Molecular Therapy
Nucleic Acids
Original Article

Anti-cancer Drug Response Prediction Using Neighbor-Based Collaborative Filtering with Global Effect Removal

Hui Liu,[1] Yan Zhao,[1] Lin Zhang,[1] and Xing Chen[1]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China

Patients of the same cancer may differ in their responses to a specific medical therapy. Identification of predictive molecular features for drug sensitivity holds the key in the era of precision medicine. Human cell lines have harbored most of the same genetic changes found in patients' tumors and thus are widely used in the research of drug response. In this work, we formulated drug-response prediction as a recommender system problem and then adopted a neighbor-based collaborative filtering with global effect removal (NCFGER) method to estimate anti-cancer drug responses of cell lines by integrating cell-line similarity networks and drug similarity networks based on the fact that similar cell lines and similar drugs exhibit similar responses. Specifically, we removed the global effect in the available responses and shrunk the similarity score for each cell line pair as well as each drug pair. We then used the K most similar neighbors (hybrid of cell-line-oriented and drug-oriented) in the available responses to predict the unknown ones. Through 10-fold cross-validation, this approach was shown to reach accurate and reproducible outcomes of drug sensitivity. We also discussed the biological outcomes based on the newly predicted response values.

INTRODUCTION

Cancer subtypes differ in chemotherapeutic response and thus may require different medical treatment. The relationships between molecular features and clinical drug responses lay the foundation for optimizing drug therapies based on a patient's genomic context.[1] Therefore, it has been a major challenge to accurately predict the anti-cancer drug response based on the patient's molecular and clinical profiles in the era of precision medicine. On the one hand, it is crucial for clinicians to make decisions in the choice of most effective and least toxic therapeutic regimen. On the other hand, the identification of a drug-sensitive biomarker is essential for cancer medicine. The emerging of high-throughput drug-screening technologies has enabled many studies to conduct large-scale experiments on cultured human cell line panels, which greatly improved systematical elucidation of the response mechanism of anti-cancer drugs. Several attempts to construct predictive models for drug response have made use of some datasets. For instance, NCI-60 was a panel of human cell lines originally derived from human cancers spanning nine different tissues of origin.[2] The other two recent consortiums, GDSC (Genomics of Drug Sensitivity in Cancer)[3] and CCLE (Cancer Cell Line Encyclopedia)[4] have analyzed around 1,500 cancer cell lines and their genomic profiles against 280 drugs, providing the concentration required for 50% of cellular growth inhibition (IC^{50}) or activity area as drug-response measurement. All studies provide genome-wide profiling of multiple cancer cell lines and drug-sensitivity data based on established anti-cancer drugs against the cell lines. However, the sensitivity levels for most cell line-drug pairs are still unknown, and it needs to be achieved by a time- and cost-effective way for potential personalized medicine.[5]

Currently, most commonly used methods for drug response prediction are multivariate linear regression (least absolute shrinkage and selection operator [LASSO] and elastic net regularizations) and nonlinear regression (e.g., neural networks and some kernel-based methods).[3,4,6–9] Daemen et al.[10] identified response-associated molecular features, such as measurements of copy number aberrations, mutations, gene and isoform expression, promoter methylation, as well as protein expression in breast cancer by least-squares-support vector machines and random forest algorithms. Staunton et al.[11] first developed a weighted voting classification model on NCI-60 basal gene expression data to predict anti-cancer drug sensitivity. Gene signatures of 232 drugs from 6,817 genes were created to predict a binary (sensitive or resistant) response. Riddick et al.[12] built an ensemble regression model using random forest to predict *in vitro* drug response from a signature of basal-gene expression. The random forest regression model for each drug was built between gene-expression signatures for the cell lines and the corresponding IC^{50} values for the exact drug for unknown response prediction. Cortes-Ciriano et al.[13] proposed a simultaneous machine-learning modeling of chemical and cell-line information for response prediction. Ammad-ud-din et al.[14] adopted a kernelized Bayesian matrix factorization (KBMF) method to predict the drug

Received 22 June 2018; accepted 18 September 2018;
https://doi.org/10.1016/j.omtn.2018.09.011.

Correspondence: Xing Chen, School of Information and Control Engineering, China University of Mining and Technology, No.1, Daxue Road, Xuzhou, Jiangsu 221116, China.
E-mail: xingchen@amss.ac.cn

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1. Comparisons of Both Methods Obtained under 10-Fold Cross-Validation on GDSC Dataset

Methods	Drug-Averaged PCC_S/R	Drug-Averaged RMSE_S/R	Drug-Averaged PCC	Drug-Averaged RMSE
NCFGER				
Hybrid	0.81 (±0.11)	1.42 (±0.29)	0.73 (±0.13)	1.18 (±0.24)
Cell-line based	0.76 (±0.14)	1.49 (±0.24)	0.67 (±0.15)	1.29 (±0.21)
Drug-based	0.75 (±0.12)	1.60 (±0.43)	0.66 (±0.14)	1.32 (±0.34)
SRMF				
	0.71 (±0.15)	1.73 (±0.46)	0.62 (±0.16)	1.43 (±0.36)

responses by integrating genomic and chemical properties in addition to drug target information. Zhang et al.[15] proposed a dual-layer cell-line drug network (DLN) model, which integrated both cell-line similarity network data and drug similarity network data, to predict the missing drug response of a given cell line. Kim et al.[16] developed a network-based classifier method for predicting sensitivity of cell lines to anti-cancer drugs from transcriptome data. Wang et al.[17] proposed a similarity-regularized matrix factorization (SRMF) method for drug-response prediction, which incorporates similarities of drugs and of cell lines simultaneously. Stanfield et al.[18] proposed a heterogeneous network-based method to predict the interaction between cell line-drug pairs. They classified the interaction between each cell line-drug pairs into sensitive and resistant and thus turned the prediction problem into classification. Suphavilai et al. have proposed a matrix factorization based recommender system (CaDRReS) method, which considers essential genes for drug-response prediction.[19]

Regarding the fact that similar cell lines and similar drugs exhibit similar drug responses,[15] the prediction of unknown drug response can be considered as a typical recommender system (RS).[20] Typically, in a RS, there is a set of users and a set of items. Each user u rates a set of items by some values. The RS attempts to profile user preferences and tries to model the interaction between users and items, which is exactly analog for drug-response prediction. The cell lines correspond to users while drugs correspond to items. Thus, we proposed an RS technique, neighborhood-based collaborative filtering with global effects removal (NCFGER), for drug-response prediction, which incorporates similarities of drugs and of cell lines in additional to the known drug response simultaneously. To demonstrate its effectiveness, we compared NCFGER with SRMF, which has been proved to show higher performance than typical similarity-based methods KBMF and DLN. The evaluation metrics are also averaged Pearson correlation coefficient (PCC) and averaged root-mean-square error (RMSE) over all drugs. The results on GDSC and CCLE drug-response datasets by 10-fold cross validation showed that NCFGER performed dramatically better than SRMF in terms of drug-averaged PCC and RMSE. NCFGER has also been applied to impute unknown response values in the GDSC dataset for detailed biological meaningful presentation.

RESULTS

Measurements of Prediction Performance

The prediction performance of our method was evaluated using PCC and RMSE between predicted and observed drug responses for each drug. A higher PCC and lower RMSE indicate a better prediction performance of a method. For comparison, PCC and RMSE of the sensitive and resistant cell lines of each individual drug, which were denoted as PCC_S/R (PCC between predicted and observed responses of sensitive and resistant cell lines) and RMSE_S/R (RMSE between predicted and observed responses of sensitive and resistant cell lines) were also evaluated, respectively. For each drug, the $\log IC^{50}$ values were split into quantiles, with cell lines in the first and fourth representing drug-sensitive (-resistant) and -resistant (-sensitive) cell lines, respectively, which followed the same definition in Wang et al.[17] Therefore, we evaluated four measures, drug-averaged PCC, drug-averaged RMSE, drug-averaged PCC_S/R, as well as drug-averaged RMSE_S/R over all drugs, respectively.

Similarity in Response Helps Improve the Prediction Performance

We first conducted 10-fold cross-validation to evaluate the prediction performance in the GDSC (https://www.cancerrxgene.org/) and CCLE (https://www.broadinstitute.org/ccle) datasets to evaluate different similarity definition. For each dataset, the drug-response entries were divided into 10 folds randomly with almost the same size. Each time, 1 fold was used as the test set, while the remaining 9 folds were used as the training sets. The prediction was repeated 10 times such that each fold acted as a test set once. The whole cross-validation was run 100 times for each dataset, and the prediction performance was compared with the best state-of-the-art method, SRMF.

As is shown in Tables S1 and S2, the performance of hybrid NCFGER with *RPCC* and *MRPCC* were better than that of *COEF* similarity, which indicates that the similarity exhibited in drug-response values can better improve the drug-response prediction performance. It was consistent with the result concluded in L.Z (unpublished data). Thus, we used *MRPCC* similarity in the following study.

10-Fold Cross-Validation Test on GDSC and CCLE Drug-Response Datasets

We also conducted 10-fold cross-validation on GDSC and CCLE drug-response datasets to investigate the performance of cell-line-based NCFGER, drug-based NCFGER, and hybrid NCFGER.

Tables 1 and S3 showed the comparison results obtained by four methods, hybrid NCFGER, cell-line-based NCFGER, drug-based NCFGER, and SRMF in the GDSC and CCLE datasets. As shown in Table 1, all NCFGER methods outperformed SRMF in all metrics in the GDSC dataset, while hybrid NCFGER attains the best performance. The drug-averaged PCC_S/R obtained by our method is 0.81, which is 14.42% higher than that of SRMF. The drug-averaged RMSE_S/R obtained by our method is 1.42, which is 17.92% lower than that obtained by SRMF.

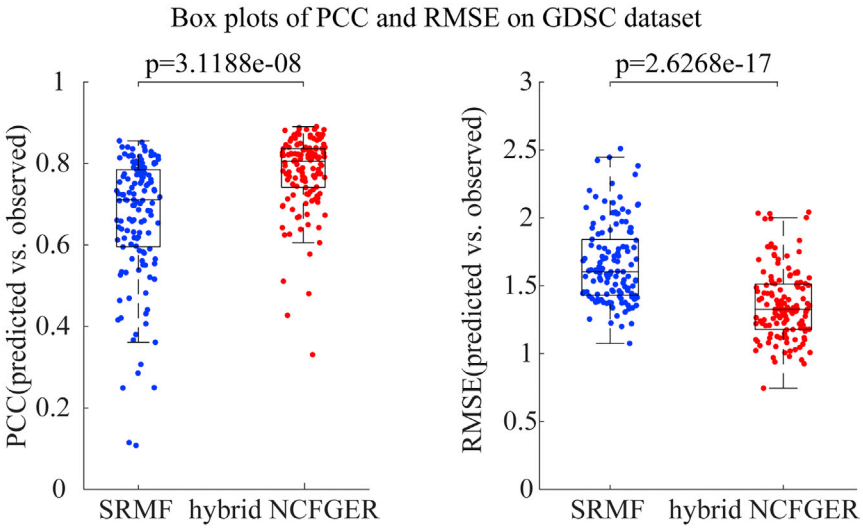

Figure 1. Boxplots of PCC and RMSE Based on SRMF and NCFGER

As shown in Table S3, all NCFGER methods also outperform SRMF in all metrics in the CCLE dataset, and hybrid NCFGER still attains the best performance. The drug-averaged PCC_S/R obtained by our method is 0.89, which is 14.40% higher than SRMF. The drug-averaged RMSE_S/R obtained by our method is 0.47, which is 36.08% lower than that obtained by SRMF.

Considering the results drawn in above, we focused on hybrid NCFGER for further analysis if there was no additional explanation. The cell-line-based NCFGER method depended on the most similar K drugs that had responses in the same cell line, which means the drug structure contributed to the prediction performance improvement from the respect of RS. Since neither cell-line-based nor drug-based NCFGER could outperform the hybrid method, both cell-line gene expression profile and drug structure contributed to the prediction performance improvement. If we only use one part of them, we will miss some information that helps to predict the drug responses.

Figures 1 and S1 also show the boxplots of both methods with respect to PCC and RMSE for each drug in the CCLE and GDSC datasets, respectively. Either PCC or RMSE averaged for each drug over the 100 times of cross-validation from hybrid NCFGER is better than that of SRMF.

Furthermore, we also investigated the prediction performance of drug target genes in specific pathways. As is known, phosphatidylinositol 3-kinase (PI3K) and extracellular signal-related kinase (ERK) signaling pathways have been identified as promising therapeutic targets for cancer therapy, which makes it meaningful to consider the prediction performance of drug responses for their targeting genes in these pathways[21] (Figures 2 and S2). From the perspective of PCC and RMSE, NCFGER performs better than SRMF in both PI3K and ERK pathways.

Drug-Cancer Association Validation Based on Predicted Responses in GDSC Dataset

Using our proposed method mentioned in the above sections, we trained NCFGER method on all available response IC^{50} and used it to predict the missing responses in GDSC dataset. Following Wang et al.,[17] we focused on an epidermal growth factor receptor (EGFR)-family inhibitor, lapatinib, where more than half of response values (342/652) were missing and a cyclin-dependent kinases (CDKs) 4 and 6 inhibitor PD-0332991, where nearly 10% of response values (62/652) were missing. Based on the definition that a mutated gene fulfills any of these criteria following Garnett et al.[7]: a coding sequence variant in the cancer gene, a total copy number = 0 (homozygous deletion) or $\geq$ 8 (amplification), we grouped the unassayed cell lines based on their EGFR mutation profiles and found that the EGFR mutated cell lines were significantly more sensitive to lapatinib. EGFR and ERBB2 amplification was shown to be associated with sensitivity to lapatinib, which has been licensed for the treatment of HER2-positive breast cancer clinically[22,23] (Figure 3A). This prediction happened to coincide with that in assayed cell lines. Similar fact was observed with predicted response of ERBB2-mutated cell lines to lapatinib (Figure 3B), which is exactly consistent with previous literatures.[24] As to PD-0332991, it is an inhibitor of upstream cyclin-dependent kinases (CDKs) 4 and 6, while CDKN2A-mutated cells have been known to have enhanced requirement for signaling through the CDK4/6-pRb signaling pathway. The predicted results show that CDKN2A-mutated cell lines were more sensitive to PD-0332991 (Figure 3C). This prediction was not only consistent with that in assayed cell lines, but also in agreement with previously a published study.

The newly predicted drug responses combined with existing drug responses were able to detect novel drug-cancer gene associations as well, which is consistent with previous literatures (Figure 4). For example, the oncogene BRAF has been found significantly associated with enhanced and selective sensitivity to mitogen-activated protein kinase (MEK) inhibitor PD-0325901 ($p = 3.70e-11$ for available responses; $p = 1.08e-12$ for combination of predicted and available responses),[25] which was approached with the combination of newly predicted drug responses and known responses versus available responses themselves. Therefore, it is important to complete the unknown observations of drug response matrix such that we can unveil the new drug-sensitivity mechanism better. Also, based on the combined newly predicted drug responses and available observations versus available observations themselves, fibroblast growth factor

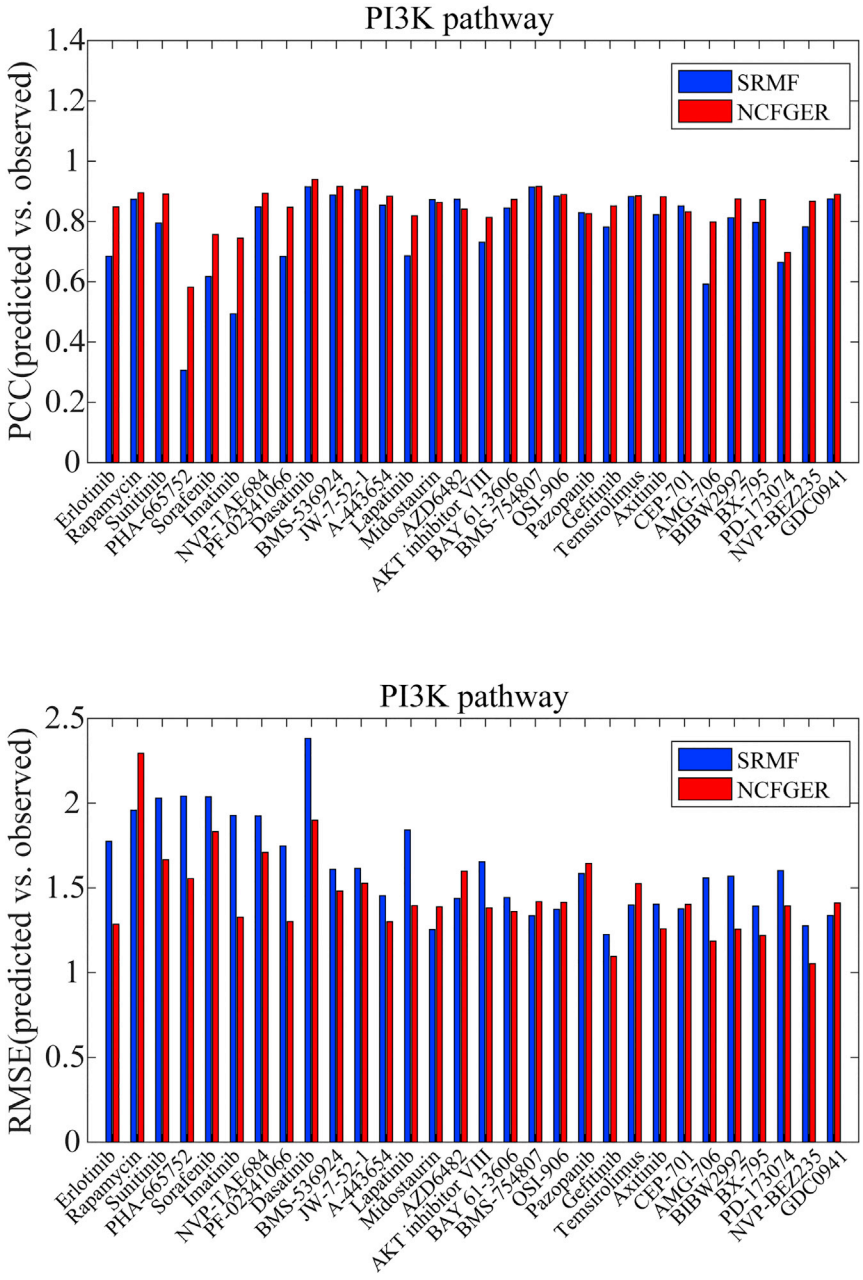

Figure 2. Prediction Performance Comparisons of SRMF and NCFGER for Drug-Targeting Genes in PI3K Pathway with Respect to PCC as Well as RMSE in CCLE Dataset

receptor 2 (FGFR2)-mutated cell lines were exquisitely sensitive to PD-173074, which has been known to prevent signaling, at nanomolar levels, through FGFR2–5 (p = 0.7e-2 for available observations; p = 0.22e-2 for combination of predicted and available observations).[26–29]

The association between the presence of inactivating mutations in tumor-suppressor genes and several drugs, which in some aspect provide insight into the interaction between tumor suppressors and the cellular mechanism in mediating drug sensitivity, were also remained in the predicted drug response (Figure 5). For instance, mutation of TP53, an important regulator of apoptosis and cell cycle arrest in response to cellular stress, confers resistance to Nutlin-3a (p = 3.53e-39 for known response; p = 5.79e-39 for combination of known and unknown response), which is an inhibitor of the mouse double minute 2 homolog (MDM2) E3-ligase that negatively regulates p53 protein levels. Just like TP53, mutation inactivation of RB1, a key repressor of cell cycle progression in normal cells, confers resistance to PD-0332991 (p = 3.33e-17 for known response; p = 5.51e-16 for combination of known and unknown response).

We also investigated the genes that are significantly differentially expressed in sensitive and resistant cell lines, as shown in Figure 6. The threshold p value was set to 0.05, and the threshold fold change was set to 1.5. Significantly differentially expressed genes further went through gene ontology enrichment analysis by DAVID[30] (https://david.ncifcrf.gov/) with default parameter settings. Finally, differentially expressed genes were found to be related to the PI3K-Akt signaling pathway (p = 7.9e-4), ECM-receptor interaction (p = 3.3e-10), and small-cell lung cancer (p = 1.5e-4).

DISCUSSION

In this study, we developed a collaborative filtering-based method, NCFGER, to estimate the response of cancer cell lines to drug treatments for IC_{50} values as well as activity area in GDSC and CCLE datasets, respectively. This method is the hybrid of cell line-based and drug-based collaborative filtering techniques, thereby incorporating similarity in responses from the perspective of cell lines and drug structures, similarity in cell-line gene expression profile, as well as similarity in drug chemical fingerprint. It also applies a global effect removal to preprocess the available drug response and a shrinkage operation on both cell-line similarity network and drug similarity network to avoid the bias caused by unknown responses. 10-fold cross-validation showed that the drug-response similarity can better improve the drug-response prediction performance in comparison with the cell-line gene expression and drug chemical

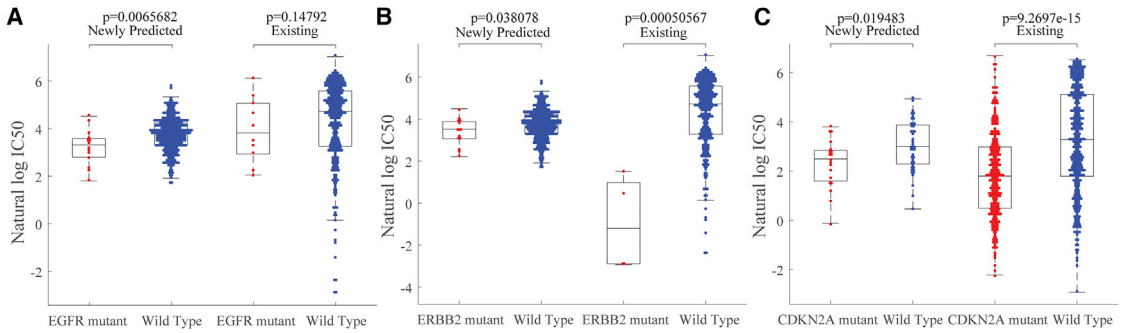

Figure 3. The Association of Drug Sensitivity and Cancer Gene Mutations Were Consistent for Predicted Response Data

EGFR (A) and ERBB2 (B) mutated cell lines were more sensitive to the drug lapatinib, while CDKN2A (C) mutated cell lines were more sensitive to drug PD-0332991.

structure similarity. 10-fold cross-validation also showed that the hybrid NCFGER algorithm consistently outperformed SRMF, suggesting that the hybrid NCFGER are more predictive of anti-cancer drug response. We also used the hybrid NCFGER to predict the unknown drug-response values in the GDSC dataset. In combination with the available observations, the results of gene mutation association with drug response, inactivating mutation association in tumor gene suppressors with drugs were all consistent with previous findings.

Compared to the existing drug-response prediction method, our NCFGER method is based on the relatively dependable neighbors (both cell lines and drugs) and dependable similarity score shrinkage technique during response prediction, so the sparsity has less influence on the prediction performance. Furthermore, the prediction performance was not seriously affected by either similarity of cell-line gene expression or of drug; thus, the input for NCFGER could be quite simple.

Despite these encouraging aspects, NCFGER suffers from the following limitations, which we hope to address in the future. First, from the respect of the cell line, construction of NCFGER relies on gene-expression profile data only, and we hope to integrate somatic mutation information and epigenetic status in the future. Some pathway-related information or other dynamic information may also help improve the predictive performance of drug response; thus, it might be better to integrate omics data later. The neighbor-based collaborative filtering framework highly depends on the selection of neighbors. We may start the incorporation of other information, such as pathway-related information or other dynamic information, with the integration of this information for neighbor selection at the first step. We can also treat the drug-response prediction as a classification problem, which would be easier to incorporate the other available information. Second, the measure metrics PCC and RMSE were regarded as two main measures for drug-response prediction projects. But different data may have different magnitudes in the drug-response value for their common drugs. Therefore, we can also seek a better measure for drug-response prediction problems.

MATERIALS AND METHODS

Problem Formulation

In this paper, we adopted a powerful collaborative filtering method to predict anti-cancer drug responses in cell lines. The primary idea comes from the basic hypothesis that similar cell lines are sensitive to similar drugs.

Before discussion, we first defined the notational conventions. Each dataset mentioned above is constructed as three matrixes. The response matrix is about m cell lines and n drugs, arranged in an $m \times n$ matrix: $R = \{r_{ui}\}_{1 \le u \le m, 1 \le i \le n}$. The cell-line similarity matrix is about the PCC of gene-expression profile between each of the m cell lines, arranged in an $m \times m$ matrix: $COEF_c = \{COEF_{uv}\}_{1 \le u \le m, 1 \le v \le m}$. The subscript c refers to cell line. The drug-similarity matrix is about the Jaccard similarity score of the PubChem fingerprint between each of the n drugs, arranged in an $n \times n$ matrix: $COEF_d = \{COEF_{ij}\}_{1 \le i \le n, 1 \le j \le n}$. The subscript d refers to drug. To distinguish between the two similarity matrixes, the special indexing letters are reserved: for cell lines u and v, and for drugs i, j, and k. Our goal is to predict unknown elements in R based on the known ones, as well as similarity matrixes $COEF_c$ and $COEF_d$.

NCFGER adopted the typical neighborhood-based CF method for drug-response prediction. The original form, which has been shared by virtually all earlier CF systems, is the user-oriented approach.[31] Its analogous alternative is the item-oriented approach.[32] They have been two state-of-the-art techniques for RS. However, the utilization of a user similarity matrix or item similarity matrix only always results in poor prediction accuracy due to the sparseness of preference data. Thus, a hybrid collaborative filtering model is often preferred by combining user-oriented CF and item-oriented CF together. In this way, both user similarity and item similarity are considered for missing value prediction. In this paper, we will focus on the user-oriented approach in the methods introduction section, and simple user-oriented CF (cell-line-based), item-oriented CF (drug-based), as well as hybrid over user-oriented CF and item-oriented CF (hybrid) were all implemented for comparison. The hybrid method simply takes the average of scores predicted from both user-oriented and item-oriented CF methods.

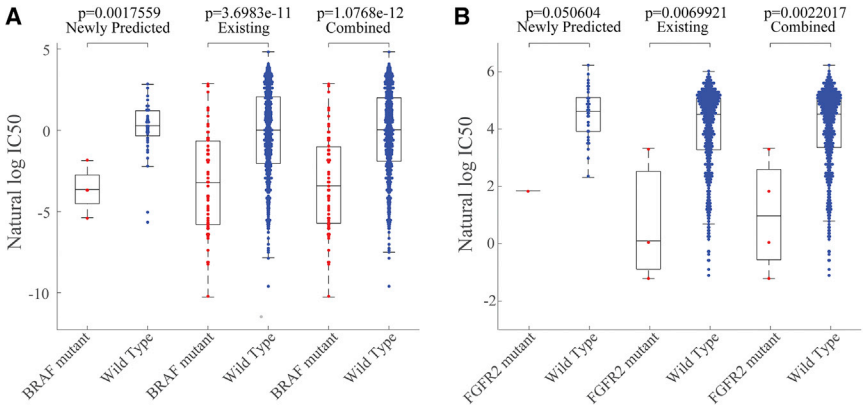

Figure 4. The New Drug-Cancer Association Was Also Observed Based on the Combination of Observed Responses and Newly Predicted Responses

The new drug-cancer association was also observed based on the combination of observed responses and newly predicted responses. (A) BRAF is found to be more significantly associated with PD-0325901 with a combination of newly predicted and known responses. (B) FGFR2 is also found to be more significantly associated with PD-173074.

The cell-line-oriented method followed an improved neighborhood based the collaborative filtering method to estimate the unknown response $\widehat{r}_{ui}$, as is shown in Figure 7. First, there may be large cell line and drug effects—i.e., systematic tendencies for some cell lines to have higher responses than other cell lines and for some drugs to get higher responses than others. This is exactly what we call "global effects" in the RS area. Thus, we adopted the normalization step that helps to remove the "global effects."

Preprocessing by Global Effects Removal

Our strategy is to estimate one "effect" at a time, in sequence (i.e., the overall mean of the response IC^{50} value, the main effect of cell line, the main effect of drug, etc.). At each step, we used residuals from the previous step as the dependent variable for the current step. Thus, after each step of effect removal, the r_{ui} refers to residuals, rather than original IC^{50} values.

Let x_{ui} be the explanatory variable of interest corresponding to cell line u and drug i. For cell-line and drug main effects, the x_{ui}'s are identically 1. For other global effects, x_{ui} is centered for each cell line by subtracting the mean of x_{ui} for the exact cell line. In each case, the model is defined as

$$r_{ui} = \theta_u x_{ui} + \varepsilon. \quad \text{(Equation 1)}$$

With sufficient responses for cell line u, the unbiased estimator of θ is

$$\widehat{\theta}_u = \frac{\sum_i r_{ui} x_{ui}}{\sum_i x_{ui}^2}. \quad \text{(Equation 2)}$$

However, the estimator is somehow unreliable, since some values of $\widehat{\theta}_u$ may be based on very few known responses. To avoid this circumstance, each individual value of $\widehat{\theta}_u$ is shrunk toward a common value from a Bayesian perspective. To be more specific, the true θ_u is supposed to follow a normal distribution. And a slightly simpler estimator used to calculate θ_u is multiplying Equation 2 by the factor defined as

$$\frac{n_u}{n_u + \alpha_1} \quad \text{(Equation 3)}$$

where n_u is the number of responses of cell line u and α_1 is a constant, which was set to 3 by cross-validation.

Similarity Definition

The similarity matrixes are required for identification of K nearest neighbors. The original similarity of cell lines $COEF_{c_{uv}}$ is drawn based on the PCC of gene-expression profiles between cell lines u and v, while that of drug $COEF_{d_{ij}}$ was drawn based on the Jaccard coefficient of drug chemical fingerprint between drugs i and j. However, to some extent, the similarity between cell lines u and v can also be shown from the perspective of drug response. Thus, in this paper, we investigated the different similarity definitions for drug-response prediction. To be more specific, the similarity of cell lines can be defined based on gene expression profile's PCC ($COEF_c$), the correlation coefficient between their response IC^{50} value ($RPCC_c$), as well as the multiplication of $COEF_c$ and $RPCC_c$, which is indicated as $MRPCC_c$ in the following. The exact $MRPCC_c$ similarity measure is defined in Equation 4:

$$MRPCC_{c_{uv}} = COEF_{c_{uv}} \times RPCC_{c_{uv}} \quad \text{(Equation 4)}$$

where $RPCC_{c_{uv}}$ is calculated as the PCC between the response IC^{50} values of cell lines u and v.

In the same way, the similarity between drugs i and j can be defined based on a drug chemical fingerprint's Jaccard coefficient ($COEF_d$), the PCC between response IC^{50} values of drugs i and j ($RPCC_d$), as well as the multiplication of $COEF_d$ and $RPCC_d$ ($MRPCC_d$) defined in Equation 5:

$$MRPCC_{d_{ij}} = COEF_{d_{ij}} \times RPCC_{d_{ij}}. \quad \text{(Equation 5)}$$

In order to avoid the bias caused by the different level of support (different number of known responses) for each drug, the similarity matrixes are further shrunk by multiplying $|U(i,j)|/(|U(i,j)+\alpha_2|)$ for some small α_2, where $U(i,j)$ is the set of cell lines that have responses to both drugs i and j. The index of i and j here will be changed to u and v for cell lines u and v. α_2 was set to 1 by cross-validation in this paper.

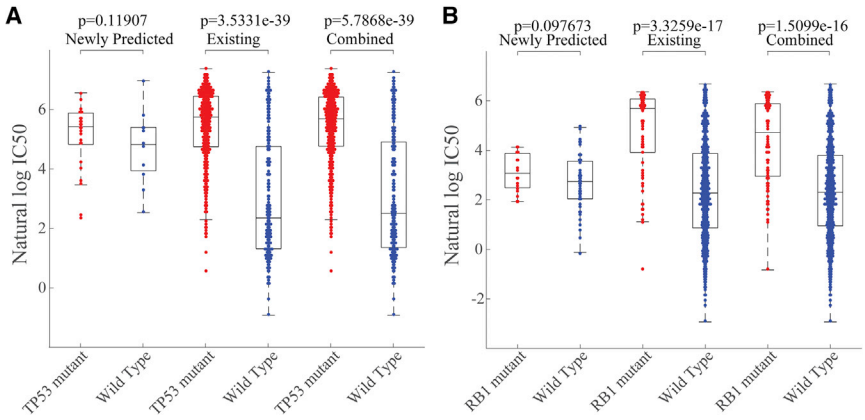

Figure 5. The Association between Inactivating Mutations in Tumor-Suppressor Genes and Drugs Were Also Obtained Based on the Combination of Observed Responses and Newly Predicted Responses

The association between inactivating mutations in tumor suppressor genes and drugs were also obtained based on the combination of observed responses and newly predicted responses. (A) TP53 is found to be significantly resistant to Nutlin-3a based on the combination of known and unknown response. (B) RB1 is also found to be significantly resistant to PD-0332991.

NCFGER

After global effects removal, we can turn to predict the unknown response IC^{50} value for cell line u of drug i, which is $\widehat{r}_{ui}$.

Among all drugs that have response values in cell line u, we resort to a set of K cell lines $N(u;i)$ that tend to have the most similar response in u and that actually have response to drug i (i.e., r_{vi} is known for each $v \in N(u;i)$). K is set to 10 in our experiments. The similarity rank is measured based on the shrunk similarity matrix $MRPCC_c$ and $MRPCC_d$, respectively.

Based on the selected set of K neighbors, the interpolation weights $\{w_{ij} \mid j \in N(i;u)\}$, which enable the best prediction of unknown response, can be reached by

$$\widehat{r}_{ui} \leftarrow \sum_{j \in N(i;u)} w_{ij} r_{uj}. \qquad \text{(Equation 6)}$$

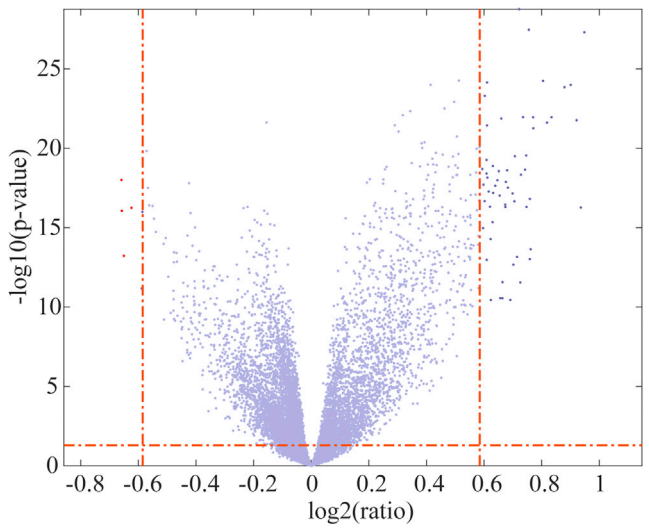

Figure 6. Volcano Plot of Gene Expression Profiles in Sunitinib-Sensitive and -Resistant Cell Lines

Thus, $w \in \mathbb{R}^K$. The interpolation weights can be solved by the definition of linear system:

$$Aw = b. \qquad \text{(Equation 7)}$$

It actually models the relationships between drug i and its neighbors through a least-squares problem of

$$\min_w \sum_{v \neq u} \left(r_{vi} - \sum_{j \in N(i;u)} w_{ij} r_{vj} \right)^2. \qquad \text{(Equation 8)}$$

For each pair of drugs i and j, we compute

$$\overline{A}_{ij} = \frac{\sum_{v \in U(i,j)} r_{vi} r_{vj}}{|U(i,j)|} \text{ and} \qquad \text{(Equation 9)}$$

$$\overline{b}_j = \frac{\sum_{v \in U(i,j)} r_{vj} r_{vi}}{|U(i,j)|}. \qquad \text{(Equation 10)}$$

Then the best estimator $\widehat{A}$ and $\widehat{b}$ for A and b are further improved based on the fact that the averages represented in Equations 9 and 10 may differ by orders of magnitude in terms of the number of cell lines included in the average.

Thus, the corresponding $K \times K$ matrix $\widehat{A}$ and the vector $\widehat{b} \in \mathbb{R}^K$ is defined as

$$\widehat{A}_{jk} = \frac{|U(j,k)|\,\overline{A}_{jk} + \beta \cdot avg}{|U(j,k)| + \beta} \text{ and} \qquad \text{(Equation 11)}$$

$$\widehat{b}_j = \frac{|U(i,j)|\,|\,\overline{b}_j + \beta \cdot avg}{|U(i,j)| + \beta} \qquad \text{(Equation 12)}$$

where avg denotes a baseline value, which is defined by taking the average of all possible $\overline{A}_{jk}$ values. It is obvious that β controls the extent of the shrinkage.

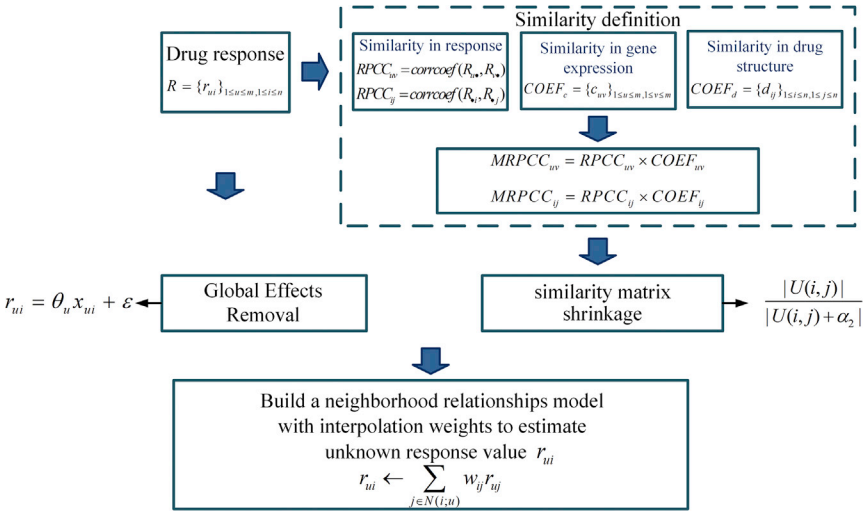

Figure 7. The Workflow of NCFGER

Thus, *w*'s are achieved by a non-negative quadratic optimization method and are used to predict r_{ui} following Equation 1. The final estimated IC^{50} value of $\widehat{r}_{ui}$ should be recovered with those removed global effects.

The drug-oriented method is the analog alternative to cell-line-oriented method. Based on the above two methods, the hybrid method got its prediction score based on the mean operation.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and three tables and can be found with this article online at https://doi.org/10.1016/j.omtn.2018.09.011.

AUTHOR CONTRIBUTIONS

X.C. conceived the project, designed the experiments, analyzed the results, revised the paper, and supervised the project. L.Z. and H.L. developed the prediction method, designed and implemented the experiments, analyzed the results, and wrote the paper. Y.Z. implemented the experiments, analyzed the results, and revised the paper.

CONFLICTS OF INTEREST

The authors have no conflicts of interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant numbers 61772531 and 61501466) and the Fundamental Research Funds for the Central Universities (grant number 2014QNA84).

REFERENCES

1. Ding, Z., Zu, S., and Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics *32*, 2891–2895.
2. Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.
3. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.
4. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.
5. Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci. Rep. *8*, 3355.
6. Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al.; NCI DREAM Community (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. *32*, 1202–1212.
7. Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.
8. Heiser, L.M., Sadanandam, A., Kuo, W.L., Benz, S.C., Goldstein, T.C., Ng, S., Gibb, W.J., Wang, N.J., Ziyad, S., Tong, F., et al. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc. Natl. Acad. Sci. USA *109*, 2724–2729.
9. Menden, M.P., Iorio, F., Garnett, M., McDermott, U., Benes, C.H., Ballester, P.J., and Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS ONE *8*, e61318.
10. Daemen, A., Griffith, O.L., Heiser, L.M., Wang, N.J., Enache, O.M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J.E., Griffith, M., et al. (2013). Modeling precision treatment of breast cancer. Genome Biol. *14*, R110.
11. Staunton, J.E., Slonim, D.K., Coller, H.A., Tamayo, P., Angelo, M.J., Park, J., Scherf, U., Lee, J.K., Reinhold, W.O., Weinstein, J.N., et al. (2001). Chemosensitivity prediction by transcriptional profiling. Proc. Natl. Acad. Sci. USA *98*, 10787–10792.
12. Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., and Fine, H.A. (2011). Predicting in vitro drug sensitivity using Random Forests. Bioinformatics *27*, 220–224.
13. Cortés-Ciriano, I., van Westen, G.J., Bouvier, G., Nilges, M., Overington, J.P., Bender, A., and Malliavin, T.E. (2016). Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics *32*, 85–95.
14. Ammad-ud-din, M., Georgii, E., Gönen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., Poso, A., and Kaski, S. (2014). Integrative and personalized QSAR

analysis in cancer by kernelized Bayesian matrix factorization. J. Chem. Inf. Model. *54*, 2347–2359.

15. Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X.S. (2015). Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. PLoS Comput. Biol. *11*, e1004498.

16. Kim, S., Sundaresan, V., Zhou, L., and Kahveci, T. (2016). Integrating Domain Specific Knowledge and Network Analysis to Predict Drug Sensitivity of Cancer Cell Lines. PLoS ONE *11*, e0162173.

17. Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer *17*, 513.

18. Stanfield, Z., Coşkun, M., and Koyutürk, M. (2017). Drug response prediction as a link prediction problem. Sci. Rep. *7*, 40321.

19. Supahvilai, C., Bertrand, D., and Nagarajan, N. (2018). Predicting Cancer Drug Response Using a Recommender System. Bioinformatics. Published online June 1, 2018. https://doi.org/10.1093/bioinformatics/bty452.

20. Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. *17*, 734–749.

21. Asati, V., Mahapatra, D.K., and Bharti, S.K. (2016). PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents: Structural and pharmacological perspectives. Eur. J. Med. Chem. *109*, 314–341.

22. Petrelli, F., Ghidini, M., Lonati, V., Tomasello, G., Borgonovo, K., Ghilardi, M., Cabiddu, M., and Barni, S. (2017). The efficacy of lapatinib and capecitabine in HER-2 positive breast cancer with brain metastases: A systematic review and pooled analysis. Eur. J. Cancer *84*, 141–148.

23. Zhao, M., Howard, E.W., Parris, A.B., Guo, Z., Zhao, Q., Ma, Z., Xing, Y., Liu, B., Edgerton, S.M., Thor, A.D., and Yang, X. (2017). Activation of cancerous inhibitor of PP2A (CIP2A) contributes to lapatinib resistance through induction of CIP2A-Akt feedback loop in ErbB2-positive breast cancer cells. Oncotarget *8*, 58847–58864.

24. Santra, T., Roche, S., Conlon, N., O'Donovan, N., Crown, J., O'Connor, R., and Kolch, W. (2017). Identification of potential new treatment response markers and therapeutic targets using a Gaussian process-based method in lapatinib insensitive breast cancer models. PLoS ONE *12*, e0177058.

25. Solit, D.B., Garraway, L.A., Pratilas, C.A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J.M., She, Y., Osman, I., et al. (2006). BRAF mutation predicts sensitivity to MEK inhibition. Nature *439*, 358–362.

26. Byron, S.A., Chen, H., Wortmann, A., Loch, D., Gartside, M.G., Dehkhoda, F., Blais, S.P., Neubert, T.A., Mohammadi, M., and Pollock, P.M. (2013). The N550K/H mutations in FGFR2 confer differential resistance to PD173074, dovitinib, and ponatinib ATP-competitive inhibitors. Neoplasia *15*, 975–988.

27. Pardo, O.E., Latigo, J., Jeffery, R.E., Nye, E., Poulsom, R., Spencer-Dene, B., Lemoine, N.R., Stamp, G.W., Aboagye, E.O., and Seckl, M.J. (2009). The fibroblast growth factor receptor inhibitor PD173074 blocks small cell lung cancer growth in vitro and in vivo. Cancer Res. *69*, 8645–8651.

28. Stavridis, M.P., Lunn, J.S., Collins, B.J., and Storey, K.G. (2007). A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification. Development *134*, 2889–2894.

29. Koziczak, M., Holbro, T., and Hynes, N.E. (2004). Blocking of FGFR signaling inhibits breast cancer cell proliferation through downregulation of D-type cyclins. Oncogene *23*, 3501–3508.

30. Huang, D.W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. *8*, R183.

31. Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 230–237.

32. Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. *7*, 76–80.

OMTN, Volume *13*

Supplemental Information

Anti-cancer Drug Response Prediction Using Neighbor-Based Collaborative Filtering with Global Effect Removal

Hui Liu, Yan Zhao, Lin Zhang, and Xing Chen

Supplemental Figures

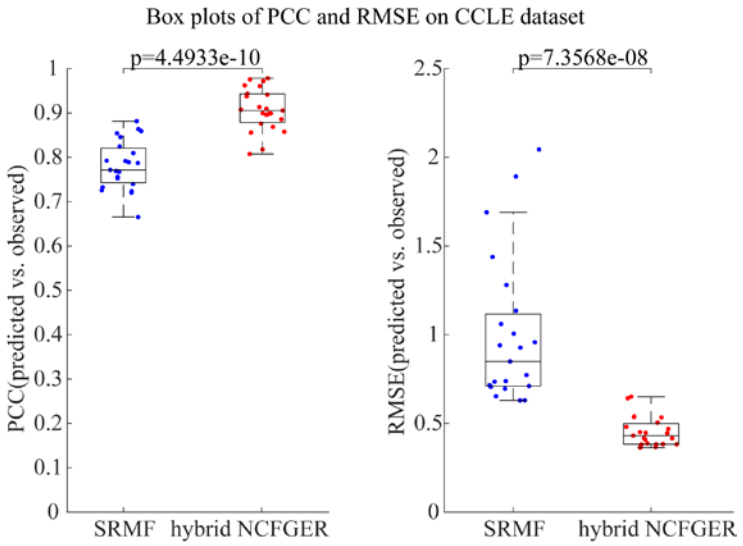

Fig. S1. Box plots of SRMF and our proposed method on CCLE dataset with respect to different evaluation metrics.

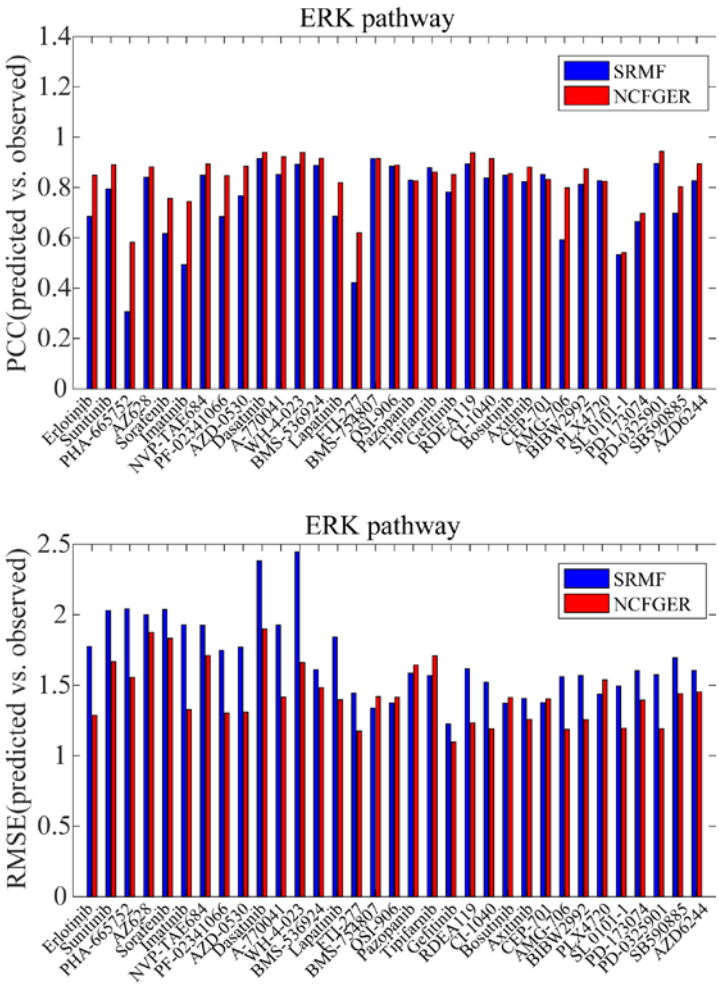

Fig.S2. Prediction performance comparisons of SRMF and NCFGER for the drug targeting genes in ERK pathway with respect to PCC as well as RMSE in CCLE dataset.

Supplemental Tables

Table S1. The comparison results between hybrid NCFGER with different similarity definition and SRMF obtained under 10-fold cross validation on GDSC dataset

Methods		Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R	Drug-averaged PCC	Drug-averaged RMSE
NCFGER	*COEF*	0.78(±0.13)	1.58(±0.34)	0.70(±0.14)	1.27(±0.26)
	RPCC	0.81(±0.11)	1.42(±0.29)	0.73(±0.13)	1.18(±0.24)
	MRPCC	0.81(±0.11)	1.42(±0.29)	0.73(±0.13)	1.18(±0.24)
SRMF		0.71(±0.15)	1.73(±0.46)	0.62(±0.16)	1.43(±0.36)

Table S2. The comparison results between hybrid NCFGER with different similarity definition and SRMF obtained under 10-fold cross validation on CCLE dataset

Methods		Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R	Drug-averaged PCC	Drug-averaged RMSE
NCFGER	*COEF*	0.89(±0.05)	0.47(±0.09)	0.85(±0.07)	0.39(±0.07)
	RPCC	0.91(±0.05)	0.45(±0.08)	0.86(±0.06)	0.38(±0.06)
	MRPCC	0.91(±0.05)	0.45(±0.08)	0.86(±0.06)	0.38(±0.06)
SRMF		0.78(±0.07)	0.74(±0.23)	0.71(±0.09)	0.57(±0.18)

Table S3. The comparison results of both methods obtained under 10-fold cross validation on CCLE dataset

Methods		Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R	Drug-averaged PCC	Drug-averaged RMSE
NCFGER	Hybrid	0.91(±0.05)	0.45(±0.08)	0.86(±0.06)	0.38(±0.06)
	cellline-based	0.85(±0.08)	0.51(±0.06)	0.78(±0.09)	0.47(±0.08)
	Drug-based	0.80(±0.07)	0.63(±0.20)	0.73(±0.09)	0.52(±0.16)
SRMF		0.78(±0.07)	0.74(±0.23)	0.71(±0.09)	0.57(±0.18)