GraphSynergy: Network Inspired Deep Learning Model for Anti–Cancer Drug Combination Prediction

Jiannan Yang
City University of Hong Kong
Zhongzhi Xu
The University of Hong Kong
William Wu
Chinese University of Hong Kong https://orcid.org/0000-0002-5662-5240
Qian Chu (✉ qianchu@tjh.tjmu.edu.cn)
Tongji Hospital https://orcid.org/0000-0001-8192-7630
Qingpeng Zhang (✉ qingpeng.zhang@cityu.edu.hk)
City University of Hong Kong https://orcid.org/0000-0002-6819-0686

Article

Keywords: cancer treatment, topological positions, protein modules, cancer cell lines, protein-protein interaction, synergistic combinations of drugs, Graph Convolutional Network for Drug Synergy

Posted Date: March 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-315189/v1

Version of Record: A version of this preprint was published at Journal of the American Medical Informatics Association on September 2nd, 2021. See the published version at https://doi.org/10.1093/jamia/ocab162.

GraphSynergy: Network Inspired Deep Learning Model for Anti–Cancer Drug Combination Prediction

Jiannan Yang[1]**, Zhongzhi Xu**[2]**, William Ka Kei Wu**[3]**, Qian Chu**[4*]**, and Qingpeng Zhang**[1*]

[1]School of Data Science, City University of Hong Kong, Hong Kong S.A.R. of China

[2]Hong Kong Jockey Club Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong S.A.R. of China

[3]Department of Anaesthesia and Intensive Care, Chinese University of Hong Kong, Hong Kong S.A.R. of China

[4]Department of Thoracic Oncology, Tongji Hospital, Huazhong University of Science and Technology, Wuhan, China

[*]Correspondence to Qingpeng Zhang (qingpeng.zhang@cityu.edu.hk) and Qian Chu (qianchu@tjh.tjmu.edu.cn)

ABSTRACT

Compared with monotherapy, anti-cancer drug combination can provide effective therapy with less toxicity in cancer treatment. Recent studies found that the topological positions of protein modules related to the drugs and the cancer cell lines in the protein-protein interaction (PPI) network may reveal the effects of drugs. However, due to the size of the combinatorial space, identifying synergistic combinations of drugs from PPI network is computationally difficult. To address this challenge, we propose an end-to-end deep learning framework, namely Graph Convolutional Network for Drug Synergy (GraphSynergy), to make synergistic drug combination predictions. GraphSynergy adapts a spatial-based Graph Convolutional Network component to encode the high-order structure information of protein modules targeted by a pair of drugs, as well as the protein modules associated with a specific cancer cell line in the PPI network. The pharmacological effects of drug combinations are explicitly evaluated by their therapy and toxic scores. By introducing an attention component to automatically allocate contribution weights to the proteins, we show the ability of GraphSynergy to capture the pivotal proteins that play a part in both PPI network and biomolecular interactions between drug combinations and cancer cell lines. Experiments on two latest drug combination datasets demonstrate that GraphSynergy outperforms the state-of-the-art in predicting synergistic drug combinations. This study sheds light on using machine learning to discover effective combination therapies for cancer and other complex diseases.

Introduction

Drug combination therapy has shown great promises to improve the efficacy and extend the duration of response in the treatment of complex diseases[1], such as cancers[2], human immunodeficiency virus (HIV)[3], and cardiovascular diseases[4]. However, identifying synergistic drug combinations is challenging because the combinatorial space of drugs is huge and the effects can be adverse[5]. Therefore, effective identification of potential synergistic drug combinations for specific cancer diseases is in a pressing need, which can minimize the unexpected adverse effects and maximize the synergistic benefits.

Traditional drug combination identification is usually based on clinical experience. With the development of High-Throughput Screening (HTS)[6] technology, researchers may discover synergistic combinations by experiments at the cost of huge money and time spending. Machine learning methods offer the opportunity to efficiently explore the large combinatorial space. Existing models, such as random forest and support vector machine, mainly focus on the drug's chemical features or biological targets[7–9] of a specific cancer. Recent deep learning models such as DeepSynergy[10] introduces the cancer genomic information to make predictions for multiple types of cancers. A latest model, namely AuDNNsynergy[11], integrates multiomics data by introducing three auto-encoders. These methods are all based on the assumption that drugs with similar chemical structures have similar treatment effects. This assumption does not take into consideration the complex biological interactions among the proteins related to drugs and diseases, and thus lacks the ability to explicitly capture the toxic effects resulted by combining drugs.

Most anti-cancer drugs work with specific proteins related to cancer cells in the Protein-Protein Interaction (PPI) network[12]. Recent network science studies provide the evidence that the topological relations between drugs and diseases in the PPI network play an essential role in drug identification[13–15]. More specifically, (a) An effective drug should target the proteins within or near the corresponding disease module[14]; (b) Two drugs with synergistic effects should target complementary (non-overlapping) proteins to prevent the toxicities brought by over-exposure[15]. The potential predictability of considering the drug-drug and drug-disease relationships in the PPI network was demonstrated[15]. However, these network science methods only focus on the topological distance between proteins directly associated with drugs and diseases, while ignoring the local connections (formed by neighboring proteins) and the global structure of the PPI network. In addition, existing network science approaches[16] treat each protein homogeneously, whereas recent studies reveal that several proteins have dominant contribution to the progression of cancers[17, 18].

To address these challenges, we propose a novel end-to-end machine learning framework, namely the Graph Convolutional Network for Drug Synergy (GraphSynergy), to identify synergistic drug combinations for specific cancer cell lines from the perspective of the molecular mechanism (i.e. biological interactions between proteins) in the PPI network. GraphSynergy introduces a Graph Convolutional Network (GCN)[19, 20] component to learn the rich topological information of drug and disease modules in the PPI network by extending neighbor aggregation to deeper layers. At each layer, GraphSynergy utilizes an attention component[21] to determine the contribution of each protein and uses them to guide the aggregation. Then, we define two scores to explicitly evaluate two pharmacological characteristics of drugs: a joint *therapy score* measured by the similarity between the drug combinations and the cancer cell lines; a *toxic score* measured by the similarity between two drugs. Experiments on two latest drug combination datasets demonstrate that GraphSynergy outperforms the state-of-the-art in predicting synergistic drug combinations, and reveals the pivotal proteins that have been verified in both biological mechanism and PPI network. This represents a novel conceptual and methodological advancement in synergy prediction from the perspective of biomolecular interactions.

The contributions are threefold. First, to the best of our knowledge, GraphSynergy is the first attempt to identify synergistic drug combinations from the PPI network with deep learning methods. Second, GraphSynergy adapts graph embedding methods to capture the complex topological relations between drug combinations and cell lines, and to derive explicit joint therapy and toxic effects of drug combinations. Third, pivotal proteins can be identified to enhance the explainability of the prediction.

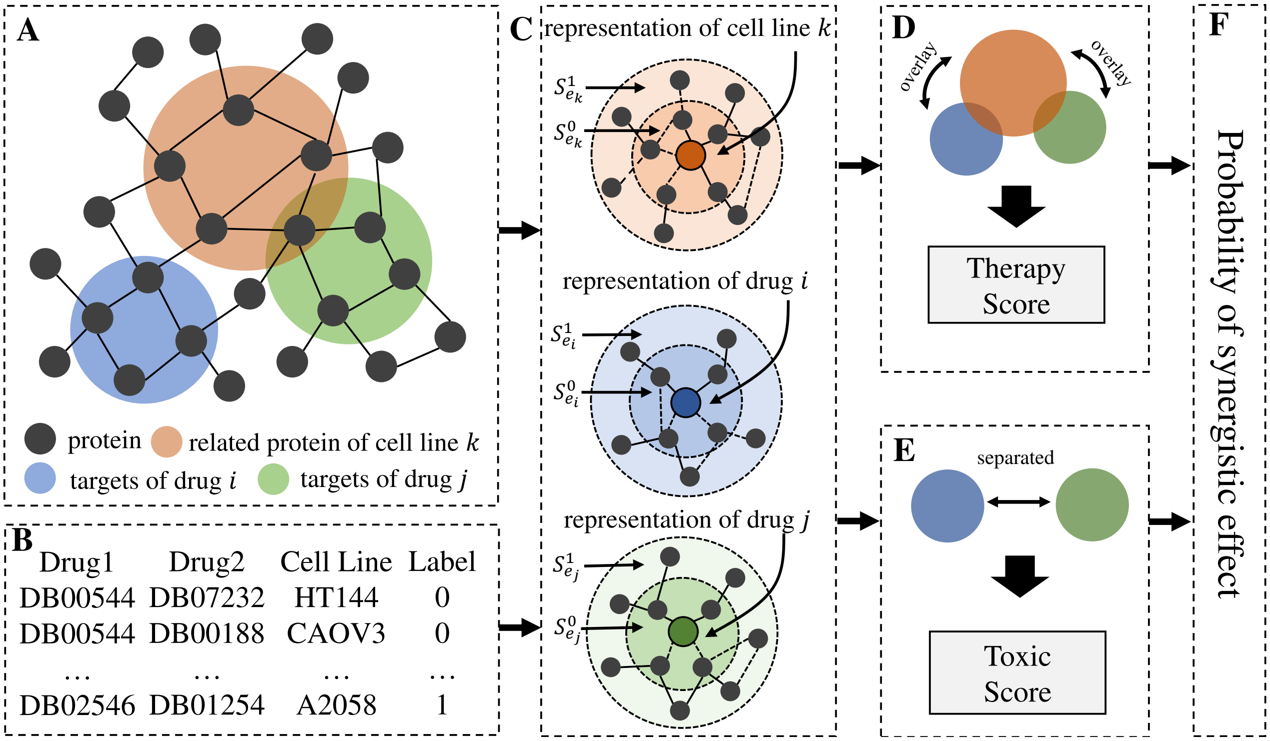
Results

GraphSynergy outperforms classic and state-of-art methods.

The framework of GraphSynergy is illustrated in Figure 1. GraphSynergy takes a combination of drug i and drug j with a cell line k as input and outputs the predicted probability that the drug combination is synergistic to the corresponding cell line. For each entity e in the input (i, j, k), its directed connected proteins S_e^0 are extracted from the drug(cell line)-protein associations A as target field, and then extended along edges in the PPI network G to form a radiant field $S_e^k(k = 1,2,...,H)$, which contains the proteins that are within k hops away from entity e. The radiant field captures the local interactions between the proteins that might play a role but are not directly targeted by the entity (either a drug or a cancer cell line). The union of target and radiant field $S_e^k(k = 0,1,2,...,H)$ are defined as the interaction field of entity e and these fields are fed into the aggregation layer iteratively to obtain the latent representation of entity e. An attentive mechanism is introduced to characterize the heterogeneous effect of proteins in the interaction field. The representations of drug i, drug j, and cell line k denote positions of their target modules in the PPI network, and they are combined to estimate the synergistic effect of the drug combination for the cancer cell line. Based on the relationships between the pharmacological effects with the topological positions of the related protein modules of drugs and cell lines, the synergistic effect is explicitly measured by two scores: *therapy score* and *toxic score*. The details of GraphSynergy is elaborated in the Methods.

We formulate the synergistic drug combination prediction problem as a binary classification problem. The performance is evaluated by five metrics: accuracy (ACC), recall, area under receiver operating characteristic curve (AUC-ROC), area under precision-recall curve (AUC-PR), and F1 score on two large recent-developed drug combination datasets: DrugCombDB[22] and Oncology-Screen[23]. DrugCombDB contains 69,436 drug combinations among 764 unique drugs and 76 unique cell lines. Oncology-Screen is a much smaller data which includes 4176 drug combinations among 21 unique drugs and 29 unique cell lines. To illustrate the superiority of GraphSynergy, we compare the performance of GraphSynergy with a number of classic and state-of-the-art baselines. The compared methods include Network Proximity[15], Matrix Factorization (GraRep[24]), Random Walk (DeepWalk[25] and Node2Vec[26], Deep Neural Network (DeepSynergy[10], and Graph Convolutional Network (KGNN[27] and GCN[19]). In general, as shown in Table 1, GraphSynergy significantly outperforms all baselines on both datasets. This demonstrates that GraphSynergy can predict the synergistic drug combinations for cancers well by capturing the topological relations between drug combination and cell line within the PPI network.

More specifically, as a model specific-designed for the Drug-Drug Interaction prediction task, KGNN has a similar graph embedding module to GraphSynergy, but it underperforms GraphSynergy significantly (e.g. at least 12.47% AUC-ROC reduction on DrugCombDB), indicating that incorporating the topological



Drug1	Drug2	Cell Line	Label
DB00544	DB07232	HT144	0
DB00544	DB00188	CAOV3	0
…	…	…	…
DB02546	DB01254	A2058	1

Figure 1. The framework of the GraphSynergy. **A.** The PPI network and the related proteins of cell lines and targeted proteins of drugs. **B.** The drug-drug-cell-line combination matrix, where the label 1 denotes the synergistic effect and 0 for the antagonistic effect. **C.** The graph aggregation layer of GraphSynergy. **D.** and **E.** show the idea of the designs of therapy score and toxic score, respectively, where the orange, blue and green round blocks denote the positions of protein modules of cell lines and two drugs, respectively. **F.** The output of GraphSynergy is the probability of a combination of drugs having the synergistic effect on a specific cell line.

proximity to characterize the therapy and toxicity effects can greatly improve the prediction power. Network Proximity (NP), which is based on proximity measures (e.g. z-score or separation score), shows mild prediction ability, but its performance is the least compared with other machine-learning-based models, indicating that simple proximity measures cannot fully capture the complex high-dimensional relations between drugs and cancer cell lines. The performances of random-walk-based models are nearly identical on both datasets. They also outperform the matrix-factorization-based model (GraRep). This is because that deep walks can capture sufficient structural information on these relatively smaller datasets. If the networks are large (such as social and bibliographic networks), GCN-based methods could have performed better than random-walk-based methods. KGNN performs better than GCN, indicating that incorporating the attention mechanism is helpful in learning the relations between drugs and cancer cell lines. DeepSynergy is the only model designed specifically for drug combination identification. Its

Model	DrugCombDB					Oncology-Screen				
	ACC	Recall	AUC-ROC	AUC-PR	F1	ACC	Recall	AUC-ROC	AUC-PR	F1
NP	0.6741	0.1752	0.4750	0.4459	0.2427	0.4547	0.0635	0.4887	0.5352	0.1117
GraRep	0.6667	0.5812	0.7282	0.6896	0.6162	0.6627	0.7667	0.7225	0.7020	0.7019
DeepWalk	0.6741	0.5964	0.7359	0.6979	0.6275	0.6830	0.7403	0.7427	0.7648	0.7225
Node2Vec	0.6735	0.5956	0.7358	0.6963	0.6268	0.6818	0.7403	0.7428	0.7649	0.7218
DeepSynergy	0.6846	0.6256	0.7461	0.7296	0.6434	0.6898	0.7179	0.7664	0.7696	0.7052
GCN	0.6733	0.6015	0.7211	0.6905	0.6264	0.6615	0.7375	0.7146	0.7096	0.6926
KGNN	0.6602	0.6751	0.7241	0.6916	0.6441	0.7080	0.7562	0.7820	0.7718	0.7280
GraphSynergy	**0.7545**	**0.7184**	**0.8351**	**0.8160**	**0.7281**	**0.7635**	**0.8008**	**0.8450**	**0.8497**	**0.7792**

Table 1. The performance of GraphSynergy compared with baselines. Each experiment is repeated 5 times, and the average performance is reported. The standard deviation scores are omitted due to the space limitation. Note that GraphSynergy presented here is with the transformation matrix for therapy score computation.

performance is in the midst of GCN and KGNN, indicating that the incapability of capturing the graphical information might have hindered its prediction power.

Prediction performance for various drugs, cancer cell lines and tissues

The prediction performance is highly dependent on the related proteins of drugs and cell lines. We further investigated the prediction performance for each individual drug and cell line on DrugCombDB dataset, which covers a wide spectrum of synergistic/antagonistic effects. The performance is represented by the ROC-AUC value between all the observed and the predicted combinations in the test set. Figure 2.A shows that the performance varies across different cell lines and drugs. GraphSynergy achieves great prediction ability (ROC-AUC is larger than 0.75) for more than 85% of drugs and 90% of cell lines.

The number and degree of related proteins vary dramatically among drugs and cell lines. We calculated the Pearson correlation between the number of related proteins and the prediction performance, as well as the Pearson correlation between the average degree and the ROC-AUC of predictions ((Figure 2.B) and Figure 2.C). The correlations are insignificant for both drugs and cell lines, indicating that simple statistics of the network connectivity are not predictive. In contrast, the complex biological mechanism and unique pharmacological properties can be well captured by the proposed GraphSynergy model.

We further calculated the prediction performance across different tissues, shown in Figure 2.D. In general, the performance of GraphSynergy didn't vary dramatically across different tissues, showing the generalization power of GraphSynergy which can be introduced to different tissues. More specifically,

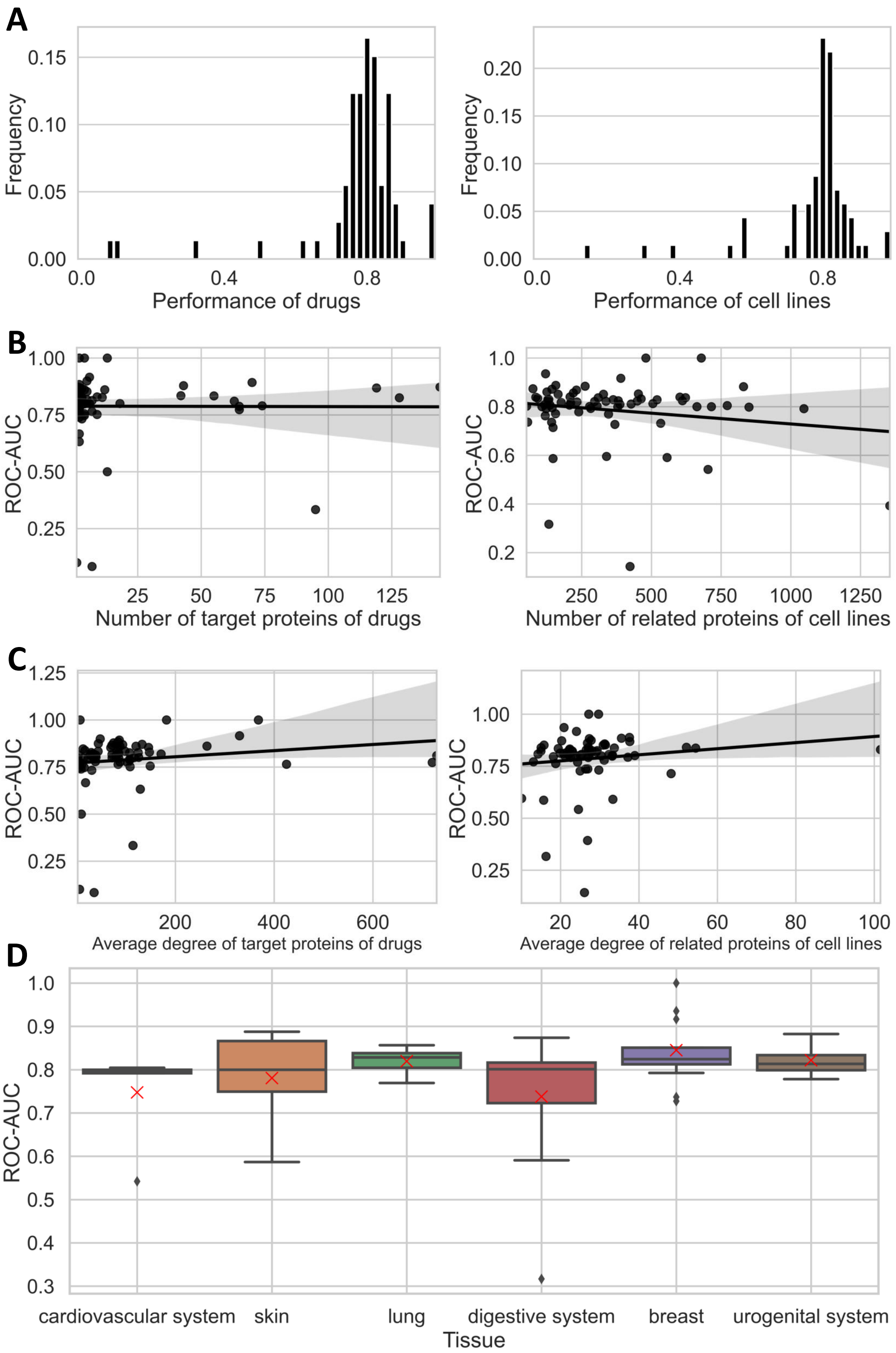


Figure 2. Prediction performance in view of different drugs, cell lines and tissues on DrugCombDB dataset. **A.** The distribution of the prediction performance for drugs and cell lines. The *x*-axis represents the value of the ROC-AUC. The *y*-axis represents the frequency of the corresponding ROC-AUC value. **B.** The distribution of the number of related proteins of drugs and cell lines with their corresponding ROC-AUC values. **C.** The distribution of the average degree of related proteins of drugs and cell lines with their corresponding ROC-AUC values. In both **B.** and **C.**, the lines are the fitting lines with grey shades denoting the fitting error. **D.** The tissue-specific distribution of the ROC-AUC values for all cell lines.

GraphSynergy achieved a high ROC-AUC value with small variance in lung (median, 0.8284), breast (median, 0.8243), and urogenital system (median, 0.8135).

Interpreting the pivotal proteins

GraphSynergy applies an attention mechanism to automatically allocate contribution weights to the proteins that are related/targeted by drugs and cell lines. The weight represents the contribution to the synergistic prediction of GraphSynergy. We define the pivotal proteins as the top 10 proteins by the contribution weight for each drug and cell line. In this section, we examine the roles played by these pivotal proteins in the connectivity of the PPI network, as well as their biological mechanisms.

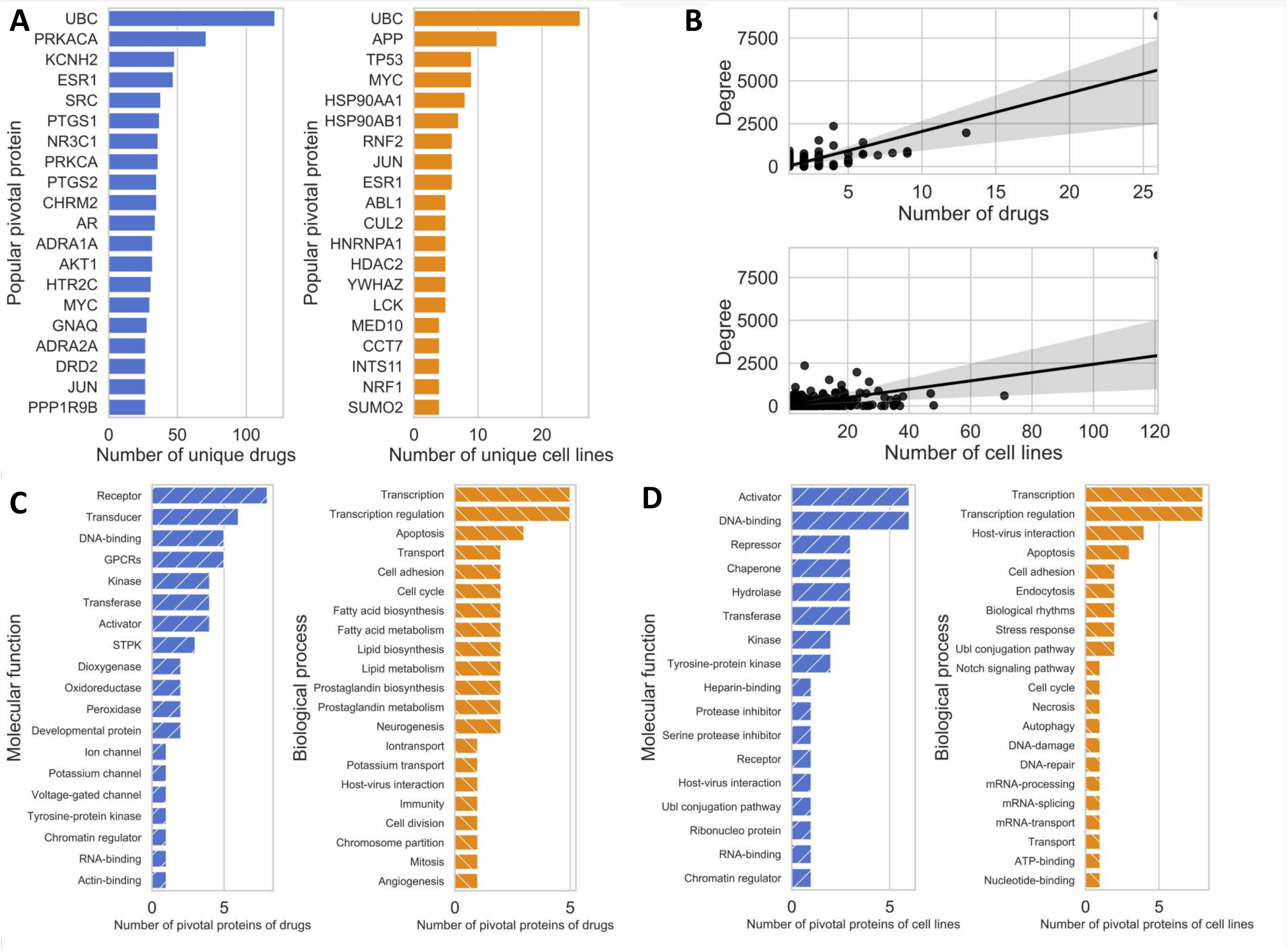


Figure 3. **A** shows the top 20 most frequent pivotal proteins among all the drugs and cell lines. **B** is the relationships between the occurrence frequencies among all the drugs (cell lines) with the degree of each pivotal protein, where the lines are the fitting lines with the grey shades denoting the fitting error. **C** and **D** show the molecular function and biological process for the top 20 most frequent pivotal proteins of drugs and cell lines, respectively. Note that in the molecular function plot, GPCRs and STPK are the abbreviations for G-protein coupled receptor and Serine/threonine-protein kinase, respectively.

Figure 3.A and Figure 3.B present the top 20 proteins measured by the frequency of being the pivotal protein for drugs and cell lines, respectively. They are noticeably difference, with only four common proteins: UBC, ESR1, JUN, and MYC. They represent the most important proteins in the progress of some cancers. For instance, UBC which encodes ubiquitin C contributes to the regulation of many celluar events, such as innate immunity, DNA repair and kinase activity[28,29] through the ubiquitin-proteasome pathway, and some recent researches found a synthetic lethal relationship between UBB and UBC that has potential to be exploited as a therapeutic strategy to fight these devastating cancers[30]. Having a closer look at the molecular functions and biological processes of these popular pivotal proteins, we found differences between those for drugs and those for cell lines (Figure 3.C-F.). Specifically, the popular pivotal proteins for drugs are more often acted as receptor and transducer, while those for cell lines tend to be activator and DNA-binding in view of molecular functions. Considering roles in the biological process, the popular proteins for both drugs and cell lines tend to play a role in transcription and transcription regulation.

Furthermore, we examine if the the network connectivity is associated with the pivotal role by calculating the Pearson correlation between the degree and the frequency of pivotal proteins in drugs (3.G) and in cell lines(3.H), respectively. They exhibit positive correlations in both drugs (0.7947, p-value < 0.0001) and cell lines (0.5585, p-value < 0.0001). These findings indicate that the proteins that interact with many others tend to be the pivotal proteins. This is aligned with the previous findings that high degree of protein is associated with its importance in the biological mechanism in human body.[31].

To intuitively demonstrate the explainability of GraphSynergy, we select two clinically-verified drug combinations in the testing set (Figure 4): the synergistic combination of Pemetrexed with Crizotinib on cell line NCIH322; the synergistic combination of Pemetrexed with Gefitinib on cell line NCIH522. All three drugs are identified to be effective for the treatment to non-small cell lung cancer (NSCLC)[32–34]. And some recent studies verified Pemetrexed in combination with Crizotinib or Gefitinib could both provide significant survival benefit for patients with NSCLC[35,36]. For both two combinations, GraphSynergy generate positive predictions, and even though the combination of Pemetrexed with Gefitinib on NCIH522 in the experiment data (DrugCombDB) is antagonistic, GraphSynergy can still identify their pharmacological effect which is verified in clinical trail stduy[36].

In Figure 4, for each protein in the interaction field of one drug or cell line, we directly obtain the normalized contribution weights. We only visualize the proteins with significant contribution weights (i.e. more than 0.001) and the proteins connecting to them (marked with purple color). Two cell lines NCIH322 and NCIH522 of non-small-cell lung cancer (NSCLC) are marked with dark orange and red, respectively, and the proteins related to them are marked with corresponding lighter colors. The color intensity is proportional to the distance. Similarly, three drugs (Crizotinib, Pemetrexed, and Gefitinib) and their related proteins are marked with blue, green and indigo. From Figure 4, we notice that even though the targets of these three drugs do not directly connect to the targets of NCIH322 or NCIH522, GraphSynergy still generates positive predictions based on their influential fields (indirectly connected proteins). Among hundreds of related proteins of NSCLC (NCIH322 and NCIH522), BNIP3L, TRIM26, AP4M1, RPL23 and so on are given a high priority by GraphSynergy. Recent clinical and experimental evidence showed the close relationships of such proteins with NSCLC. For example, TRIM26 was decreased in NSCLC and overexpression of TRIM26 inhibited NSCLC cell growth by suppressing PI3K/AKT pathway, which suggested that TRIM26 could be as a potential target for the treatment of NSCLC[37]; the loss of BNIP3L regulation through p53 under hypoxia facilitates microenvironmental adaptation and may be a key step in tumor development[38]. On the other hand, for these three drugs, MET, SBK1, IRAK3, ATIC, IKBKE and so on are given high importance, which are all connected to the related proteins of two cell lines. Previous studies show that these proteins are highly related to the progression of NSCLS. For instance, MET is found to be a promising therapeutic target in advanced NSCLC which can initiate and maintain tumor transformation, promote cell proliferation, survival, tumor invasion and angiogenesis when signals are abnormally activated[39]; Target IKBKE has shown to be a strategy to eradicate EGFR-TKI–resistant NSCLC[40].

Discussion

Experiments show that GraphSynergy can accurately identify synergistic anti-cancer combinations by capturing the biological mechanisms of proteins related to drugs and cancer cell lines in the PPI network. Among these three measures of therapy score, GraphSynergy with transformation matrix (GraphSynergy-tm) performs best, indicating that transformation matrix better captures the relations among the representations of drugs and cancer cell lines. As for the parameter sensitivity, we examine the effects of H,

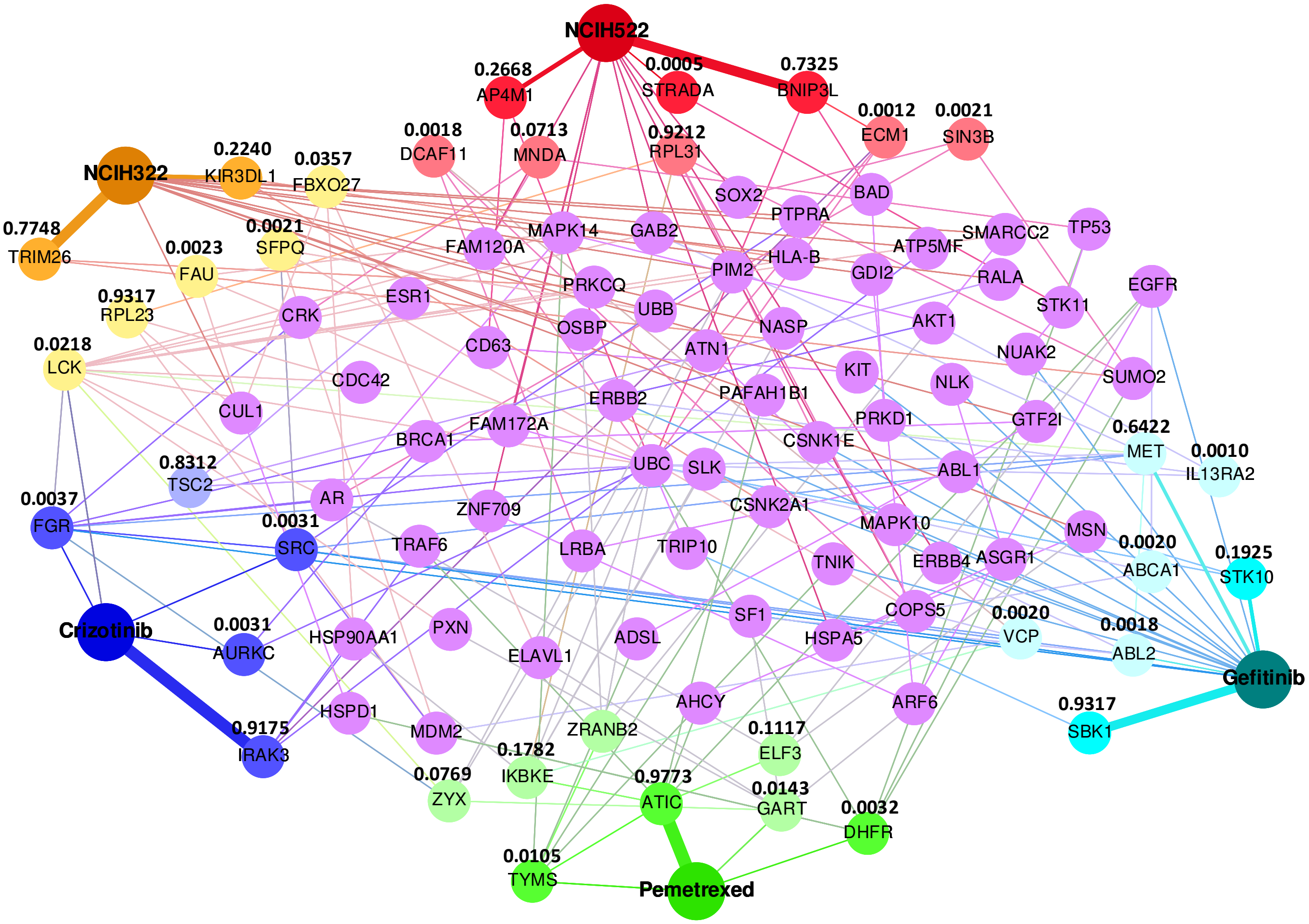


Figure 4. Visualization of contribution weights w.r.t. the related proteins for synergistic drug combination verified by clinical trials. Pivotal proteins are marked with their contribution weights above.

the depth in the interaction fields, and $\hat{S}$, the sample size of neighbors in each layer. Figure 5 reports the performance given other parameters fixed. In terms of H, GraphSynergy achieves the best performance when $H \geq 2$, indicating that a moderate depth is sufficient to capture the proteins that are most relevant to drug and cell line targets. In terms of $\hat{S}$, we find that $\hat{S} = 128$ yields the best performance, indicating that GraphSynergy needs to sample a representative subset of proteins in each layer to capture the complex relations between proteins in the PPI network.

GraphSynergy is inspired by the recent advances in network medicine[14, 15]. Based on a GCN framework, GraphSynergy further enhance the prediction power of network medicine, and the results verify that the topological positions of protein modules related to the drugs and cell lines in the PPI network are associated with the effects of drugs or drug combinations. By explicitly considering the therapy score and toxic score simultaneously, GraphSynergy outperform both network proximity method and state-of-the-art

Metrics	GraphSynergy-wip	GraphSynergy-mp	GraphSynergy-tm	GraphSynergy-tm with external features
ACC	0.7348(±0.0025)	0.7371(±0.0019)	**0.7545**(±0.0015)	0.7460 (±0.0045)
Recall	0.6859(±0.0039)	0.6920(±0.0030)	**0.7184**(±0.0015)	0.7077(±0.0117)
AUC-ROC	0.8144(±0.0022)	0.8161(±0.0019)	**0.8351**(±0.0008)	0.8215(±0.0043)
AUC-PR	0.7942(±0.0021)	0.7965(±0.0013)	**0.8160**(±0.0011)	0.7979(±0.0064)
F1	0.7041(±0.0029)	0.7080(±0.0023)	**0.7281**(±0.0015)	0.7154 (±0.0052)

Table 2. The performance of variations of GraphSynergy on DrugCombDB dataset. Exch experiment is repeated 5 times, and the average performance is reported. "wip", "mp", and "tm" denotes weighed inner product, max pooling and transformation matrix for therapy score in GraphSynergy, respectively.

deep learning models (Table 1).

Different from other deep-learning based approaches[10,16] (e.g. DeepSynergy) which are based on the detailed biological descriptions of cell lines (e.g. genomic profiles) and chemical descriptors of drugs, GraphSynergy only requires the related proteins modules of drugs and cell lines in the PPI network. Additionally, we examined if incorporating the genomic profiles and chemical descriptors into GraphSynergy would further improve the performance of GraphSynergy. Specifically, we followed the preprocessing steps of DeepSynergy and used the features of drugs and cell lines as the initial embeddings. As shown in Table 2, we found that these additional genomic and chemical features could not improve the performance of GraphSynergy. These results indicate that the molecular mechanisms introduced by the PPI network can better describe the relationships between drugs and cell lines even without any genomic and chemical features. However, similar to other existing approaches, GraphSynergy can only predict synergistic pharmacodynamic interactions whereas the effects of a drug on the absorption, distribution, metabolism and excretion of another drug (pharmacokinetic interactions), which are important for informing co-prescription of drugs in the clinical settings, is largely ignored.

To summarize, the superior performance of GraphSynergy suggests that the combination of network science knowledge with deep learning methods could be a valuable tool for the discovery of efficacious anti-drug combinations. With enough biological knowledge of gene expressions or target proteins in the PPI network, GraphSynergy is able to accurately identify identify novel combination therapies for multiple complex diseases. This study sheds light on using machine learning to discover effective combination therapies for cancer and other complex diseases.

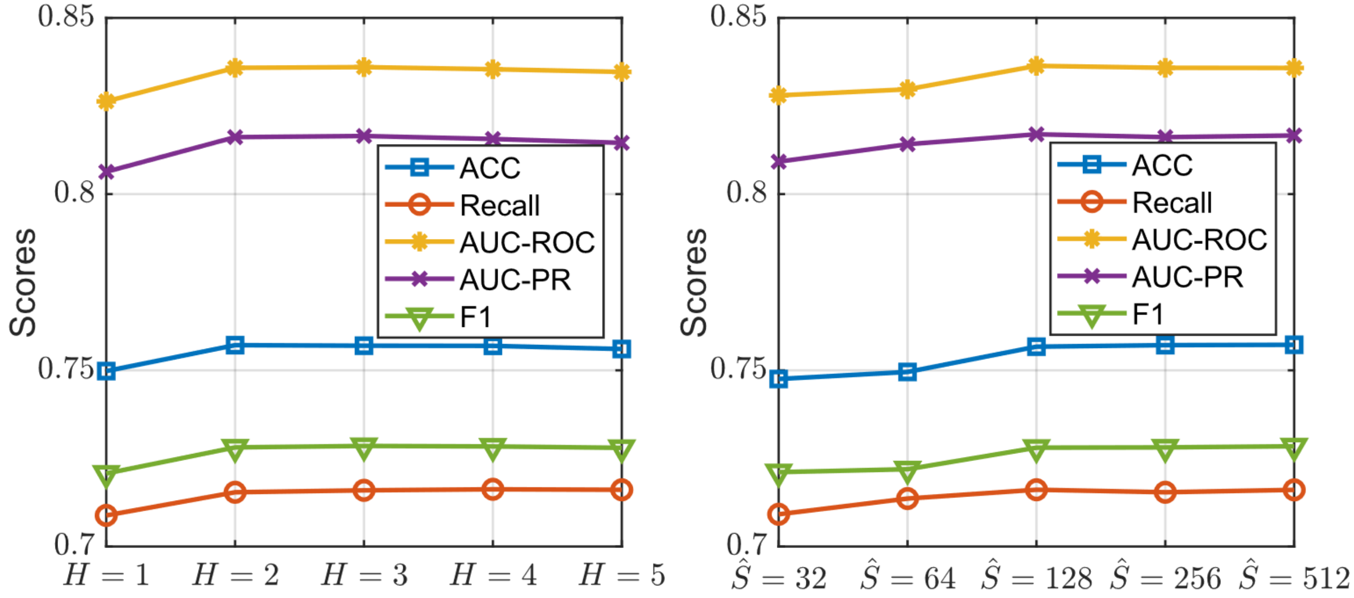


Figure 5. Results of GraphSynergy with respect to the depth in the interaction fields H and the depth in the interaction fields $\hat{S}$.

Methods

Datasets

Protein-Protein Interaction (PPI) Network. We use the comprehensive human interactome network generated by [15], which assembled by 15 commonly used databases with experimental evidence. The PPI network contains 217,160 interactions connecting 15,790 unique proteins. Each protein is mapped to its coding genes based on GeneCards.

Drug-protein Associations. The drug-protein associations are based on FDA-approved or clinically investigational drugs[15], which has 4,428 drugs and 2,256 human proteins connected by 15,051 associations.

Cell-protein Associations. The cell-protein associations dataset is harvested from the Cancer Cell Line Encyclopedia[41], which has 18,022 genes mapped to their coding proteins in the PPI network, 1,035 cancer cell lines, and 74,9551 associations.

Drug-drug-cell Associations. We evaluate GraphSynergy on two anti-cancer drug combination datasets: (a) **DrugCombDB**[22], a database with the largest number of drug combinations to date. (b) **Oncology-Screen**, an oncology screening dataset[23]. In both datasets, the drugs and cancer cell lines are mapped to

Datasets	Drugs	Cell lines	Positive pairs	Negative pairs	Target proteins of drugs	Related proteins of cell lines
DrugCombDB	764	76	31623	37813	6.92	364.86
Oncology-Sreen	21	29	2257	1919	19.28	306.24

Table 3. Basic statistics for the two datasets. All the values denote the number of corresponding entity and the numbers of related proteins of drugs and cell lines are the average number.

drug-protein associations and cell-protein associations, respectively, and the synergistic score above zero indicates the synergistic effect. The basic statistics of these two datasets are summarized in Table 3.

GraphSynergy

Problem Formulation

We formulate the synergistic drug combination prediction problem based on the PPI network as follows:

Drug-Combination-Cell-Line table. Given a set of N_d drugs and a set of N_c cancer cell lines[1], the effect of drug combinations on a specific cell line is defined as table $Y \in (0,1)^{|N_d| \times |N_d| \times |N_c|}$, where $|N_d|$ and $|N_c|$ denote the number of drugs and the number of cell lines, respectively. In the table, $y_{i,j,k} = 1$ $(i, j \in N_d, k \in N_c, i \neq j)$ indicates that the combination of drugs i and j is synergistic to cell line k; otherwise $y_{i,j,k} = 0$ denotes that the drug combination is antagonism (adverse) to this cell line.

Protein-Protein-Interaction network. We denote a PPI network as $G = (P, E)$, where P is the set of proteins and E is the set of interactions among these proteins. The interactions are physical contacts of high specificity established between two proteins.

Drug-Protein associations and Cell-Line-Protein associations. We integrate the associations between drugs and proteins and the associations between cell lines and proteins together as $A = \{e \rightarrow p | e \in (N_d \cup N_c), p \in P\}$. For any entity e (drug or cell line), it has multiple target proteins p existed in the PPI network G. The topological relations between drugs and cell lines are described by both their related proteins and the interactions between these proteins in the PPI network.

Given the drug-combination-cell table Y, the PPI network G, and the entity (drug or cell line)-protein associations A, we aim to predict whether the combination of drug $i (i \in N_d)$ and drug $j (j \in N_d)$ is

[1] Cell lines are cultures of human/animal cells that can be propagated repeatedly. Cell lines are commonly used to study the biology of cancer and to test cancer treatments.

synergistic to the cell line $k(k \in N_c)$. A prediction function $\hat{y}_{i,j,k} = \mathscr{F}(i,j,k|Y,G,A,\Theta)$ is formulated, where $\hat{y}_{i,j,k}$ denotes the probability that the combination of drugs i and j will be synergistic to the cell line k, and Θ denotes the model parameters to be learned.

Interaction Fields

Previous studies usually focus on the molecular mechanisms of the proteins directly targeted by drugs and diseases. Recent studies found out that some proteins act as mediators in the progression of drug efficacy or disease progressing[42]. Meanwhile, many drug-protein and cell-protein associations or the PPI network are still not yet discovered[14]. For example, Figure 6 shows part of proteins related to Tamoxifen, an anti-cancer drug with several targets: EBP, EBPL, ESR2, DHCR7, and HTR6. Although protein BAZ1A is not directly targeted by Tamoxifen, a recent study shows that BAZ1A plays an important role in the progression of non-small-cell lung cancer (NSCLC)[43]. Specifically, the knockdown of BAZ1A results in senescence-associated phenotypes in A549 cell lines, which is a main tissue cell of NSCLC. Subsequently, clinical trail studies[44] verified the therapy effect of Tamoxifen to NSCLS, indicating protein BAZ1A may play an important role in the therapy of Tamoxifen by interacting with the related proteins of NSCLS.

Inspired by the spatial-based GCN approach[19], we define the set of H-hop relevant neighbors $S_e^h(h = 0,1,2,...H)$ of entity e as its interaction fields, which contains the target filed S_e^0 (direct targets of entity e) and the radiant field $S_e^h(h = 1,2,...H)$ (indirectly connected proteins that may also play a role in the therapy mechanism). Note that the size of S_e^k may vary significantly across entities, thus we uniformly sample a fixed-size subset $\tilde{S}$ instead of using all the proteins at each layer.

Neighborhood Aggregation

This section illustrates how to aggregate the information of proteins at each layer in the interaction field for one entity.

Contribution Propagation Inspired by the attention mechanism[21], we utilize the inner product (defined as $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for simplicity) to compute the protein's contribution to the effect of the drug or the cell line. Given the representation of entity e and its interaction field S_e^h, each protein $p \in S_e^h$ is assigned a contribution weight:

$$\pi_p^e = g(e,p), \tag{1}$$

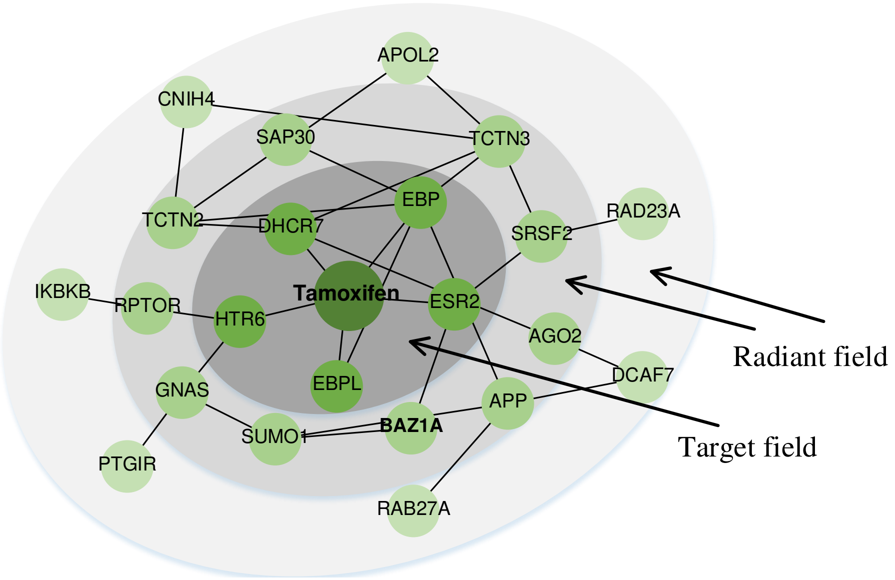


Figure 6. Illustration of target field and radiant field for drug Tamoxifen. The color from dark to light denotes the distance between these proteins and drug node from near to far. Each grey shade area represents one interaction field.

where $e \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$, d is the dimension of the representations. π_p^e denotes the contribution of a protein p to the effect of an entity e and can be regarded as the similarity of e and p.

After obtaining the contribution of each protein in one layer, we compute the linear combination of the proteins in this layer weighted by their corresponding contributions:

$$I_{S_e^h} = \sum_{p \in S_e^h} \hat{\pi}_p^e p, \tag{2}$$

where $I_{S_e^h}$ is the representation of layer h and $\hat{\pi}_p^e$ is the normalized contribution score:

$$\hat{\pi}_p^e = \frac{\exp\left(\pi_p^e\right)}{\sum_{p \in S_e^h} \exp\left(\pi_p^e\right)}. \tag{3}$$

Aggregation Layer By updating the representation of entity e in Eq. (1) using the representation $I_{S_e^h}$ of interaction field S_e^h, we can repeat the procedure of contribution propagation to deeper layers in order to obtain entity e's multiple-hops' representations $\{I_{S_e^0}, I_{S_e^1}, \ldots, I_{S_e^H}\}$. The problem now is how to aggregate the representations of different layers. Previous studies[27] experimented three types of aggregators $aggre : \mathbb{R}^d \times \mathbb{R}^d \ldots \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ on a similar structure, including: $aggre_{sum} = W \cdot \sum_{h=0}^{H} I_{S_e^h} + b$, $aggre_{concat} = W \cdot concat(I_{S_e^0}, I_{S_e^1}, \ldots, I_{S_e^H}) + b$, and $aggre_{neighbor} = W \cdot I_{S_e^H} + b$. Their experiments showed that $aggre_{concat}$ performs best. GraphSynergy adopts it and the final representation of entity e is:

$$\hat{e} = W_{agg} \cdot concat(I_{S_e^0}, I_{S_e^1}, ..., I_{S_e^H}) + b_{agg}, \tag{4}$$

where $\hat{e}$ is the final representation of entity e, W_{agg} and b_{agg} are the aggregation weight and bias, respectively.

Therapy Score and Toxic Score

Given the final representations $\hat{e}_i$, $\hat{e}_j$ and $\hat{e}_k$ of two drugs and one cell line, we aim to predict whether the drug combination is synergistic to this cell line. GraphSynergy uses the inner product of entities' representations to measure the similarity between two entities. The higher the similarity, the more overlap between the two corresponding protein modules. Therefore we define two scores as follows:

Therapy Score. The drugs targeting proteins that are within or near the proteins in the disease module are found to be more effective in treating the disease[14]. Thus, we use three methods (denoted as $\Gamma : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$) to compute the therapy score s_p, which is evaluated by the similarity between the final representations of drug pairs and cell line.

- *Weighted inner product* first computes the inner products of two drugs with the cell line separately, and then takes the weighted sum as the therapy score.

  $$\Gamma_{wip}(\hat{e}_i, \hat{e}_j, \hat{e}_{c_k}) = \alpha(\hat{e}_i \odot \hat{e}_k) + \beta(\hat{e}_j \odot \hat{e}_k) \tag{5}$$

  where α and β are the weights, respectively.

- *Max pooling* utilizes the element-wise maximum of the representations of two drugs as the combined drug representation, and then compute the the inner product of the combined drug representation and the cell line representation.

  $$\Gamma_{mp}(\hat{e}_i, \hat{e}_j, \hat{e}_k) = \max(\hat{e}_i, \hat{e}_j) \odot \hat{e}_k \tag{6}$$

- *Transformation matrix* concatenates the representations of the two drugs, and then compute the inner product of the concatenated drug representation and the cell line representation.

  $$\Gamma_{tm}(\hat{e}_i, \hat{e}_j, \hat{e}_k) = (W_\Gamma \cdot concat(\hat{e}_i, \hat{e}_j) + b_\Gamma) \odot \hat{e}_k, \tag{7}$$

  where W_Γ and b_Γ are the weight and bias, respectively.

Toxic Score. It has been recognized that we should avoid the combination of drugs that are overlapping in the PPI network to prevent the toxicities[15]. Thus, GraphSynergy prefers the pair of drugs that are dissimilar. The toxic score is computed as the inner product of the two drugs' representations.

$$s_n = \Psi(\hat{e}_i, \hat{e}_j) = \hat{e}_i \odot \hat{e}_j \tag{8}$$

Synergistic Drug Combination Prediction

The synergistic drug combination prediction problem can be viewed as a binary classification task. Given the representations of drug i, drug j, and cell line k, the synergistic probability $\hat{y}_{i,j,k}$ is the evaluated by the difference between the therapy score s_p and the toxic score s_n.

$$\hat{y}_{i,j,k} = \sigma(s_p - s_n), \tag{9}$$

where σ is the *sigmoid* function.

Given a set of drug-drug-cell pairs and the ground-truth, we formulate the following loss function for GraphSynergy:

$$\mathbb{L} = \sum_{(i,j,k)\in Y(i\neq j)} \mathscr{L}(y_{i,j,k}, \hat{y}_{i,j,k}) + \frac{\lambda_1}{2}\|\mathbb{E}\|_2^2 + \frac{\lambda_2}{2}\|\Theta\|_2^2, \tag{10}$$

where $\mathscr{L}(y_{i,j,k}, \hat{y}_{i,j,k}) = -y_{i,j,k}\log\hat{y}_{i,j,k} - (1-y_{i,j,k})\log(1-\hat{y}_{i,j,k})$ is the binary-cross-entropy loss. The second and third terms are the regularizers to prevent over-fitting, where $\mathbb{E}$ is the embedding matrix for all items.

The model size and time complexity of GraphSyner are $(|N_d| + |N_c| + |N_p|) \times d + (H+2) \times d^2 + d$ and $\mathscr{O}(nHd^2)$ (where n is the number of drug-drug-cell combinations), respectively. Please refer to the supplementary materials for detailed complexity analysis.

Baselines

We compare the performance of GraphSynergy with the a number of classic and state-of-the-art baselines based on Network Proximity, Matrix Factorization, Random Walk, DNN, and GCN.

(a) **Network Proximity (NP)**:[15] utilized two measures (z-score and separation score) in the PPI network to quantify the proximity between the proteins targeted by the drugs and the diseases.

(2) **GraRep**: GraRep[24] is a matrix-factorization-based approach which integrates global structural information to learn the representations of graph.

(3) **DeepWalk** and **Node2Vec**: Random walk-based methods are commonly used for link prediction and knowledge representation. We adopt DeepWalk[25] and Node2Vec[26].

(4) **DeepSynergy**: DeepSynergy[10] applied a normalization strategy to account for input data heterogeneity, and conical layers to model drug synergies with a DNN framework.

(5) **GCN** and **KGNN**: We choose a representative spatial-based GCN methods[19] and a state-of-art knowledge-graph-based method: KGNN[27], which is specially designed for DDI prediction task.

Experiments Setup

GraphSynergy is implemented with Python 3.7, PyTorch 1.6.0, NumPy 1.19.1 and scikit-learn 0.23.2. The batch size, learning rate, the depth of the interaction field H, and the sample size of neighbors in each layer $\hat{S}$ are set to 512, 10^{-3}, 2, and 128, respectively, for both two datasets. The ratio of training, validation, and test set is 7: 1: 2. The hyper-parameters are determined by optimization AUC-ROC on the validation set. Specifically, for DrugCombDB dataset, the dimension of embeddings d, and the regularizer weights λ_1 and λ_2 are 64, 10^{-6}, and 10^{-4}, respectively. For Oncology-Screen dataset, d, λ_1 and λ_2 are 32, 10^{-5}, and 10^{-6} separately. All trainable parameters are optimized by Adam algorithm and the number of epoch is set to 50 for training. All the models are trained from scratch without any pre-training on a single NVIDIA T4 GPU. Please refer to the supplementary materials for more details of the parameter settings of baselines.

References

1. Sun, X., Vilar, S. & Tatonetti, N. P. High-throughput methods for combinatorial drug discovery. *Sci. translational medicine* **5**, 205rv1–205rv1 (2013).

2. Yadav, B., Wennerberg, K., Aittokallio, T. & Tang, J. Searching for drug synergy in complex dose–response landscapes using an interaction potency model. *Comput. structural biotechnology journal* **13**, 504–513 (2015).

3. De Clercq, E. The design of drugs for hiv and hcv. *Nat. reviews Drug discovery* **6**, 1001–1018 (2007).

4. Gu, L. *et al.* Treatment outcomes of transcatheter arterial chemoembolization combined with local ablative therapy versus monotherapy in hepatocellular carcinoma: a meta-analysis. *J. cancer research clinical oncology* **140**, 199–210 (2014).

5. Tol, J. *et al.* Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. *New Engl. J. Medicine* **360**, 563–572 (2009).

6. He, L. *et al.* Methods for high-throughput drug combination screening and synergy scoring. In *Cancer systems biology*, 351–398 (Springer, 2018).

7. Li, P. *et al.* Large-scale exploration and analysis of drug combinations. *Bioinformatics* **31**, 2007–2016 (2015).

8. Wildenhain, J. *et al.* Prediction of synergism from chemical-genetic interactions by machine learning. *Cell Syst.* **1**, 383–395 (2015).

9. Li, J., Tong, X.-Y., Zhu, L.-D. & Zhang, H.-Y. A machine learning method for drug combination prediction. *Front. Genet.* **11**, 1000 (2020).

10. Preuer, K. *et al.* Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* **34**, 1538–1546 (2018).

11. Zhang, T., Zhang, L., Payne, P. R. O. & Li, F. *Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models*, 223–238 (Springer US, 2021).

12. Kumar, B., Singh, S., Skvortsova, I. & Kumar, V. Promising targets in anti-cancer drug development: Recent updates. *Curr. Medicinal Chem.* **24**, 4729–4752 (2017).

13. Ma, J. *et al.* A comparative study of cluster detection algorithms in protein–protein interaction for drug target discovery and drug repurposing. *Front. pharmacology* **10**, 109 (2019).

14. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. communications* **7**, 1–13 (2016).

15. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nat. communications* **10**, 1–11 (2019).

16. Li, H., Li, T., Quang, D. & Guan, Y. Network propagation predicts drug synergy in cancers. *Cancer research* **78**, 5446–5457 (2018).

17. Chen, J. *et al.* Low expression lncrna tuba4b is a poor predictor of prognosis and regulates cell proliferation in non-small cell lung cancer. *Pathol. & Oncol. Res.* **23**, 265–270 (2017).

18. Jing, H. *et al.* Sirt2 and lysine fatty acylation regulate the transforming activity of k-ras4a. *Elife* **6**, e32436 (2017).

19. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17 (2017).

20. Wang, H., Zhao, M., Xie, X., Li, W. & Guo, M. Knowledge graph convolutional networks for recommender systems. In *The world wide web conference*, 3307–3313 (2019).

21. Bahdanau, D., Cho, K. H. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015* (2015).

22. Liu, H. *et al.* Drugcombdb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic acids research* **48**, D871–D881 (2020).

23. O'Neil, J. *et al.* An unbiased oncology compound screen to identify novel combination strategies. *Mol. cancer therapeutics* **15**, 1155–1162 (2016).

24. Cao, S., Lu, W. & Xu, Q. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 891–900 (2015).

25. Perozzi, B., Al-Rfou, R. & Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710 (2014).

26. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864 (2016).

27. Lin, X., Quan, Z., Wang, Z.-J., Ma, T. & Zeng, X. Kgnn: Knowledge graph neural network for drug-drug interaction prediction (IJCAI, 2020).

28. Rajsbaum, R. & García-Sastre, A. Unanchored ubiquitin in virus uncoating. *Science* **346**, 427–428 (2014).

29. Pickart, C. M. & Fushman, D. Polyubiquitin chains: polymeric protein signals. *Curr. opinion chemical biology* **8**, 610–616 (2004).

30. Kedves, A. T. *et al.* Recurrent ubiquitin b silencing in gynecological cancers establishes dependence on ubiquitin c. *The J. clinical investigation* **127**, 4554–4568 (2017).

31. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

32. Skricková, J. *et al.* Pemetrexed in maintenance therapy of 164 patients with advanced non-small-cell lung cancer (nsclc) (2016).

33. Giaccone, G. The role of gefitinib in lung cancer treatment. *Clin. cancer research* **10**, 4233s–4237s (2004).

34. Chuang, J. C. & Neal, J. W. Crizotinib as first line therapy for advanced alk-positive non-small cell lung cancers. *Transl. lung cancer research* **4**, 639 (2015).

35. Gandhi, L., Drappatz, J., Ramaiya, N. H. & Otterson, G. A. High-dose pemetrexed in combination with high-dose crizotinib for the treatment of refractory cns metastases in alk-rearranged non–small-cell lung cancer. *J. Thorac. Oncol.* **8**, e3–e5 (2013).

36. Cheng, Y. *et al.* Randomized phase ii trial of gefitinib with and without pemetrexed as first-line therapy in patients with advanced nonsquamous non–small-cell lung cancer with activating epidermal growth factor receptor mutations. *J. Clin. Oncol.* **34**, 3258–3266 (2016).

37. Tao, J.-L. *et al.* Overexpression of tripartite motif containing 26 inhibits non-small cell lung cancer cell growth by suppressing pi3k/akt signaling. *The Kaohsiung journal medical sciences* **36**, 417–422 (2020).

38. Mellor, H. R. & Harris, A. L. The role of the hypoxia-inducible bh3-only proteins bnip3 and bnip3l in cancer. *Cancer Metastasis Rev.* **26**, 553–566 (2007).

39. Liang, H. & Wang, M. Met oncogene in non-small cell lung cancer: mechanism of met dysregulation and agents targeting the hgf/c-met axis. *OncoTargets therapy* **13**, 2491 (2020).

40. Challa, S. *et al.* Ikbke is a substrate of egfr and a therapeutic target in non–small cell lung cancer with activating mutations of egfr. *Cancer research* **76**, 4418–4429 (2016).

41. Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

42. Gonzalez, M. W. & Kann, M. G. Protein interactions and disease. *PLoS Comput. Biol* **8**, e1002819 (2012).

43. Li, X. *et al.* Chromatin remodeling factor baz1a regulates cellular senescence in both cancer and normal cells. *Life sciences* **229**, 225–232 (2019).

44. Wen, S. *et al.* Efficacy of tamoxifen in combination with docetaxel in patients with advanced non-small-cell lung cancer pretreated with platinum-based chemotherapy. *Anti-Cancer Drugs* **27**, 447–456 (2016).

Acknowledgements (not compulsory)

To be added

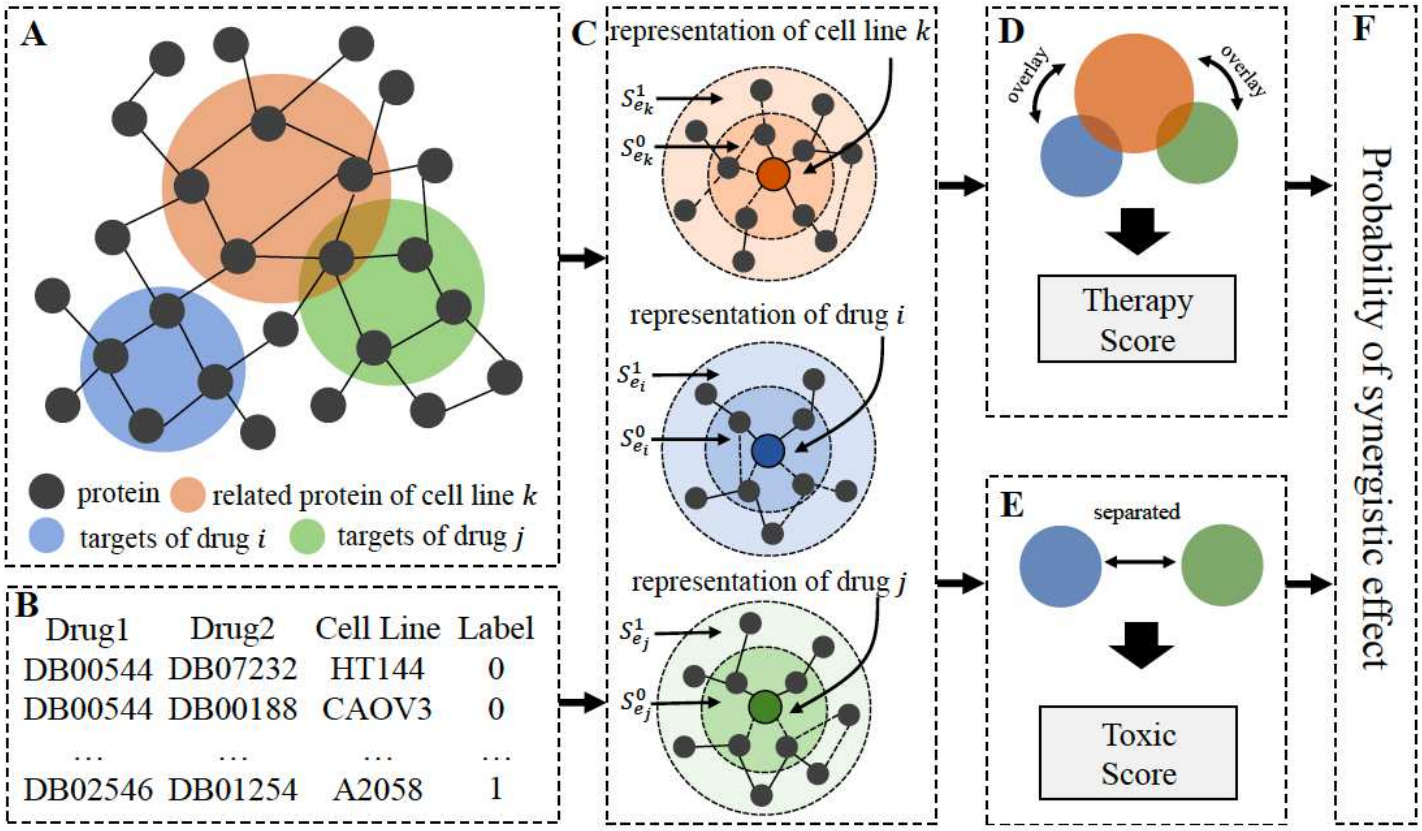
Author contributions statement

To be added

Additional information

The source code for GraphSynergy is available at `https://github.com/JasonJYang/GraphSynergy`.

Figures



Drug1	Drug2	Cell Line	Label
DB00544	DB07232	HT144	0
DB00544	DB00188	CAOV3	0
…	…	…	…
DB02546	DB01254	A2058	1

Figure 1

The framework of the GraphSynergy. A. The PPI network and the related proteins of cell lines and targeted proteins of drugs. B. The drug-drug-cell-line combination matrix, where the label 1 denotes the synergistic effect and 0 for the antagonistic effect. C. The graph aggregation layer of GraphSynergy. D. and E. show the idea of the designs of therapy score and toxic score, respectively, where the orange, blue and green round blocks denote the positions of protein modules of cell lines and two drugs, respectively. F. The output of GraphSynergy is the probability of a combination of drugs having the synergistic effect on a specific cell line.

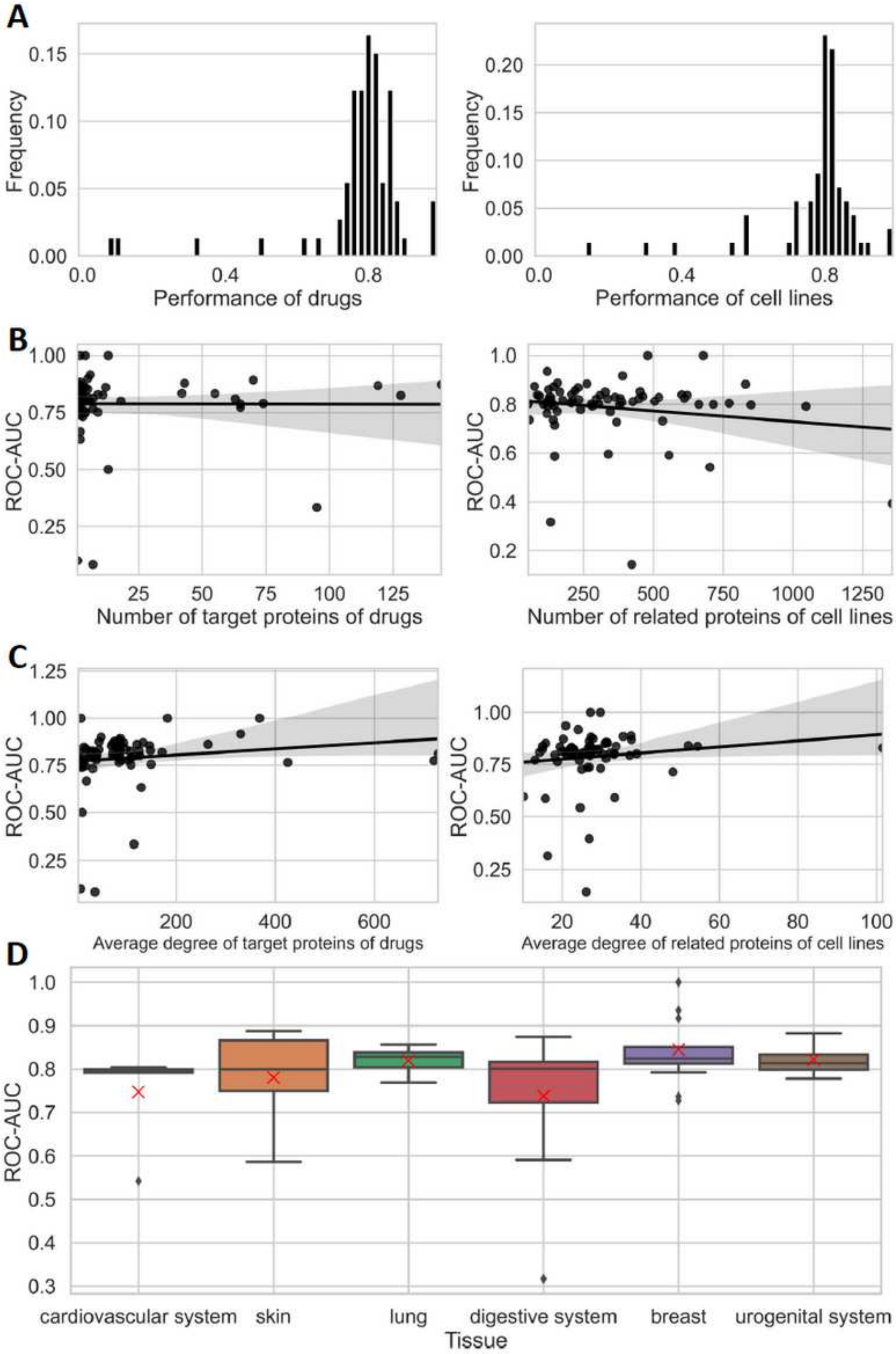


Figure 2

Prediction performance in view of different drugs, cell lines and tissues on DrugCombDB dataset. A. The distribution of the prediction performance for drugs and cell lines. The x-axis represents the value of the ROC-AUC. The y-axis represents the frequency of the corresponding ROC-AUC value. B. The distribution of the number of related proteins of drugs and cell lines with their corresponding ROC-AUC values. C. The distribution of the average degree of related proteins of drugs and cell lines with their corresponding ROC-

AUC values. In both B. and C., the lines are the fitting lines with grey shades denoting the fitting error. D. The tissue-specific distribution of the ROC-AUC values for all cell lines.

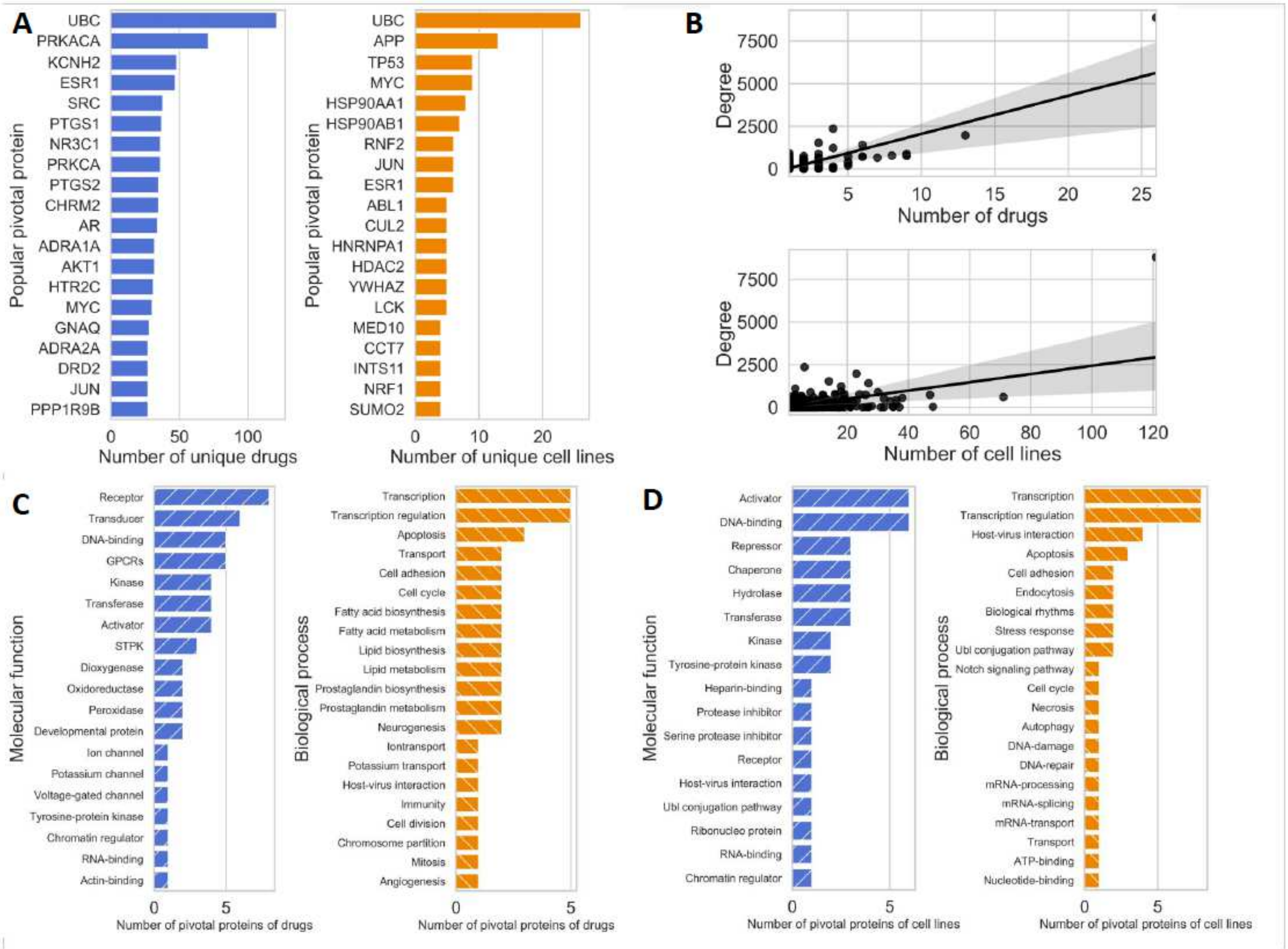


Figure 3

A shows the top 20 most frequent pivotal proteins among all the drugs and cell lines. B is the relationships between the occurrence frequencies among all the drugs (cell lines) with the degree of each pivotal protein, where the lines are the fitting lines with the grey shades denoting the fitting error. C and D show the molecular function and biological process for the top 20 most frequent pivotal proteins of drugs and cell lines, respectively. Note that in the molecular function plot, GPCRs and STPK are the abbreviations for G-protein coupled receptor and Serine/threonine-protein kinase, respectively.

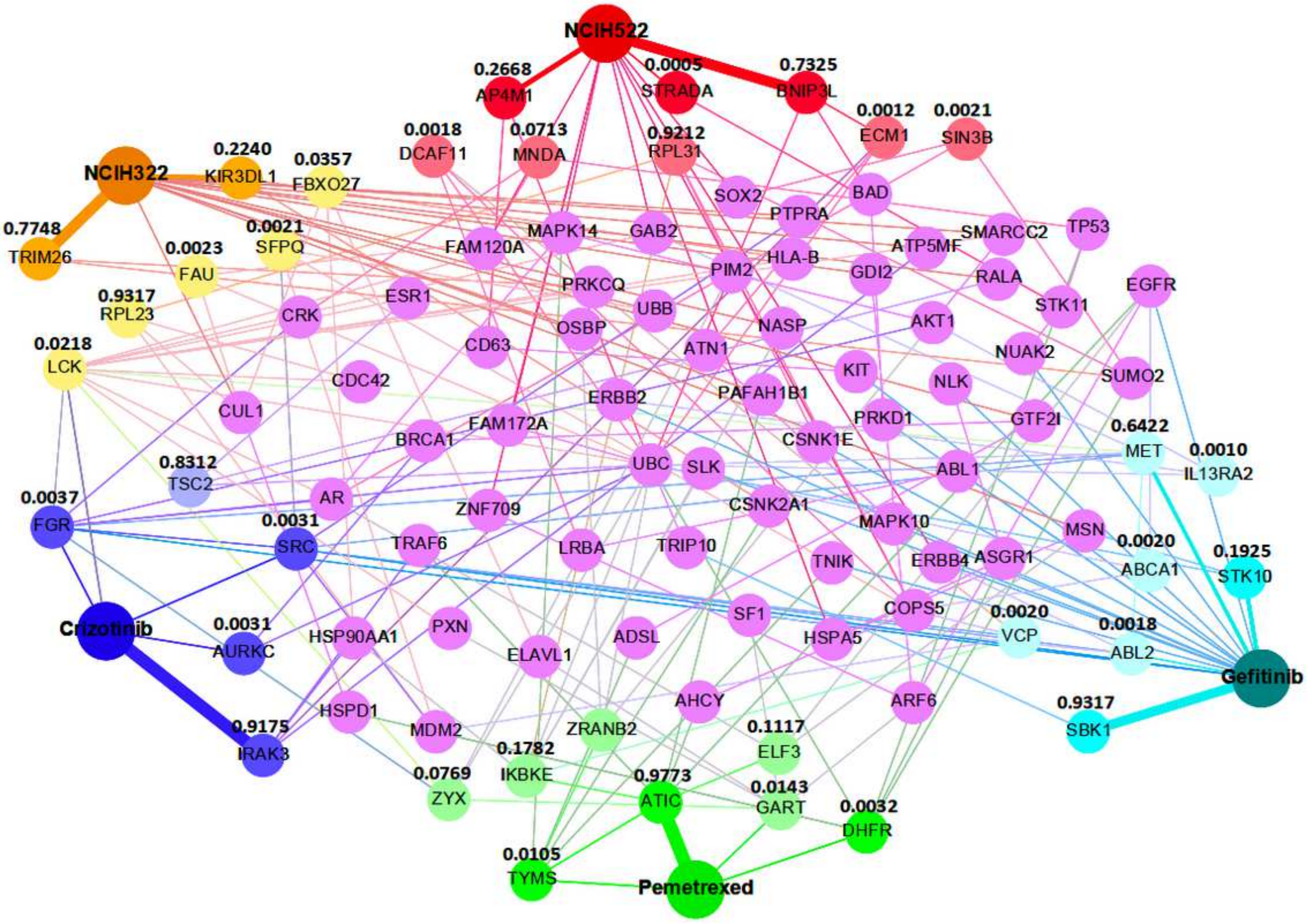


Figure 4

Visualization of contribution weights w.r.t. the related proteins for synergistic drug combination verified by clinical trials. Pivotal proteins are marked with their contribution weights above.

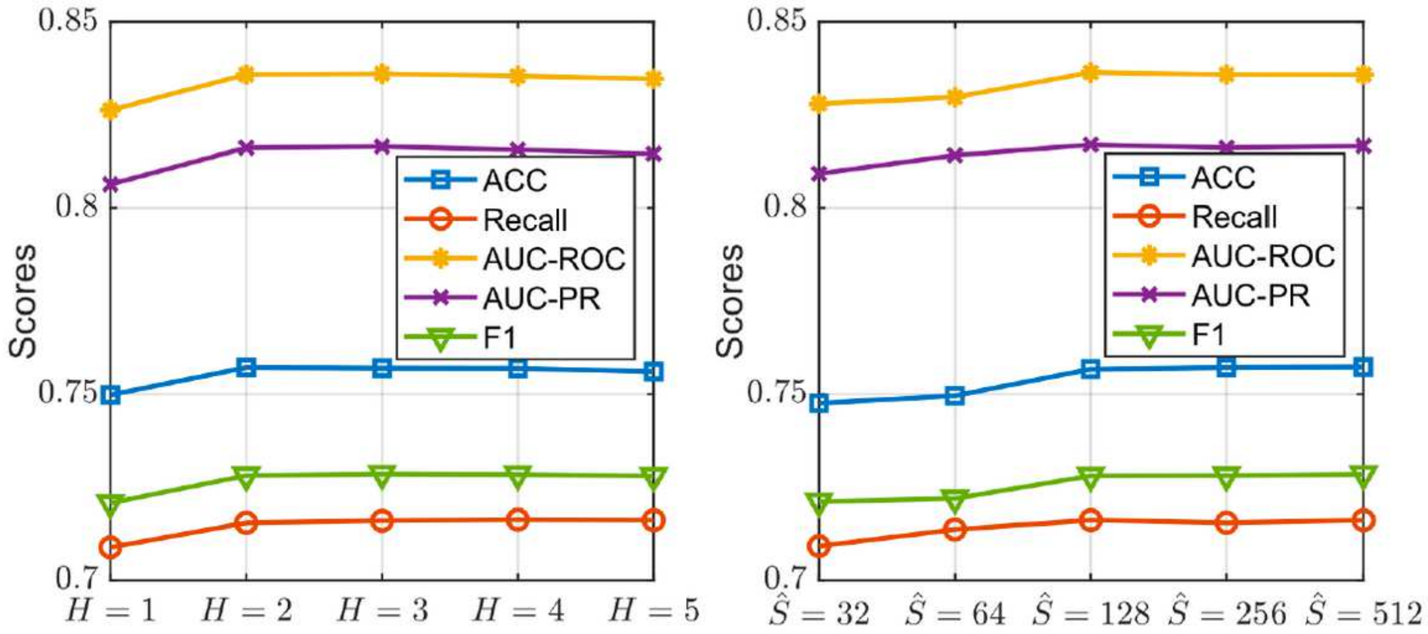


Figure 5

Results of GraphSynergy with respect to the depth in the interaction fields H and the depth in the interaction fields S.

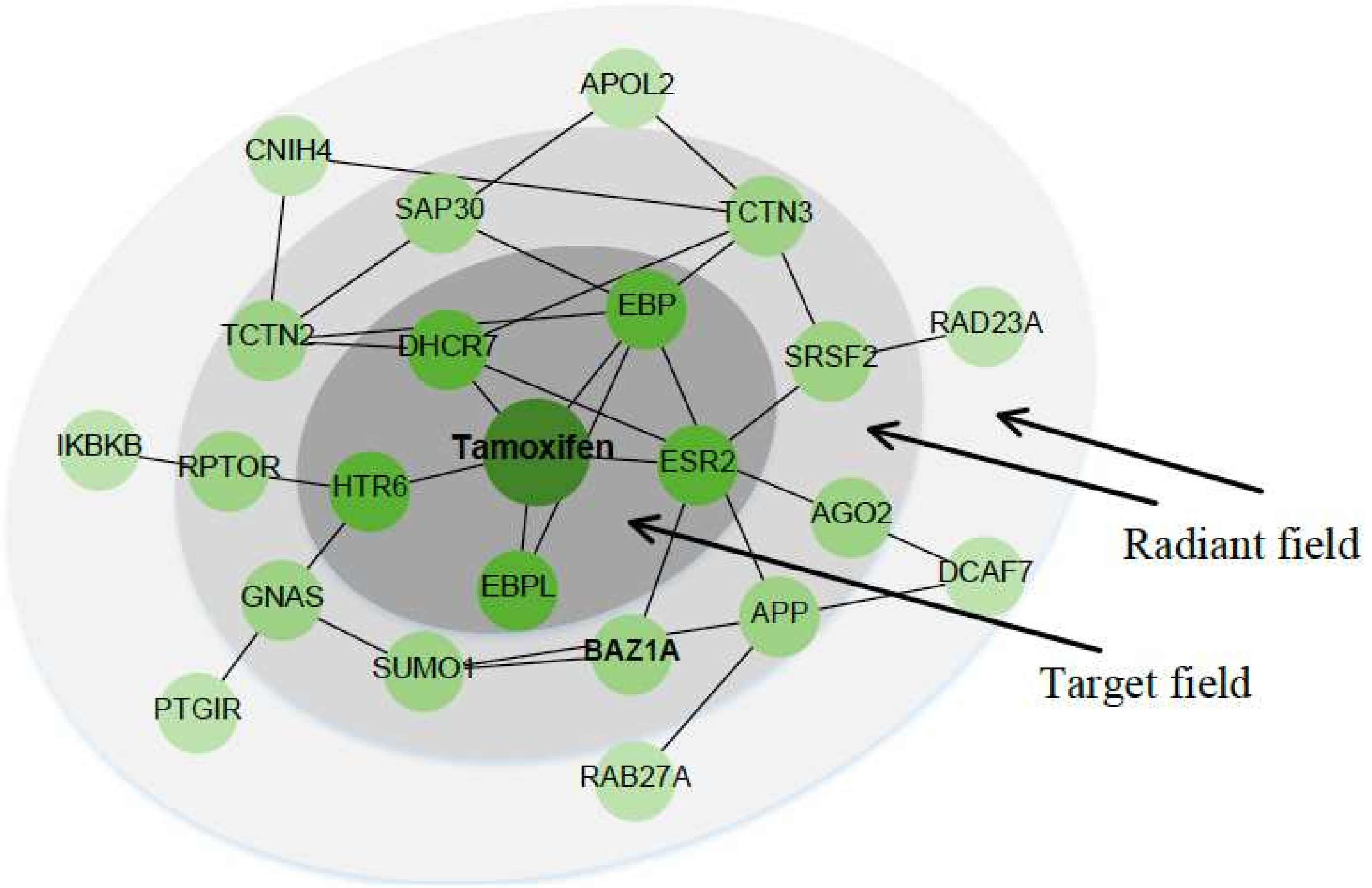


Figure 6

Illustration of target field and radiant field for drug Tamoxifen. The color from dark to light denotes the distance between these proteins and drug node from near to far. Each grey shade area represents one interaction field.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf