Prediction of Effective Drug Combinations based on Potential Drug Profiles

Changheng Li*

College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang, China
*Corresponding Author

Received: 01 January 2024/ Revised: 11 January 2024/ Accepted: 18 January 2024/ Published: 31-01-2024

Abstract— *Cancer is a great threat to the health of all mankind, and cancer monotherapy has been characterized by drawbacks such as toxicity and drug resistance. With the development of network pharmacology, multi-targeted drug combinations have become an ideal choice for cancer treatment. The dosage of combination drugs is usually lower than that of monotherapies, which has the advantages of improving efficacy, reducing toxicity, and delaying the development of drug resistance. In order to obtain better prediction results, this paper proposes a method for constructing drug potential features based on graph embedding model to predict anticancer drug combinations, establishes a control group to validate our method, and selects four performance metrics to measure the prediction performance of the model. The results show that the prediction results obtained from the drug potential features are better than the drug features. The drug potential features we designed can be used as one of the optional features for predicting drug combinations.*

Keywords— synergistic drug combinations, graph embedding, machine learning, cancer, neural network.

I. INTRODUCTION

Chemotherapy is a commonly used treatment for cancer, which often has many side effects, such as drug resistance and toxicity. With the development of modern medicine, drug combinations have become an ideal method for cancer treatment. By combining two or more anticancer drugs, drug toxicity can be reduced, drug resistance can be delayed, and efficacy can be improved. Therefore, finding synergistic combinations of drugs for specific cancer types is important to improve the efficacy of anticancer therapy[1-3].

Methods such as machine learning models offer the possibility to explore the combination space effectively. Machine learning models can quickly adapt to the ever-changing task of anticancer drug combination prediction and continuously optimize the prediction results. For example, by using machine learning models such as support vector machines, random forests, and neural networks, synergistic effects between different drugs can be effectively predicted[2, 4, 5].

In recent years, various prediction methods for anticancer drug combinations have been developed rapidly, e.g., Li et al [6]proposed a novel network propagation method based on gene-gene networks and drug-target information to simulate molecular signatures after treatment. By comparing the models of individual features, it was found that single-drug treatment data were better predictors of drug synergism than simulated molecular profiles. Janizek et al[7] predicted drug synergism using the physicochemical properties of the drug and the gene expression level of the cell line based on the XGBoost, and the results showed that 83 out of the 100 features with the highest level of evaluation were drug-based, which suggests that in the importance of drug features is higher than that of cell line features in the experiments for predicting drug combinations. Recently, Wang et al[8] proposed a new deep learning prediction model, PRODeepSyn. The model utilizes graph convolutional neural networks to integrate protein-protein interaction networks (PPIs) and histological data to construct low-dimensional embeddings of cell lines, and then constructs feed-forward neural networks with a batch normalization mechanism to compute drug synergy scores. In addition, Hu et al[9] understood the mechanism of drug synergism from the perspective of chemical-gene-tissue interactions and proposed a DTSyn model based on the mechanism of multiple attention to identify new drug

combinations, and they designed a fine-grained transformer encoder to capture chemical substructure-gene and gene-gene associations as well as a coarse-grained transformer encoder to extract chemical-chemical and chemical-cell lineage interactions, and finally designed a multilayer perceptron (MLP) to predict new drug combinations by connecting the outputs of the two transformer encoders in series as inputs to the MLP.

By summarizing the previous research, this paper understands that drug features are more important than cell line features in the experiments of predicting drug combinations, but through the research of the above scholars, there are few innovative studies on drug features, this paper constructs a set of potential features of drugs by using the graph embedding model, and puts forward a new method of predicting synergistic effects of anticancer drugs based on the constructing of potential features of drugs by the graph embedding model.

II. FEATURE ENGINEERING

2.1 Data set:

NCI-ALMANAC(National Cancer Institute - Analysis of Large-Scale Molecular Cancer Pharmacogenomic Data) is a publicly available database developed by the National Cancer Institute (NCI) to improve cancer therapeutic outcomes by identifying effective drug combinations and predicting patient response[10]. NCI-ALMANAC contains data on more than 60 cancer cell lines and more than 100 drugs, including FDA-approved drugs and experimental compounds. In this paper, we use the NCI-ALMANAC dataset as a source of anticancer drug synergy prediction data. In NCI-ALMANAC, the combinatorial effects of drugs are quantified by a method called ComboScore (a modified version of the Bliss independence model). The theoretical expectation of the effect when the effect is additive is calculated from the entire dose-response matrix considering the effects achieved by the combination of drug combinations, cell lineage metagenomes, and gains (or losses). Positive values of the ComboScore indicate that the drug combinations are synergistic, whereas negative values indicate that the drug combinations are not synergistic (those with purely additive effects obtain a ComboScore value of zero). In this paper, only drugs possessing at least one target gene were considered (68 drugs), and 59 cancer cell lines were selected through screening, with a total of 130,182 samples for model training and prediction. All NCI-60 cell line characteristics (expression, mutation, CNV, etc.) were downloaded from CellMinerCDB. Drug target information was obtained from DrugBank, and drug molecular properties were calculated using the RDKit package in Python.

2.2 Feature selection:

Gene expression profiling plays an important role in the prediction of anticancer drug combinations, and gene expression profiling can help reveal the mechanism of drug action on tumor cells. By analyzing the changes in gene expression profiles, the regulatory effects of certain anticancer drug combinations on specific signaling pathways, genes, or proteins can be understood, providing clues for understanding drug action. Referring to the studies of Janizek, RemziCelebi, and Preuer, et al [7, 11, 12], this paper decides to use the gene expression profiles as the cell lineage characteristics.

Morgan molecular fingerprints can be used to compare the similarities and correlations between different compounds, and by comparing the Morgan molecular fingerprints of anticancer drug molecules with those of other compounds, it is possible to predict possible interactions, including synergism, antagonism, etc., between them. Drug target information helps to better understand drug-target interactions. Monotherapy information may play an important role in anticancer drug combination prediction[6], and drug features may be more important than cell line features in prediction[7], In summary, in this paper, Morgan molecular fingerprints, drug target information, and monotherapy information are selected as drug features. In this paper, the selected drug features are utilized for comparison experiments with the designed drug potential features.

2.3 Machine learning and deep learning models:

2.3.1 CatBoost:

CatBoost is a gradient boosting tree framework with fewer parameters, support for categorical variables, and high accuracy, implemented as an oblivious trees-based learner. CatBoost uses ranked boosting to counteract noisy points in the training set, thus avoiding biased gradient estimation, and thus solving the prediction bias problem. CatBoost can match any state-of-the-art machine learning algorithm in terms of performance. rivals any state-of-the-art machine learning algorithm in that it reduces

the need for much hyper-parameter tuning, reduces the chance of overfitting, and makes the model more generalizable. In addition, CatBoost can handle categorical and numerical features and supports customized loss functions.

2.3.2 Deep Neural Network:

Feedforward Neural Network (FNN), also known as Multilayer Perceptron (MLP). It consists of multiple neurons and these neurons are arranged in a hierarchical structure. The basic structure of a feed forward neural network consists of an input layer, a hidden layer and an output layer. The input layer receives external input data, the hidden layer processes the data and extracts features, and the output layer performs classification or regression prediction based on the results of the hidden layer.

2.3.3 XGBoost:

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm based on gradient boosting decision trees. The core idea of XGBoost is to iteratively train weak classifiers (usually decision trees) and combine them into a powerful model. It employs gradient boosting to effectively optimize the objective function. Specifically, XGBoost improves the model's predictions incrementally by minimizing the negative gradient of the loss function. XGBoost performs well in all kinds of machine learning tasks, including classification, regression, and sorting.

2.3.4 Logistic Regression:

Logistic Regression is a statistical learning method for classification problems. It is a generalized linear model that makes classification decisions by mapping the output of a linear regression model to a probability value and using a logistic function. Logistic Sti regression assumes that there is a linear relationship between the dependent and independent variables and uses a logistic function (also known as a sigmoid function) to map the outcome of the linear combination to a probability between [0, 1], the sigmoid function formula is shown below:

$$P(y = 1|x) = \frac{1}{1+e^{-z}} \quad (1)$$

where $P(y = 1|x)$ denotes the probability that the dependent variable y takes the value 1 when the independent variable x is given, and z denotes the outcome of the linear combination. The training process of logistic Steele regression is to solve the parameters of the model by maximum likelihood estimation. Logistic Steele regression is widely used in practical applications, especially for binary classification problems, and is often used as a benchmark comparison for other machine learning algorithms.

III. CONSTRUCTION OF POTENTIAL DRUG FEATURES

In this paper, graph embedding model is utilized to construct drug based potential features. Graph embedding model is a technique for mapping nodes in a graph to a low-dimensional vector space, also known as graph representation learning. It converts high-dimensional graph data into a low-dimensional continuous vector representation by learning the relationships and features between nodes while preserving the original graph structure information.

Here, this paper proposes a network on drug-drug interaction. In this network, drugs are nodes and drug-drug interactions are edges. Considering that there are two kinds of drug-drug interactions, synergistic or antagonistic, a set of signed directed graphs is constructed, where "+" indicates that there is a synergistic interaction between two drugs and "-" idicates the existence of antagonism between two drugs. However, the effect of anticancer combination drugs can vary in different cell lines. This is because different cell lines have different genetic backgrounds, gene expression patterns, metabolic characteristics, etc., and thus the sensitivity and responsiveness to drugs will be different. Therefore, when constructing the drug-drug interaction network in this paper, cell lines need to be included in the consideration of the network.

In this paper the drug network contains three kinds of relationships, synergistic, antagonistic and unconnected edges, where synergistic represents a positive relationship and antagonistic represents a negative relationship. Referring to the study of Xu et al[13], for a given node pair (u, v), this paper calculates the inner product of U_u^{out}(positive outward vector) and U_v^{in} (positive received vector) and calls the result positive feedback F_{uv}^{+}, and calculates the inner product of W_u^{out} (negative outward vector) and W_v^{in} (negative received vector) and calls the result negative feedback F_{uv}^{-}. Next, this paper quantifies the effects of positive and negative feedback on the drug relationship with distributions denoted as $f_a(U_u^{out}U_v^{in})$and $f_a(W_u^{out}W_v^{in})$, where:

$$f_a(x) = \frac{p_0 e^x}{1+p_0(e^x-1)} \tag{2}$$

$f_a(x)$ is the activation function and $p_0$denotes the effect of no feedback. Next this paper defines the quantization of synergy, antagonism and zero as $F_{uv}^{+}(1-F_{uv}^{-})$, $(1-F_{uv}^{+})F_{uv}^{-}$ and $(1-F_{uv}^{+})(1-F_{uv}^{-})$,respectively. The node information in this paper is represented by four vectors, U_u^{out}, U_v^{in}, W_u^{out} and W_v^{in}, and the relationship from node u to v may contain synergism, antagonism and zero, which can be expressed as $f_a(U_u^{out}U_v^{in})\left(1-f_a(W_u^{out}W_v^{in})\right)$,$\left(1-f_a(U_u^{out}U_v^{in})\right)f_a(W_u^{out}W_v^{in})$ and $\left(1-f_a(U_u^{out}U_v^{in})\right)\left(1-f_a(W_u^{out}W_v^{in})\right)$.

In this paper, the negative log-likelihood loss function is chosen as the objective function, Negative Log-Likelihood Loss Function is a common loss function used in classification models, especially in logistic regression models are often used, Negative Log-Likelihood Loss Function has a lot of advantages in classification models. In order to relate the four vectors and the three drug relationships, the objective function is defined in this paper as the following equation:

$$L = -\sum_{(u,v)\in E^p}\log\left(F_{uv}^{+}(1-F_{uv}^{-})\right) - \sum_{(u,v)\in E^n}\log((1-F_{uv}^{+})F_{uv}^{-}) - \sum_{(u,v)\in E^{non}}\log\left((1-F_{uv}^{+})(1-F_{uv}^{-})\right) \tag{3}$$

Where E^p denotes the set of positively connected edges, E^n denotes the set of negatively connected edges, and E^{non} denotes the set of node pairs with no connected edges.

In order to minimize the objective function, this paper chooses gradient descent as the optimization function, which is a commonly used optimization algorithm to minimize the loss function and find the optimal parameters. The basic idea of gradient descent is to gradually reduce the value of the loss function by iteratively updating the parameters. Equation 3 consists of three components three components, the distribution for the synergistic, antagonistic and null relationships. In this paper, we use gradient descent to update the three vectors of one node in each iteration while fixing the vectors of all other nodes, so the problem becomes a convex optimization problem. The specific formulation is as follows:

$$L_{(u)} = -\sum_{v\in N_{out}^{p}(u)}\log\left(F_{uv}^{+}(1-F_{uv}^{-})\right) - \sum_{v\in N_{in}^{p}(u)}\log\left(F_{uv}^{+}(1-F_{uv}^{-})\right) - \sum_{v\in N_{out}^{n}(u)}\log((1-F_{uv}^{+})F_{uv}^{-}) - \sum_{v\in N_{in}^{n}(u)}\log((1-F_{uv}^{+})F_{uv}^{-}) - \sum_{v\in N_{out}^{non}(u)}\log\left((1-F_{uv}^{+})(1-F_{uv}^{-})\right) - \sum_{v\in N_{in}^{non}(u)}\log\left((1-F_{uv}^{+})(1-F_{uv}^{-})\right) \tag{4}$$

Where $N_{out}^{p}(u)$,$N_{out}^{n}(u)$,$N_{out}^{non}(u)$ denotes the set of three types of nodes with synergistic, antagonistic and zero relationships that node u outputs to node v. $N_{in}^{p}(u)$,$N_{in}^{n}(u)$,$N_{in}^{non}(u)$ denotes the set of three types of nodes with synergistic, antagonistic and zero relationships that node v inputs to u.

The dataset in this paper contains a total of 59 cell lines, which need to be divided into 59 copies, with each cell line corresponding to one copy of the dataset, which eliminates the interference of cell lines on the drug-drug interaction network. Next, the drug-drug interaction network needs to be established for each of these 59 copies of data. In this paper, the dimension of the potential features is set to be 16 dimensions, so the 16-dimensional potential features of each drug in the 59 cell lines are obtained. Since the data contained in these potential features are abstract and topological information of the network structure, it is difficult to capture the significance represented by each column, so this paper decided to merge the potential features of each drug in these 59 cell lines. Eventually, the dimension of the potential features for each drug is 944 dimensions, which is a 1 × 944 vector.

IV. PREDICTION RESULTS AND ANALYSIS

4.1 Experimental setup

In order to make the model generalizable to unseen datasets, this paper uses Stratified k-fold cross-validation to test the model. The advantage of stratified k-fold cross-validation is that it can better handle unbalanced datasets and ensure that the samples of each category are adequately represented during training and testing. This helps to reduce bias problems due to category imbalance and provides a more reliable assessment of model performance. Through stratified 5-fold cross-validation, this paper also adjusts the parameters of each model to obtain the optimal model.

Drug synergy prediction can be classified into two categories: regression task and classification task. In this paper, synergy prediction is considered as a classification task. Therefore, this paper chooses to binarize the synergy score, i.e., synergy is 1

and antagonism is 0. In this paper, the prediction thresholds of the models are optimized through hierarchical 5-fold cross validation to achieve the optimal equilibrium, and the final threshold is set to 10.

4.2 Comparison of predictive performance

In this paper, four performance metrics are used to measure the prediction performance of the model, which are the area under the receiver operating characteristic curve (ROC AUC), the area under the precision recall curve (PR AUC), the mean square error (MSE) and the Pearson correlation coefficient (PCC). The above performance metrics were calculated by five hierarchical cross-folding, and the average of the five results was taken as the final performance evaluation result. The predicted results are as follows:

TABLE 1
PERFORMANCE RESULTS OF FOUR MODELS (CELL LINE CHARACTERISTICS AND DRUG CHARACTERISTICS)

Model	ROC AUC	PR AUC	MSE	Pearson
CatBoost	0.9217	0.4651	0.1365	0.5335
Deep Neural Networks	0.9118	0.3876	0.1441	0.4761
XGBoost	0.8856	0.3601	0.1439	0.4552
Logistic Regression	0.8505	0.1945	0.1534	0.3101

TABLE 2
PERFORMANCE RESULTS OF FOUR MODELS (CELL LINE FEATURES AND DRUG POTENTIAL FEATURES)

Model	ROC AUC	PR AUC	MSE	Pearson
CatBoost	0.9258	0.4937	0.1348	0.5497
Deep Neural Networks	0.9178	0.4065	0.1459	0.4760
XGBoost	0.8943	0.3690	0.1430	0.4635
Logistic Regression	0.8657	0.2460	0.1507	0.3583

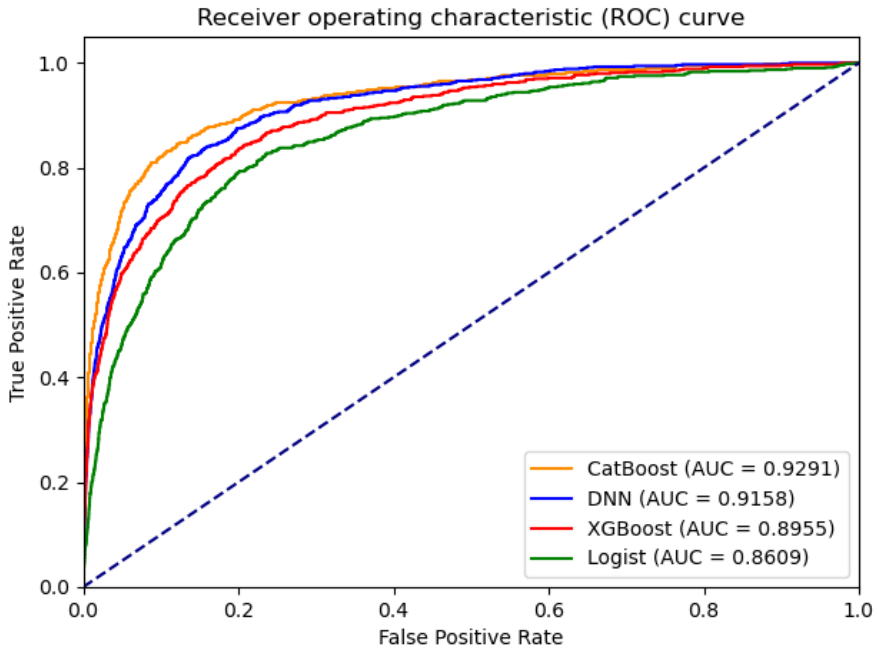

FIGURE 1: ROC curves of four models (cell line characteristics and drug characteristics)

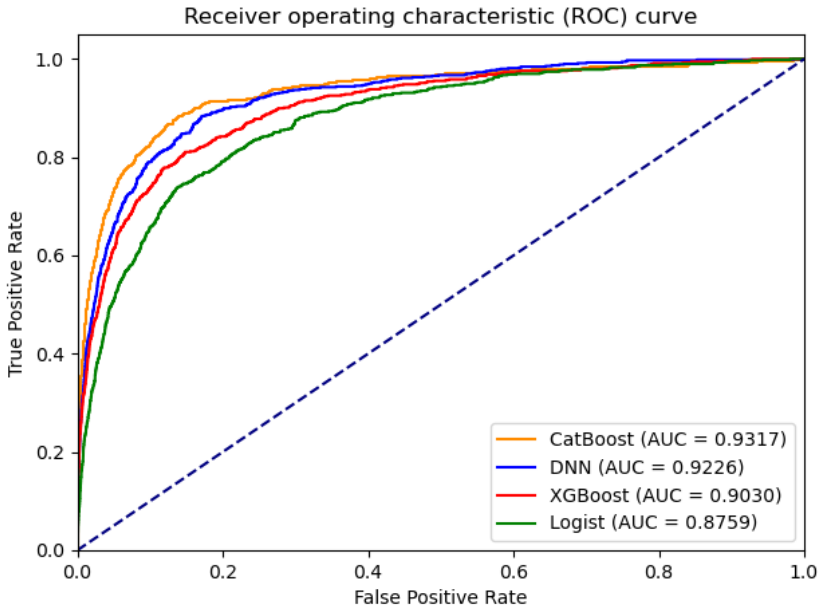

FIGURE 2: ROC curves of four models (cell line characteristics and drug potential characteristics)

Since Morgan molecular fingerprints, drug target information and monotherapy information have been outstanding in drug combination prediction, this paper firstly selects these three features together with cell line gene expression profiles as input features, and then utilizes popular machine learning algorithms for prediction respectively. Using different models for prediction and comparing the performance indexes of the above evaluated models, CatBoost, Deep Neural Networks, XGBoost and Logistic Regression algorithms finally show good prediction effects. The specific indexes are shown in Table 1, and the ROC curve is shown in Figure 1. Afterwards, in order to reflect the role of potential features of drugs in predicting effective drug combinations, this paper utilizes the potential features of drugs designed by the graph embedding model to replace the drug features to join, keep the parameters unchanged, and then carry out the same operation as above, and then the obtained indicators are shown in Table 2, and the ROC curves are shown in Fig. 2.

4.3 Result analysis

From the comparison of Tables 1 and 2, it can be seen that CatBoost and Deep Neural Networks have the best prediction effect, while XGBoost and Logistic Regression have poorer prediction effect when using cell line features and drug features for prediction. After using cell line features and drug potential features for prediction, the performance of the quasi-four models are improved to different degrees. In the case of ROC AUC, XGBoost and Logistic Regression improved more significantly, by 0.0087 and 0.0179, respectively. In the PR AUC, CatBoost and Logistic Regression are improved by 0.0283 and 0.0515 respectively, and in the MSE, CatBoost and Logistic Regression are reduced by 0.0087 and 0.0179 respectively. In the item of PCC, CatBoost and Logistic Regression improved more obviously. In summary, the drug potential features proposed in this paper are significantly better than the drug features and can be used as one of the optional features for predicting drug combinations.

V. CONCLUSION

The use of combinatorial drugs can undoubtedly help people to treat complex diseases, while the use of computer technology, histology and network technology helps people to discover new drug combinations and is therefore a proven means. The method greatly reduces the scope of the search and is safer and more reliable in a small area before experimental tests are carried out. Discovering new reliable features is also one of the keys to accurate prediction, and the drug potential features designed in this paper play an important role in improving the prediction accuracy, and among the four models, the contribution of drug potential features to the prediction accuracy is significantly higher than that of drug features. The Morgan molecular fingerprints, drug target information and monotherapy information used in this paper are basically classical and have been used by previous authors, so their credibility can be guaranteed.

The principle of constructing drug potential features in this paper lies in the need for a large amount of drug synergy information, and the accuracy can be further increased if enough drug synergy information is available. In addition, the drug

potential features constructed in this paper can also be utilized in the migration learning method, using a large number of drug-drug interactions in the data set, to extract the potential information to construct drug potential features, which can be applied to other prediction tasks that lack cell line information or drug feature information, which is also one of the future research directions. Finally, the models used in this paper are supervised models, and it is believed that the prediction accuracy will be further improved if semi-supervised models or other more advanced algorithms are utilized.

REFERENCES

[1] G. Chevereau and T. Bollenbach, "Systematic discovery of drug interaction mechanisms," *Molecular systems biology,* vol. 11, no. 4, p. 807, 2015.

[2] J. Jia *et al.*, "Mechanisms of drug combinations: interaction and network perspectives," *Nature Reviews Drug Discovery,* 2009.

[3] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery. A comprehensive review," *Pharmacology & therapeutics,* vol. 138, no. 3, pp. 333-408, 2013.

[4] L. He *et al.*, "Methods for High-Throughput Drug Combination Screening and Synergy Scoring," *Cold Spring Harbor Laboratory,* 2016.

[5] J. O'Neil *et al.*, "An unbiased oncology compound screen to identify novel combination strategies," *Molecular cancer therapeutics,* vol. 15, no. 6, pp. 1155-1162, 2016.

[6] H. Li, T. Li, D. Quang, and Y. Guan, "Network propagation predicts drug synergy in cancers," *Cancer research,* vol. 78, no. 18, pp. 5446-5457, 2018.

[7] J. D. Janizek, S. Celik, and S. I. Lee, "Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine," *Cold Spring Harbor Laboratory,* 2018.

[8] X. Wang *et al.*, "PRODeepSyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network," *Briefings in Bioinformatics,* vol. 23, no. 2, p. bbab587, 2022.

[9] J. Hu *et al.*, "DTSyn: a dual-transformer-based neural network to predict synergistic drug combinations," *Briefings in Bioinformatics,* vol. 23, no. 5, p. bbac302, 2022.

[10] S. L. Holbeck *et al.*, "The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity," *Cancer Research,* 2017.

[11] R. Celebi, O. Bear Don't Walk, R. Movva, S. Alpsoy, and M. Dumontier, "In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data," *Scientific Reports,* vol. 9, no. 1, pp. 1-10, 2019.

[12] K. Preuer, R. P. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer, "DeepSynergy: predicting anti-cancer drug synergy with Deep Learning," *Bioinformatics,* vol. 34, no. 9, pp. 1538-1546, 2018.

[13] P. Xu, W. Hu, J. Wu, and B. Du, "Link prediction with signed latent factors in signed social networks," in *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2019, pp. 1046-1054.

[14] CHEVEREAU G, BOLLENBACH T. Systematic discovery of drug interaction mechanisms [J]. Molecular systems biology, 2015, 11(4): 807.

[15] AMZALLAG A, RAMASWAMY S, BENES C H. Statistical assessment and visualization of synergies for large-scale sparse drug combination datasets [J]. BMC bioinformatics, 2019, 20(1): 1-15.

[16] LUKAČIŠIN M, BOLLENBACH T. Emergent gene expression responses to drug combinations predict higher-order drug interactions [J]. Cell Systems, 2019, 9(5): 423-33. e3.

[17] LAKHTAKIA R, BURNEY I. A historical tale of two Lymphomas: part II: non-Hodgkin lymphoma [J]. Sultan Qaboos University Medical Journal, 2015, 15(3): e317.

[18] PROPERZI M, MAGRO P, CASTELLI F, et al. Dolutegravir-rilpivirine: first 2-drug regimen for HIV-positive adults [J]. Expert Review of Anti-Infective Therapy, 2018, 16(12): 877-87.

[19] DAVIES G, BOEREE M, HERMANN D, et al. Accelerating the transition of new tuberculosis drug combinations from Phase II to Phase III trials: New technologies and innovative designs [J]. PLoS medicine, 2019, 16(7): e1002851.

[20] MENDEN M P, WANG D, MASON M J, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen [J]. Nature communications, 2019, 10(1): 2674.

[21] YI L, ZHOU L, LUO J, et al. Circ-PTK2 promotes the proliferation and suppressed the apoptosis of acute myeloid leukemia cells through targeting miR-330-5p/FOXM1 axis [J]. 2021, 86: 102506.

[22] DUQUESNE J, BOUGET V, COURNEDE P H, et al. Machine learning identifies a profile of inadequate responder to methotrexate in rheumatoid arthritis [J]. Rheumatology, 2023, 62(7): 2402-9.