Quantitative Biology 2021, 9(4): 426–439
https://doi.org/10.15302/J-QB-021-0259

RESEARCH ARTICLE

Interpretable prediction of drug-cell line response by triple matrix factorization

Xiao-Ying Yan[1,2], Shao-Wu Zhang[1,*], Siu-Ming Yiu[3], Jian-Yu Shi[4,*]

[1] Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2] College of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
[3] Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China
[4] School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: zhangsw@nwpu.edu.cn; jianyushi@nwpu.edu.cn

Received November 23, 2020; Revised January 24, 2021; Accepted February 22, 2021

***Background*: One of the challenges in personalized medicine is to determine specific drugs and their dosages for patient individuals who are undergoing a common disease. The technique of cell lines provides a safe approach to capture the drug responses of patient individuals when given specific drugs with varied dosages. However, it is still costly to determine drug responses in cells w.r.t dosages by biological assays. Computational methods provide a promising screening to infer possible drug responses in the cells of patient individuals on a large scale. Nevertheless, existing computational approaches are insufficient to interpret the underlying reason for drug responses.**

***Methods*: In this work, we propose an interpretable model for analyzing and predicting drug responses across cell lines. The proposed model bridges drug features (*e.g.,* chemical structure fingerprints), cell features (*e.g.,* gene expression profiles), and drug responses across cells (measured by IC50) by a triple matrix factorization (TMF), such that the underlying reason for drug responses in specific cells is possibly interpreted.**

***Results*: The comparison with state-of-the-art computational approaches demonstrates the superiority of our TMF. More importantly, a case study of drug responses in lung-related cell lines shows its interpretable ability to find out highly occurring drug substructures, crucial mutated genes, as well as significant pairs between substructures and mutated genes in terms of drug sensitivity and resistance.**

***Conclusion*: TMF is an effective and interpretable approach for predicting cell lines responses to drugs, and can dig out crucial pairs of chemical substructures and genes, which uncovers the underlying reason for drug responses in specific cells.**

Keywords: drug response; drug sensitivity; drug resistance; triple matrix factorization

Author summary: Personalized medicine aims to identify the cause at the molecular level for a given patient and then tailor treatment for individual. Triple matrix factorization (TMF) algorithm can bridge chemical substructures of drugs and genes of cell lines to the associated response values by using the bi-projection matrixΘ, which is an effective and interpretable approach for predicting cell lines responses to drugs.

INTRODUCTION

Personalized medicine aims to identify the cause at the molecular level for a given patient and then tailor treatment for individual [1]. However, the identification of drug response patterns needs many cancer patients under the treatment of numerous drugs, it is often not feasible in cost and in morality. Recently, the technique of cultured human cell lines in place of patient samples wins a lot of concerns [2]. With this technique, the responses of cell lines to drugs are easily determined by log-transformed IC50 values, which is defined as the drug concentration required for 50% inhibition of cell growth. For example, NCI-60 project generates an ensemble of 60

cancer cell lines and their responses to more than 100,000 chemical compounds with a low cost [3]. More studies, such as Cancer Cell Line Encyclopedia (CCLE) [4] and Genomics of Drug Sensitivity in Cancer (GDSC) [5], have been developed to facilitate pharmacogenomic researches by integrating diverse genomic data (*e.g.*, gene expression, chromosomal copy number variation). These studies generate the response data involving more drugs and more cell lines. Therefore, it is promising to develop computational approaches to preliminarily screen response data across cell lines on a large scale. More importantly, computational approaches can provide a new sight to elucidate the response mechanism of anti-cancer drugs [6].

Current computational approaches for predicting drug-cell responses can be roughly classified into three types: regression-based, network inference-based and matrix factorization-based. By assuming that drugs are independent, regression-based approaches [6–9] apply Elastic Net or LASSO to build a regression between gene expressions and response values in terms of cell lines w.r.t drugs. However, drugs often are depended because of 'me-too' phenomenon. In addition, they neglect drug features.

By integrating the information about drugs and cell lines to construct a drug-cell line heterogeneous network, network inference-based approaches use direct neighbors or higher-order neighbors of nodes to predict drug-cell line responses. For example, DLN [10] constructs a weighted prediction model based on direct neighbors in both a drug similarity network model and a cell line similarity network model. By integrating gene expressions, drug targets, protein-protein interactions (PPI) and chemical structure information together to build a heterogeneous network, HNMDRP [11] applies Information Flow-based algorithm [12] on this network to predict drug responses of cell lines. By integrating response pattern network and molecular profile network, NRL2DRP [13] utilizes network representing learning to extract representation vectors from the response network, and applies SVM to predict drugs-cell lines responses. However, as the construction of a network requires a similarity matrix, derived from drug/cell line features, these approaches cannot confirm, which feature contributes to understand or even uncover the mechanism of drug responses.

Matrix factorization-based approaches (*e.g.*, KBMF [14] and SRMF [15]) map drug properties and cell line properties into a common latent space, and suppose that there is a higher probability of sensitivity response association between drugs and cell lines if they are closer to each other in this latent space. KBMF [14] integrates multiple drug/cell line similarity matrices by multi-kernel learning, and apply Kernelized Bayesian Matrix Factorization to predict response associations between drugs and cell lines. Based on the suppose that similar drugs and similar cell lines exhibit similar drug responses, SRMF [15] constructs a share latent space by the prior knowledge on similarities of drugs and cell lines, and proposes a similarity-regularized matrix factorization for such a predicting task. However, like network inference-based approaches, because existing matrix factorization-based approaches require similarity matrices (or networks), they discard some information among features, which contribute to interpret drug responses.

In this present paper, we propose a novel Triple Matrix Factorization (TMF) to improve the prediction of drug responses and uncover significant features of both drugs and cell lines, which contribute to drug responses. TMF associates drug features, cell line features with observed drug-cell line responses by a bi-projection matrix, of which each entry indicates how important a pair of a drug feature and a cell line feature contributes to a corresponding response.

RESULTS AND DISCUSSION

Datasets

Recommended by other state-of-the-art approaches [11,13–16], we collected the response entries of drugs to cell lines from the Genomics of Drug Sensitivity in Cancer project (GDSC) [5], which consists of 256 test drugs and a panel of 1001 cancer cell lines in total. Since aiming to dig out both drug chemical substructures and genomic properties which may contribute to understand drug responses to cell lines, we only kept the drugs having PubChem fingerprint and the cell lines having genomic profiles except for their response entries. Thus, our dataset includes 183 drugs, 962 cell lines and 142,451 response entries between them.

To be fed into our approach, the response entries were organized into a 183×962 response matrix Y, which contains partially observed (80.9%) response entries valued by IC50. In addition, the terms of drug sensitivity and drug resistance are concerned in specific scenarios. Usually, the lower IC50, the more sensitivity of a cell line to a given drug; the larger IC50, the more resistance. In order to fit these scenarios, these response entries can be binarized by a specific threshold in terms of drug [6,11]. Moreover, these drugs were represented by PubChem fingerprints [17] and the cell lines are characterized by a large scale of genomic expression profiles, which can be downloaded from http://www.cancerrxgene.org/. In accordance, the drugs and the cell lines were stacked into a 183×881 drug feature matrix F_d and a 962×16383 cell line feature matrix F_c.

Preprocessing

Before comparing our approach to other state-of-the-art predictive approaches, we investigate how both the normalization and the parameters influence the predicting results. Because drug response values vary significantly causes computational difficulties, they should be normalized before they are handled further. We consider two different normalization strategies. The first one is presented in this work (Formula 1), which consider the fact that the definition of drug sensitivity/resistance is specific to drug themselves but not to cell lines. The comparison to the raw responses shows that all the response values of the drugs fall into roughly same numerical ranges (Fig. 1). The drugs, which target the PI3K/MTOR signaling pathway, was taken as an illustration. For example, before the normalization, the minima, maxima, mean and standard deviation of response values for GSK2126458 are -8.2429, 1.6761, -2.772 and 1.891, respectively, while the statistical values of PF-4708671 are -0.3539, 6.898, 3.553 and 1.348, respectively. In contrast, after the normalization, the minima, maxima, mean and standard deviation of response values for GSK2126458 are -2.893, 2.352, 0 and 1, respectively, while the statistical values of PF-4708671 are -2.899, 2.482, 0 and 1, respectively.

Another normalization strategy was presented by [15], which maps all the response entries into $[-1,1]$ by

$$y_{ij} = \frac{y_{ij}}{\max(abs(y))},$$

where $\max(abs(y))$ is the absolute maxima among all response values.

After this normalization, all the response values of the drugs fall into roughly the same distribution as the original response values but in range $[-1,1]$ (Fig. 2). The minima, maxima, mean and standard deviation of the normalization response values for GSK2126458 are -0.6439, 0.1309, -0.2165 and 0.1477, respectively, while the statistical values of PF-4708671 are -0.0276, 0.5388, 0.2775 and 0.1053, respectively.

We further compared the predictions on the raw response matrix (denoted as *Y1*), the response matrix normalized by [15] (denoted as *Y2*), and the response matrix normalized by this work (denoted as *Y*) (Table 1). The results show that normalization strategies can improve the performance significantly, and our strategy can obtain better *PCC* and *PCC_S/R* values.

Moreover, during performing the prediction, we investigated how the parameters of TMF, including the latent dimension k and four regularization parameters λ_u, λ_v,λ_d,λ_c, influence the results. Since both large *PCC* and small *RMSE(E)* reflect a good prediction, we combined them into $\rho 1 = \frac{PCC}{E}$ and $\rho 2 = \frac{PCC\_S/R}{E\_S/R}$ as the metrics to find the best value of a parameter. As descripted in the section of "Compared with other state-of-the-art approaches", $\rho 1$ is the drug-averaged quotient of *PCC* divided by *E* based all the response values; $\rho 2$ is the drug-averaged quotient of *PCC_S/R* divided by *E_S/R* based the response values of sensitive and resistant cell lines.

First, with fixing the values of four regularization parameters with 1, the influence of k to the performance of TMF is tested by tuning its value from a value list of {1, 2, 3, 4, 5, 10, 15, 20,...,100}. The predicting performance of TMF against k (measured by$\rho 1$, $\rho 2$) shows that $\rho 1$ and $\rho 2$

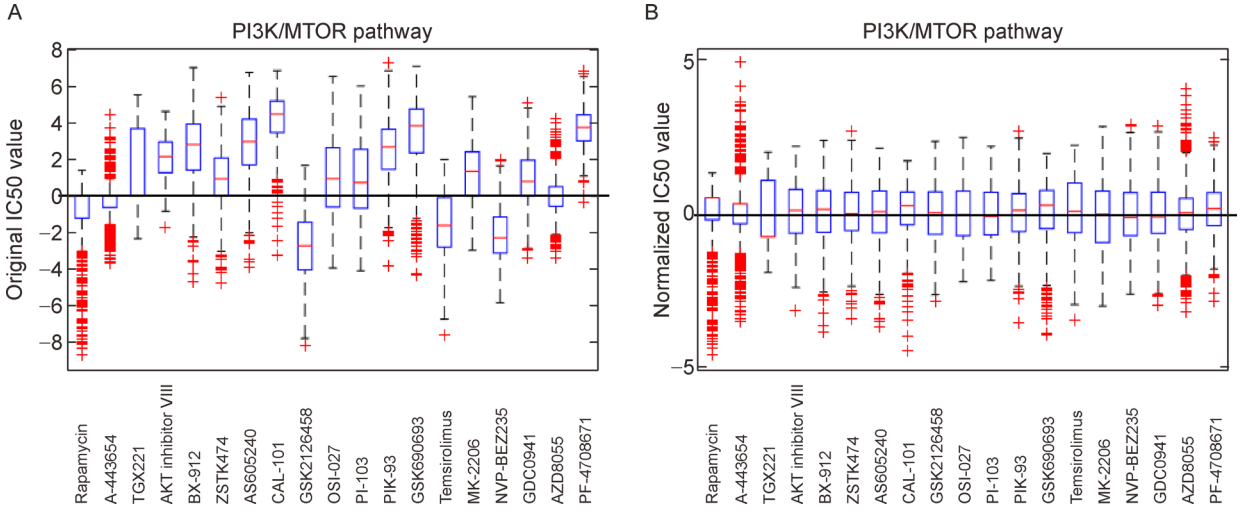

Figure 1. Box plot of normalized IC50 value for part drugs in GDSC by using Formula 1. (A) Before normalized. (B) After normalized. Each box response for a drug, and there are five values to depict the response value about a drug, they are maximum value, 75th percentile, 50th percentile, 25th percentile and minimum value. The red line in box is for median, and the long black line is for the means of all drugs.

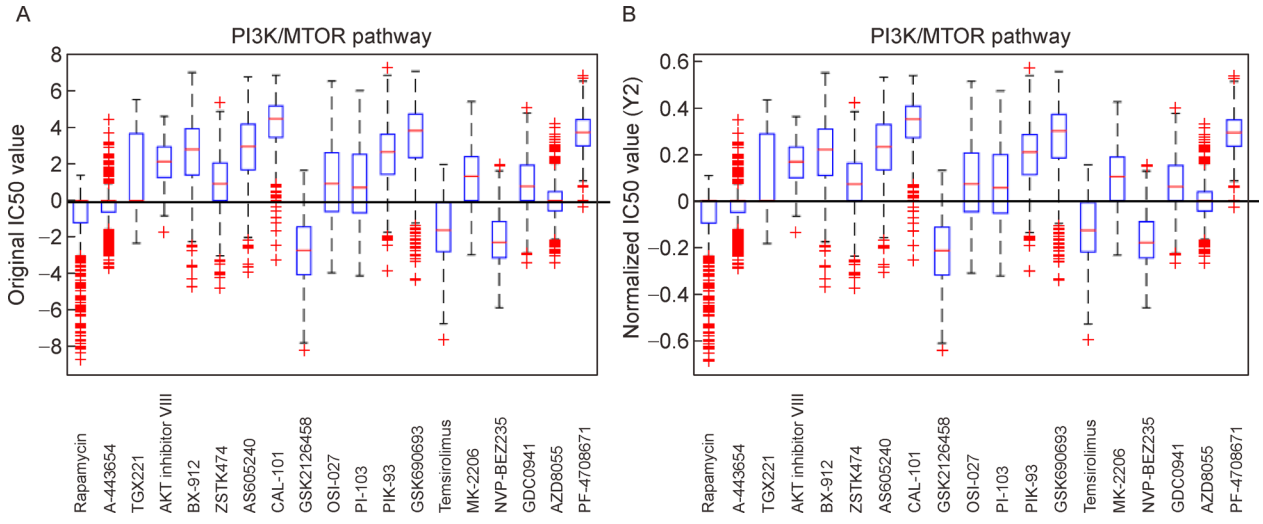

Figure 2. Box plot of normalized IC50 value for part drugs in GDSC by using strategy in Ref [17]. (A) Before normalized. (B) After normalized.

Table 1 Results of TMF_y1, TMF and TMF_y2 on GDSC dataset with 10-CV test

Normalization	*PCC_S/R*	*PCC*
Y1	0.52 (±0.01)	0.43 (±0.01)
Y2	0.49 (±0.005)	0.65 (±0.009)
Y	0.80 (±0.006)	0.72 (±0.005)

rise quickly when $0 < k \leqslant 20$, while $\rho 1$ and $\rho 2$ rise slowly when $20 < k \leqslant 50$. When $k > 50$, $\rho 1$ and $\rho 2$ are almost no change (Fig. 3). Thus, we picked up $k = 50$ accounting for the best prediction performance from test as the value of the latent dimension in our TMF. For different issues and different datasets, the value of k may be different.

For those regularization parameters, the values of λ_u, λ_v, λ_d and λ_c were tuned from the list {0.005, 0.05, 0.5, 1}. As λ_u and λ_v are the regularization parameters for latent response matrices of drugs and cell lines, respectively. While λ_d and λ_c are the regularization parameters for regression coefficient matrices of drugs and cell lines, respectively. For the convenience of calculation, we set same value for λ_u and λ_v, λ_d and λ_c. In terms of $\rho 1$, the performance w.r.t. the searching grid

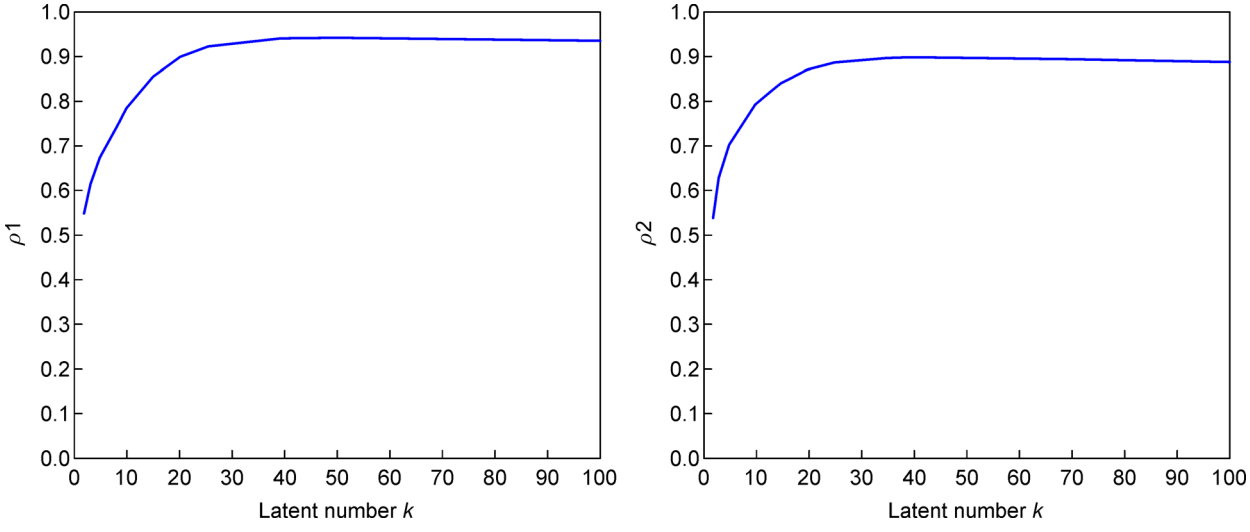

Figure 3. The relationship between the parameter k and $\rho 1$, $\rho 2$ for TMF algorithm on GDSC dataset. $\rho 1$ is the drug-averaged quotient of *PCC* divided by *RMSE* based all the response values for a series of k; ρ_2 is the drug-averaged quotient of *PCC* divided by *RMSE* based the response values of sensitive and resistant cell lines for a series of k.

Table 2 The performance of TMF in the case of tuning λ_u,λ_v,λ_d,λ_c on GDSC dataset

λ_u,λ_v	λ_d λ_c			
	1	0.5	0.05	0.005
1	0.9986	1.0402	1.0326	1.0207
0.5	0.9967	1.0387	1.0241	1.0070
0.05	0.9921	1.0346	1.0127	0.9832
0.005	0.9912	1.0341	1.0111	0.9793

Note: The performance is measured by $\rho 1$. The regularization parameters λ_u,λ_v,λ_d,λ_c are tuned from the list {0.005, 0.05, 0.5, 1}. $\rho 1$ is the drug-averaged quotient of *PCC* divided by E based all the response values for a series of λ_u,λ_v,λ_d,λ_c.

listed in Table 2 shows the trivial differences. Besides, we also test different values of λ_u, λ_v, λ_d and λ_c, the results are shown in Supplementary Table S1. From the results, we can see that with different values of λ_u and λ_v, the prediction performance is almost same. With different values of λ_d and λ_c, the prediction performance is also almost same. Here, we set $\lambda_u = 1$, $\lambda_v = 1$, $\lambda_d = 1$ and $\lambda_c = 1$.

Analysis of data noises and incompleteness

This section shows two experiments to investigate how TMF is robust to noisy data and incomplete data respectively. First, to simulate noisy data, we added the Gaussian noise N to the original response matrix Y, such that Y is changed as $Y\_noise = Y + (noise\_level) * N$, where *noise_level* is a scale value which is tuned from 0.1 to 0.5 with the interval of 0.05. Since both *PCC* and small root mean squared error (*RMSE*(*E*)) reflect a good prediction, we used $\rho 1 = \frac{PCC}{E}$ as the metric to evaluate the algorithm performance. The performance against the noise level (shown in Fig. 4) illustrates that the performance is robust to the noise when *noise_level* $< 20\%$, but degrades significantly when *noise_level* $> 20\%$.

Secondly, for the analysis of incomplete data, we masked the observed values in the response matrix Y from 0% to 30% with the interval of 5%. Because there are only 80.9% observed values in the original Y, thus the real rate of incompleteness is from 19% to 49% accordingly. The performance against the incomplete rate (shown in Fig. 5) illustrates that TMF is robust to the incomplete data.

In summary, the proposed TMF is robust to the noises and incompleteness of data.

Compared with other state-of-the-art approaches

With our normalization strategy and the best parameter, we compared the performance of TMF with two state-of-the-art approaches of KBMF [14] and SRMF [15]. The former is a kernelized Bayesian matrix factorization algorithm, which used four kinds of information for drugs, and three kinds of information for cell lines. Each feature is introduced as a similarity matrix, then three main parts are included in KBMF algorithm: (1) kernel-based nonlinear dimensionality reduction; (2) multiple kernel learning; (3) Bayesian matrix factorization. SRMF is a similarity-regularized matrix factorization, which incorporated similarities of drugs and of cell lines simultaneously. For KBMF and SRMF, the low dimensionality K was set as 45, three regularization parameters λ_l, λ_d, λ_c of SRMF were selected from $\{2^{-3}, ..., 2^3\}$, $\{2^{-5}, ..., 2^1, 0\}$ and $\{2^{-5}, ..., 2^1, 0\}$ respectively.

Both require similarity matrices as approach inputs. Thus, to make a fair comparison with them under 10-CV, we calculate pairwise similarities between drugs (denoted as $S_{dd} = (s_{ij}^d)_{n \times n}$) by Jaccard similarity based on PubChem fingerprints, and pairwise similarities between cell lines (denoted as $S_{cc} = (s_{ij}^c)_{m \times m}$) by Pearson correlation based on their gene expression profiles [11]. The comparison of drug response prediction shows that TMF is significantly superior to both KBMF and SRMF in terms of Pearson correlation coefficient (*PCC*), *E*, *PCC_S/R* and *E_S/R* (Table 3).

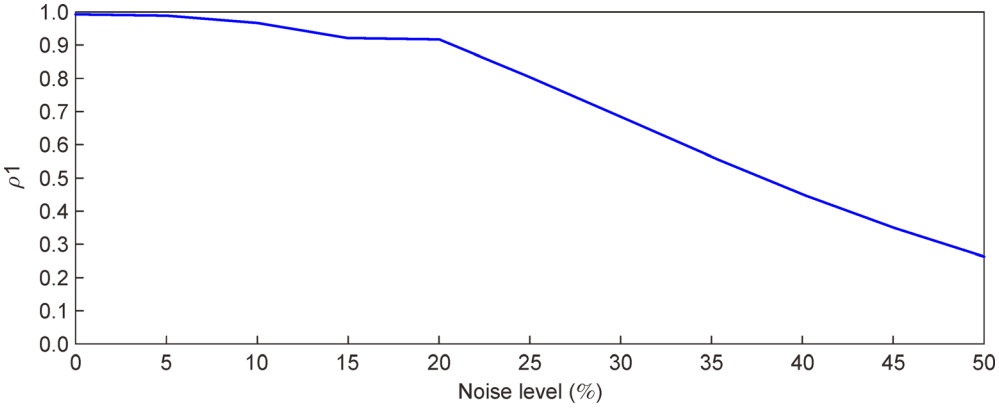

Figure 4. The performance of TMF against the noise level.

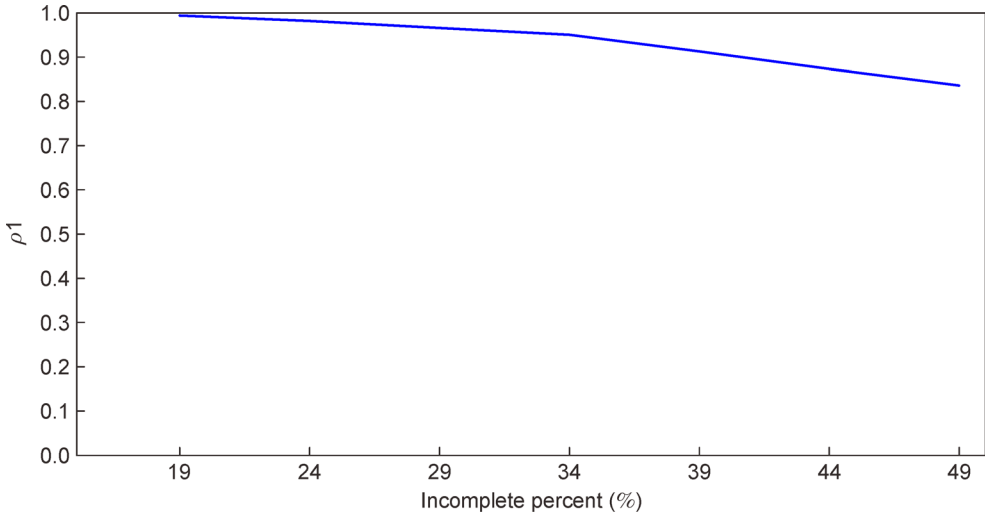

Figure 5. The performance of TMF against the incomplete rate.

Table 3 Results of KBMF, SRMF and TMF on GDSC dataset with 10-CV test

Methods	*PCC_S/R*	*E_S/R*	*PCC*	*E*
KBMF	0.59 (±0.14)	2.00 (±0.51)	0.49 (±0.14)	1.59 (±0.42)
SRMF	0.73 (±0.008)	1.33 (±0.01)	0.60 (±0.009)	1.49 (±0.01)
TMF	**0.80** (±0.006)	**0.93** (±0.003)	**0.72** (±0.005)	**0.72** (±0.002)

As a case study, we selected the response prediction of 19 drugs targeting genes in the PI3K/MTOR pathway. The results indicate that TMF obtained higher *PCC* (Fig. 6) and lower *RMSE* (Fig. 7) for these 19 drugs. In addition, four out of the drugs (AR-42,CUDC-101, Belinostat and CAY10603) were selected to illustrate the correlation between observed response values and predicted values (Fig. 8).

In addition, we adopted an extra drug response dataset, CCLE (Cancer Cell Line Encyclopedia), consisting of 23 drugs and 491 cancer cell lines, to further test the performance of our TMF. The comparison results on CCLE show that TMF is superior to both KBMF and SRMF in terms of both *PCC* and *PCC_S/R* (Supplementary Table S2).

Except for matrix factorization-based methods, we also compared TMF with one regression based methods MVLR [7] and two network inference-based approaches

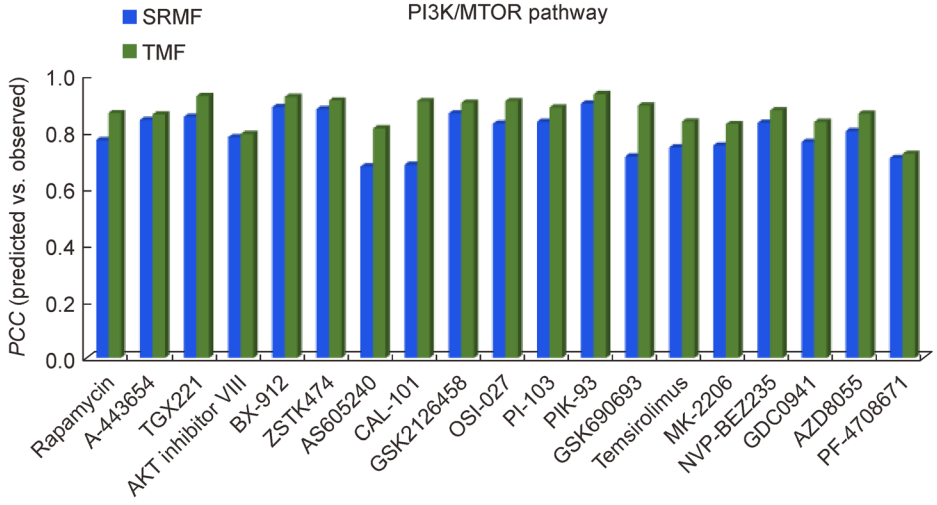

Figure 6. *PCC* result comparisons of SRMF and TMF for drugs targeting genes in the PI3K/MTOR pathway.

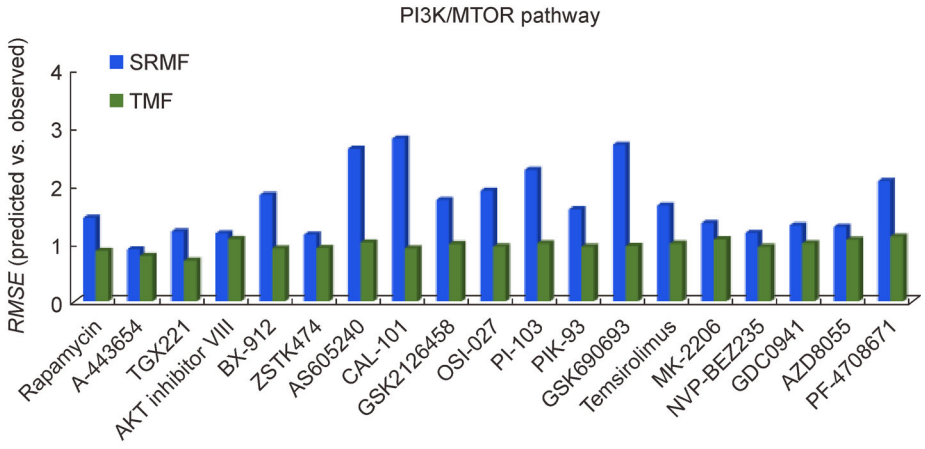

Figure 7. ***RMSE*** **result comparisons of SRMF and TMF for drugs targeting genes in the PI3K/MTOR pathway.**

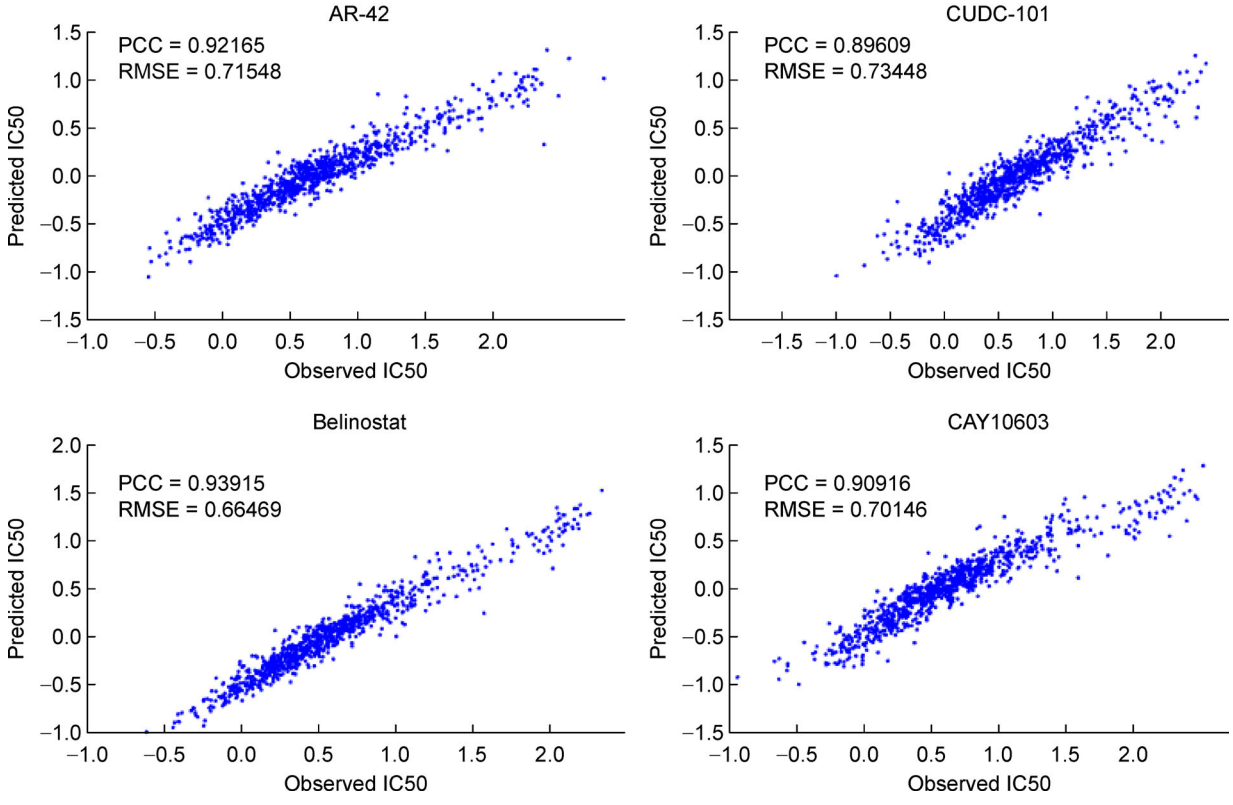

Figure 8. Scatter plots of observed and predicted drug responses for four drugs AR-42, CUDC-101, Belinostat and CAY10603.

of HNMDRP [11] and NRL2DRP [13]. MVLR provides a Bayesian multi-view multi-task linear regression while both HNMDRP and NRL2DRP only treat the current task as a classification problem. HNMDRP constructs a heterogeneous network and utilizes Information Flow-based algorithm to predict drug response. NRL2DRP applies Network Representation Learning to perform the prediction.

To make a fair comparison, we adopted the same datasets, the same cross-validation schemes and the same performance metrics as those in these methods. Thus, a small set extracted from GDSC, Leave-one-out cross-

validation (LOOCV) and *PCC* were used in the comparison with MVLR, while GDSC, 5-fold cross-validation (5CV) and the area under the receiver operating characteristic curve (AUC) were uses in the comparison with both HNMDRP and NRL2DRP. In terms of *PCC* on the small set extracted from GDSC, TMF achieved 0.6870, whereas MVLR achieved 0.375. In terms of AUC, TMF achieves 0.8256, whereas HNMDRP and NRL2DRP achieve 0.7391 and 0.7882 respectively. Obviously, TMF outperforms other approaches significantly. The comparison shows again that TMF outperforms all of MVLR, HNMDRP and NRL2DRP.

Case study of interpretable property of TMF

Both gene expression and drug responses are usually tissue-specific [12]. Our dataset contains 19 tissue types of cell lines, of which the most major tissue is lung. Thus, we selected the cell lines about lung tissue and the corresponding drugs for a case study to identify the dominant feature pairs, which contribute to understand how cell lines response to drugs in terms of drug sensitivity and drug resistance. Totally, there are 191 lung-related cell lines and 177 associated drugs. On the other side, mutation genes are associated with sensitivity to drugs [6], so we download the cell line mutation genes from COSMIC [18], and use them to explain the important feature genes for cell lines.

In the case study, we attempt to answer three questions in terms of cell line feature and/or drug features as follows: (1) why a cell line shows a significant response to a drug; (2) what the substructures frequently occurring in all drugs are and why a set of drugs triggers the response of a specific cell line in terms of drug feature (substructures); (3) what the important mutated genes across all the cell lines are and why a set of cell lines show significant responses to a specific drug.

Technically, the drug-cell line response matrix of lung tissue is denoted as $Y_1 = (y_{ij})_{177\times191}$ (here, subscript 1 is used to differ from the above drug-cell line response matrix Y), in which, $y_{ij} = +1$ and $y_{ij} = -1$ represent sensitive and resistant responses respectively. Accordingly, the drug feature matrix and the cell feature matrix are denoted as a 177×881 matrix F^{d1} and a 191×16383 matrix F^{c1}.

After performing TMF on, we obtained the regression coefficient matrix B^*_{d1} and B^*_{c1}, which depict the correlation between the feature matrix and the latent response matrix. Sequentially, three significant matrices can be derived from B^*_{d1} and B^*_{c1} , including the bi-projection matrix $\Theta^*_1 = (B^*_d(B^*_{c1})^T)_{p\times q}$, the drug projection matrix $\Theta_{d1} = \Theta^*_1 F^T_{c1}$, and the cell line projection matrix $\Theta_{c1} = (F_{d1}\Theta^*_1)^T$. The following sections shall exhibit how these matrices contribute to answer the abovementioned questions.

Significant feature pairs

We answered the first question by the entries in $\Theta^*_1 = (B^*_d(B^*_{c1})^T)_{p\times q}$, which build bridge between the feature pairs and the response values. Each positive/negative entry θ_{ij} in Θ indicate how well the corresponding pair of the i-th drug feature and the j-th cell line feature contribute to the sensitive/resistant drug-cell line response respectively. As the feature number of drugs and cell lines are very large, most entries in Θ^*_1 is very small. We define the entries as the important entry, whose absolute values are greater than 1 and two subscripts account for the significant feature pair of a drug and a cell line respectively. As an illustration, the top positive {'O = C-O-C:C':'GUCA1A'} indicates 'O = C-O-C:C' is the most frequent substructure in the drugs inducing the sensitive response of the lung-related cell lines which contains the significant expression of gene 'GUCA1A'. In contrast, the top negative feature pair {'N#C-C = C': 'PAGE5'} indicates N#C-C = C' is the most frequent substructure in the drugs triggering the resistive response of the lung-related cell lines which contains the significant expression of gene 'PAGE5'.

Significant drug substructure features

We answered the second question by the drug projection matrix $\Theta_{d1} = \Theta^*_1 F^T_{c1}$, each column represents a substructure. The question is split into two parts to answer. One is the substructures frequently occur in all drugs (denoted as FD1). The other is the important substructure features frequently occur in drug groups, which are sensitive or resistant to a specific cell line (denoted as FD2).

(1) To find out FD1, we counted the occurrence of each substructure in positive entries and negative entries of Θ_{d1} respectively. Highly occurring substructures in positive entries (Table 4) and those in negative entries (Table 5) play an important role for designing the drugs, to which lung-related cell lines are sensitive and resistant respectively. For example, the drugs having'C(~N)(:C) (:C)'and 'O-C-C-C-C' are sensitive and resistant to lung-related cell lines respectively.

(2) As each column in Θ_{d1} is for a cell line, the entry θ^d_{ij} represents the importance of the i-th substructure of drugs on their response with cell line c_j. Again, positive values are for sensitive responses while negative values are for resistant responses. Based on this, the cell line LU-65 was taken as an example. There are 176 drugs resistant to LU-65, whereas GSK-1904529A is the only drug sensitive to LU-65. According to the values of the entries accounting

Table 4 Highly occurring substructures of drugs for sensitive to Lung tissue-related cell lines

Rank	Substructures	Group of PubChem fingerprint
1	> = 32 C	G1:Hierarchic element counts
2	C(~N)(:C)(:C)	G4:Simple atom nearest neighbors
3	N(~C)(:C)(:C)	G4:Simple atom nearest neighbors
4	> = 8 N	G1:Hierarchic element counts
5	N-S-C:C	G6:Simple SMARTS pattern
6	O = C-N-C = O	G6:Simple SMARTS pattern
7	N = C-C-[#1]	G6:Simple SMARTS pattern
8	N-C = N-[#1]	G6:Simple SMARTS pattern
9	C(-N)(= C)	G5 Detailed atom neighborhoods
10	N#C-C = C	G6 Simple SMARTS patterns

for LU-65 in Θ_{d1}, the top 5 substructures (*i.e.* '> = 8N', 'O = C-C:C-O', '> = 3 saturated or aromatic nitrogen-containing ring size 6', '> = 32 C' and 'CC1CCC(O)CC1') are extracted. We found that all of them appear in the chemical structure of GSK-1904529A, but seldom occur in those of resistant drugs.

Therefore, it is believed that these five substructures are the most important substructure in the drugs inducing the response to cell line LU-65. Moreover, FD2 contributes to identify drug features, which can guide to design drugs having a good efficacy for a specific cell line.

Significant gene features

We answer the third question by the cell line projection matrix $\Theta_{c1} = (F_{d1}\Theta^*)^T$, where each row reflects the importance of a gene. Again, it can be answered by two parts. One is the important genes, which are mutation genes in all cell lines (denoted as FC1); the other is the significant mutation genes in some cell lines, which are sensitive or resistant to a specific drug (denoted as FC2).

(1) To find out FC1, we counted important genes via positive entries of Θ_{c1} and found out PCDH8, BMP8B, PAGE5, TCEANC, MKRN3, PABPC4, ZIC3, CCDC39, TMEM163 and HIST1H2AI. Similarly, by negative entries of Θ_{d1}, we picked up ten important genes, including HCN2, SIX1, C17orf50, PRSS57, OR51D1, PNLIP, DDX25, PTPRZ1, SIX4 and SLFN11. These two groups of genes can be the indicator of lung-related cell lines tending to have drug sensitivity and drug resistance respectively.

(2) As each column in Θ_{c1} is for a drug, the entry θ_{ij}^c represents the importance of the i-th gene on their response with drug d_j. Again, positive values are for sensitive response, while negative values are for resistant response. Based on this, we select five drugs (Sunitinib, Crizotinib, CGP-082996, CMK and GW843682X) as an example. After analysis, we found these five drugs all sensitive to cell line LC-2-ad but resistant to 63 cell lines. They are DMS-114, NCI-H187, LU-139, SBC-1, *etc*.

After extracting the important genes (PCDH8, GRIA4 and CDO1) for these drugs from Θ_{c1}, we find that these genes all mutate in cell line LC-2-ad, but PCDH8 is not mutated in 63 resistant cell lines, GRIA4 mutated in only 4 cell lines(4/63) and CDO1 mutated in only 3 cell lines (3/63). Therefore, we believe that FC2 can genetically explain why some cell lines are sensitive to a specific drug but others are resistant to it. On the other hand, it also explains why LC-2-ad is sensitive to these five drugs and other cell lines are resistant to these drugs.

CONCLUSIONS

To obtain an effective and interpretable approach for predicting cell lines responses to drugs, we bridge chemical substructures of drugs, genes of cell lines, and drug-cell line responses by a novel model, Triple Matrix Factorization (TMF). The comparison with other state-of-the-art matrix approaches under 10-CV demonstrate its superiority in terms of both *PCC* and *RMSE*. More importantly, this proposed approach can dig out crucial pairs of chemical substructures and genes, which uncover the following questions: (1) how often does a pair of substructure-gene occur in sensitive responses or resistant responses? (2) What are the substructures shared by drugs, to which cell lines are sensitive/resistant & what are the key substructures of drugs if a cell line is sensitive/resistant to them? (3) What are the important genes across cell lines & what are the important genes of different cell lines sensitive/resistant to a common drug? For example, we've found that the drugs containing these five substructures (> = 8N, O = C-C:C-O, > = 3 saturated or aromatic nitrogen-containing ring size 6, > = 32 C and CC1CCC(O)CC1) are sensitive to cell-line LU-65. Otherwise, resistant to it. On the other side, cell lines which contain three mutated genes (PCDH8, GRIA4 and CDO1) are sensitive to the drugs, including Sunitinib, Crizotinib, CGP-082996, CMK and GW843682X. On the other side, though TMF accommodates so far only three types of information matrices, including drug feature matrix, cell line feature matrix and "drug-cell line" response matrix, it can integrate more features into itself by concatenating them horizontally into one combined feature matrix or by using multi-kernel learning feature fusion etc. In conclusion, our TMF is an effective and interpretable approach to predict drug-cell line responses. It is anticipated that the extension of TMF to integrate more heterogeneous features of drugs and cell lines can improve the prediction and be more helpful to understand why cell lines are sensitive or resistant to drugs.

MATERAILS AND METHODS

Problem formulation

A set of response values between n drugs $D=\{d_1, d_2,...,d_n\}$ and m cell lines $C=\{c_1,c_2,...,c_m\}$ are organized into a partially observed response matrix $Y=(y_{ij})_{n\times m}$, $y_{ij}\in\mathbb{R}$, in which $y_{ij}\neq 0$ represents the observed (or known) response value of cell line c_j to drug d_i, otherwise, $y_{ij}=0$ represents the response value between them are unknown .

In the scenario that one is only interested in whether a cell line is sensitive or resistant to a drug, a drug-specific threshold can be used to split the real-valued responses into two classes: sensitivity and resistance. Formally, a binary response matrix $A=(a_{ij})_{n\times m}$ can be transformed from Y, in which $a_{ij}=+1$ represents that a sensitive association between d_i and c_j while $a_{ij}=-1$ represents a resistant association. Usually, we define c_j is sensitive to d_i if $y_{ij} < th_i$, where th_i is the given threshold specific to d_i. Otherwise, c_j is resistant to d_i.

Moreover, suppose that each drug can be described as a p-dimensional feature vector and each cell line can be described as a q-dimensional feature vector. Then, all the drug feature vectors and all the cell line feature vectors are stacked to form a feature matrix $F^d_{n\times p}$ and $F^c_{m\times q}$ respectively.

In order to investigate the relationship between drug features (*e.g.*, chemical substructures fingerprint), cell-line features (*e.g.*, gene expressions) and drug-cell line responses, we develop a new triple matrix factorization (TMF) to predict unknown drug-cell line responses. The whole procedure of predicting drug responses consists of the following four steps: (1) constructing corresponding drug and cell-line feature matrices from chemical structures of D and genomic expressions of C respectively; (2) normalizing the drug-cell line response matrix Y, in order to eliminate the calculation error caused by greatly varied numerical ranges of response values; (3) applying TMF to obtain a bi-projection matrixΘ which bridges $F^d_{n\times p}$, $F^c_{m\times q}$ and Y; (4) reconstructing Y by $F^d_{n\times p}$, $F^c_{m\times q}$ and Θ to compute predicted response values for unobserved drug-cell line pairs. Figure 9 shows the flowchart of the presented approach.

Feature extraction and response normalization

The PubChem fingerprint list consisting of 881 chemical substructures [17] was used to encode drug chemical structures. Formally, drug d_i is defined as a p-dimensional binary vector: $f_{d_i}=[s_{i1}, ..., s_{il}, ..., s_{ip}]$, where $p=881$, $s_{il}=1$ if the l-th substructure in PubChem fingerprint list exists in drug d_i, otherwise $s_{il}=0$.

Table 5 Highly occurring substructures of drugs for resistant to Lung tissue-related cell lines

Rank	Substructures	Group of PubChem fingerprint
1	O=C-C-C-C-O	G6:Simple SMARTS pattern
2	O-C-C-C-C	G6:Simple SMARTS pattern
3	>=16 O	G1:Hierarchic element counts
4	>=5 any ring size 6	G2:Rings in a canonic ESSSR ring set
5	C-C-N-C-C	G6:Simple SMARTS pattern
6	>=2 Cl	G1:Hierarchic element counts
7	N-C:C:C-N	G6:Simple SMARTS pattern
8	>=4 N	G1:Hierarchic element counts
9	O-C-C-C-C-C(C)-C	G6:Simple SMARTS pattern
10	O=C-C=C	G6 Simple SMARTS patterns

Since genomic expression helps understanding the responding behaviors of cell lines to drugs [19], it is used to characterize cell lines here. Let $G=[gene_1, ..., gene_k, ..., gene_q]$ be the set of genes. Cell line c_i can be represented as a q-dimensional vector: $f_{c_i}=[g_{i1}, ..., g_{ik}, ..., g_{iq}]$ where g_{ik} is the expression value of k-th gene in G for c_i. After downloading the corresponding genes related to the cell lines from GDSC dataset [5], we obtain 16,383 genes and their expressions.

The responses of cell lines to drugs obtained by experiment are typically measured by IC50 values [10], which is the drug concentration required for 50% inhibition of cell growth [20]. However, we observed that the numerical ranges of response values vary greatly. For example, there are 19 drugs, which target pathway (*e.g.*, PI3K/MTOR) crucial to many aspects of cell growth and survival in both physiological and pathological conditions (*e.g.*, cancer) [21]. The statistics of the responses across all the cell lines to them illustrate the issue of greatly varied numerical value ranges (Fig. 3). Because this issue surely causes calculation errors, we utilized Z-score to normalize all the response values for each drug as follows

$$Z\text{-score}_{ij}=\frac{y_{ij}-\overline{y_i}}{\sigma_i} \quad (1)$$

where y_{ij} is the observed response value of c_j to drug d_i, $\overline{y_i}$ and σ_i are the mean and standard deviation of all response values for drug d_i respectively.

Triple matrix factorization (TMF)

A triple matrix factorization (TMF) is proposed to associate drug features, cell line features with drug-cell line responses and can be formulated as follows:

$$Y\approx F_d\Theta F_c^T \quad (2)$$

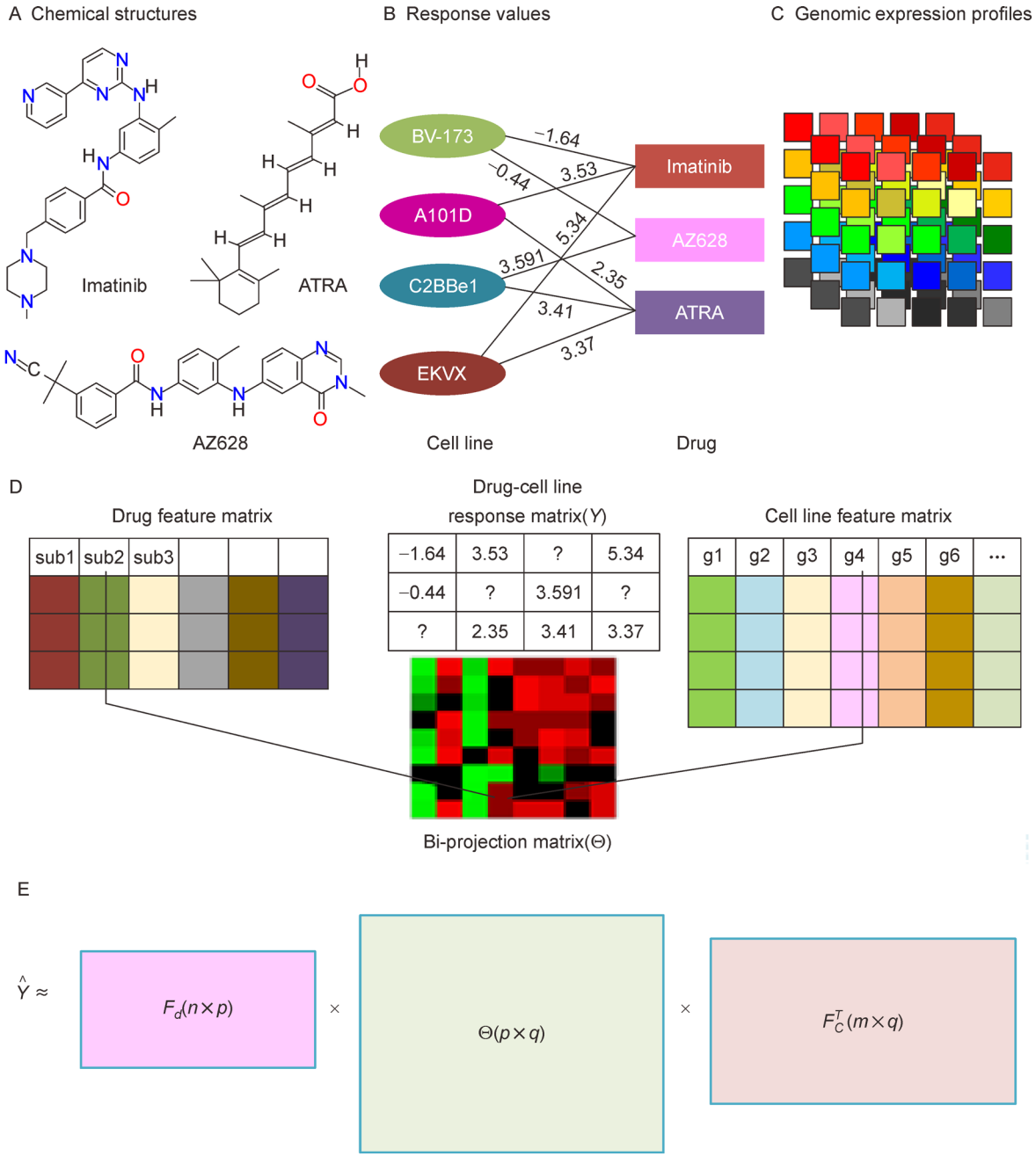

Figure 9. The illustration of triple matrix factorization(TMF). (A) Drug chemical structure feature data; (B) Known drug-cell line response network where the edge weight is the response value; (C) Cell line genomic expression feature data; (D) Construction of drug feature matrix $F^{d}_{n\times p}$, cell line feature matrix $F^{c}_{m\times q}$, drug-cell line response matrix Y and bi-projection matrix; (E) The drug-cell line response matrix is represented as a triple matrix factorization(TMF).

where $\Theta(\theta_{ij})$ is a $p \times q$ bi-projection matrix and its entry $\theta_{ij} \in R$ indicates the importance of the pair of drug feature f_{d_i} and cell line feature f_{c_j} for drug responses. It builds the bridge between feature pairs and responses. If $F_d(n \times p)$ and $F_c\,(m \times q)$ are inverse matrix, Θ can be directly solved by $\Theta = (F_d)^{-1} Y (F_c^T)^{-1}$. However, the number of drug features (881) and that of cell line features (16383) are bigger than the number of drugs and that of cell lines respectively. For example, GDSC contains 256 drugs and 1001 cell lines [5]; the CCLE has only 24 drugs and 504 cell lines [4]. In this condition, Θ cannot be solved by $\Theta = (F_d)^{-1} Y (F_c^T)^{-1}$.

In addition, Θ also cannot be solved by $\Theta = (F_d^T F_d)^{-1} F_d^T Y F_c (F_c^T F_c)^{-1}$, as there exists the dependency between the features of drugs or cell lines, which causes the multicollinearity issue. Thus, $F_d^T F_d$ and $F_c^T F_c$ are all nearly singular and their inverse matrices are ill-conditioned.

Thus, we consider numerical solution of Θ by adopting the low-rank decomposition of Y as follows to solve this matrix factorization,

$$\arg\min_{U,V} \|W \circ (Y - UV^T)\|_F^2 \tag{3}$$

where $U_{n\times k}$ and $V_{m\times k}$ ($k \ll \text{rank}(Y)$) are latent response matrices for drugs and cell lines respectively. They represent the latent topological attributes of drug and cell line, and jointly reflect the underlying response space. The $n \times m$ weight matrix W is introduced to distinguish observed drug-cell line response pairs from unobserved (incomplete) pairs, in which $w_{ij} = 1$ if the response value between drug d_i and cell line c_j is observed, otherwise $w_{ij} = 0$. In addition, $W \circ Z$ denotes the element-wise product of matrices W and Z.

Inspired by the optimization process in the form works [22,23], we adopt the following bi-linear regression model to bridge responses with both drug features and cell line features. In other words, we build a linear regression between drugs' latent response properties U and their input features F_d, as well as a linear regression between cell lines' latent response properties V and their input features F_c as follows:

$$\mathrm{U} \approx F_d B_d \tag{4}$$

$$\mathrm{V} \approx \mathrm{F}_c B_c \tag{5}$$

where $B_d(p \times k)$ and $B_c(q \times k)$ are the regression coefficient matrices of drugs and cell lines respectively. These two constrains can join into the above objective function by Lagrange Multiplier. Meanwhile, we adopt *L2*-regularization to ensure U, V, B_d and B_c smooth [24]. Therefore, the origin optimization problem can be formularized as follows,

Algorithm TMF

Input: drug-cell line observed response matrix Y, drug feature matrix F_d, cell line feature matrix F_c, parameter k, λ_u, λ_v, λ_d, λ_c and max_iter

Output: predicted response value matrix $\hat{Y}$

TMF($\hat{Y}$, Y, F_d, F_c, k, λ_u, λ_v, λ_d, λ_c)

Step1. Generate random matrices as the initialization of U, V, B_d, B_c

Step2. For t = 1, ..., *max_iter*

Update U, V, B_d and B_c iteratively by the solutions of the partial derivatives $\frac{\partial J}{\partial U} = 0$, $\frac{\partial J}{\partial V} = 0$, $\frac{\partial J}{\partial B_d} = 0$ and $\frac{\partial J}{\partial B_c} = 0$ in turn as follows:

$U = (WYV + F_d B_d)(WV^T V + I + \lambda_u I)^{-1}$

$V = (WY^T U + F_c B_c)(WU^T U + I + \lambda_v I)^{-1}$

$B_d = (F_d^T F_d + \lambda_d I)^{-1}(F_d^T U)$.

$B_c = (F_c^T F_c + \lambda_c I)^{-1}(F_c^T V)$

Step3. If not converged, go back to Step 2.

Step4. Calculate the bi-projection matrix by $\Theta^* = B_d^*(B_c^*)^T$

Step5. Obtain the predicted response value between drug d_i and cell line c_j, by $\hat{Y}_{i,j} = F_i^d \Theta^* F_j^c$

Figure 10. The description of the algorithm TMF. It takes Y, F_d,F_c , the parameters k,λ_u,λ_v,λ_d,λ_c and the iteration times *max_iter* as the inputs, then performs drug-cell line response prediction by using Θ^*, which is obtained by iteratively updating the values of matrices U, V, B_d and B_c.

$$\{\mathrm{U}^*,\mathrm{V}^*,\mathrm{B}_d^*,\mathrm{B}_c^*\} = \mathrm{argmin}\mathrm{J}(\mathrm{U},\mathrm{V},\mathrm{B}_d,\mathrm{B}_c) \quad (6)$$

where

$$\mathrm{J} = W\|Y - UV^T\|_F^2 + \|U - F_d B_d\|_F^2 + \|V - F_c B_c\|_F^2 + \lambda_u\|U\|_F^2 + \lambda_v\|V\|_F^2 + \lambda_d\|B_d\|_F^2 + \lambda_c\|B_c\|_F^2 \quad (7)$$

Here, λ_u, λ_v, λ_d and λ_c are positive regularization coefficients. The above objective function can be minimized using the iterative update algorithm [25], which guarantees the convergence by iteratively solving $\frac{\partial \mathrm{J}}{\partial U} = 0$, $\frac{\partial \mathrm{J}}{\partial V} = 0$, $\frac{\partial \mathrm{J}}{\partial B_d} = 0$, $\frac{\partial \mathrm{J}}{\partial B_c} = 0$, and updating U, V, B_d and B_c. The procedure of TMF algorithm is listed in Fig. 10.

Performance evaluations

To evaluate the performance of predicting drug responses, k-fold cross-validation (10-CV) is usually adopted on the collected dataset. According to k-CV, all observed response entries are randomly divided into 10 subsets having nearly equal size. The whole procedure of k-CV contains 10 rounds, of which each round selects one subset in turn as the testing set and unites the remaining nine subsets as the training set. This procedure was repeated 20 times under different random seeds to evaluate the predicting performance.

As former works mentioned that the correlation between observed and predicted response values across all the drugs are an optimistic measure of drug response prediction [10,15], we adopt the drug-specific correlation, which contains two measuring metrices, Pearson correlation coefficient (*PCC*) and Root mean squared error (*RMSE*) as follows:

$$PCC(d_i) = \frac{\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)(\hat{y}_{ij} - \overline{\hat{y}}_i)}{\sqrt{\sum_{j=1}^{m}(y_{ij} - \overline{y})^2 \sum(\hat{y}_{ij} - \overline{\hat{y}}_i)^2}}, \quad (8)$$

$$RMSE(d_i) = \sqrt{\frac{\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2}{m}}, \quad (9)$$

where y_{ij} and $\hat{y}_{ij}$ are the observed and predicted response values w.r.t. d_i and c_j respectively, $\overline{y}_i$ and $\overline{\hat{y}}_i$ are their corresponding averages for d_i across all the cell lines. It is expected that a good prediction has both a great *PCC* and a small *RMSE*.

In addition, considering the fact that the sensitive and resistant cell lines of each drug are valuable to unveil mechanisms of drug actions, we also compute *PCC* and *RMSE* from sensitive and resistant cell lines for each drug, and they were denoted as *PCC_S/R* and *RMSE_S/R* [10,15]. According to all the IC50 response values of a drug, the cell lines in the first and the fourth quartiles represent sensitive cell lines and resistant cell lines respectively [16,26].

$$PCC\_S/R(d_i) = \frac{\sum_{j \in C_i^S \cup C_i^R}(y_{ij} - \overline{y}_i)(\hat{y}_{ij} - \overline{\hat{y}}_i)}{\sqrt{\sum_{j \in C_i^S \cup C_i^R}(y_{ij} - \overline{y}_i)^2 \sum(\hat{y}_{ij} - \overline{\hat{y}})^2}}, \quad (10)$$

$$RMSE\_S/R(d_i) = \sqrt{\frac{\sum_{j \in C_i^S \cup C_i^R}(y_{ij} - \hat{y}_{ij})^2}{|C_i^S \cup C_i^R|}}, \quad (11)$$

where $C_i^S \cup C_i^R$ is the cell line set, in which the elements are sensitive or resistant to drug d_i, and $|.|$ is the number of elements in the set.

Finally, these four metrics of all drugs are averaged respectively as the final evaluation metrics, denoted as *PCC*, *E*, *PCC_S/R* and *E_S/R*.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-021-0259.

ACKNOWLEDGEMENTS

This work has been supported by the National Natural Science Foundation of China (Nos. 61872297 and 61873202) as well as by Shaanxi Provincial Key R&D Program, China (No. 2020KW-063).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xiao-Ying Yan, Shao-Wu Zhang, Siu-Ming Yiu and Jian-Yu Shi declare that they have no competing interests.

All procedures performed in studies were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

REFERENCES

1. Mirnezami, R., Nicholson, J. and Darzi, A. (2012) Preparing for precision medicine. N. Engl. J. Med., 366, 489–491
2. Mcdermott, U., Sharma, S. V., Dowell, L., Greninger, P., Montagut, C., Lamb, J., Archibald, H., Raudales, R., Tam, A., Lee, D., *et al.* (2007) Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. Proc. Natl. Acad. Sci., USA 104, 19936–19941
3. Shoemaker, R. H. (2006) The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer, 6, 813–823
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. 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
6. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., *et al.* (2016) A landscape of pharmacogenomic interactions in cancer. Cell, 166, 740–754
7. Ammad-Ud-Din, M., Khan, S. A., Wennerberg, K. and Aittokallio, T. (2017) Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics, 33, i359–i368
8. Geeleher, P., Cox, N. J., Huang, R. S. (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol.,15, R47
9. 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
10. 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
11. Zhang, F., Wang, M., Xi, J. (2018) A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci. Rep., 8, 3355
12. Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet., 47, 569–576
13. Yang, J., Li, A., Li, Y., Guo, X. and Wang, M. (2019) A novel approach for drug response prediction in cancer cell lines via network representation learning. Bioinformatics, 35, 1527–1535
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. 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
16. Zhang, L., Chen, X., Guan, N. N., Liu, H. and Li, J. Q. (2018) A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. Front. Pharmacol., 9, 1017
17. Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S. and Zhang, J. (2017) PubChem BioAssay: 2017 update. Nucleic Acids Res., 45, D955–D963
18. Forbes, S. A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C. Y., Jia, M., Ewing, R., Menzies, A., *et al.* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res., 38, D652–D657
19. Chen, J. and Zhang, S. (2016) Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics, 32, 1724–1732
20. Sebaugh, J. L. (2011) Guidelines for accurate EC50/IC50 estimation. Pharm. Stat., 10, 128–134
21. Porta, C., Paglino, C. and Mosca, A. (2014) Targeting PI3K/Akt/mTOR Signaling in Cancer. Front. Oncol., 4, 64
22. Shi, J. Y., Zhang, A. Q., Zhang, S. W., Mao, K. T. and Yiu, S. M. (2018) A unified solution for different scenarios of predicting drug-target interactions via triple matrix factorization. BMC Syst. Biol., 12, 136
23. Shi, J. Y., Huang, H., Li, J. X., Lei, P., Zhang, Y. N., Dong, K. and Yiu, S. M. (2018) TMFUF: a triple matrix factorization-based unified framework for predicting comprehensive drug-drug interactions of new drugs. BMC Bioinformatics, 19, 411
24. Guan, N., Tao, D., Luo, Z., Yuan B. (2011) Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent . IEEE Trans. Image Process., 20, 2030–2048
25. Lee, D. D. and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401, 788–791
26. Marcotte, R., Sayad, A., Brown, K. R., Sanchez-Garcia, F., Reimand, J., Haider, M., Virtanen, C., Bradner, J. E., Bader, G. D., Mills, G. B., *et al.* (2016) Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell, 164, 293–309

