Journal of Computing and Security

Research Article, 2025, Volume 12, Number 2 (pp. 93–105)

http://www.jcomsec.org

Improving Drug Response Prediction using Dual Similarity Regularization

Ali Reza Ebadi [a] Ali Soleimani [b,*] Seyed Jafar Hashemizade Haghighi [c]

[a] *Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran.*
[b] *Department of Computer Engineering, ShQ.C., Islamic Azad University, Shahr-e Qods, Iran.*
[c] *Department of Computer Engineering, Malard Branch, Islamic Azad University, Tehran, Iran.*

ARTICLE INFO.

Article history:
Received: 29 March 2025
Revised: 01 November 2025
Accepted: 28 December 2025
Published Online: 03 January 2026

Keywords:
Personal Medicine, Matrix Factorization, Anticancer Drug Response Prediction, Therapeutic.

ABSTRACT

Personalized medicine aims to identify effective anticancer therapies tailored to individual patients, a core goal of precision oncology. Despite significant advances, achieving reliable and accurate drug response prediction remains challenging due to the complexity and heterogeneity of pharmacogenomic data. Motivated by the principle that similar cell lines exhibit similar responses to similar drugs, we propose an enhanced matrix factorization framework incorporating a novel dual similarity regularization strategy. The proposed Dual Similarity-Regularized Matrix Factorization (DSRMF) model constrains the latent representations of cell lines and drugs to preserve biological and chemical similarity relationships, ensuring that similar entities occupy proximate positions in the latent space while dissimilar ones remain distant. The model integrates two-dimensional (2D) and three-dimensional (3D) chemical structural features to construct a refined drug similarity matrix and was trained and validated on processed datasets from the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE). Experimental results demonstrate that DSRMF achieves robust predictive performance, with an average Pearson correlation coefficient (PCC) of approximately 0.96 and a root mean square error (RMSE) of 0.30, indicating a strong correlation between predicted and observed drug responses. These findings confirm that incorporating dual similarity regularization and heterogeneous biological information enhances both predictive accuracy and interpretability. Overall, DSRMF advances drug response modeling and provides a scalable framework for integrating multi-dimensional biological data to improve personalized cancer treatment strategies.

1 Introduction

Accurately predicting how patients or cell lines respond to specific therapeutic agents represents a cornerstone of precision medicine. Drug response prediction (DRP) models aim to personalize treatment strategies by integrating genomic, transcriptomic,

* Corresponding author.
Email address: al.soleimani@iau.ac.ir (A. Soleimani)
https://dx.doi.org/10.22108/jcs.2025.144573.1159
ISSN: 2322-4460

and chemical information, thereby enhancing therapeutic efficacy while minimizing adverse effects. Despite significant progress in this domain, achieving reliable and generalizable prediction remains challenging due to the intrinsic high dimensionality, sparsity, and heterogeneity of pharmacogenomic data.

Recent advances in deep learning have shown considerable promise for modeling the complex, nonlinear relationships between molecular features and drug responses. Approaches such as deep neural networks, graph convolutional networks (GCNs), and attention-based architectures have demonstrated strong predictive performance by leveraging multi-omics profiles and detailed drug descriptors. However, many of these methods underexploit the rich relational structure embedded in the data—particularly the similarities among drugs and among cell lines—which can provide valuable biological context and improve model generalization, especially in predicting unseen drug–cell line interactions.

To overcome these limitations, we introduce a novel framework for drug response prediction based on Dual Similarity Regularization (DSR). The proposed approach incorporates prior biological knowledge through predefined similarity matrices for both drugs (e.g., chemical structure and target profiles) and cell lines (e.g., gene expression and mutation signatures). These similarity matrices are embedded within the learning process as regularization constraints, encouraging the latent representations of drugs and cell lines to preserve biologically meaningful relationships. By explicitly enforcing this relational structure, the model achieves enhanced predictive accuracy and improved interpretability.

Specifically, we extend a matrix factorization-based formulation by introducing dual regularization terms that align the learned latent spaces with their corresponding similarity graphs for drugs and cell lines. Additional graph-based constraints, such as Laplacian regularization, are applied to promote local smoothness and structural coherence within the latent embeddings. This dual similarity regularization acts as an inductive bias, guiding the model toward biologically consistent predictions even in scenarios with limited or noisy data, and ultimately providing a more robust and interpretable framework for precision oncology applications.

Our contributions are as follows:

- We propose a deep learning framework that integrates dual similarity regularization into drug response prediction.
- We demonstrate how domain-specific similarity structures improve both prediction accuracy and biological plausibility.
- We conduct extensive evaluations and ablation studies to validate the model's effectiveness and interpretability.

2 Problem Statement

Predicting drug response in cancer cell lines is a fundamental challenge in precision medicine. Experimental determination of drug sensitivity, typically quantified by metrics such as IC or AUC, is expensive, time-consuming, and infeasible for all possible drug–cell line combinations. Consequently, computational approaches are required to predict missing responses in drug cell line matrices, enabling personalized treatment planning and drug discovery.

The main challenges in this prediction task include:

1. **Sparsity:** Most drug–cell line response matrices are incomplete, with many missing entries.
2. **High-dimensionality:** Cell line profiles (gene expression, mutations) and drug descriptors (chemical structure) are high-dimensional.
3. **Heterogeneity:** Drugs and cell lines differ significantly, making it difficult to generalize predictions.
4. **Need for similarity integration:** Similar drugs tend to have similar effects, and similar cell lines tend to respond similarly to drugs. Ignoring these similarities can reduce prediction accuracy.

3 Background and Related Works

Personalized medicine has emerged as a transformative approach in modern healthcare, aiming to tailor medical treatments to the unique molecular and genomic profiles of individual patients. This strategy enables the selection of the most effective therapeutic interventions with minimal adverse effects. Particularly in oncology, where cancer exhibits substantial heterogeneity in drug response, personalized treatment is essential to improving patient outcomes.

Cancer remains one of the leading causes of mortality worldwide, with a complex and highly variable response to treatment. In this context, precision medicine has proven especially promising. Researchers are actively developing computational frameworks that integrate genomic features of cancer cell lines to predict their sensitivity to anticancer drugs. These models are often trained using large-scale pharmacogenomic datasets such as the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC).

Studies in drug response prediction typically fall into two major categories. The first category focuses on identifying biomarkers—genomic or molecular indicators that predict a cell line's sensitivity to a specific drug. For instance, [1] employed molecular-level variations to forecast drug response, while others [2–6] utilized machine learning techniques such as elastic net regularization and random forests to identify predictive genomic biomarkers.

The second category involves predictive modeling using large-scale gene expression data. In [7], the Kernelized Bayesian Matrix Factorization (KBMF) method was used to model drug response by capturing complex nonlinear relationships between drugs and cell lines. Similarly, [8] proposed a Weighted Graph-Regularized Matrix Factorization (WGRMF) algorithm, where both drug similarity and cell similarity graphs are incorporated. This method uses neighborhood-based similarity to enhance latent feature learning and demonstrates improved prediction accuracy using GDSC data.

In another comparative study [9], three different deep learning models were benchmarked against random forests (RF) and K-nearest neighbors (KNN). The combination of RF, gene expression (GE), and residual KNN achieved the best performance, especially in scenarios involving outlier removal and small sample sizes. In [10], a private linear regression model was introduced for drug sensitivity prediction. In [11], the Kernelized Bayesian Multitask Learning (KBMTL) framework was proposed to share a low-dimensional subspace across tasks and improve forecasting performance.

A number of multi-task learning approaches have also been explored. For example, [12] applied multitask models to the CCLE**,** CTD^2, and NCI60 databases to improve generalizability across datasets. In [13], a network-based method called GloNetDRP was used to integrate drug–drug and drug–cell line similarities through a heterogeneous graph, allowing drug responses to be inferred not only from nearest neighbors but also from broader structural similarity.

Several additional models have leveraged similarity-based reasoning. In [14], a drug similarity model was proposed in which the drug sensitivity profile of a new compound was inferred based on its structural similarity to known drugs using CCLE and GDSC data. The Network-Based Classifier (NBC) introduced in [15] assessed sensitivity across tumor types using apoptotic gene expression and clinical dose predictions. In [16], PRECISE, a domain adaptation method, was employed to align molecular profiles between human tumors and preclinical models. The Support Vector Machine (SVM) method combined with recursive feature elimination was used in [17] to classify drug-sensitive and drug-resistant cell lines based on genomic features.

In [18], the consensus p-median clustering method was applied to infer drug responses on tumor-derived cell lines, while in [19], a Bayesian network was constructed using a selected set of heterogeneous genes to model drug response using the NCI60 dataset.

Together, these studies underscore the importance of integrating biological similarity, molecular features, and advanced machine learning techniques to improve drug response prediction. However, many existing approaches either rely solely on gene expression or fail to simultaneously model both drug–drug and cell–cell similarities. Our proposed model addresses this gap by introducing a Dual Similarity-Regularized Matrix Factorization (DSRMF**)** framework that jointly incorporates drug and cell line similarity into the prediction process-improving both biological coherence and predictive accuracy.

Previous models, such as Similarity-Regularized Matrix Factorization (SRMF), have been widely used for predicting anticancer drug responses based on the structural similarity of drugs and the gene expression profiles of cancer cell lines. However, SRMF has certain limitations: (1) the chemical similarity data used were often incomplete or insufficiently validated, (2) many models relied on limited 2D structural features, omitting key dimensional and stereochemical information, and (3) numerical instability, such as overflow issues, occasionally affected prediction accuracy, particularly in large or sparse matrices.

To address these limitations, we built upon SRMF by developing a Dual Similarity-Regularized Matrix Factorization (DSRMF) model. Unlike SRMF, our method integrates both drug–drug similarity and cell line–cell line similarity into a unified framework, thus capturing biological and chemical relationships more comprehensively. To enhance the accuracy of the drug similarity matrix, we synthesized data from the CCLE**,** GDSC**,** and NCBI PubChem repositories. Drugs were represented using **their** SDF (Structure Data Format) files, and a detailed two-dimensional (2D) structural analysis was conducted using PaDEL software to extract fingerprint-based features. This allowed us to overcome prior limitations in chemical feature extraction and extend the usable drug set.

One notable challenge we encountered was the limited availability of anticancer drugs with fully specified chemical structures in the public datasets. We addressed this by curating and filtering drugs with valid PubChem Compound IDs (CIDs) and generating consistent SDF-based descriptors, which improved the

robustness of our drug similarity computation.

To evaluate the performance of our DSRMF model, we used two key metrics: the Pearson Correlation Coefficient (PCC) to assess correlation between predicted and true IC50 values, and the Root Mean Square Error (RMSE) to measure prediction error. Our results show that DSRMF consistently outperformed prior SRMF-based approaches, achieving lower RMSE and higher PCC. In addition, our model was able to infer missing drug response values in the GDSC dataset, offering a more complete and biologically coherent prediction matrix.

These findings demonstrate that incorporating dual similarity regularization and chemically enriched descriptors significantly improves prediction performance, paving the way for more accurate and scalable drug sensitivity modeling in cancer research.

Predicting drug response has become a critical task in precision medicine, aiming to tailor treatments to individual patients based on molecular and genomic profiles. Traditional approaches such as linear regression or kernel methods [1, 2] have limited ability to capture complex, nonlinear patterns in high-dimensional data. Recent advancements in deep learning have significantly improved prediction accuracy by learning expressive representations from heterogeneous data sources.

3.1 Deep Learning for Drug Response Prediction

Deep neural networks (DNNs) have been widely used to predict drug sensitivity based on omics features of cell lines or patients. For example, DeepDR [20] uses a multi-layer feedforward neural network to learn a nonlinear mapping from genomic features to drug response scores. Other models, such as CDRscan [21] and GraphDRP [22], incorporate structural information of drugs and cellular features via convolutional and graph-based architectures.

However, these models often overlook pairwise relationships such as similarity between drugs or between cell lines. Such relationships are important because drugs with similar chemical structures often have similar effects, and cell lines with similar genomic profiles may respond similarly to a given drug.

3.2 Incorporating Similarity via Regularization

To incorporate prior biological knowledge, several methods use graph-based regularization. For instance, **DualGCN** [23] models both cell line and drug similarity using graph convolutional networks (GCNs) on predefined similarity graphs. Similarly, GraphDRP and DrugCell [24] integrate gene ontology and pathway information into the prediction pipeline.

These methods are closely related to the dual similarity regularization framework, where regularization terms are used to ensure that latent representations of samples and drugs are consistent with similarity graphs:

This technique is used in DeepDSC [25], which improves drug screening by encouraging the learned features to respect drug and cell line similarities.

3.3 Dual Similarity-Aware Matrix Factorization

A line of work also applies matrix factorization with dual similarity constraints, particularly in recommender systems [26] and multi-view learning [27]. These approaches are adapted to the DRP setting by treating the drug-cell line interaction matrix similarly to a user-item matrix, with latent features regularized by both drug similarity (e.g., chemical structure or MoA) and cell line similarity (e.g., gene expression, mutation profiles).

Deep Neural Network Integrating Prior Knowledge (DIPK) presents a compelling design for drug response modeling. It employs self-supervised pretraining to integrate gene interaction networks, gene expression, and molecular structure, enabling the deep neural network to respect biological similarity constructs and boosting prediction accuracy—particularly on unseen cell lines and drugs. DIPK demonstrated strong performance in clinical subsamples, such as correctly identifying heightened paclitaxel response in patients with pathological complete response (pCR), highlighting the value of embedding prior knowledge. [28]

NMDP, proposed in 2025, exemplifies dual-regulation via multi-omics similarity fusion. It uses:

1. A *semi-supervised weighted SPCA* to extract key genomic features,
2. A weighted similarity network fusion approach to combine multi-omics similarities,
3. A deep learning predictor based on 1D convolutions and Kolmogorov–Arnold networks (KAN).

This pipeline effectively aligns omics spaces and regularizes the latent space according to similarity structure—analogous to dual similarity regularization—with promising results on small-sample datasets [29].

The Dual Similarity Regularization (DSR) method and Dual-Branch Deep Neural Matrix Factorization

(DB-DNMF) [30] differ primarily in methodology and feature representation. DSR uses matrix factorization with dual similarity regularization, leveraging explicit drug–drug and cell–cell similarity matrices to constrain latent factors. In contrast, DB-DNMF employs a dual-branch neural network to learn nonlinear embeddings for drugs and cell lines, capturing complex interactions. Consequently, DSR is more interpretable and computationally efficient, while DB-DNMF excels at modeling nonlinear relationships in rich, high-dimensional datasets.

The Dual Similarity Regularization (DSR) method and Drug-Specific Gene Selection via Biological Network Learning (DSGS-BNL) [31] differ fundamentally in their approach to drug response prediction. While DSR employs matrix factorization with dual similarity regularization to capture global relationships among drugs and cell lines, DSGS-BNL focuses on identifying drug-specific predictive genes using biological network information. Consequently, DSR emphasizes latent factor modeling and benefits from similarity across the entire dataset, whereas DSGS-BNL offers gene-level interpretability and mechanistic insights by highlighting the genetic determinants of response for each drug. Overall, DSR is well-suited for leveraging sparse datasets and generalizing across multiple drugs, whereas DSGS-BNL excels in elucidating drug-specific biological mechanisms.

While not explicitly framed for drug response, several recent 2024–2025 models emphasize dual similarity integration in related biomedical tasks:

- RSML GCN (2024) for drug–disease association employs reinforcement symmetric metric learning and GCNs. It integrates both drug–drug and disease–disease similarity through adjacency-aware pretraining, improving hits in "new drug" settings [32]
- Broadly in drug interaction prediction, models like SNF–HCNN (2025) and DDI Hybrid leverage similarity network fusion of multiple drug attributes (e.g., structural and biological) followed by CNN or BiLSTM modules—effectively dual similarity-enhanced deep architectures. [33]

4 Contribution

In this study, the Dual Similarity-Regularized Matrix Factorization (DSRMF) model was employed to enhance the predictive accuracy of drug responses in cancer cell lines. The model integrates both drug similarity and cell line similarity into the matrix factorization framework, based on the premise that pharmacologically similar drugs and biologically similar cell lines are likely to exhibit comparable response patterns. By incorporating similarity information, the DSRMF approach provides a robust and biologically informed strategy for drug response prediction. Beyond improving predictive performance, this framework offers valuable insights into the underlying relationships between drugs and cell lines, thereby contributing to the development of more effective and personalized therapeutic strategies in cancer treatment.

Table 1. Comparison of Shallow (Traditional) Models and Deep Learning–Based Models.

Model / Method	Category
SRMF, WGRMF, KBMTL, RF, SVR, Elastic Net, KBMF, DSRMF	Shallow models (non-neural networks; classical machine learning / matrix factorization)
CDRscan, GraphDRP, FC DNNs, SWNet, DeepIC50, Dr.VAE, DLN, etc.	Deep learning models (DNNs, CNNs, GNNs, autoencoders)

Drug response data were represented by half-maximal inhibitory concentration (IC) values, which denote the concentration of a compound required to inhibit cell viability by 50%; thus, lower IC values correspond to greater cellular sensitivity. Each cell line was characterized by its gene expression profile, whereas each drug was described by its molecular structure derived from Structure Data Format (SDF) files obtained from the NCBI PubChem repository. The processed dataset was organized into a dense response matrix comprising 604 cell lines and 97 drugs, resulting in 58,588 potential cell line–drug response pairs.

Data were sourced from two well-established pharmacogenomic resources: the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE). Specifically, GDSC release 5.0, which initially contained 790 cancer cell lines and 135 drugs, served as the primary dataset. During preprocessing, drugs lacking unique chemical identifiers—such as PubChem Compound IDs (CIDs) and corresponding SDF files—or those not chemically well-defined within the CCLE database were excluded, yielding a refined subset of 97 drugs. Similarly, duplicate or incomplete cell line entries were removed, resulting in a final collection of 604 unique cell lines for downstream analyses.

To enhance model performance, similarity information was incorporated for both cell lines and drugs. The PaDEL-Descriptor software was used to extract two-dimensional (2D) and three-dimensional (3D) molecular descriptors and to compute PubChem substructure fingerprints for all drugs. These processed

descriptors formed the basis of the drug similarity matrix, where pairwise structural similarities were quantified using the Jaccard coefficient. In parallel, a cell line similarity matrix was constructed by calculating Pearson correlation coefficients between the gene expression profiles of all cell line pairs, thereby capturing transcriptional similarity.

The resulting drug response matrix included 58,588 cell line–drug combinations, with missing entries arising from experimental limitations. For model evaluation, the dataset was randomly divided into training and testing subsets, with 70% (41,012 samples) used for training and 30% (17,576 samples) for testing. Additionally, a 10-fold cross-validation procedure was implemented on the GDSC dataset to rigorously assess the predictive performance and generalization capability of the model. This comprehensive preprocessing pipeline ensured meaningful representation of both cell lines and drugs, allowing the DSRMF framework to effectively leverage biological similarity through its dual similarity regularization mechanism.

Table 2. DSRMF on CCLE and GDSC.

Component	Setting / Value
Datasets	GDSC and CCLE drug–cell-line response (IC_{50}) [34, 35]
Normalization	Scaling to $[-1, +1]$ by dividing by the maximum absolute value
Latent Dimension (k)	44 (GDSC)
Iterations	20
Hyperparameters	α, β, and λ scanned in the range $[10^{-8}, 10^{-3}]$
Best β (drug similarity weight)	1×10^{-7}
Replicates	100 randomized runs
$PCC_{s/r}$	∼0.96 (DSRMF) vs. ∼0.71 (SRMF baseline)
$RMSE_{s/r}$	∼0.30 (DSRMF) vs. ∼0.31 (SRMF), ∼0.78 (RF/KBMF)

5 Problem Formulation

In this paper, we use the matrix factorization algorithm to predict the response to the anticancer drug. A similar framework has been adopted to predict drug targets [36]. First, we mapped the drug m and n cells into a shared low-dimensional K latent space in which the properties of each cell line drug are represented by the latent coordinates $U_i$and V_j, respectively. We call the drug response watermarks in this space Y. The purpose is to obtain an approximation of the response value of d_j to the cell line c_i, by determining their corresponding latent coordinates. Our objective function here is as follows.

Table 3. Parameters and Their Ranges.

Parameter	Value / Range
Dataset	GDSC Release 5.0
Cell lines	604
Drugs	97
Potential total pairs	$604 \times 97 = 58{,}588$
Observed response pairs	$< 58{,}588$
Cross-validation	10-fold cross-validation
Approx. per-fold test samples	∼5,859
Evaluation metrics	PCC, RMSE
Average PCC	∼0.96 (drug-averaged)
Average RMSE	∼0.12

$$\min_{u,v} \left\| W.\left(Y - UV^{\top}\right) \right\|_F^2 \tag{1}$$

Where W is the matrix weight $W_{ij} = 1$, if Y_{ij} has a certain amount of drug response. Otherwise $W_{ij} = 0$ and $\|.\|_F$ is frobenius norm. And to avoid overflow of U, V training data, L2 (Tikhonov) regularization is added to the latent variables U, V as fines.

$$\begin{aligned} \min_{U,V} \quad & \left\| W \odot \left(Y - UV^{\top}\right) \right\|_F^2 + \lambda_l \left(\|U\|_F^2 + \|V\|_F^2 \right) \\ & + \lambda_d \left\| S_d - UU^{\top} \right\|_F^2 + \lambda_c \left\| S_c - VV^{\top} \right\|_F^2 \end{aligned} \tag{2}$$

According to the work done by [16], Similar drugs and similar cell lines have similar responses to drugs.

Lin Wang et al.'s research helped to enhance the prediction accuracy of drug response by minimizing the objective functions of $\left\| S_d - UU^T \right\|$ and $\left\| S_C - VV^T \right\|_F^T$.

The primary contribution of this study lies in introducing a more stringent dual similarity regularization constraint within the matrix factorization framework. Specifically, the model enforces that the distance between the latent representations of two cell lines or two drugs is inversely correlated with their biological or chemical similarity. In other words, pairs of similar drugs or cell lines are encouraged to have closer latent representations, whereas dissimilar entities are positioned farther apart in the latent space. By embedding this principle directly into the objective function, the proposed approach strengthens the preservation of intrinsic similarity relationships, thereby im-

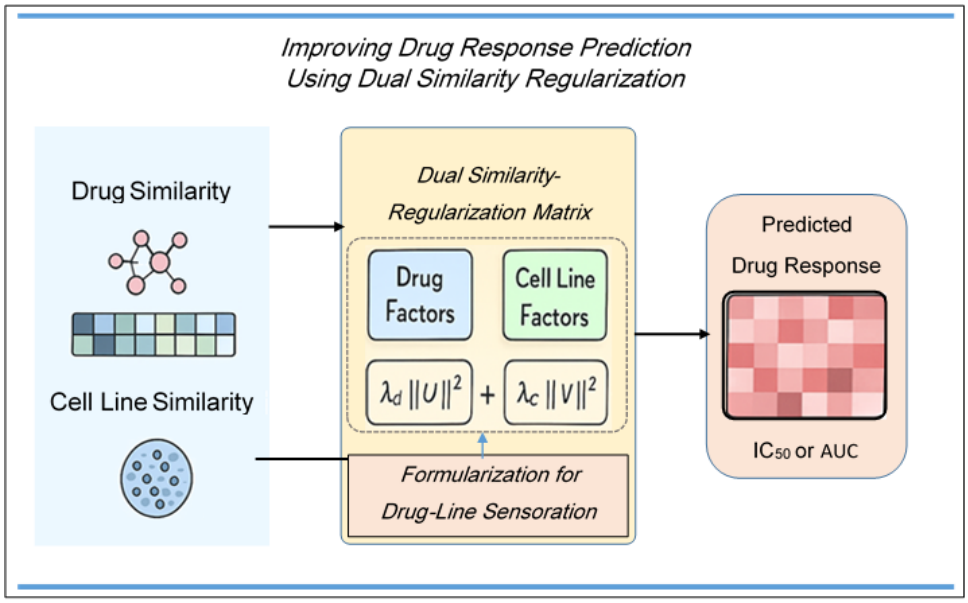

Figure 1. Block Diagram of the Proposed Method.

proving the accuracy of drug response prediction compared to existing methods.

The similarity regularization term for cell lines is formulated as follows:

$$SimReg_c = \frac{1}{8}\sum_{i=1}^{m}\sum_{j=1}^{m}\left((S_c(i,j) - e^{\gamma\|U_i - U_j\|^2}\right)^2 \quad (3)$$

And to obtain similarity regularization for drugs:

γ is to control the radius of the Gaussian function

$$SimReg_d = \frac{1}{8}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(S_d(i,j) - e^{\gamma\|V_i - V_j\|^2}\right)^2 \quad (4)$$

Finally, the constraints are added to the main objective function.

$$\begin{aligned}\min_{U,V} \quad & \left\|W.\left(Y - UV^\top\right)\right\|_F^2 + \lambda_l\left(\|U\|_F^2 + \|V\|_F^2\right) \\ & + \lambda_d\left\|S_d - UU^\top\right\|_F^2 + \lambda_c\left\|S_c - VV^\top\right\|_F^2 \\ & + \alpha\,\mathrm{SimReg}_c + \beta\,\mathrm{SimReg}_d\end{aligned} \quad (5)$$

and adjust the ratio of the similarity regularization term to compare cell lines and drugs. Both and positively influence model performance when moderately weighted. Overly high values for either or degrade performance, likely due to over-constraining the latent space and reducing flexibility. The optimal configuration in our experiments is $\alpha = 0.01$ and $\beta = 0.01$, suggesting that incorporating both content and demographic similarity improves factorization quality. The similarity regularization terms are complementary; using both yields better results than either alone.

6 Evaluation

Dataset Description: In this study, data from the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC) databases were utilized to assess similarities between cancer cell lines and drug profiles. Gene expression profiles were employed to quantify the similarity between cell lines. In addition, other molecular characteristics—such as DNA methylation, reverse-phase protein array (RPPA) data, and microRNA expression—could be incorporated to capture complementary aspects of cellular similarity. Furthermore, leveraging genomic features of cell lines, including copy number variation, pathway information, and somatic mutations, has the potential to further enhance the predictive capability of the proposed Dual Similarity-Regularized Matrix Factorization (DSRMF) model. Integrating these diverse data modalities may improve the accuracy of anticancer drug response prediction and extend the applicability of the framework to other predictive modeling tasks within computational biology.

Evaluation Metrics: The evaluation of the proposed DSRMF model was conducted using the Similarity-Regularized Matrix Factorization (SRMF) algorithm as a reference, following the methodology outlined in [19]. The predictive performance of DSRMF was compared with that of existing approaches to assess its effectiveness in estimating drug response values. The results demonstrated that

DSRMF achieved improved prediction accuracy relative to previous models. All implementations were performed using Python 3.7 (64-bit).

Measurements of prediction performance: To quantitatively evaluate model performance, two standard metrics were employed: the Root Mean Squared Error (RMSE) and the Pearson Correlation Coefficient (PCC), calculated for each drug as described in [19]. The RMSE measures the deviation between predicted and observed drug response values and is computed as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{c=1}^{n}\left(R(D,c) - \hat{R}(D,c)\right)^2} \qquad (6)$$

The value of n refers to the number of cell lines that respond to a particular drug. Also, R (D, C) and R (D, C) refer to the observed and predicted response values of the cell lines to the drugs.

The hyperparameter settings in the machine learning algorithm used in this paper, which is based on the matrix factorization method, are as follows: First, the matrix data of the response lines of the cell lines to the anticancer drugs were collected from the CCLE and GDSC databases. By dividing the values of the data in the matrix by the maximum absolute value, they are converted to values between the range [-1, 1]. Therefore, the regularization parameters λ_c, $\lambda_d$and λ_l were tested in the intervals $[2^{-6}... 2^2]$ respectively. The point is, in [19] paper, which has achieved the best result at λ_d=0, actually, the chemical structure of drugs was ignored. But in our paper, accuracy was higher in the DSRMF method at $\lambda_d = 0.0000001$ by applying chemical structure control settings to the loss function. Another difference in this paper is producing new cell line drug response data by incorporating two-dimensional and three-dimensional chemical structures in drug similarity and response, compared to the previous drug similarity in previous work. The results of applying the SRMF model to the new production data are shown in this study. The value of k, which is the low dimensionality K of the GDSC database, was set as 44, and the Iteration training was set as 20. In this research, we compared the performance of the SRMF model with the proposed DSRMF model. The average PCC and average RMSE of 100 replicates were used. The newly generated data from the CCLE and GDSC databases were used in the DSRMF model to average PCC_S / R (Pearson correlation coefficient). The sensitivity of the cell line to the anticancer drug between the observed and the predicted was approximately 96.0. Comparing with the previous SRMF model in [19], which at best was 0.71 on its data by λ_d= 0, was about 0.25. For the SRMF model, on data was about 0.95

Figure 2 compares the mean PCC_S / R between a sample of drugs, and the averaged RMSE_S /R (root mean square error) between observed and predicted cell lines' response to drugs in the DSRMF model was 0.30, and in the SRMF model was about 0.31 for the data in this Research. Compared to the previous work of [16], which had an average RMSE_S / R value of 0.78 on its own data conditions, the proposed method was about 0.47 better than the previous models in RF [37] and KBMF [38]. Also, in Figure 2, we see a comparison of the proposed method and the previous method for randomly selected anticancer drugs. And in Figure 3, we compared the previous method and the proposed method by adding a number of algorithm iterations. Table 1 shows the prediction details and the comparisons of results for different methods.

Table 4. Comparison Results of Different Methods under 10-Fold Cross-Validation on the GDSC Dataset.

Method	**Drug-averaged $\mathbf{PCC}_{S/R}$**	**Drug-averaged $\mathbf{RMSE}_{S/R}$**	**Reference**
DSRMF	0.96	0.31	Proposed
NMDP	0.92	0.47	[29]
SRMF	0.71	0.62	[19]
KBMF	0.49	–	[38]
DLN	0.55	0.44	[22]
RF	0.50	0.40	[37]

The average cell line response to drugs, which is the same IC50 value, is evaluated for SRMF and DSRMF (proposed method) and their difference with the observed values are illustrated in Figure 5 and Figure 6. The DSRMF model is more accurate than the SRMF model.

7 Conclusions

In this study, we developed a Dual Similarity-Regularized Matrix Factorization (DSRMF) model to predict the responses of cancer cell lines to anticancer drugs, using half-maximal inhibitory concentration (IC) values as the primary indicator of drug sensitivity. The proposed framework integrates data from two major pharmacogenomic resources-the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC)-to comprehensively model both cellular and chemical similarities. Cell line similarity was primarily derived from gene expression profiles; however, the framework is inherently extensible and can incorporate

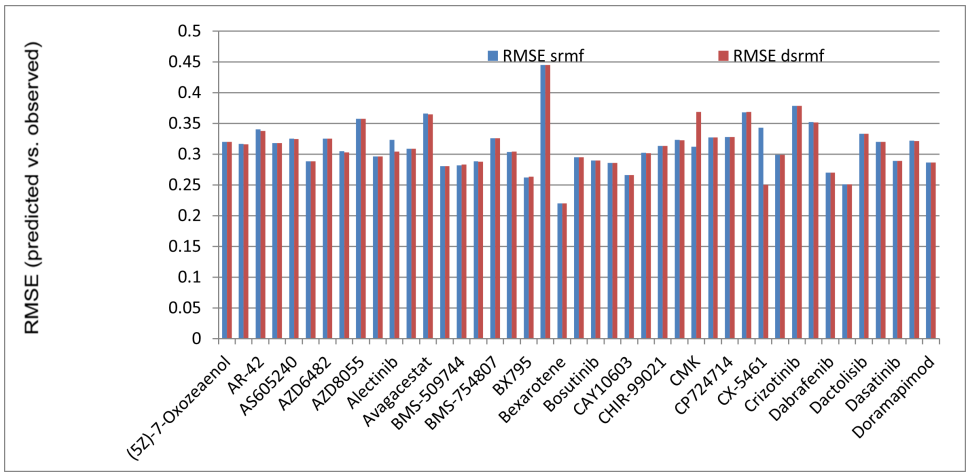

Figure 2. Comparison of the mean of PCC_S / R among the samples of drugs between the proposed method (dsrmf) and the previous method

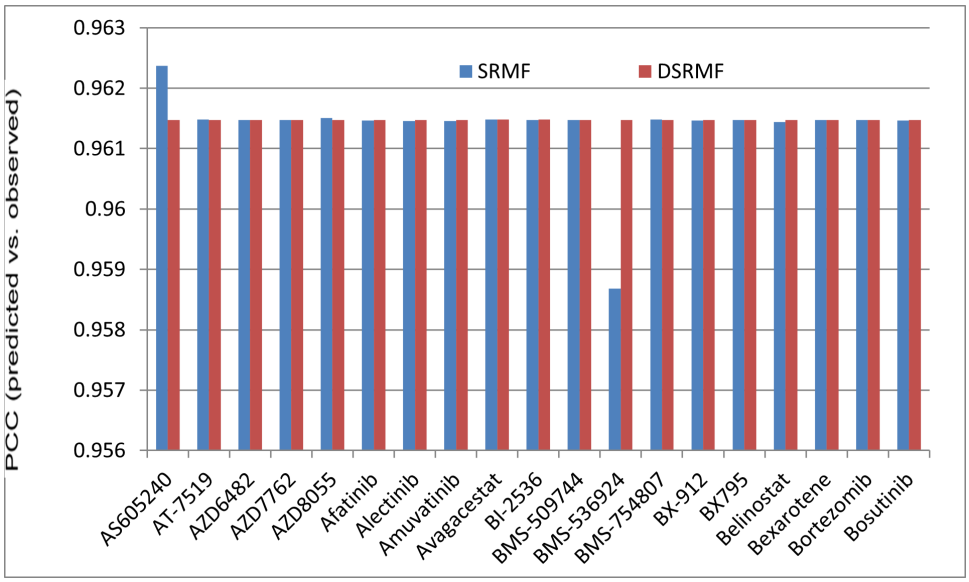

Figure 3. Comparison of the average RMSE_S / R among the samples of other drugs between the proposed method (DSRMF) and the previous method.

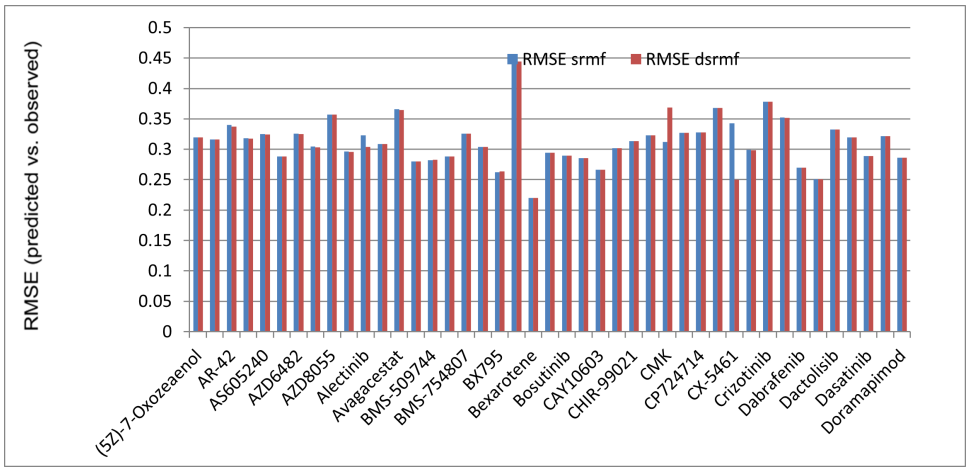

Figure 4. Comparison of the average RMSE_S / R between a sample of cell line drugs between the proposed method (dsrmf) and the previous method.

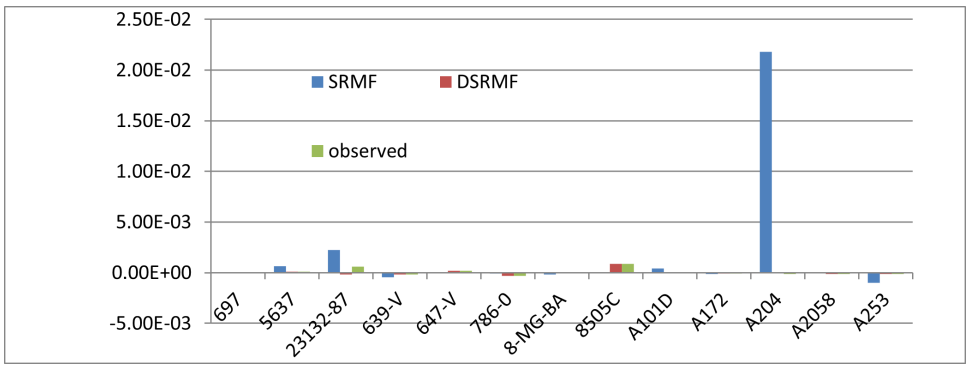

Figure 5. The average cell line response to drugs, which is the same IC50 value.

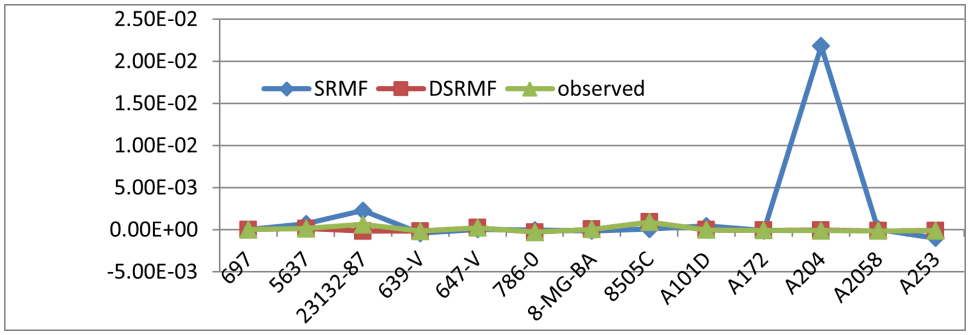

Figure 6. The average cell line response to drugs, which is the same IC50 value.

additional omics layers such as DNA methylation, microRNA expression, reverse-phase protein array (RPPA) data, and other multi-omics modalities to capture more complex biological relationships.

For the drug similarity network, a comprehensive matrix was constructed by integrating two-dimensional (2D) and three-dimensional (3D) chemical structures obtained from Structure Data Format (SDF) files available through the NCBI PubChem database. These structural representations were further enriched with chemical and molecular descriptors adopted from prior studies, thereby enhancing the precision of the drug similarity estimation.

Moreover, the DSRMF framework can be expanded to include additional genomic features—such as copy number variation (CNV), somatic mutation profiles, and pathway-level information—to further improve prediction accuracy. Owing to its modular and scalable architecture, DSRMF offers a versatile platform not only for anticancer drug response prediction but also for broader applications in pharmacogenomics, biomarker identification, and personalized therapeutic modeling.

Future works are as follows:

- Dynamic Similarity Learning
  Rather than using static similarity matrices, future work could involve learning similarity representations dynamically during training using deep metric learning or contrastive loss.
- Explain ability and Model Interpretation
  Integrate explainable AI (XAI) techniques to interpret which genes, pathways, or drug features contribute most to predicted responses, improving trust and adoption in clinical settings.
- Application to Drug Combination Prediction
  Extend the dual similarity approach to predict synergistic drug combinations, where pairwise drug–drug interaction effects are modeled jointly with similarity regularization.
- Scalability and Transferability
  Improve scalability to handle large-scale drug libraries and test transfer learning strategies to adapt models across datasets (e.g., from GDSC to CCLE or real-world data).

- Availability of data and materials

The cell line data in this article were derived from https://portals.broadinstitute.org/ccle

The drug data in this article was sourced from https://www.cancerrxgene.org/downloads/anova, and https://pubchem.ncbi.nlm.nih.gov/. Also, data can be sent via email as needed.

References

[1] P. Geeleher, N. J. Cox, R. S. Huang, Cancer biomarker discovery is improved by accounting

for variability in general levels of drug sensitivity in pre-clinical models, Genome Biology 17 (2016) 190. doi:10.1186/s13059-016-1050-9.

[2] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, et al., A landscape of pharmacogenomic interactions in cancer, Cell 166 (2016) 740–754. doi:10.1016/j.cell.2016.06.017.

[3] J. Barretina, G. Caponigro, N. Stransky, et al., The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (2012) 603–607. doi:10.1038/nature11003.

[4] M. J. Garnett, E. J. Edelman, S. J. Heidorn, et al., Systematic identification of genomic markers of drug sensitivity in cancer cells, Nature 483 (2012) 570–575. doi:10.1038/nature11005.

[5] L. C. Stetson, T. Pearl, Y. Chen, J. S. Barnholtz-Sloan, Computational identification of multiomic correlates of anticancer therapeutic response, BMC Genomics 15 (Suppl 7) (2014) S2. doi:10.1186/1471-2164-15-S7-S2.

[6] A. Basu, R. Mitra, H. Liu, S. L. Schreiber, P. A. Clemons, Rwen: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines, Bioinformatics 34 (2018) 3332–3339. doi:10.1093/bioinformatics/bty353.

[7] M. A. ud din, E. Georgii, M. Gönen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, S. Kaski, Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization, Journal of Chemical Information and Modeling 54 (2014) 2347–2359. doi:10.1021/ci500152b.

[8] N.-N. Guan, Y. Zhao, C.-C. Wang, J.-Q. Li, X. Chen, X. Piao, Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization, Molecular Therapy - Nucleic Acids 17 (2019) 164–174. doi:10.1016/j.omtn.2019.05.017.

[9] K. Matlock, C. D. Niz, R. Rahman, S. Ghosh, R. Pal, Investigation of model stacking for drug sensitivity prediction, BMC Bioinformatics 19 (Suppl 7) (2018) 71. doi:10.1186/s12859-018-2140-3.

[10] A. Honkela, M. Das, A. Nieminen, O. Dikmen, S. Kaski, Efficient differentially private learning improves drug sensitivity prediction (2018). doi:10.1186/s13062-018-0204-5.

[11] M. Gönen, A. A. Margolin, Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning, Bioinformatics 30 (2014) i556–i563. doi:10.1093/bioinformatics/btu264.

[12] H. Yuan, I. Paskov, H. Paskov, A. J. González, C. S. Leslie, Multitask learning improves prediction of cancer drug sensitivity, Scientific Reports 6 (2016) 31618. doi:10.1038/srep31618.

[13] D. Van, H. Pham, Drug response prediction by globally capturing drug and cell line information in a heterogeneous network, Journal of Molecular Biology 430 (2018) 2993–3004. doi:10.1016/j.jmb.2018.05.028.

[14] P. Shivakumar, M. Krauthammer, Structural similarity assessment for drug sensitivity prediction in cancer, BMC Bioinformatics 10 (Suppl 9) (2009) S17. doi:10.1186/1471-2105-10-S9-S17.

[15] S. Kim, V. Sundaresan, L. Zhou, T. Kahveci, Integrating domain specific knowledge and network analysis to predict drug sensitivity of cancer cell lines, PLoS One 11 (2016) e0162173. doi:10.1371/journal.pone.0162173.

[16] S. Mourragui, M. Loog, M. A. van de Wiel, M. J. T. Reinders, L. F. A. Wessels, Precise: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors, Bioinformatics 35 (2019) i510–i519. doi:10.1093/bioinformatics/btz372.

[17] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, X. Zheng, Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection (2015). doi:10.1186/s12885-015-1492-6.

[18] E. Fersini, E. Messina, F. Archetti, A p-median approach for predicting drug response in tumour cells, BMC Bioinformatics 15 (2014) 353. doi:10.1186/s12859-014-0353-7.

[19] L. Wang, X. Li, L. Zhang, Q. Gao, Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization, BMC Cancer 17 (2017) 513. doi:10.1186/s12885-017-3500-5.

[20] B. M. Kuenzi, J. Park, S. H. Fong, K. S. Sanchez, J. Lee, J. F. Kreisberg, J. Ma, T. Ideker, Predicting drug response and synergy using a deep learning model of human cancer cells, Cancer Cell 38 (2020) 672–684.e6. doi:10.1016/j.ccell.2020.09.014.

[21] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, J.-M. Shin, Cancer drug response profile scan (cdrscan): a deep learning model that predicts drug effectiveness from cancer genomic signature, Scientific Reports 8 (2018) 8857. doi:10.1038/s41598-018-27214-6.

[22] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, S. Venkatesh, Graphdta: predicting drug-target binding affinity with graph neural networks, Bioinformatics 37 (2021) 1140–1147. doi:10.1093/bioinformatics/btaa921.

[23] T. Ma, Q. Liu, H. Li, M. Zhou, R. Jiang,

X. Zhang, Dualgcn: a dual graph convolutional network model to predict cancer drug response, BMC Bioinformatics 23 (Suppl 4) (2021) 129. doi:10.1186/s12859-022-04664-4.

[24] B. M. Kuenzi, J. Park, S. H. Fong, K. S. Sanchez, J. Lee, J. F. Kreisberg, J. Ma, T. Ideker, Predicting drug response and synergy using a deep learning model of human cancer cells, Cancer Cell 38 (5) (2020) 672–684.e6. doi:10.1016/j.ccell.2020.09.014.

[25] M. Li, Y. Wang, R. Zheng, X. Shi, Y. Li, F.-X. Wu, J. Wang, Deepdsc: A deep learning method to predict drug sensitivity of cancer cell lines, IEEE/ACM Transactions on Computational Biology and Bioinformatics 18 (2021) 575–582. doi:10.1109/TCBB.2019.2919581.

[26] C. Wang, D. M. Blei, Collaborative topic modeling for recommending scientific articles, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (2011) 448–456doi:10.1145/2020408.2020480.

[27] V. Sindhwani, P. Niyogi, M. Belkin, Beyond the point cloud: from transductive to semi-supervised learning, Proceedings of the 22nd international conference on Machine learning (2005) 824–831doi:10.1145/1102351.1102455.

[28] P. Li, Z. Jiang, T. Liu, X. Liu, H. Qiao, X. Yao, Improving drug response prediction via integrating gene relationships with deep learning, Briefings in Bioinformatics 25 (2024) bbae153. doi:10.1093/bib/bbae153.

[29] R. Miao, B.-J. Zhong, X.-Y. Mei, X. Dong, Y.-D. Ou, Y. Liang, H.-Y. Yu, Y. Wang, Z.-H. Dong, A semi-supervised weighted spca- and convolution kan-based model for drug response prediction, Frontiers in Genetics 16 (2025) 1532651. doi:10.3389/fgene.2025.1532651.

[30] H. Liu, F. Wang, J. Yu, Y. Pan, C. Gong, L. Zhang, L. Zhang, Dbdnmf: A dual branch deep neural matrix factorization method for drug response prediction, PLoS Computational Biology 20 (2024) e1012012. doi:10.1371/journal.pcbi.1012012.

[31] M. Pak, D. Bang, I. Sung, S. Kim, S. Lee, Dgdrp: drug-specific gene selection for drug response prediction via re-ranking through propagating and learning biological network, Frontiers in Genetics 15 (2024) 1441558. doi:10.3389/fgene.2024.1441558.

[32] H. Luo, C. Zhu, J. Wang, G. Zhang, J. Luo, C. Yan, Prediction of drug–disease associations based on reinforcement symmetric metric learning and graph convolution network, Frontiers in Pharmacology 15 (2024) 1337764. doi:10.3389/fphar.2024.1337764.

[33] D. Kumari, D. Agrawal, A. Nema, N. Raj, S. Panda, J. Christopher, J. K. Singh, S. Behera, A study on improving drug-drug interactions prediction using convolutional neural networks, Applied Soft Computing 166 (2024) 112242. doi:10.1016/j.asoc.2024.112242.

[34] J. Barretina, et al., The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 (2012) 603–607. doi:10.1038/nature11003.

[35] W. Yang, et al., Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Research 41 (2013) D955–D961. doi:10.1093/nar/gks1111.

[36] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, X. S. Liu, Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model, PLoS Computational Biology 11 (2015) e1004498. doi:10.1371/journal.pcbi.1004498.

[37] I. Cortés-Ciriano, G. J. P. van Westen, G. Bouvier, M. Nilges, J. P. Overington, A. Bender, T. E. Malliavin, Improved large-scale prediction of growth inhibition patterns using the nci60 cancer cell line panel, Bioinformatics 32 (2016) 85–95. doi:10.1093/bioinformatics/btv529.

[38] M. A. ud din, E. Georgii, M. Gönen, T. Laitinen, O. Kallioniemi, K. Wennerberg, et al., Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization, Journal of Chemical Information and Modeling 54 (2014) 2347–2359. doi:10.1021/ci500152b.

Dr. Ali Reza Ebadi is a researcher who holds his Ph.D. degree in Computer Engineering from Islamic Azad University, Sanandaj, Iran. His research interests include data mining, bioinformatics, machine learning, Computational Biology, and Artificial Intelligence in Medical Applications.

Dr. Ali Soleimani researcher and associate professor at the university, holds a doctorate in computer science in Emerging Media from Colorado Technical University (Colorado, United States of America, 2013). Since 2013, he has been a faculty member of the Department of Computer and Engineering, Azad University, Tehran. He has over 20 years of experience in computer science. Dr. Soleimani's research blends Computer Science, Virtual Worlds, Virtual Education, Data Mining, and Machine Learning.

SeyedJafar Hashemizadehaghighi, a researcher who has over 15 years' experience in computer engineering. A highly dedicated M.Sc. graduate in Computer Engineering (Software) with a strong academic foundation and research interest in data mining, database systems, and intelligent data processing. Demonstrates a deep commitment to advancing knowledge in the areas of data-driven discovery, large-scale data management, and analytical modeling.