OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–14
https://doi.org/10.1093/bib/bbab587
Problem Solving Protocol

PRODeepSyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network

Xiaowen Wang†, Hongming Zhu†, Yizhi Jiang, Yulong Li, Chen Tang, Xiaohan Chen, Yunjie Li, Qi Liu and Qin Liu

Corresponding authors: Qin Liu, School of Software Engineering, Tongji University, Shanghai 201804, China. Tel.: +86-021-69589075; E-mail: qin.liu@tongji.edu.cn; Qi Liu, Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China. Tel.: +86-021-65980296; E-mail: qiliu@tongji.edu.cn
†These authors contribute equally to this work.

Abstract

Although drug combinations in cancer treatment appear to be a promising therapeutic strategy with respect to monotherapy, it is arduous to discover new synergistic drug combinations due to the combinatorial explosion. Deep learning technology holds immense promise for better prediction of *in vitro* synergistic drug combinations for certain cell lines. In methods applying such technology, omics data are widely adopted to construct cell line features. However, biological network data are rarely considered yet, which is worthy of in-depth study. In this study, we propose a novel deep learning method, termed PRODeepSyn, for predicting anticancer synergistic drug combinations. By leveraging the Graph Convolutional Network, PRODeepSyn integrates the protein–protein interaction (PPI) network with omics data to construct low-dimensional dense embeddings for cell lines. PRODeepSyn then builds a deep neural network with the Batch Normalization mechanism to predict synergy scores using the cell line embeddings and drug features. PRODeepSyn achieves the lowest root mean square error of 15.08 and the highest Pearson correlation coefficient of 0.75, outperforming two deep learning methods and four machine learning methods. On the classification task, PRODeepSyn achieves an area under the receiver operator characteristics curve of 0.90, an area under the precision–recall curve of 0.63 and a Cohen's Kappa of 0.53. In the ablation study, we find that using the multi-omics data and the integrated PPI network's information both can improve the prediction results. Additionally, the case study demonstrates the consistency between PRODeepSyn and previous studies.

Keywords: synergistic drug combinations, deep learning, graph convolutional network, protein–protein interaction network, omics data

Introduction

Drug combination therapy [1] is usually adopted to treat complex diseases, such as cancer [2]. Compared with monotherapy, drug combination therapy can improve the efficacy of cancer treatment [3], decrease the dose-dependent toxicity of drugs [4] and prevent the development of drug resistance [5]. However, the drug combinations will not only present synergistic effects but also may have antagonistic or additive effects [6, 7].

One of the biggest challenges in discovering new synergistic drug combinations is the combinatorial explosion. As the number of drugs increases, the size of the complete drug combination space grows rapidly. Although the mechanism of drug synergy has been explored

Xiaowen Wang is a doctoral candidate of the School of Software Engineering, Tongji University, Shanghai, China. Her research interests lie in the field of deep learning and drug combination therapy.
Hongming Zhu is an associate professor of the School of Software Engineering, Tongji University, Shanghai, China. His research interests include knowledge graph, bioinformatics and remote sensing data analysis.
Yizhi Jiang is a master candidate of the School of Software Engineering, Tongji University, Shanghai, China. His research interests lie in the field of the entity alignment between cross-lingual knowledge graphs and graph representation learning.
Yulong Li is a master candidate of the School of Software Engineering, Tongji University, Shanghai, China. His research interests lie in the field of predicting anticancer synergistic drug combinations and drug representation.
Chen Tang is a doctoral candidate of the School of Life Sciences and Technology, Tongji University, Shanghai, China. His research interests lie in the field of tumor precision medicine based on omics characterization.
Xiaohan Chen is a master candidate of the School of Life Sciences and Technology, Tongji University, Shanghai, China. Her research interests lie in the field of precision medicine based on artificial intelligence and multi-omics data.
Yunjie Li is a master candidate of the School of Life Sciences and Technology, Tongji University, Shanghai, China. His research interests lie in the field of precision medicine based on artificial intelligence and multi-omics data.
Qi Liu is a professor of the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China. He directs the Biological and Medical Big data Mining Laboratory.
Qin Liu is a professor of the School of Software Engineering, Tongji University, Shanghai, China. Her research interests include heterogeneous graph data, deep graph learning and big data in bioinformatics.

Received: September 28, 2021. **Revised:** December 20, 2021. **Accepted:** December 21, 2021

[8–10], most synergistic drug combinations are currently proposed based on clinical experience [11, 12]. Early studies [13, 14] require *in vivo* experiments. Such trial-based methods have the drawbacks of being time-consuming, labour-intensive and costly and may cause patients to receive unnecessary or even harmful treatments. The high-throughput screening technology (HTS) can test cell lines in plates efficiently [15]. However, it is not feasible to test the complete combination space with HTS, and it requires significant cost infrastructure construction [16]. Therefore, researchers turn their attention to computational methods.

The early design of computational methods is inspired by systems biology [17, 18] and the integration of biology and non-biology [19]. Several studies treat the biological system as a black box and establish statistical models. But it is difficult for them to fit complex nonlinear biological processes. Methods based on the explicit model [20–23] try to simulate the influence of drug combinations on the biological network to optimize the combination plan. However, they are only applicable to specific targets, pathways, diseases or cell lines.

The boom of artificial intelligence, including machine learning (ML) and deep learning (DL), has promoted the proposal of new computational methods. A comparative advantage is that these methods can simulate complex nonlinear processes. The typical pattern of these methods is to first construct features for cell lines and drugs and then perform prediction with ML or DL models. Studies before 2018 adopt various of ML models to predict the synergy of anticancer drug combinations, including Random Forest [24, 25], Logistic Regression [26], XGBoost[27] and Extremely Randomized Trees [28]. Web tools like H-RACS provide convenient services for predicting drug combinations with ML [29]. In recent years, the publication of multiple large-scale synergy datasets [30, 31] provides a data basis for applying DL methods in drug combination prediction. The applied DL models mainly include Deep Neural Network (DNN) [32, 33] and Residual Neural Network [34]. There is also a DL method that focuses on the model's interpretability [35]. Besides, special technologies such as Ensemble Learning [36, 37], Transfer Learning [38] and Tensor factorization [39] are also introduced into the field of drug combination prediction. Readers could refer to [40] for a more comprehensive review of ML and DL methods.

Cell line features play an indispensable role in discovering synergistic drug combinations, because the drug combination that has been validated on one cell line may not be effective on another [41]. The omics data are commonly used in ML and DL methods to construct cell line features [26–29, 32–35]; however, the biological networks are rarely considered. Li's group proposes a network propagation strategy that simulates post-treatment cell line features based on the drug targets and a mouse gene interaction network [25]. But the propagation strategy considers only the direct linkage between target and non-target genes, which could be too simplistic to make full use of the interactions between genes. This is the only study that attempts to construct cell line features with network data to our knowledge. Since the network-based treatment is argued to have various potential biological and clinical applications [42], the absence of network data in constructing cell line features may prevent the further development of synergistic drug combination prediction methods.

To fill the gaps in the application of network data to construct cell line features, this study proposes a novel method named PRODeepSyn for predicting anticancer synergistic drug combinations. PRODeepSyn integrates the protein–protein interaction (PPI) network data with the omics data using the deep learning model graph convolutional network (GCN) [43] to predict anticancer synergistic drug combinations. GCN is designed specifically for graph-structured data such as biological networks, but only few works [44, 45] use GCN to extract informative features of biological networks in the domain of drug combination prediction yet, and none of them consider about cell line features. Specifically, PRODeepSyn extracts the topological structure of PPI network data and omics data to construct low-dimensional dense embeddings for cell lines with GCN. The GCN model is trained with semi-supervised learning. Besides, PRODeepSyn uses drug molecular fingerprints and descriptors to construct the drug features. Finally, PRODeepSyn predicts the synergy scores of drug combinations using DNN that has the Batch Normalization mechanism. On the O'Neil dataset [30], we have verified that PRODeepSyn outperforms other state-of-the-art methods, including both ML and DL methods. We have also performed the ablation study and sensitivity analysis to present more details of PRODeepSyn. The case study also proves that the predictions of PRODeepSyn are consistent with many previous studies. Overall, PRODeepSyn is expected to be a satisfactory prediction method of anticancer synergistic drug combinations.

Materials and methods

Synergy dataset

A large-scale synergy dataset published by O'Neil *et al.* [30] is used to train and evaluate our model. This dataset covers experiment results of 583 different drug combinations on 39 cancer cell lines from 7 tissues. Each experiment was repeated four times with a 4 × 4 dosage regimen, and the cell growth rate relative to the control group after 48 hours was measured. The results of single drug screening on the same cell lines were released simultaneously. Preuer *et al.* [32] integrated this dataset by computing Loewe Additivity Value [46]. The integrated dataset contains 23 062 samples, each consisting of two drugs and one cell line, as well as the corresponding synergy score. All samples are divided into five disjoint folds with equivalent quantity concerning drug combinations, which means the same drug combination does

not exist between folds. Therefore, the divided samples can be used to evaluate the ability of methods to predict novel drug combinations with cross-validation.

Drug features

As shown in Figure 1A, in order to represent the structural and physicochemical properties of drugs, the molecular fingerprint and descriptors are used to construct the feature vector for each drug. Preuer *et al*. [32] provide the SMILES expression of each drug contained in the O'Neil dataset. We adopt RDKit [47] to compute the fingerprint and descriptors for each drug based on their SMILES expressions. Firstly, we generate the Morgan fingerprint [48] with a radius of 2 for each drug and represent it as a 256-dimensional binary-valued vector. Afterwards, we obtain 200 descriptors of each drug that compose a real-valued vector. At last, we concatenate two types of feature vectors mentioned above and filter out features with zero variance. Finally, 253-dimensional Morgan fingerprint together with 163-dimensional descriptors are retained as features of each drug. That is, the final feature vectors of drugs are 416-dimensional. We use the z-score normalization method to preprocess the drug features to eliminate the possible impact of the scale of the features.

Cell line features

PRODeepSyn integrates three types of heterogeneous cell line features containing gene expression data, gene mutation data and interactions between gene expression products to construct embeddings for cell lines. The gene expression data is downloaded from the ArrayExpress database (accession number: E-MTAB-3610) [49]. A total of 3739 informative genes are first summarized with the Factor Analysis for Robust Microarray Summarization method [50] and then processed by the z-score normalization method. Gene mutation data of cell lines are obtained from the COSMIC cell lines project [51]. We remove data whose mutation type is *coding silent* or *unknown* and retain the mutation data of 12 333 genes for 39 cell lines. The gene mutation data of each cell line is represented as a 12 333-dimensional binary-valued vector. According to whether the cell line is mutated on a gene, the corresponding element of the vector is 0 or 1. Interactions between gene expression products are collected from the PPI network contained in the STRING database [52]. We ignore the interactions whose combined scores are lower than 0.7 in STRING and retain a total of 839 522 interactions between 17 161 proteins. We associate the nodes in the PPI network with gene expression data and gene mutation data via gene identifiers and symbols. More details for linking data are provided in the Supplementary Material.

PRODeepSyn

In this paper, we propose a novel DL method named PRODeepSyn to predict synergy scores of drug combinations on cell lines. The overview of our method is shown in Figure 1. Drug features are constructed as described above, whereas cell line embeddings are constructed with StateEncoder that integrates the PPI network with multi-omics data. Afterwards, the Predictor of PRODeepSyn utilizes the constructed drug features and cell line embeddings to predict synergy scores. More details about PRODeepSyn are introduced below.

Cell line embeddings

The high-dimensional sparse omics data are widely used to construct feature vectors of cell lines. Instead of directly using omics data as model input, filtering essential genes [53, 54] or using AutoEncoder to reduce the dimensionality of omics data [55, 56] can substantially reduce the number of model parameters and improve computational efficiency. PRODeepSyn also aims to construct low-dimensional dense embeddings for cell lines. The significant difference is that PRODeepSyn integrates the information of the PPI network into cell line embeddings. We expect such embeddings to represent cell line states that are predictive to drug synergy scores. We name the expression level of a certain gene or its mutation result as the explicit state of the gene. Inspired by the fact that the same gene expresses differently among different cell lines [57], PRODeepSyn assumes that the gene explicit state results from the interaction between the cell line-independent gene hidden state and the cell line state. When the gene expression level and mutation result are known, we need to answer two questions: 1) how to construct the gene hidden state? 2) How to solve the state of the cell line through the explicit state and hidden state of genes?

Construction of gene hidden states

For the first question, PRODeepSyn leverages the GCN model [43] to construct gene hidden states based on the PPI network, considering that the interactions between proteins expressed by genes are important in drug combination therapy [25]. The PPI network is a typical kind of graph-structured data where proteins are nodes and the interactions are edges. For graph-structured data, both the node's properties and the network's topology are of great significance. GCN is a graph representation learning model that can generate low-dimensional dense embeddings for nodes while retaining the information of nodes and the network topology. PRODeepSyn constructs the embeddings of the nodes in the PPI network as gene hidden states with GCN.

We first agree on symbols to explain the principle of the GCN model more clearly. Graph G consists of the vertex set V and the edge set E, i.e., $G = (V, E)$. The feature matrix of nodes $X \in \mathbb{R}^{N\times K}$, where N is the number of nodes, and K is the number of features. GCN aims to obtain the node embedding matrix $Z \in \mathbb{R}^{N\times F}$, where F is the dimension of embedding space. The adjacency matrix of an undirected graph is denoted as $A \in \mathbb{R}^{N\times N}$, in which the element $A_{i,j}$ is 1 when there exists an edge between node i and j; otherwise, it is 0. The degree matrix

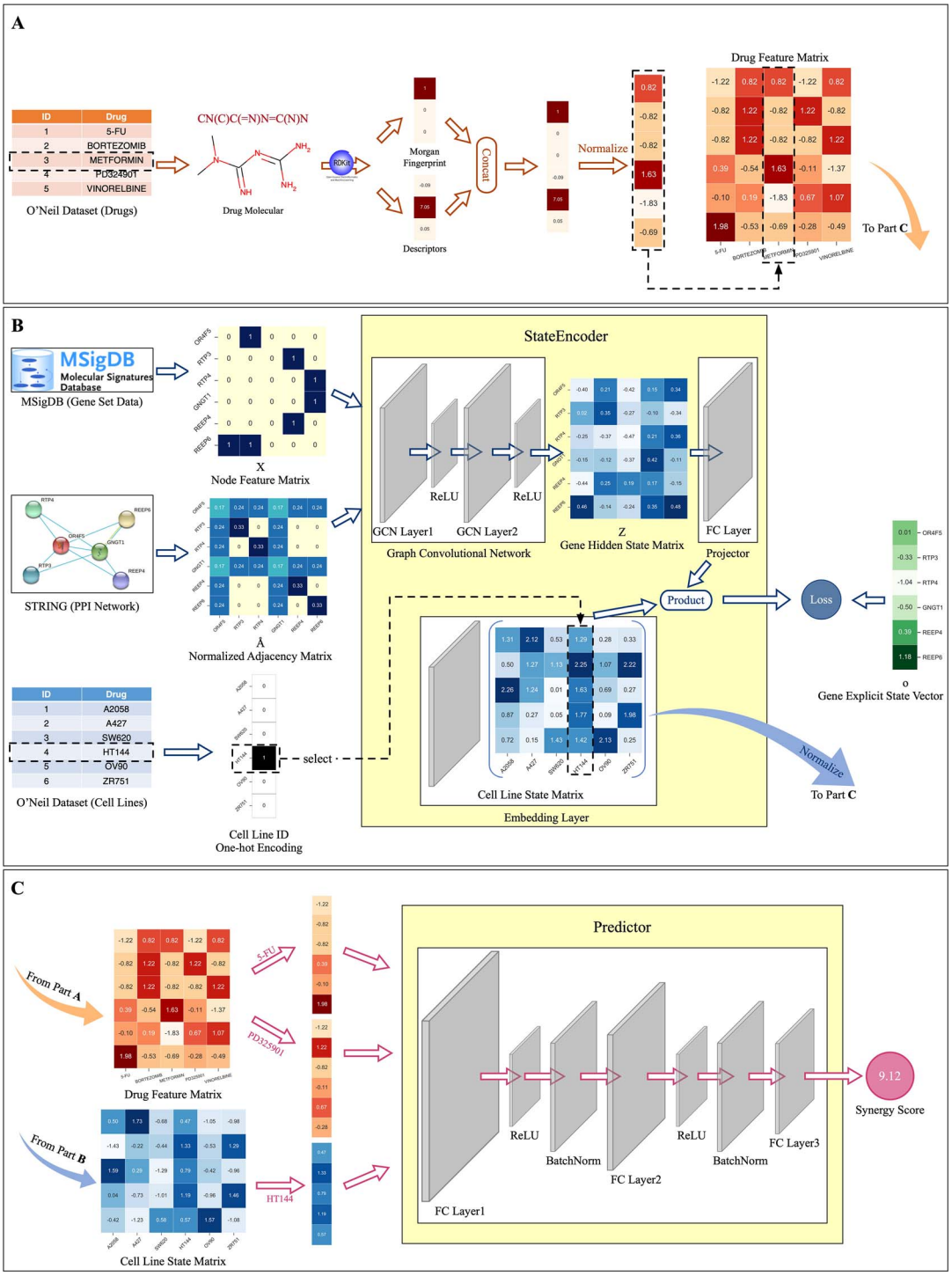

Figure 1. Overview of PRODeepSyn. (**A**) PRODeepSyn constructs drug features with their Morgan fingerprints and descriptors using RDKit based on drugs' SMILES expressions. The fingerprints and descriptors are concatenated as the final feature vectors of drugs. (**B**) PRODeepSyn extracts the hidden state matrix of genes from the PPI network with the GCN and regards the omics data as the explicit state vector of genes. The dot product between the projected gene hidden state matrix and the cell line embedding is applied to approximate the gene explicit state vector. (**C**) PRODeepSyn predicts the synergy score of a drug combination on a certain cell line using the DNN with the batch normalization mechanism.

of the graph is a diagonal matrix D, where the element $D_{i,i}$ on the diagonal is equal to the degree of node i ($D_{i,i} = \sum_j A_{i,j}$).

GCN aggregates the information of the neighbours of the node and the node itself as the node's information by defining the convolution operation on the graph. With multiple convolutional layers, GCN generates the embeddings of a node with information of its multi-hop neighbours and itself. The output of the l-th layer $H^{(l)}$ is

$$H^{(l)} = \sigma(\hat{A}H^{(l-1)}W^{(l)}),\ \hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \quad (1)$$

where $\tilde{A} = A + I$ and I is the identity matrix, $W^{(l)}$ is the parameter matrix of layer l, σ is a nonlinear activate

function. GCN adds the self-loop to each node by adopting $\tilde{A}$ rather than the original adjacency matrix A. $\hat{A}$ is the normalized adjacency matrix. For the GCN model with L layers, $H^{(0)} = X$, $H^{(L)} = Z$.

It can be seen from Equation (1) that the input required by GCN except for $\tilde{A}$ is the feature matrix of nodes. We use the position gene set, motif gene set and immunological signature gene set from the Molecular Signatures Database (MSigDB) [58] to construct node features referring to a previous study [59]. A total of 971 gene sets containing >95% of the genes corresponding to proteins in the PPI network are selected to generate 972-dimensional binary-valued feature vectors for nodes. Only when the first 971 elements of the feature vector are all 0 will the last element be 1. If multiple genes are corresponding to the same protein, the OR operation is applied for aggregation.

In PRODeepSyn, the GCN model with two graph convolutional layers is adopted. We set $W^{(1)} \in \mathbb{R}^{972 \times 2e_{gene}}$, $W^{(2)} \in \mathbb{R}^{2e_{gene} \times e_{gene}}$, where e_{gene} is the dimensionality of gene hidden state vectors. Each of the graph convolutional layers uses ReLU as the activate function. The i-th row of GCN's output matrix Z represents the hidden state of the i-th gene.

Solution of cell line states

PRODeepSyn answers the second question via the StateEncoder as illustrated in Figure 1**B**. It initializes the embedding matrix $C \in \mathbb{R}^{e_{cell} \times M}$ randomly, where M is the number of cell lines, and e_{cell} is the embedding dimension of cell line states. The gene expression level and gene mutation results of the cell line are the explicit states of genes, which are noted as O^{Exp} and O^{Mut}, respectively. In the j-th cell line, the relationship among the explicit state of genes $\mathbf{o}_j^t, t \in \{Exp, Mut\}$, the cell line state $\mathbf{c}_j$, and the hidden state matrix of genes Z is modelled as:

$$\mathbf{o}_j^t = f(Z) \cdot \mathbf{c}_j \tag{2}$$

where the project transformation f transforms the hidden state matrix of genes into the cell line space, thereby it could interact with the cell line state vector through the dot product to present the explicit state vector of genes in the cell line.

The StateEncoder of PRODeepSyn consists of a GCN model and a project model *Projector*. The structure of GCN is as described in Section 2.4.2. The Projector contains a fully connected layer to simulate the project transformation f, whose output dimension is e_{cell}.

The completed PPI network contains 17 161 nodes. However, neither of the two types of gene explicit states, the gene expression data and the gene mutation data, could be totally matched to the nodes. Therefore, we define the semi-supervised loss function as:

$$L_t = \sum_j^M \sum_i^{N^t} (o_{i,j}^t - \hat{o}_{i,j}^t)^2 \tag{3}$$

where N^t is the number of genes corresponding to the nodes in the PPI network. We retain 3384 genes for the expression data and 10 707 genes for the mutation data, i.e., $N^{Exp} = 3384, N^{Mut} = 10707$. The predicted explicit state of the i-th gene of the j-th cell line $\hat{o}_{i,j}^t$ is calculated with Equation (2).

By calculating the loss and back-propagating the gradient, PRODeepSyn can solve the hidden state matrix Z of genes and the state matrix C of cell lines simultaneously. Different states solved with different gene explicit states of the same cell line will be first concatenated and then z-score normalized as the final embedding of the cell line.

Prediction

After constructing drug features and cell line embeddings, we design a DNN termed Predictor for predicting synergy scores of drug combinations on cell lines. The structure of Predictor is shown in Figure 1C. It receives the features of two drugs and the embedding of one cell line as input and predicts the corresponding synergy score. It has three fully connected (FC) layers, among which the first two FC layers use the ReLU activate function and are followed by the Batch Normalization layer. We set the number of neurons in the second FC layer to half of the first one. The last FC layer contains only one neuron, which represents the synergy score predicted by the model. The loss function for training Predictor is the mean square error loss.

Experimental setup

Method comparison

In order to present the ability of PRODeepSyn to predict the synergy scores of new drug combinations, we compare PRODeepSyn with other advanced methods on the large dataset released by O'Neil *et al.* [30] using a 5-fold nested cross-validation. PRODeepSyn is compared with two DL methods including DeepSynergy [32] and AuDNNSynergy [33] and four ML methods including Elastic Net [60], Support Vector Regression (SVR) [61], Random Forest [62] and XGBoost [63]. The experiment results of DL methods are obtained from their original papers, whereas the results of ML methods are obtained using the same input as PRODeepSyn. Detailed settings for the compared ML methods are described in the Supplementary Material. Although it is oversimplified to treat the prediction of synergistic drug combinations as a classification task [64], we also evaluate the performance of PRODeepSyn and other methods on the task to compare with the previous methods more comprehensively. Referring to the practice of Preuer *et al.* [32], we only consider samples with a synergy score higher than 30 as positive samples, and all others are negative samples. The predicted synergy scores are also binarized with 30 as the threshold. We exclude the work of Jiang *et al.* [44] because their results are obtained with a 10-fold cross-validation, and the divided dataset has not been released.

Metrics

For the regression task, we adopt the mean square error (MSE) as the main evaluation metric. We also report the root mean square error (RMSE) and the Pearson correlation coefficient (PCC) between the predictions and the ground truth. Since we use a 5-fold cross-validation for experiments, we report each evaluation metric's mean and SD on the 5-fold data. For MSE, we also report the 95% confidence interval. Metrics for the classification task include the area under the receiver operator characteristics curve (ROC-AUC), the area under the precision–recall curve (PR-AUC), accuracy (ACC), precision (PREC) and the Cohen's Kappa.

Global settings

In StateEncoder, we set the dimension of the gene hidden state vector $e_{gene} = 128$, and the dimension of the cell line state vector $e_{cell} = 384$. When training the Predictor, the optimal hyperparameters come from grid search. We mainly adjust the hidden layer size and the learning rate of the model. The number of neurons of the first FC layer is chosen from {2048, 4096, 8192}, and the learning rate is chosen from {0.00001, 0.0001, 0.001}. We adopt the mini-batch method for training, and the size of each batch is 512. The maximum number of epochs per training is 500. More implementation details are described in the Supplementary Material.

During training, we adopt the Early Stopping technology to prevent the model from overfitting. When the model's loss on the validation set does not decrease for 50 consecutive epochs, the training will be terminated. In the 5-fold nested cross-validation, we use 4-folds to search for the optimal hyperparameters in the inner loop, where 3-folds consist of the inner training set and the other fold is the validation set. In the outer loop, we divide the 4-folds into the outer training set and the outer validation set at a ratio of 9:1 randomly, and the other 1-fold is used as the test set. The model is trained on the outer training set with the searched optimal hyperparameters and predicts on the test set after training. The process is repeated five times to ensure every fold of data is selected as the test set for exact one time. The drug combinations in the test set are not included in the training data, which can be used to evaluate the generalization ability of the model.

Results

Method comparison

The experiment results of the comparison between PRODeepSyn and other methods on the regression task are summarized in Table 1, which involves two DL methods and four ML methods. As the related data could not be found in the manuscript of DeepSynergy, we do not report the 95% confidence interval of DeepSynergy's MSE. PRODeepSyn achieves the lowest MSE and RMSE and the highest PCC. Its MSE is 229.49, which is 10.18% less than DeepSynergy's, 4.82% less than AuDNNSynergy's and 22.56% less than XGBoost's. PRODeepSyn attains the PCC of 0.75. Figure 2 illustrates the correlation between the prediction results and the ground truth of all data points. The straight line in blue is the figure of the function between the predicted score and the ground truth fitted using the least squares regression. The slope of the straight line is 1.02, and the bias is 0.53 (*P*-value < 1e-5). Both the PCC of PRODeepSyn and the fitted function indicate a strong linear correlation between the prediction results of PRODeepSyn and the ground truth. The experiment results show that PRODeepSyn outperforms other state-of-the-art methods.

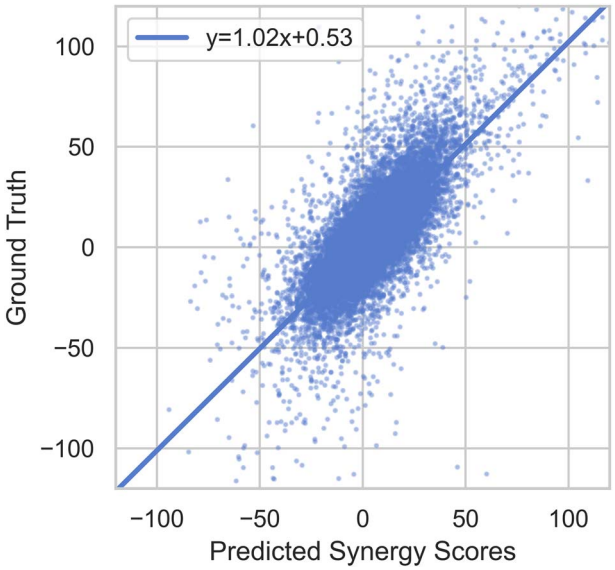

Figure 2. Scatter plot of the predicted synergy scores and the ground truth. The straight line in blue is the figure of the function fitted using the least squares regression, whose slope is 1.02 and bias is 0.53 (*P*-value < 1e-5).

Considering that many previous studies treat the prediction task as a classification task, we have further carried out related experiments to facilitate comparison and analysis. The results of each method on the classification task are summarized in Table 2. It is noteworthy that all methods have almost the same accuracy, whereas their values of other metrics vary, resulting from the high ratio of the negative samples in test data. Therefore, we consider ROC-AUC and PR-AUC relatively fair metrics [65], and PR-AUC is better than ROC-AUC on imbalanced dataset [66]. PRODeepSyn achieves the best PR-AUC and the very similar ROC-AUC to the best one. We have not paid attention to the imbalanced distribution of classes in the training dataset because it is not the focus of our work. Nonetheless, PRODeepSyn has a comprehensive advantage over other methods on the classification task.

Predictions aggregated by tissue

As shown in Figure 3, we visualize the prediction results and the ground truth of PRODeepSyn according to the tissue type of the cell line. Figure 3**A** shows the distribution of the ground truth and the predicted scores given by PRODeepSyn in the cell lines of each tissue. In the seven tissues, most of the real and predicted synergy scores are

Table 1. Results of method comparison on the regression task

Type	Method	MSE	RMSE	Confidence Interval	PCC
DL	PRODeepSyn	**229.49 ± 42.81**	**15.09 ± 1.37**	[176.34, 282.64]	**0.75 ± 0.02**
DL	DeepSynergy	255.49	15.91 ± 1.56	–	0.73 ± 0.04
DL	AudnnSynergy	*241.12 ± 43.52*	*15.46 ± 1.44*	[187.09, 295.15]	*0.74 ± 0.03*
ML	Elastic Net	418.06 ± 53.99	20.41 ± 1.30	[351.03, 485.09]	0.45 ± 0.02
ML	SVR	325.91 ± 54.75	17.99 ± 1.48	[257.94, 393.89]	0.63 ± 0.02
ML	Random Forest	312.75 ± 41.35	17.65 ± 1.13	[261.41, 364.08]	0.64 ± 0.03
ML	XGBoost	296.34 ± 46.37	17.16 ± 1.31	[238.77, 353.90]	0.66 ± 0.02

Values of MSE, RMSE and PCC are mean values ± 1 SD. The best and second best performance are shown in bold and with italic, respectively.

Table 2. Results of method comparison on the classification task

Type	Method	ROC-AUC	PR-AUC	ACC	PREC	Kappa
DL	PRODeepSyn	*0.90 ± 0.03*	**0.63 ± 0.05**	**0.93 ± 0.01**	0.72 ± 0.06	**0.51 ± 0.03**
DL	DeepSynergy	*0.90 ± 0.03*	0.59 ± 0.06	0.92 ± 0.03	0.56 ± 0.11	0.51 ± 0.04
DL	AudnnSynergy	**0.91 ± 0.02**	*0.63 ± 0.06*	**0.93 ± 0.01**	0.72 ± 0.06	0.51 ± 0.04
ML	Elastic Net	0.78 ± 0.04	0.34 ± 0.09	*0.92 ± 0.01*	0.61 ± 0.33	0.14 ± 0.08
ML	SVR	0.88 ± 0.02	0.54 ± 0.05	**0.93 ± 0.01**	**0.80 ± 0.04**	0.32 ± 0.02
ML	Random Forest	0.87 ± 0.02	0.54 ± 0.04	**0.93 ± 0.01**	*0.74 ± 0.02*	0.36 ± 0.04
ML	XGBoost	0.87 ± 0.02	0.56 ± 0.05	**0.93 ± 0.01**	0.74 ± 0.04	0.41 ± 0.04

Values of all metrics are mean values ±1 SD. The best and second best performance are shown in bold and with italic, respectively.

concentrated in the range of [-50, 75]. For negative samples, the distribution of prediction results of PRODeepSyn is similar to the ground truth. For positive samples, PRODeepSyn tends to give more conservative prediction results. For the prostate, the difference between the predicted synergy scores of PRODeepSyn and the ground truth is relatively apparent. We believe that this is related to the small number of cell lines belonging to this tissue in the dataset and the scattered distribution of synergy scores. Figure 3**B** summarizes the PCC between the predicted scores and the ground truth on each cell line, and the color of the bar shows the tissue to which the cell line belongs. Among them, COLOR320DM has the lowest PCC, 0.60, whereas UWB1289 has the highest PCC, 0.87. Among the 39 cell lines, the PCC on only four cell lines is lower than 0.65, whereas the PCC on 16 cell lines is higher than 0.75. One possible reason for the lower PCC of COLO320DM than other cell lines is that on this cell line there are couples of samples whose drug combinations are the same but the synergy scores are very different. For example, the two synergy scores of GEMCITABINE and MK-8776 on COLO320DM are 20.30 and 44.82, which will even be classified as opposing samples. The same problem also occurs on EC2, A375 and several other cell lines with lower PCC. In Figure 3**C**, we further aggregate PCC according to the tissue type. The median values of PCC of all tissues are higher than 0.70 except pleura. The correlation between the prediction results of PRODeepSyn and the ground truth for cell lines belonging to the ovary is the strongest. Overall, in different tissues, there is a strong correlation between the prediction results given by PRODeepSyn and the ground truth and no obvious association between PCC and tissue types is found. PRODeepSyn has the potential value of predicting anticancer synergistic drug combinations in the cell lines belonging to various tissues.

Ablation study

Compared with DL methods that only use omics data to construct cell line features, PRODeepSyn integrates the PPI network with multi-omics data into cell line embeddings. To inspect the contribution of using multi-omics data and integrating the PPI network information to the prediction, we compare PRODeepSyn with several variants that include:

- **PRODeepSyn-GE.** PRODeepSyn-GE is the variant of PRODeepSyn that only using the gene expression data as the gene explicit state when solving the state of the cell line.
- **PRODeepSyn-MUT.** PRODeepSyn-MUT is the variant of PRODeepSyn that only using the gene mutation data as the gene explicit state when solving the state of the cell line.
- **PRODeepSyn-RandomZ.** Instead of using GCN to extract the PPI network information to construct the gene hidden state matrix Z, the variant PRODeepSyn-RandomZ solves cell line states with the randomly initialized matrix with trainable elements.
- **PRODeepSyn-AE.** This variant first compresses the gene expression data and the gene mutation data into the e_{cell}-dimensional space separately and then concatenates them as the state vectors of cell lines.

Table 3 summarizes the results of the ablation study. Compared with PRODeepSyn-GE and PRODeepSyn-MUT, PRODeepSyn has a lower MSE. Although the improvement is not significant, using multi-omics data are helpful to improve the prediction results. Compared

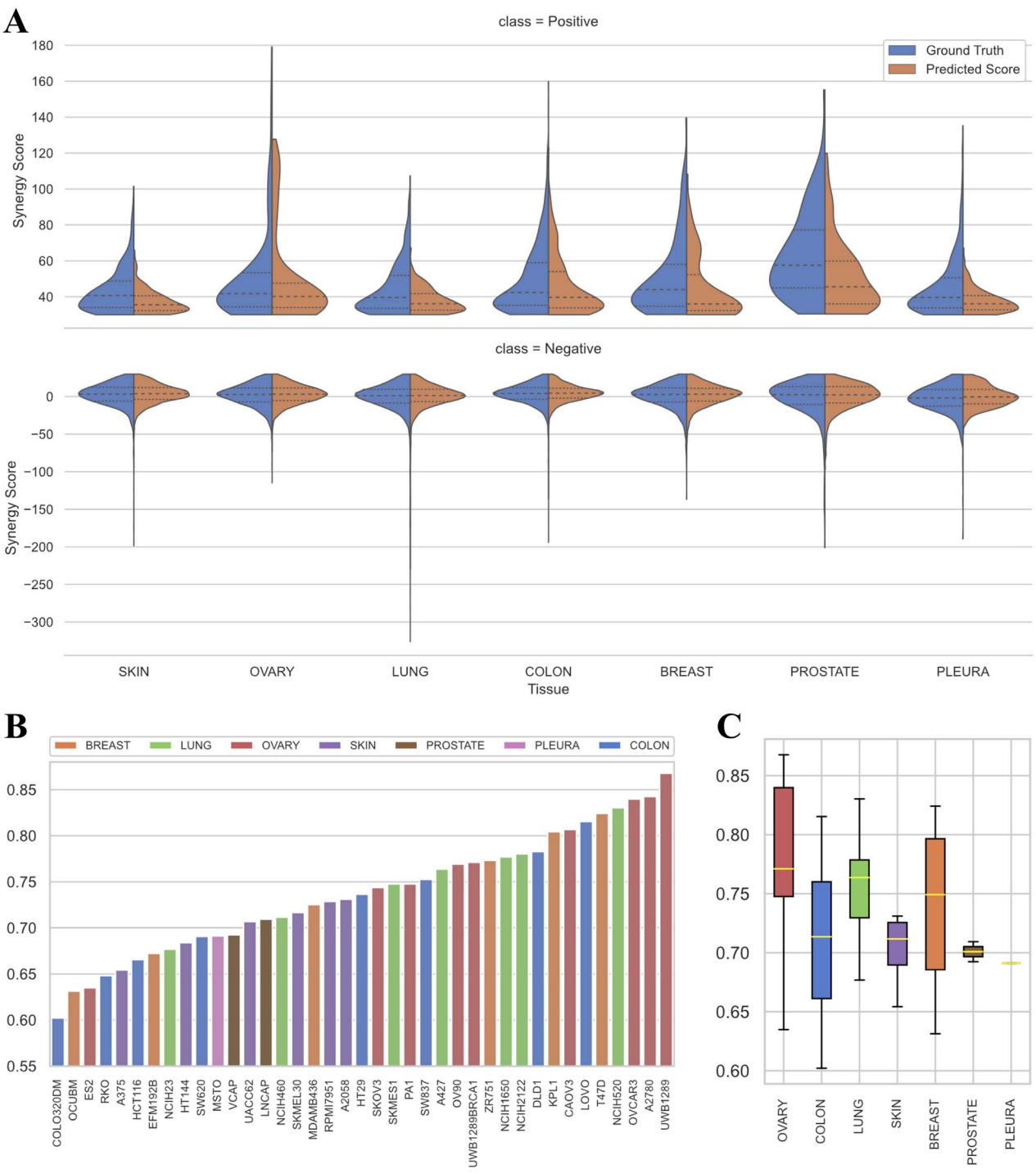

Figure 3. Predictions aggregated by tissue. (**A**) Comparison between the distribution of the ground truth and the predicted synergy scores aggregated by tissue. The results of positive samples and negative samples are plotted in the upper part and the lower part, respectively. (**B**) The PCC values of each cell line. The color of the bar represents the tissue of the cell line. (**C**) The boxplot of the PCC values of cell lines aggregated by tissue. The yellow horizontal line in each box indicates the median.

with PRODeepSyn-RandomZ, using the gene hidden state matrix computed from the PPI network can obtain a lower MSE. In PRODeepSyn-RandomZ, the trained matrix could also represent a kind of gene hidden states since its interaction with cell line state vectors can fit the gene explicit states well. However, PRODeepSyn-RandomZ performs worse than PRODeepSyn, indicating that the lack of PPI network information may reduce the performance. Besides, PRODeepSyn outperforms PRODeepSyn-AE that does not integrate the PPI network information. The superiority of PRODeepSyn over PRODeepSyn-RandomZ and PRODeepSyn-AE presents that the integration of the PPI network information could improve the prediction results.

Table 3. Results of the ablation study

Method	MSE	RMSE	Confidence Interval	PCC
PRODeepSyn	**229.49 ± 42.81**	**15.09 ± 1.37**	[176.34, 282.64]	**0.75 ± 0.02**
PRODeepSyn-GE	231.14 ± 44.08	15.14 ± 1.41	[176.42, 285.86]	**0.75 ± 0.02**
PRODeepSyn-MUT	231.36 ± 41.01	15.15 ± 1.30	[180.45, 282.27]	0.75 ± 0.03
PRODeepSyn-RandomZ	240.29 ± 40.94	15.45 ± 1.28	[189.46, 291.12]	**0.75 ± 0.02**
PRODeepSyn-AE	238.11 ± 46.41	15.36 ± 1.46	[180.49, 295.73]	0.74 ± 0.03

Values of MSE, RMSE and PCC are mean values ± 1 SD. The best performance is shown in bold.

Sensitivity analysis

PRODeepSyn utilizes the embeddings of genes obtained from the PPI network when solving the state vectors of cell lines. In order to explore the impact of the dimension of the gene's embedding space and the dimension of the cell line's embedding space on the final prediction, we conduct the sensitivity analysis. We select the dimension of gene's embedding space e_{gene} from {32, 64, 128, 256, 512}, while select the dimension of cell line's embedding space e_{cell} from {128, 320, 384, 448, 512}. In each experiment, we only modify one of e_{gene} or e_{cell} and keep the other hyperparameter as its original value.

Figure 4 shows the fluctuation of the MSE and PCC with the change of embedding dimension of the gene hidden state or the cell line. It can be found from the 1st column of Figure 4 that the prediction results hardly change with the embedding dimension of the gene hidden state. Since the increment of the embedding's dimension increases the computational complexity while bringing no substantial improvement in the prediction results, it is recommended to embed the hidden state of genes into a 128-dimensional space. The 2nd column of Figure 4 indicates that the prediction results keep stable when the embedding dimension of the cell line changes. Since embedding the cell line into a 384-dimensional space achieves the lowest MSE and the highest PCC, and the dimensionality of the space is moderate, choosing 384 as the embedding dimension of cell lines is recommended. Detailed results are provided in Supplementary Tables S2 and S3. Overall, the disturbance of the embedding dimension of the gene hidden state and the cell line has slight impact on the prediction results of PRODeepSyn.

Explanation of Cell line embeddings

As we expected cell line states to play a significant role in determining whether the drug combinations are synergistic on the cell line, we are curious about whether the same drug combination behave similarly on cell lines with similar state vectors. Here we define the synergy distance $Dist(C_i, C_j)$ between cell line C_i and cell line C_j as:

$$Dist(C_i, C_j) = \frac{1}{B_{ij}} \sum_{k=1}^{B_{ij}} \mathrm{std}(s_{ik}, s_{jk}) \tag{4}$$

where B_{ij} is the total number of drug combinations that has been tested on both C_i and C_j, $s_{*,k}$ is the synergy score of the k-th drug combination on cell line C_*, and $\mathrm{std}(\cdot,\cdot)$ is the function to calculate SD. In the O'Neil dataset, the number of common drug combinations on two arbitrary cell lines is always 583, so $B_{i,j}$ could be treated as a constant when comparing distances among different pairs of cell lines. We adopt one of the best dimensionality reduction methods t-SNE [67] to visualize cell line embeddings in a two-dimensional space.

Figure 5 shows the results of t-SNE. The color of each point represents the cell line's tissue. The distances between many pairs of cell lines in Figure 5 agree with their synergy distances. For example, in the bottom border of Figure 5, the synergy distances from SW837 to DLD1, COLO320DM and OCUBM are 9.95, 12.29 and 13.32, respectively, which are consistent with their relative Euclidean distances in Figure 5. Another example is that in the top border of Figure 5, the synergy distances from UWB1289 to UWB1289BRCA1, A2780, SW620 and LNCAP are 9.02, 11.28, 12.38 and 20.49, respectively. The analysis shows that in Figure 5 the far a cell line from UWB1289, the larger the synergy distance will be. The consistency of the synergy distance and the Euclidean distance between cell lines helps PRODeepSyn outperform other methods.

Case study

We analyze the prediction results of PRODeepSyn and find that many cases are consistent with previous studies. For example, Gil-Martin *et al.* [68] tested the therapeutic effect of BEZ-235 and Paclitaxel on breast cancer patients. The experiment did not obtain any evidence that this drug combination had a synergistic effect, and the subjects suffered from various adverse reactions. The prediction results given by PRODeepSyn are consistent with this experiment. On breast cancer cell lines OCUBM and EFM192B, the synergy scores predicted by PRODeepSyn are -3.00 and -13.67, respectively. In addition, Lara *et al.* [69] argued that the therapeutic effect of the combination of MK-2206 and Erlotinib on patients with non-small cell lung cancer (NSCLC) is worthy of further exploration. We check the prediction results of PRODeepSyn for three NSCLC cell lines included in the dataset, namely SKMES1, NCIH460, NCIH520, which are 38.36, 21.88 and 19.75, respectively. Since 30 is the threshold for distinguishing strong synergy from weak synergy [32], the results also indicate that the combination of MK-2206 and Erlotinib is likely to show a synergistic effect in the treatment of NSCLC.

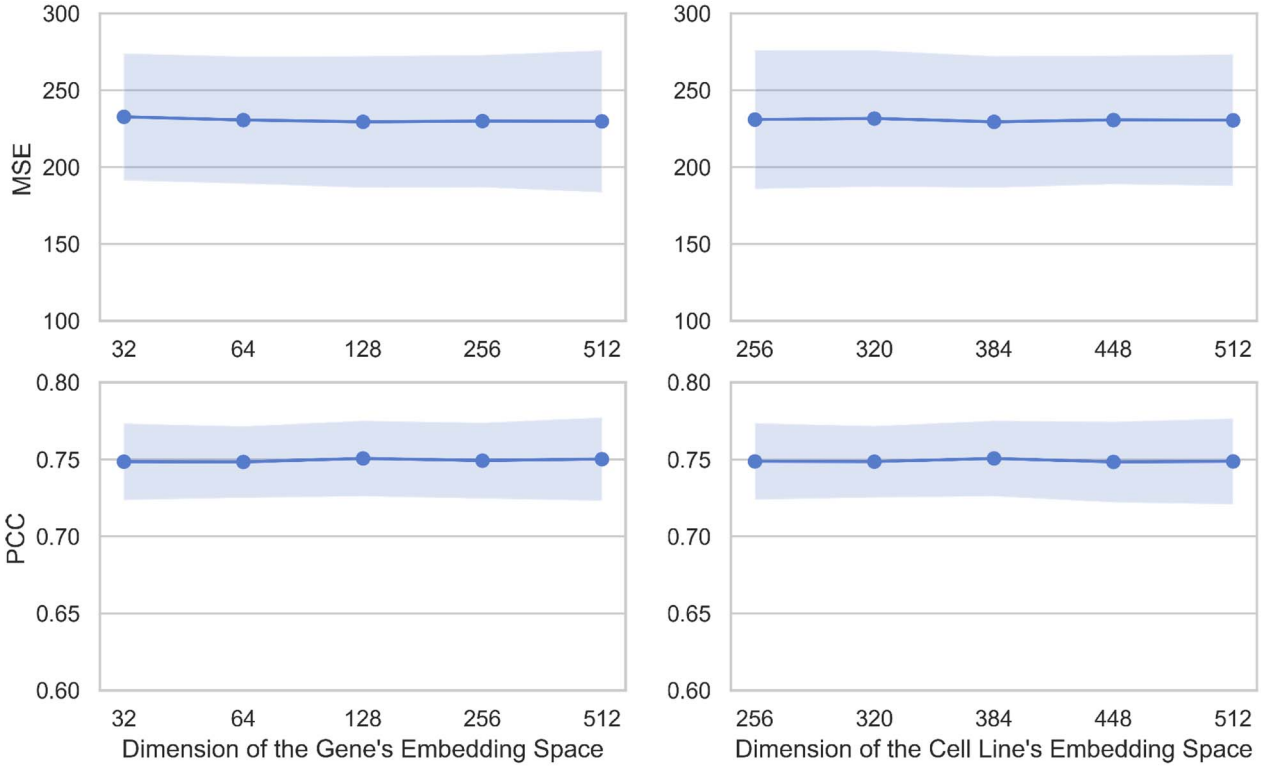

Figure 4. Results of sensitivity analysis. MSE and PCC of PRODeepSyn remain stable when the dimensions of gene hidden states or the dimensions of cell line embeddings change. The regions in light blue indicate the range of the mean values ± 1 SD.

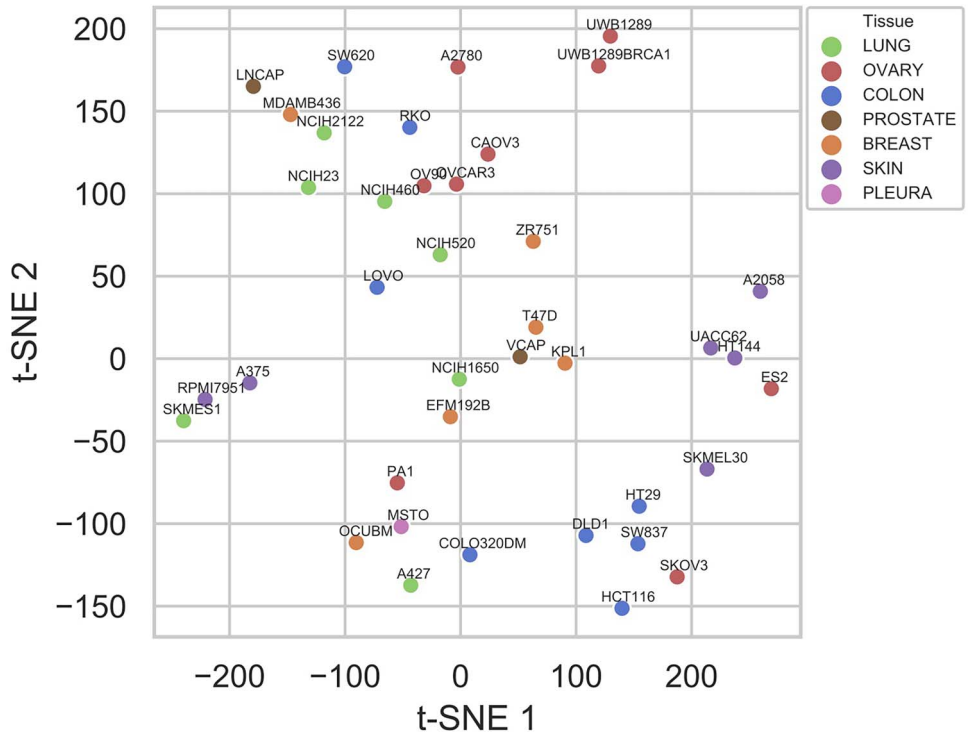

Figure 5. Visualization of cell line embeddings in 2-dimensional space using t-SNE.

Moreover, Wang *et al.* [70] conducted *in vivo* and *in vitro* experiments on the combined dosage regimen of 5-FU and BEZ-235, in which RKO and HCT116 were selected for *in vitro* experiments. The experiment results indicate that this combined dosage regimen can lead to PUMA-dependent tumour inhibition. For RKO and HCT116 cell lines, PRODeepSyn gives the predicted synergy scores of 18.41 and 20.44, respectively. Furthermore,

Wisinski *et al*. [71] confirmed that the combination of MK-2206 and Lapatinib could be tolerated with a higher dose than monotherapy. They conducted *in vitro* experiments on HCT-15 to evaluate the mechanism of this drug interaction. PRODeepSyn gives higher predicted synergy scores on the DLD1, HT29 and LOVO cell lines, which have the same disease as HCT-15, which are 47.74, 32.23 and 45.96, respectively. Another example is that Berndsen *et al*. [72] found that the combination of Erlotinib and Dasatinib can lead to growth retardation of colon cancer cells. This study conducted *in vitro* experiments on SW620, DLD1 and HT29 cell lines. The prediction results given by PRODeepSyn are consistent with this study, and the predicted synergy scores are 15.22, 42.36 and 33.46, respectively. As can be seen from the cases we mentioned above, the prediction results obtained by PRODeepSyn are consistent with many previous *in vivo* and *in vitro* studies. Therefore, PRODeepSyn has practical application value.

Discussion and conclusion

In this study, we propose a new DL method, PRODeepSyn, to model the potential relationship between drug combinations and cell lines to achieve the purpose of predicting anticancer synergistic drug combinations. PRODeepSyn constructs drug features with drugs' fingerprints and descriptors and the low-dimensional dense embeddings for cell lines by integrating the PPI network with omics data. Specifically, PRODeepSyn first models the gene explicit state as the interaction of the gene hidden state and cell line state, where the gene explicit state comes from omics data, the gene hidden state is extracted from the PPI network using GCN, and the cell line state is a randomly initialized embedding. The interaction is represented using vector inner product. Then PRODeepSyn updates the cell line embedding and the GCN model simultaneously by optimizing the semi-supervised loss function. Theoretically, PRODeepSyn integrates genomics data, transcriptomics data and the relationship between proteins into the final cell line embeddings, benefiting from the powerful ability of GCN to extract the structural features of the PPI network. The embeddings with multi-level information have been proved to be significantly helpful to predict the synergy scores of drug combinations in our experiments. Besides, PRODeepSyn uses DNN with the Batch Normalization mechanism to predict the synergy score of the combination of two drugs on a certain cell line. The reason for using Batch Normalization is that it can reduce the dependence of the DL model on initialized parameters, accelerate convergence and enhance generalization ability, which further reduces the labor cost and time overhead of building a DL model. Meanwhile, PRODeepSyn has the ability to construct embeddings for new cell lines, which improves its scalability. After preprocessing the multi-omics data of a new cell line, the cell line embedding could be obtained by training with the pre-trained and frozen gene hidden state matrix in StateEncoder.

We conduct experiments on a large public dataset, and the results present the superiority of PRODeepSyn to other state-of-the-art methods. In addition, on cell lines belonging to different tissues, there is a strong linear correlation between the prediction results of PRODeepSyn and the ground truth. Therefore, PRODeepSyn has a wide range of applications. Furthermore, we find that the prediction results of PRODeepSyn are consistent with many previous studies through the case study. The results of sensitivity analysis show that PRODeepSyn is not sensitive to the dimensions of gene and cell line embeddings, which can reduce the workload of optimizing PRODeepSyn. Overall, PRODeepSyn shows the ability to identify anticancer synergistic drug combinations beyond other computational methods. Notably, in the ablation study, we find that using multi-omics data and integrating the PPI network can improve the prediction results, which may have implications for the study of the synergistic mechanism of drug combinations.

PRODeepSyn still has some shortcomings. We find that PRODeepSyn gives more conservative prediction results for drug combinations that should have high synergy scores, which may result from the concentration of synergy scores near 0 in the training set. The problem is expected to be solved with the publication of more experiment data. Another problem is that in addition to the PPI network, other graph-structured data, such as drug–target interaction network and drug–drug interaction network, are also of great significance for the study of drug combination therapy, which have not been included in PRODeepSyn yet. Besides, PRODeepSyn can only predict the synergistic combination of two drugs at present. Exploring more generic methods for predicting the synergy scores of drug combinations consisting of more than two drugs is the direction of our follow-up efforts.

In summary, our findings suggest that the DL method integrating biological network has great advantages in discovering anticancer synergistic drug combinations and also provides a possible reference for studying the synergistic mechanism of anticancer drug combinations. PRODeepSyn is expected to become a powerful tool for the prescreening of anticancer synergistic drug combinations.

Key Points

- PRODeepSyn integrates the PPI network and multi-omics data into cell line embeddings using the GCN.
- PRODeepSyn achieves the best performance compared with six advanced methods on predicting synergistic drug combinations.
- PRODeepSyn's predictions are consistent with many previous *in vivo* and *in vitro* studies.
- PRODeepSyn is expected to become a powerful tool for the prescreening of anticancer synergistic drug combinations.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The data underlying this article are available in the article and in its online Supplementary Material. The code of PRODeepSyn, the training data, and the prediction results are available at https://github.com/TOJSSE-iData/PRODeepSyn.

Author contributions statement

X.W., H.Z. and Y.J. designed and implemented the model, conducted the experiments, analyzed the results and wrote the manuscript. Y.L., C.T. , X.C. and Y.L. reviewed the manuscript. Q.L. and Q.L. directed the conception of the model and experiments and the writing of the manuscript.

Funding

National Key Research and Development Program of China (grant nos 2017YFC0908500, 2016YFC1303205); National Natural Science Foundation of China (grant nos 31970638, 61572361); Shanghai Natural Science Foundation Program(grant no. 17ZR1449400); Shanghai Artificial Intelligence Technology Standard Project (grant no. 19DZ2200900); Major Program of Development Fund for Shanghai Zhangjiang National Innovation Demonstration Zone (ZJ2018-ZD-004); Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai; 2019-nCov Emergency Research Project of Zhejiang University; Fundamental Research Funds for the Central Universities.

Conflict of interest

All authors have no conflict of interest to disclose.

References

1. Mokhtari RB, Homayouni TS, Baluch N, *et al.* Combination therapy in combating cancer. *Oncotarget* 2017;**8**(23):38022.
2. Chou T-C. Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer Res* 2010;**70**(2):440–6.
3. Sicklick JK, Kato S, Okamura R, *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the i-predict study. *Nat Med* 2019;**25**(5):744–50.
4. Sun X, Vilar S, Tatonetti NP. High-throughput methods for combinatorial drug discovery. *Sci Transl Med* 2013;**5**(205):205rv1–1.
5. Liu J, Gefen O, Ronin I, *et al.* Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science* 2020;**367**(6474):200–4.
6. Foucquier J, Guedj M. Analysis of drug combinations: current methodological landscape. *Pharmacol Res Perspect* 2015;**3**(3):e00149.
7. Yadav B, Wennerberg K, Aittokallio T, *et al.* Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Comput Struct Biotechnol J* 2015;**13**:504–13.
8. Siddiqui-Jain A, Bliesath J, Macalino D, *et al.* Ck2 inhibitor cx-4945 suppresses DNA repair response triggered by DNA-targeted anticancer drugs and augments efficacy: mechanistic rationale for drug combination therapy. *Mol Cancer Ther* 2012;**11**(4):994–1005.
9. White NJ, Olliaro PL. Strategies for the prevention of antimalarial drug resistance: rationale for combination chemotherapy for malaria. *Parasitol Today* 1996;**12**(10):399–401.
10. Tallarida RJ. An overview of drug combination analysis with isobolograms. *J Pharmacol Exp Ther* 2006;**319**(1):1–7.
11. Gayvert KM, Aly O, Platt J, *et al.* A computational approach for identifying synergistic drug combinations. *PLoS Comput Biol* 2017;**13**(1):e1005308.
12. Jonker DM, Visser SAG, Van Der Graaf PH, *et al.* Towards a mechanism-based analysis of pharmacodynamic drug–drug interactions in vivo. *Pharmacol Ther* 2005;**106**(1):1–18.
13. Li MC, Whitmore WF, Golbey R, *et al.* Effects of combined drug therapy on metastatic cancer of the testis. *JAMA* 1960;**174**(10):1291–9.
14. Muss HB, White DR, Richards F, *et al.* Adriamycin versus methotrexate in five-drug combination chemotherapy for advanced breast cancer. a randomized trial. *Cancer* 1978;**42**(5):2141–8.
15. Astashkina A, Mann B, Grainger DW. A critical evaluation of in vitro cell culture models for high-throughput drug screening and toxicity. *Pharmacol Ther* 2012;**134**(1):82–106.
16. Macarron R, Banks MN, Bojanic D, *et al.* Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* 2011;**10**(3):188–95.
17. Lamb J, Crawford ED, Peck D, *et al.* The connectivity map: using gene-expression signatures to connect small molecules. *Genes Dis Sci* 2006;**313**(5795):1929–35.
18. Janes KA, Albeck JG, Gaudet S, *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* 2005;**310**(5754):1646–53.
19. Feala JD, Cortes J, Duxbury PM, *et al.* Systems approaches and algorithms for discovery of combinatorial therapies. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**(2):181–93.
20. Araujo RP, Petricoin EF, Liotta LA. A mathematical model of combination therapy using the EGFR signaling network. *Biosystems* 2005;**80**(1):57–69.
21. Nelander S, Wang W, Nilsson B, *et al.* Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol* 2008;**4**(1):216.
22. Giuliano AE, Connolly JL, *et al.* Breast cancer-major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* 2017;**67**(4):290–303.
23. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. *Nat Commun* 2019;**10**(1):1–11.
24. Li X, Yingjie X, Cui H, *et al.* Prediction of synergistic anticancer drug combinations based on drug target network and drug induced gene expression profiles. *Artif Intell Med* 2017;**83**:35–43.
25. Li H, Li T, Quang D, *et al.* Network propagation predicts drug synergy in cancers. *Cancer Res* 2018;**78**(18):5446–57.
26. Low YS, Daugherty AC, Schroeder EA, *et al.* Synergistic drug combinations from electronic health records and gene expression. *J Am Med Inform Assoc* 2017;**24**(3):565–76.

27. Celebi R, Oliver Bear Don't Walk, Movva R, *et al.* In-silico prediction of synergistic anti-cancer drug combinations using multi-omics data. *Sci Rep* 2019;**9**(1):1–10.
28. Jeon M, Kim S, Park S, *et al.* In silico drug combination discovery for personalized cancer therapy. *BMC Syst Biol* 2018;**12**(2):59–67.
29. Yan X, Yang Y, Chen Z, *et al.* H-RACS: a handy tool to rank anticancer synergistic drugs. *Aging (Albany NY)* 2020;**12**(21):21504.
30. O'Neil J, Benita Y, Feldman I, *et al.* An unbiased oncology compound screen to identify novel combination strategies. *Mol Cancer Ther* 2016;**15**(6):1155–62.
31. Holbeck SL, Camalier R, Crowell JA, *et al.* The national cancer institute almanac: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res* 2017;**77**(13):3564–76.
32. Preuer K, Lewis RPI, Hochreiter S, *et al.* Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 2018;**34**(9):1538–46.
33. Zhang T, Zhang L, Payne PRO, *et al.* Synergistic drug combination prediction by integrating multiomics data in deep learning models. In: *Translational Bioinformatics for Therapeutic Development*. New York, NY, United States: Springer, 2021, 223–38.
34. Xia F, Shukla M, Brettin T, *et al.* Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* 2018;**19**(18):71–9.
35. Zhang H, Feng J, Zeng A, *et al.* Predicting tumor cell response to synergistic drug combinations using a novel simplified deep learning model. In: *AMIA Annual Symposium Proceedings*, Vol. **2020**. American Medical Informatics Association, 2020, 1364.
36. Ding P, Yin R, Luo J, *et al.* Ensemble prediction of synergistic drug combinations incorporating biological, chemical, pharmacological, and network knowledge. *IEEE J Biomed Health Inform* 2019;**23**(3):1336–45.
37. Singh H, Rana PS, Singh U. Prediction of drug synergy score using ensemble based differential evolution. *IET Syst Biol* 2019;**13**(1):24–9.
38. Kim Y, Zheng S, Tang J, *et al.* Anticancer drug synergy prediction in understudied tissues using transfer learning. *J Am Med Inform Assoc* 2021;**28**(1):42–51.
39. Sun Z, Huang S, Jiang P, *et al.* Dtf: deep tensor factorization for predicting anticancer drug synergy. *Bioinformatics* 2020;**36**(16):4483–9.
40. Fan K, Cheng L, Li L. Artificial intelligence and machine learning methods in predicting anti-cancer drug combination effects. *Brief Bioinform* 2021.
41. Meng J, Dai B, Fang B, *et al.* Combination treatment with MEK and AKT inhibitors is more effective than each drug alone in human non-small cell lung cancer in vitro and in vivo. *PloS One* 2010;**5**(11):e14124.
42. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**(1):56–68.
43. Kipf TN, Welling M. . Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016.
44. Jiang P, Huang S, Zhenyuan F, *et al.* Deep graph embedding for prioritizing synergistic anticancer drug combinations. *Comput Struct Biotechnol J* 2020;**18**:427–38.
45. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**(13):i457–66.
46. Loewe S. The problem of synergism and antagonism of combined drugs. *Arzneimittelforschung* 1953;**3**:285–90.
47. Landrum G, *et al.* Rdkit: cheminformatics and machine learning software. *RDKIT ORG* 2013.
48. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc* 1965;**5**(2):107–13.
49. Iorio F, Knijnenburg TA, Vis DJ, *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**(3):740–54.
50. Hochreiter S, Clevert D-A, Obermayer K. A new summarization method for affymetrix probe level data. *Bioinformatics* 2006;**22**(8):943–9.
51. Tate JG, Bamford S, Jubb HC, *et al.* Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2019;**47**(D1): D941–7.
52. Szklarczyk D, Gable AL, Lyon D, *et al.* String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**(D1):D607–13.
53. Jiang L, Chen H, Pinello L, *et al.* Giniclust: detecting rare cell types from single-cell gene expression data with GINI index. *Genome Biol* 2016;**17**(1):1–13.
54. Andrews TS, Hemberg M. M3drop: dropout-based feature selection for scRNAseq. *Bioinformatics* 2019;**35**(16):2865–7.
55. Chiu Y-C, Chen H-IH, Zhang T, *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics* 2019;**12**(1):143–55.
56. Li M, Wang Y, Zheng R, *et al.* Deepdsc: a deep learning method to predict drug sensitivity of cancer cell lines. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.
57. Gurdon JB. Transplanted nuclei and cell differentiation. *Sci Am* 1968;**219**(6):24–35.
58. Liberzon A, Subramanian A, Pinchback R, *et al.* Molecular signatures database (msigdb) 3.0. *Bioinformatics* 2011;**27**(12): 1739–40.
59. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, 1025–35.
60. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology* 2005;**67**(2):301–20.
61. Drucker H, CJC B, Kaufman L, *et al.* Support vector regression machines. *Adv Neural Information Process Syst* 1997;**9**:155–61.
62. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.
63. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, United States: Association for Computing Machinery (ACM), 2016, 785–94.
64. Pahikkala T, Airola A, Pietilä S, *et al.* Toward more realistic drug–target interaction predictions. *Brief Bioinform* 2015;**16**(2):325–37.
65. Jeni LA, Cohn JF, Torre FDL. Facing imbalanced data–recommendations for the use of performance metrics. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. Piscataway, NJ, United States: IEEE, 2013, 245–51.
66. Davis J, Goadrich M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, United States: Association for Computing Machinery (ACM), 2006, 233–40.
67. Van der Maaten L, Hinton G. Visualizing data using t-sne. *J Mach Learn Res* 2008;**9**(11):2579–605.
68. Gil-Martin M, Fumoleau P, Isambert N, *et al. Abstract p2–16-22: A dose-finding phase LB study of bez235 in combination with paclitaxel*

in patients with her2-negative, locally advanced or metastatic breast cancer. 2013;**73**(24 Supplement):P2–16–22.
69. Lara PN, Longmate J, Mack PC, *et al.* Phase ii study of the akt inhibitor mk-2206 plus erlotinib in patients with advanced non–small cell lung cancer who previously progressed on erlotinib. *Clin Cancer Res* 2015;**21**(19):4321–6.
70. Wang H, Zhang L, Yang X, *et al.* Puma mediates the combinational therapy of 5-fu and nvp-bez235 in colon cancer. *Oncotarget* 2015;**6**(16):14385.
71. Wisinski KB, Tevaarwerk AJ, Burkard ME, *et al.* Phase i study of an akt inhibitor (mk-2206) combined with lapatinib in adult solid tumors followed by dose expansion in advanced her2+ breast cancer. *Clin Cancer Res* 2016;**22**(11):2659–67.
72. Berndsen RH, Swier N, van Beijnum JR, *et al.* Colorectal cancer growth retardation through induction of apoptosis, using an optimized synergistic cocktail of axitinib, erlotinib, and dasatinib. *Cancer* 2019;**11**(12):1878.