References

[1] 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 research 41 (D1) (2012) D955–D961.

[2] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, S. H. Bryant, Pubchem: a public information system for analyzing bioactivities of small molecules, Nucleic acids research 37 (suppl_2) (2009) W623–W633.