mathematics

MDPI

Article

Graph Convolutional Network for Drug Response Prediction Using Gene Expression Data

Seonghun Kim [1,†], Seockhun Bae [1,†], Yinhua Piao [2] and Kyuri Jo [1,*]

[1] Department of Computer Engineering, Chungbuk National University, Cheongju 28644, Korea; tjdgns4560@naver.com (S.K.); dinky24@chungbuk.ac.kr (S.B.)
[2] Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Korea; yinhuapark@naver.com
* Correspondence: kyurijo@chungbuk.ac.kr
† These authors contributed equally to this work.

Abstract: Genomic profiles of cancer patients such as gene expression have become a major source to predict responses to drugs in the era of personalized medicine. As large-scale drug screening data with cancer cell lines are available, a number of computational methods have been developed for drug response prediction. However, few methods incorporate both gene expression data and the biological network, which can harbor essential information about the underlying process of the drug response. We proposed an analysis framework called DrugGCN for prediction of Drug response using a Graph Convolutional Network (GCN). DrugGCN first generates a gene graph by combining a Protein-Protein Interaction (PPI) network and gene expression data with feature selection of drug-related genes, and the GCN model detects the local features such as subnetworks of genes that contribute to the drug response by localized filtering. We demonstrated the effectiveness of DrugGCN using biological data showing its high prediction accuracy among the competing methods.

Keywords: neural network; graph convolutional network; spectral graph theory; drug response; bioinformatics

Citation: Kim, S.; Bae, S.; Piao, Y; Jo, K. Graph Convolutional Network for Drug Response Prediction Using Gene Expression Data. *Mathematics* **2021**, *9*, 772. https://doi.org/10.3390/math9070772

Academic Editors: Basil Papadopoulos, Gustavo Santos-García and Moo Chung

Received: 31 January 2021
Accepted: 30 March 2021
Published: 2 April 2021

1. Introduction

Cancer is a disease driven by the accumulation of somatic mutations. Mutations on specific genes that are called cancer driver genes can affect the transcription of the genes and cause the differential expression of the genes. As many of the driver genes are signaling molecules that control the expression of downstream genes, their differential gene expressions may impinge on the cell and contribute to the hallmarks of cancer such as sustained cell proliferation and resistance to cell death [1,2].

In the early stages of the analysis on differential gene expression in cancer, several research works focused on comparative studies between normal and cancer cells [1,3]. Since the era of precision medicine or personalized medicine, however, the analysis of the differential gene expression among individual patients has become popular, as researchers have observed heterogeneity for immune responses induced by the same cancer therapy due to the diverse genetic background of individuals [4,5]. A recent study suggested that only around 5% of patients benefit from precision oncology [6], which highlights the importance of improving the prediction accuracy of drug response.

As the number of molecular data from cancer patients increases, several large-scale database have been created including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [7,8]. Although TCGA and the ICGC provide multi-platform genomic profiling data across cancer types, these databases do not include a large number of patient records with drug response or responses to multiple drugs as the data were collected from patients (donors). On the other hand, the Genomics of Drug Sensitivity in Cancer (GDSC) [9] database provides large-scale drug screening results for

266 drugs on around 1000 human cancer cell lines that can be utilized to learn and predict drug responses from gene expression by computational methods.

With a large number of cell line data, various computational methods based on machine learning have been proposed to predict drug response [10,11]. Drug response prediction is one of the supervised learning problems. Computational models are trained to compute a drug response value (output) of cell lines (m samples) with a genomic profile (n input features). Depending on the type of method, learning can be performed using data of the d entire drugs at once or for each drug separately. The genomic profile of cell lines is usually given as a matrix $C_{m\times n}$, and the responses to d drugs are given as a matrix $R_{m\times d}$. The purpose of learning here is to predict the response accurately when a new cell line is given for a known drug. Various types of genomic profiles of cell lines can be provided, among which the gene expression profile is the most frequently used one [10], owing to its representability of the cellular state and the amount of data released. Drug responses in $R_{m\times d}$ are often measured as the half maximal inhibitory concentration (IC_{50}). IC_{50} represents the amount of drug needed to inhibit a cell's activity by 50%. Another measure is the Area Under the concentration response Curve (AUC), where a lower AUC implies lesser effectiveness of the drug.

Reference [10] reviewed several computational methods for drug response prediction that utilized gene expression data including linear regression methods and their generalizations. The linear regression model learns a response function $r(x) = w \cdot x$ with the coefficient vector $w \in \mathbb{R}^n$ where $x \in \mathbb{R}^n$ is a genomic profile vector of a cell line. Kernelized Ridge Regression (KRR) maps x to a vector in a high-dimensional feature space with a kernel function and computes w in the image vector space [12]. Reference [13] recently developed a method called Kernelized Ranked Learning (KRL) that recommends the k most sensitive drugs for each cell line, rather than the response value itself [13]. Response-Weighted Elastic Net (RWEN) is based on the linear regression model, but it incorporates additional weights to find a coefficient vector w that results in a good prediction for cell lines with low drug responses [14]. As shown in KRR and KRL, feature engineering such as feature selection or extraction can help improve the prediction accuracy. Recent studies on feature engineering included kernel principal component analysis integrated with bootstrapping [15] and rough set theory [16].

When gene expression is used as the genomic profile of a cell line, information on the relationship among genes can be incorporated into the drug response prediction process. The relationships among biological entities such as genes or proteins are usually represented as a biological network. STRING [17] is one of the databases that provides a Protein-Protein Interaction (PPI) network that corresponds to the gene-gene relationship. As biological processes in a cell are operated by certain groups of genes with interactions like binding or regulation, we can assume that expression values of genes located close to each other in a network may affect the cellular state of a cell line together, thereby contributing to the drug response. However, the computational models for drug response prediction mentioned above do not incorporate the prior knowledge of the biological network, which may enhance the prediction accuracy.

Deep learning with neural networks has shown remarkable achievements compared to the traditional machine learning methods in the field of drug development such as drug-drug interaction extraction [18], drug target prediction [19,20], drug side effect detection [21], and drug discovery [22]. For drug response prediction, a number of methods have been developed as well, each of which utilizes different input data for prediction [11]. Multi-omics Late Integration (MOLI) [23] is a deep learning model that uses multi-omics data including gene expression, Copy Number Variation (CNV), and somatic mutations to characterize a cell line. Three separate subnetworks of MOLI learn representations for each type of omics data, and a final network uses concatenated features and classifies the response of a cell as a responder or non-responder. Reference [24] proposed a deep autoencoder model for representation learning of cancer cells from input data consisting of gene expression, CNV, and somatic mutations. The latent variables learned from the

deep autoencoder are used to train an elastic net or support vector machine to classify the response. Those methods share two characteristics in common: the integration of multiple input data (multi-omics) and binary classification of the drug response. Although the integration of multiple types of omics data can improve the learning of the status of the cell lines, it might limit the availability of the method for testing on different cell lines or patients as the model requires additional data other than gene expression. Furthermore, a certain threshold of the IC_{50} values should be set before binary classification of the drug response, which may vary depending on the experimental condition such as drug or tumor types.

The Convolutional Neural Network (CNN) is one of the neural network models adopted for drug response prediction [11]. The CNN has been actively used for image, video, text, and sound data due to its strong ability to preserve the local structure of data and learn hierarchies of features [25]. Twin Convolutional Neural Network for drugs in SMILES format (tCNNS) [26] takes a one-hot encoded representation of drugs and feature vectors of cell lines as the inputs for two encoding subnetworks of a One-Dimensional (1D) CNN. One-hot encodings of drugs in tCNNS are derived from Simplified Molecular Input Line Entry System (SMILES) strings that describe the chemical structure of a drug compound. Binary feature vectors of cell lines represent 735 mutation states or CNVs of a cell. Cancer Drug Response profile scan (CDRscan) [27] proposes an ensemble model composed of five CNNs, each of which predicts the IC_{50} values from the binary encoding of the genomic signature (mutation) and the drug signature (PaDEL-descriptors [28]). KekuleScope [29] adopts transfer learning, which uses a pre-trained CNN on ImageNet data. The pre-trained CNN is trained with images of drug compounds represented as Kekulé structures to predict the drug response. Recently, several algorithms have been proposed to extend CNNs for data on irregular or non-Euclidean domains represented as graphs [30–32]. Reference [33] proposed a method to predict drug response called GraphDRP, which integrates two subnetworks for drug and cell line features, similar to tCNNS [26]. Instead of one-hot encoding, GraphDRP uses a molecular graph to represent the drug structure converted from the SMILES string, and the Graph Convolutional Network (GCN) model from [32] is used to learn the features of drugs. Along with GraphDRP, there have been a number of approaches to use graphs to represent the structural properties of drug compounds for drug development and discovery [34].

Although the aforementioned CNN models incorporate a number of features in the input data, they do not include gene expression values in the genomic features because gene expression cannot be described as 1D binary sequences [26,27] or images [29] used in those CNN models. However, gene expression is known to be the most informative data type for drug response prediction [35,36], whereas mutation and the CNV profiles of cell lines added little to the performance in a comparative study [10]. Furthermore, most of the regression-based methods that utilize gene expression data in the prediction do not consider interactions between genes [12–14]. Recent studies successfully introduced a GCN model to use gene expression data for subtype classification of cancer [37,38], and a similar model can be transferred into the problem of drug response prediction. Thus, we propose an analysis framework, DrugGCN, for drug response prediction that can leverage gene expression data and network information using a GCN. DrugGCN constructs an undirected graph from a PPI network and maps gene expression values to each vertex (gene) as graph signals, which will be learned by the GCN to predict the drug response such as the IC_{50} or AUC. In addition, DrugGCN incorporates the feature selection process to use genes that can possibly improve the prediction accuracy using prior knowledge. The main contributions of DrugGCN are as follows:

- We propose a novel framework for drug response prediction that uses a GCN model learning genomic features of cell lines with a graph structure, which is the first approach to our knowledge.

- DrugGCN generates a gene graph suitable for drug response prediction by the integration of a PPI network and gene expression data and the feature selection process of genes with high predictive power.
- DrugGCN with localized filters can detect the local features in a biological network such as subnetworks of genes that contribute together to the drug response, and its learning complexity is suitable for biological networks with a huge number of vertices and edges.
- The performance of the proposed approach is demonstrated by a GDSC cell line dataset, and DrugGCN shows high prediction accuracy among the competing methods.

2. Materials and Methods

Figure 1 shows the proposed DrugGCN framework. DrugGCN introduces a two-step approach for drug response prediction as follows: (1) construction of an input graph with signals and (2) learning a graph convolutional network with localized filters. The following sections describe the methods and evaluation criteria for the comparative analysis in Section 3.

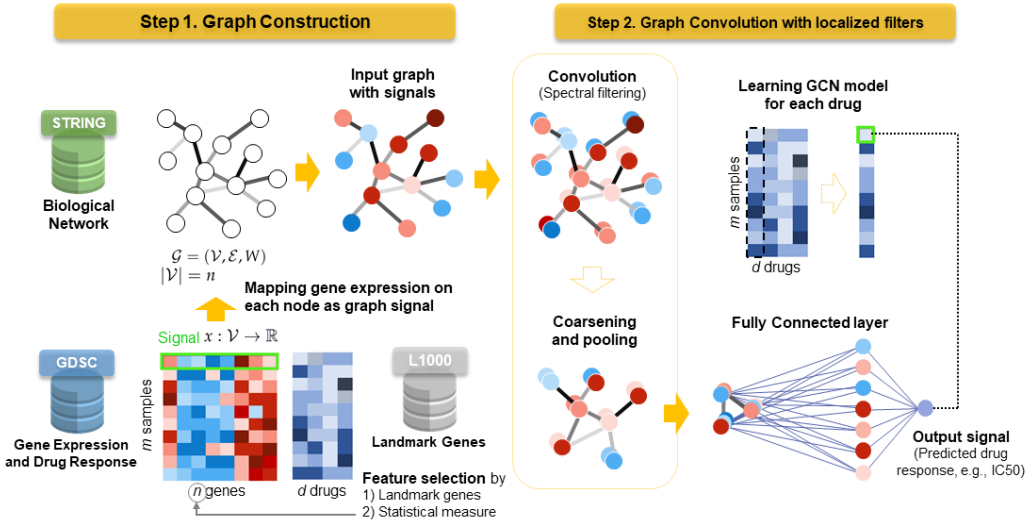

Figure 1. Drug Graph Convolutional Network (DrugGCN) framework for drug response prediction.

2.1. Graph Construction for Drug Response Prediction

DrugGCN utilizes three biological databases, GDSC [9], L1000 [39], and STRING [17]. GDSC dataset retrieved from [10] includes gene expression matrix $C_{734\times8046}$ of 8046 genes in 734 cell lines and drug response matrix $R_{734\times201}$ of 201 drugs for the same cell lines with two measures, common logarithms of the IC_{50} and AUC. Drugs with missing responses for 80% or more cell lines were removed from the initial 266 drugs.

2.1.1. Feature Selection on Gene Expression Data

To select effective features that potentially contribute to drug response, two criteria can be used in DrugGCN: prior knowledge on landmark genes or statistical measurement from the given data. A list of landmark genes is derived from Library of Integrated Network-Based Cellular Signatures (LINCS) L1000 project by the National Institutes of Health (NIH) [39]. In the LINCS L1000 project, nine-hundred seventy-eight landmark genes are selected as the assay targets of large-scale experiments on multiple time points, doses, perturbagens, and cell lines due to their representability of human genes. The landmark genes are known to be widely expressed across different cellular contexts and have good predictive power for inferring the expression of the other 22,268 human genes, based on the model built from Gene Expression Omnibus (GEO) data. After the feature selection on the GDSC dataset, six-hundred sixty-three common genes with landmark genes remained, leading to the final gene expression matrix $C_{734\times663}$. As another criterion

for feature selection, DrugGCN calculated the 1000 most variable genes from the GDSC dataset according to the variance of gene expression across cell lines. High variances of genes imply that the genes may be dependent on the heterogeneous cellular states and responsive to perturbation such as drug treatment. The criteria of feature selection can be selected by the user configuration in the DrugGCN framework.

2.1.2. Graph Construction with Gene Expression Data

The input of the GCN model is an undirected input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ where $\mathcal{V}$ is a set of vertices, $\mathcal{E}$ is a set of edges, and W is a weighted adjacency matrix. We assumed that the vertices and edges of the graph indicate genes and interactions between genes, respectively. To construct the gene graph, biological network information was retrieved from the STRING database [17], which provides the known and predicted PPI corresponding to gene-gene interactions. The original STRING database contains 11,759,454 interactions among 19,566 genes for Homo sapiens, reduced as 25,069 interactions with 663 landmark genes after feature selection. STRING also provides the weights of the interactions that reflect the amount of available evidence of the interaction between two genes, which is used to construct a weighted adjacency matrix. A signal $x : \mathcal{V} \to \mathbb{R}$ is defined for each gene in the graph as a row vector $x \in \mathbb{R}^{663}$ of $C_{734\times663}$ that represents the gene expression values in a certain cell line.

2.2. Localized Filters for the Graph Convolutional Network

To capture the localized patterns of the gene expression profile, convolution is performed on the input graph $\mathcal{G}$ with signal x. The graph Laplacian matrix L of $\mathcal{G}$ is defined as $L = D - A$ where D is the degree matrix and A is the adjacency matrix of the graph. With normalized Laplacian matrix $L = I_n - D^{-1/2}WD^{-1/2}$, L has a complete set of orthonormal eigenvectors $\{u_l\}_{l=0}^{n-1} \in \mathbb{R}^n$ and corresponding eigenvalues $\{\lambda_l\}_{l=0}^{n-1}$. The graph Laplacian can be diagonalized as $L = \mathrm{U}\Lambda\mathrm{U}^T$ on the Fourier basis $\mathrm{U} = [u_0, \ldots, u_{n-1}] \in \mathbb{R}^{n\times n}$ where $\Lambda = diag([\lambda_0, \ldots, \lambda_{n-1}]) \in \mathbb{R}^{n\times n}$. The graph Fourier transform of the signal $x \in \mathbb{R}^n$ is defined as $\hat{x} = \mathrm{U}^T x \in \mathbb{R}^n$, with its inverse as $x = \mathrm{U}\hat{x}$ [40].

Based on the Fourier inversion theorem, the convolution operator on graph $*_{\mathcal{G}}$ can be defined in the Fourier domain as $x *_{\mathcal{G}} y = \mathrm{U}((\mathrm{U}^T x) \odot (\mathrm{U}^T y))$ where $\odot$ is the element-wise Hadamard product. A signal x filtered by g_θ is:

$$y = g_\theta(L)x = g_\theta(\mathrm{U}\Lambda\mathrm{U}^T)x = \mathrm{U}g_\theta(\Lambda)\mathrm{U}^T x. \tag{1}$$

where $g_\theta(\Lambda)$ is a diagonal matrix $diag([g_\theta(\lambda_0), \ldots, g_\theta(\lambda_{n-1})])$. That is, only $g_\theta(\Lambda)$ is used to define convolution filters as U is determined by the input graph. Reference [31] suggested a polynomial filter $g_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k \Lambda^k$ [31] that is localized in K hops from the central vertex where the parameter $\theta \in \mathbb{R}^K$ is a vector of polynomial coefficients. To reduce the learning complexity of the polynomial filter from $\mathcal{O}(n^2)$ to $\mathcal{O}(K|\mathcal{E})$, a fast filtering by Chebyshev expansion [41] is introduced [31]. A filter is parameterized as the expansion:

$$g_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\Lambda}), \tag{2}$$

where the parameter $\theta \in \mathbb{R}^K$ is a vector of Chebyshev coefficients and $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_n$. $T_k(\tilde{\Lambda})$ is the Chebyshev polynomial of order k that can be computed by the recurrence relation $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0 = 1$ and $T_1 = x$. The filtering operation is learned by the backpropagation algorithm with a cost of $\mathcal{O}(K|\mathcal{E}|F_{in}F_{out}S)$ where F_{in} and F_{out} are the number of input and output convolution filters, respectively [31].

Coarsening of the graph is performed by the Graclus multilevel clustering algorithm, which is a greedy algorithm to compute coarser versions of a graph by picking a vertex i and its neighbor j that maximize the local normalized cut [42]. Graph signals are then

coarsened as well by pooling with a structure of a balanced binary tree as proposed in [31]. Among the pooling methods, max pooling is selected empirically in DrugGCN.

2.3. Evaluation Criteria

In the following section, the performance of DrugGCN and other algorithms is evaluated by four metrics: Root Mean Squared Error (RMSE), Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC), and Normalized Discounted Cumulative Gain (NDCG). The metrics are based on the difference between the observed drug responses $y = (y_i)_{i=1}^m$ (IC_{50} or AUC) and the predicted drug responses $\widehat{y} = (\widehat{y}_i)_{i=1}^m$ where m is the number of cell lines.

RMSE is defined as the square root of the mean squared error, which is the average squared difference between the true and predicted responses by a method:

$$RMSE(y, \widehat{y}) = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{m}}. \tag{3}$$

PCC attempts to measure if there is a linear correlation between two variables. The PCC of y and $\widehat{y}$ is defined as follows where $\mu(X)$ is the mean of a random variable X:

$$PCC(y, \widehat{y}) = \frac{\sum_i \left(\widehat{y}_i - \mu(\widehat{y})\right)\left(y_i - \mu(y)\right)}{\sqrt{\sum_i \left(\widehat{y}_i - \mu(\widehat{y})\right)^2}\sqrt{\sum_i \left(y_i - \mu(y)\right)^2}}. \tag{4}$$

The SCC between two variables is equivalent to the PCC between the rank values of the two variables. That is, the SCC assesses whether the two variables are monotonically related. Assume that $r(y)$ and $r(\widehat{y})$ are the rank vector of y and $\widehat{y}$, respectively. SCC is defined as:

$$SCC(y, \widehat{y}) = PCC(r(y), r(\widehat{y})). \tag{5}$$

The Discounted Cumulative Gain (DCG) [43] is a measure of ranking quality.

$$DGG(y, \widehat{y}) = \sum_{i=1}^{m} \frac{2^{-y_i} - 1}{log_2(r(-\widehat{y}_i) + 1)}. \tag{6}$$

To only consider the highest k scores in the ranking, m can be set as k. The NDCG is used to normalize the DCG by the ideal $DCG(y, y)$, as the DCG can be affected by the number of scores.

$$NDCG(y, \widehat{y}) = \frac{DCG(y, \widehat{y})}{DCG(y, y)}. \tag{7}$$

3. Results and Discussion

3.1. Performance Evaluation on the GDSC Dataset

To assess the prediction accuracy of the proposed framework, gene expression and drug response data from the GDSC database including 734 cell lines and 201 drugs were used. Pre-processed GDSC data were retrieved from the paper by Chen and Zhang [10] where genes with little variation in expression and drugs with missing responses in 80% or more cell lines were removed. We generated four sets of data with two different gene sets (L1000 and Var1000) and two types of drug response values (IC_{50} and AUC), referred to as the L1000-IC50, L1000-AUC, Var1000-IC50, and Var1000-AUC datasets. The L1000-IC50 and L1000-AUC datasets contained 663 genes that commonly exist in 978 landmark genes in L1000 and 8046 genes in the GDSC dataset. The Var1000-IC50 and Var1000-AUC datasets included the top 1000 genes with the largest variance of gene expression values among the entire cell lines, which might suggest that those genes can be expressed differently to generate diverse cellular statuses and consequent drug responses.

The computational models compared with DrugGCN were KRR, KRL [13], RWEN [14], a Multi-Layer Perceptron (MLP), and a Bagging Regressor (BR). The methods using gene expression data without considering the interaction between genes were selected to compare the prediction accuracy and to analyze the contribution of gene-gene interaction information utilized in DrugGCN. As for KRR, MLP, and BR, we used Python functions implemented in the `scikit-learn` library [44]. The hyperparameters of the methods were set as default values except the regularization parameter λ in KRL. λ was tuned with a line search with $\lambda \in 1.0, 0.1, 0.01, \ldots, 1 \times 10^{-6}$ as described in the original paper and finally selected as one. The hyperparameters of DrugGCN are described in Appendix A Table A1. In the learning process, a three-fold cross-validation with a 50% training set, a 25% validation set, and a 25% test set was carried out.

Figure 2 shows the result of the six methods with four measures, RMSE, PCC, SCC, and NDCG, using the L1000-IC50 dataset. k for NDCG was tested as $k \in 1, 2, 5, 10, 20, 50$. As KRL calculates the predicted ranking of drugs to recommend for a given cell line rather than the exact drug response values for all potential drugs, KRL was not included in the plots of the RMSE results. As in Figure 2, the top three methods showed consistently high performance in all measures in the order of DrugGCN, BR, and KRR. The ranks of the other three methods changed depending on the measures.

To compare the result of the six methods using a statistical test, the RMSE, PCC, and SCC values for each drug predicted by the six methods were ranked from one (the smallest for RMSE and the largest for PCC and SCC) to six. The rank vectors of each method were generated with the ranks that each method gained for 201 drugs. If a method always predicted the drug response most correctly among the six methods, the rank vector would be $(1, 1, \ldots, 1)$. On the contrary, the worst rank vector would be $(6, 6, \ldots, 6)$ for a method whose prediction was always the worst. We performed the one-sided Wilcoxon signed-rank test [45], which compares two matched measurements and assesses whether their means differ, comparing pairs of rank vectors of DrugGCN and the other five methods. The p-values estimated by the Wilcoxon signed-rank test are shown in Table 1. As expected, in Figure 2, the mean rank of DrugGCN is significantly higher (smaller in number) than the other methods.

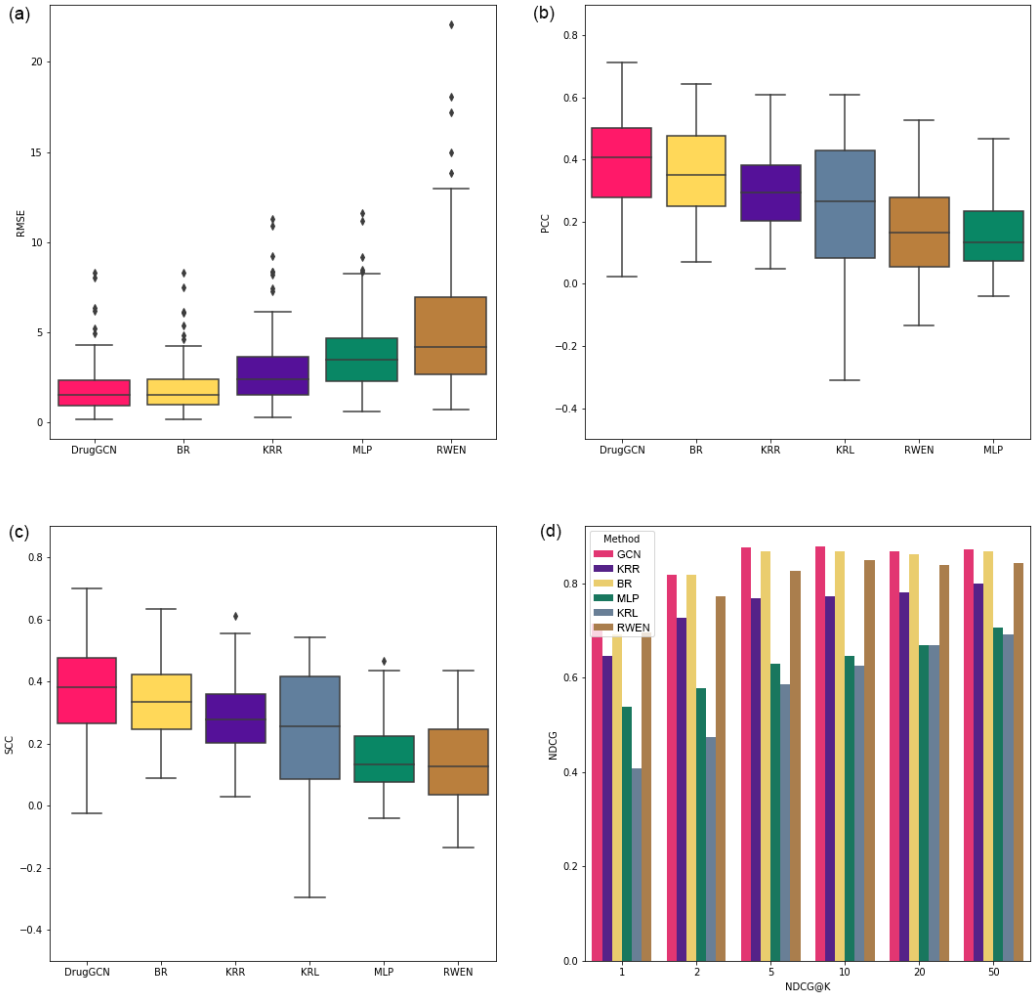

Figure 2. Performance evaluation results of the six methods from L1000-IC50 dataset using the (**a**) RMSE, (**b**) PCC, (**c**) SCC, and (**d**) Normalized Discounted Cumulative Gain (NDCG). KRR, Kernelized Ridge Regression; BR, Bagging Regressor; KRL, Kernelized Ranked Learning; RWEN, Response-Weighted Elastic Net.

Table 1. *p*-values of the one-sided Wilcoxon signed-rank test using the L1000-IC50 dataset comparing the ranks of DrugGCN and the other five methods.

Measure	BR	KRR	KRL	MLP	RWEN
RMSE	1.48×10^{-10}	9.44×10^{-38}	-	6.99×10^{-37}	2.77×10^{-37}
PCC	9.57×10^{-15}	5.03×10^{-28}	2.93×10^{-31}	1.38×10^{-35}	3.89×10^{-35}
SCC	5.77×10^{-12}	1.07×10^{-25}	1.31×10^{-28}	2.11×10^{-35}	9.28×10^{-35}

For the remaining three datasets, DrugGCN showed consistently high performance as shown in Appendix A Figures A1–A3 and Tables A2–A4. In most tests of the Wilcoxon signed-rank test from the RMSE, PCC, and SCC measures, DrugGCN predicted better than the other methods except two tests with KRL and BR, where the mean ranks of DrugGCN, KRL, and BR were nearly the same as in Appendix A Figure A1c and Figure A2a. In the NDCG results of the three other datasets, BR showed the highest accuracy.

In summary, DrugGCN and BR were the first and second methods that showed good prediction power for the GDSC datasets. KRR-KRL and RWEN-MLP showed similar results in the 3rd-4th and 5th-6th places. The high performance of DrugGCN and BR presumably stemmed from the power of network information and ensemble learning, respectively. From this result, we can suggest an ensemble model integrated with GCN, for which similar

approaches have been made for other biological problems such as subtype classification of cancer [37,38]. Lee et al. [38] constructed an ensemble model from multiple GCNs of smaller subnetworks of PPI (biological pathways) with hundreds of genes, known to regulate a biological process together. In addition, the GCN model can be extended with other types of networks rather than PPI such as molecular graphs of drugs [33] and biological networks from multi-omics data [23], which can provide additional information for drug response prediction. As in [23,33], different types of input features can be integrated into a multi-modal neural network, and the contribution of each drug or genomic features can be measured and compared with the augmented model structure.

3.2. Case Study: ERK MAPK Signaling-Related Drugs

We investigated the L1000-IC50 dataset to uncover how the network of genes affected the accuracy of drug response prediction. First, drugs were classified into 23 drug groups depending on their target pathways, as defined by [10]. For each accuracy measure, two-hundred one drugs can be ranked from one (a drug predicted with the smallest RMSE or largest PCC or SCC) to 201. The ranks of drugs in the same drug group were collected, and a box plot showing the distribution of the ranks of each drug group was generated, as shown in Figures 3 and 4. Figure 3 shows the distribution of ranks when ranks from all measures were put together, and Figure 4 is the individual box plots for the RMSE, PCC, and SCC. The overall prediction accuracy was high in a group of drugs targeting the ERK MAPK signaling pathway.

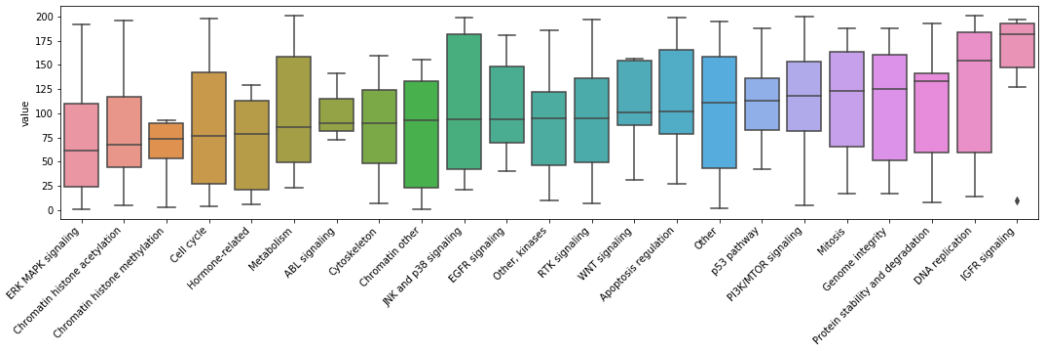

Figure 3. Box plot of ranks for each drug group predicted by DrugGCN using the L1000-IC50 dataset. Results from the RMSE, PCC, and SCC are combined.

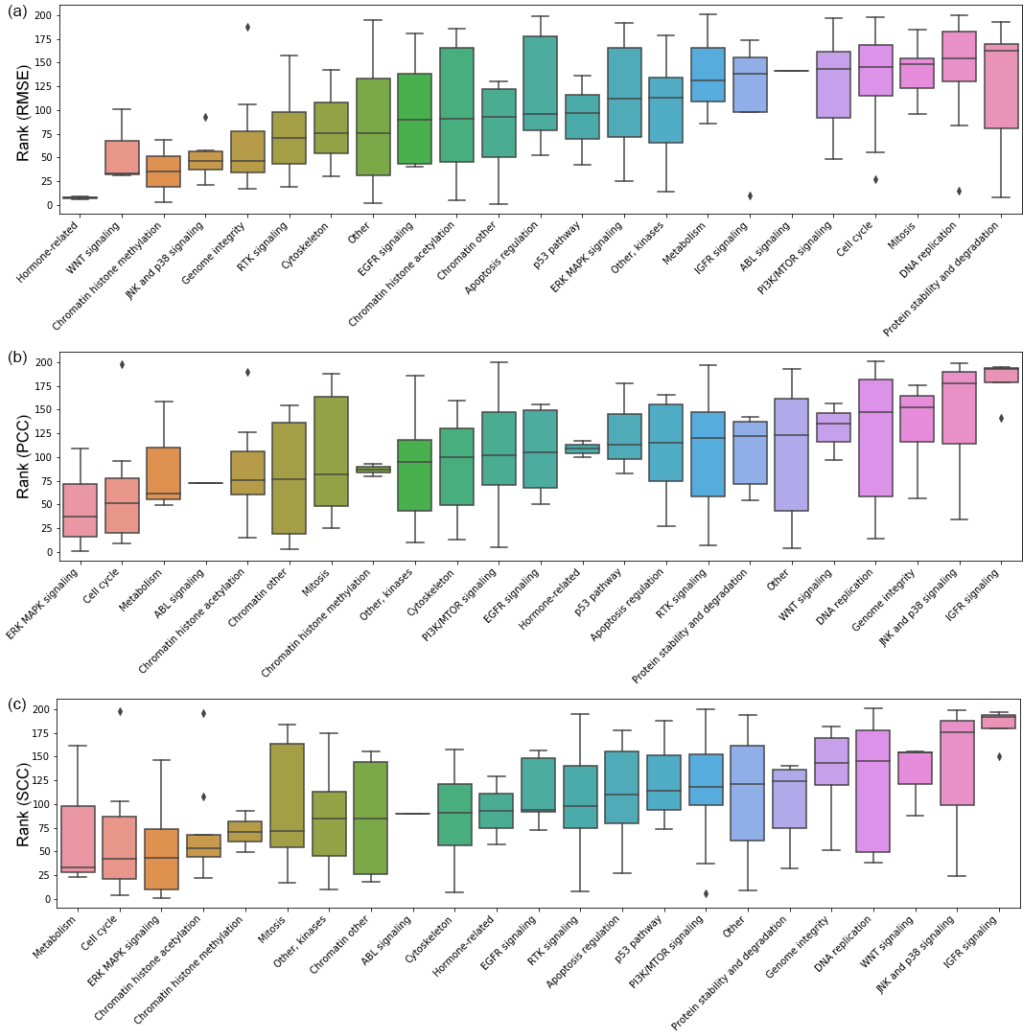

Figure 4. Box plot of ranks for each drug group predicted by DrugGCN using the L1000-IC50 dataset. Ranks of drug groups are calculated from the (**a**) RMSE, (**b**) PCC, and (**c**) SCC.

We assumed that high prediction accuracy of a drug response was derived by the information from the highly connected genes forming a subnetwork in the PPI network. On the contrary, lower prediction accuracy may result from the important genes scattered in the network, even though those drugs with high and low prediction accuracy have similar roles in a cell. To verify the hypothesis, we selected two drugs from the "ERK MAPK signaling pathway" group with the highest rank (refametinib, rank one) and the lowest rank (PLX-4720, rank 109) in terms of the PCC. For refametinib and PLX-4720, fifty genes that potentially contributed to predict the IC50 were selected by the correlation between their gene expression values and the IC50. That is, the gene expression of 50 genes had a relatively strong (positive or negative) correlation with the IC50 values across cell lines, among 663 genes. With those two sets of 50 genes, we performed a pathway enrichment analysis that assessed the ratio of query genes found together in the subnetworks of the PPI network called biological pathways. The DAVID platform [46] provided over-representation analysis, one of the pathway enrichment analyses [47] that conducts statistical tests on query genes based on the hypergeometric distribution.

The pathway analysis results of refametinib and PLX-4720-related genes are shown in Tables 2 and 3, respectively. Fifty genes having a high correlation with refametinib were enriched (p-value under 0.05) in a number of biological pathways including MAPK signaling itself; even four pathways remained as significantly enriched after FDR correction of the p-values. On the other hand, genes selected for PLX-4720 had three or zero enriched pathways

with *p*-values or FDR corrected *p*-values under 0.05. The results from pathway enrichment analysis supported our hypothesis that the gene network contributes to the prediction of drug response.

Table 2. Pathway enrichment test results with 50 genes for refametinib. QueryG indicates the number of query genes found in the pathway, and PathwayG is the number of total genes in the pathway.

Pathway	QueryG	PathwayG	%	*p*-Value	FDR
hsa05212:Pancreatic cancer	6	65	12.2449	7.65×10^{-6}	9.11×10^{-4}
hsa05210:Colorectal cancer	5	62	10.20408	1.38×10^{-4}	0.008185
hsa05220:Chronic myeloid leukemia	5	72	10.20408	2.46×10^{-4}	0.009764
hsa05200:Pathways in cancer	8	393	16.32653	0.001218	0.036235
hsa04380:Osteoclast differentiation	5	131	10.20408	0.002345	0.054046
hsa05214:Glioma	4	65	8.163265	0.002725	0.054046
hsa05166:HTLV-I infection	6	254	12.2449	0.004409	0.074948
hsa05205:Proteoglycans in cancer	5	200	10.20408	0.010509	0.156321
hsa04068:FoxO signaling pathway	4	134	8.163265	0.019992	0.26434
hsa04010:MAPK signaling pathway	5	253	10.20408	0.023081	0.264767
hsa05223:Non-small cell lung cancer	3	56	6.122449	0.024474	0.264767
hsa05206:MicroRNAs in cancer	5	286	10.20408	0.034222	0.32247
hsa04917:Prolactin signaling pathway	3	71	6.122449	0.037938	0.32247
hsa05218:Melanoma	3	71	6.122449	0.037938	0.32247
hsa04062:Chemokine signaling pathway	4	186	8.163265	0.046265	0.367033

Table 3. Pathway enrichment test results with 50 genes for PLX-4720. QueryG indicates the number of query genes found in the pathway, and PathwayG is the number of total genes in the pathway.

Pathway	QueryG	PathwayG	%	*p*-Value	FDR
hsa05010:Alzheimer's disease	4	168	8	0.027233	1
hsa05120:Epithelial cell signaling in Helicobacter pylori infection	3	67	6	0.028047	1
hsa04912:GnRH signaling pathway	3	91	6	0.049052	1

4. Conclusions

In this study, DrugGCN, a computational framework for drug response prediction, was proposed. DrugGCN incorporated PPI network and gene expression data into the GCN model to detect the local features in graphs by localized filtering. The effectiveness of DrugGCN was tested with four GDSC datasets, and it showed high prediction accuracy in terms of the RSME, PCC, and SCC. In the case study of ERK MAPK signaling-related drugs, we discovered supporting evidence of the hypothesis that the high accuracy of DrugGCN was due to the genes forming a subnetwork in the PPI network that provided much information to predict cellular states and consequent drug responses.

The prediction accuracy of DrugGCN can be further improved in terms of the current limitations pertaining to the model structure and input features as described below. Among the competing methods of DrugGCN, the bagging regressor showed high performance with the support of the ensemble model. An ensemble model consisting of multiple GCN was proposed in [38] for the cancer subtype classification problem where the original graph of the biological network was divided into smaller subnetworks with hundreds of genes using prior knowledge on which genes cooperated with each other for a certain biological process, such as biological pathways from the KEGG database [48]. The prediction accuracy of the DrugGCN model can be improved with the aforesaid ensemble model, as we showed the predictive power of subnetworks in the case study of ERK MAPK signaling.

The DrugGCN model also can be extended with additional genomic features from different omics data or drug features such as the chemical structures of drug compounds. In particular, the structural properties of drug compounds have been used for drug development and discovery as the form of a graph [34], which can be easily integrated into the

GCN model. As in the similar models [23,26], learning from multiple types of features can be implemented with multiple GCN models, the learned representations of which are then concatenated and put into fully connected layers.

Author Contributions: Conceptualization, K.J.; methodology, K.J.; software, S.K. and Y.P.; validation, S.K., S.B. and K.J.; formal analysis, K.J.; investigation, S.K. and S.B.; data curation, S.B. and S.K.; writing—original draft preparation, K.J. and S.B.; writing—review and editing, K.J. and Y.P.; visualization, S.B.; supervision, K.J.; project administration, K.J.; funding acquisition, K.J. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1G1A1003558).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: GDSC data used in the paper are retrieved from https://github.com/Jinyu2019/Suppl-data-BBpaper (accessed on 30 October 2020) [10]. Python codes of DrugGCN framework are available at https://github.com/BML-cbnu/DrugGCN.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Hyperparameters of DrugGCN used for the GDSC dataset.

Hyperparameters
GCN Layer = 2
Number of Kernels = [20, 10]
Kernel Size = [40, 20]
Polling size = [4, 4]
FC Size = [128, 1]
Reg = 0
Dropout = 1
l_rate = 0.02
Momentum = 0.9
Decay_rate = 0.95
Batch_size = 4

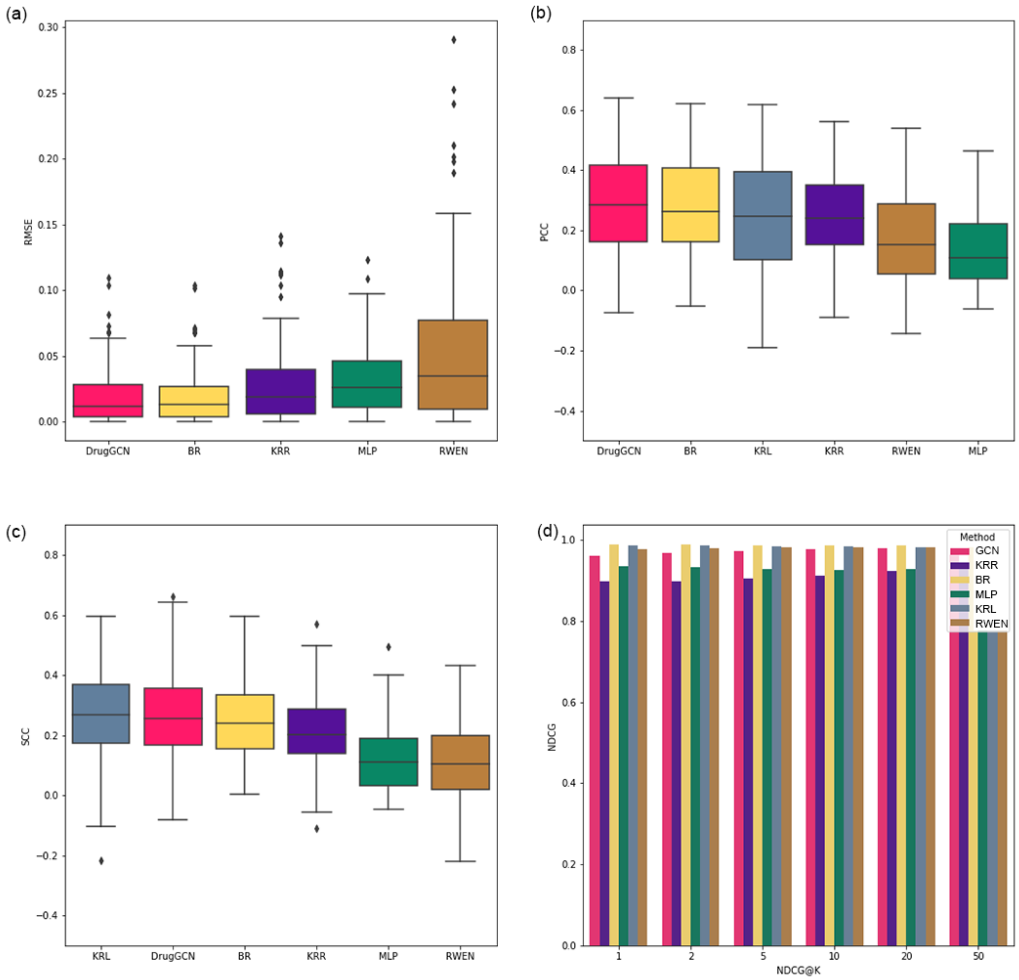

Figure A1. Performance evaluation results of six methods from the L1000-AUC dataset using the (**a**) RMSE, (**b**) PCC, (**c**) SCC, and (**d**) NDCG.

Table A2. *p*-values of the one-sided Wilcoxon signed-rank test using the L1000-AUC dataset comparing the ranks of DrugGCN and the other five methods.

Measure	BR	KRR	KRL	MLP	RWEN
RMSE	0.0038	9.91×10^{-35}	-	4.55×10^{-46}	1.99×10^{-36}
PCC	0.0082	1.43×10^{-8}	4.09×10^{-9}	1.01×10^{-32}	3.73×10^{-23}
SCC	0.0025	1.71×10^{-8}	0.5151	2.75×10^{-30}	3.28×10^{-26}

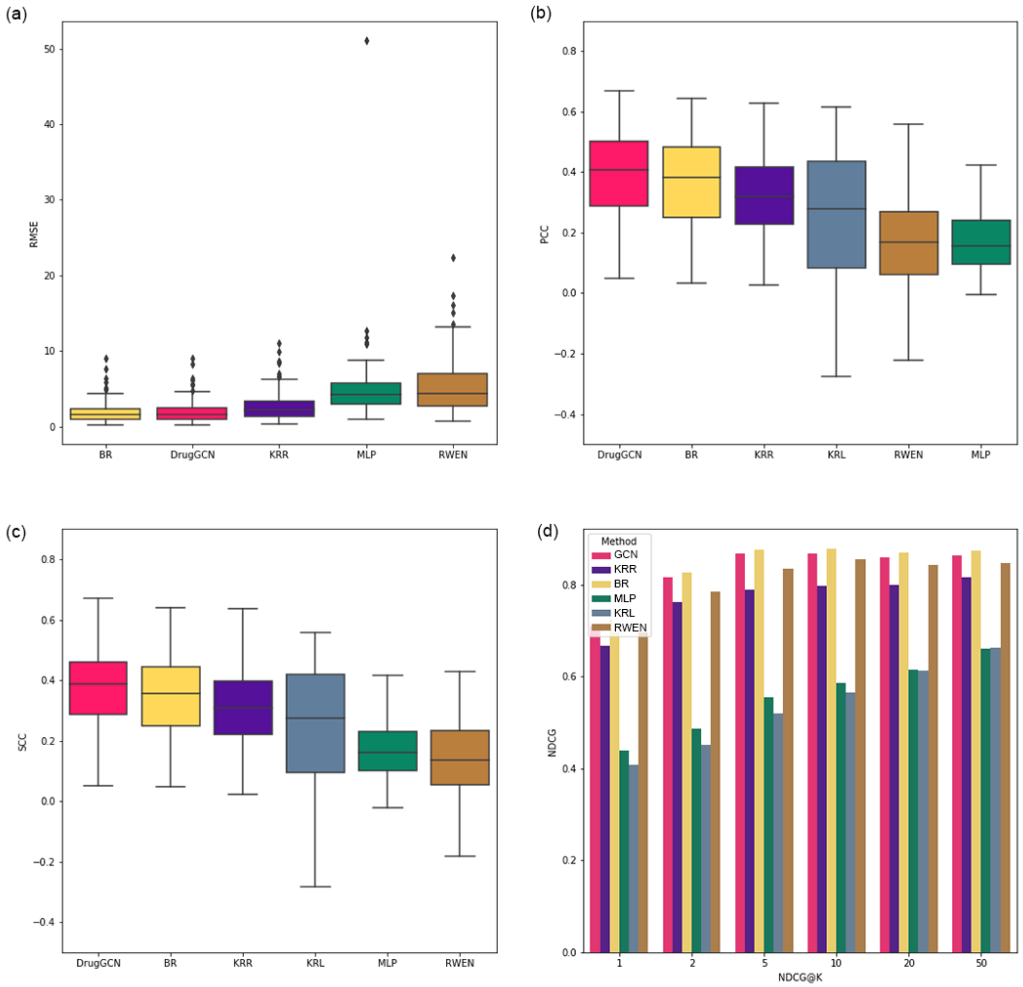

Figure A2. Performance evaluation results of the six methods from the Var1000-IC50 dataset using the (**a**) RMSE, (**b**) PCC, (**c**) SCC, and (**d**) NDCG.

Table A3. *p*-values of the one-sided Wilcoxon signed-rank test using the Var1000-IC50 dataset comparing the ranks of DrugGCN and the other five methods.

Measure	BR	KRR	KRL	MLP	RWEN
RMSE	0.9914	4.87×10^{-37}	-	4.59×10^{-36}	2.34×10^{-37}
PCC	3.66×10^{-14}	2.39×10^{-29}	4.10×10^{-26}	1.31×10^{-35}	1.48×10^{-35}
SCC	6.01×10^{-13}	1.18×10^{-27}	1.03×10^{-25}	1.31×10^{-35}	9.06×10^{-36}

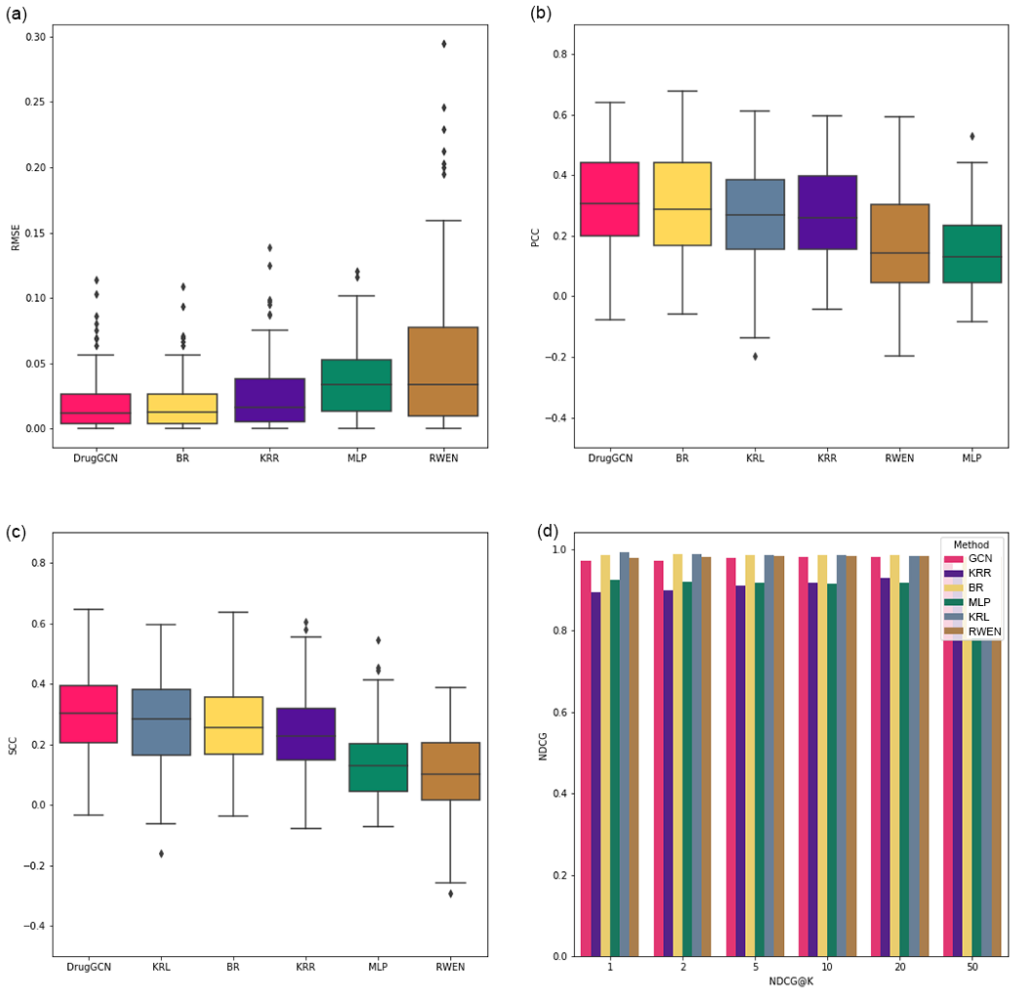

Figure A3. Performance evaluation results of the six methods from the Var1000-AUC dataset using the (**a**) RMSE, (**b**) PCC, (**c**) SCC, and (**d**) NDCG.

Table A4. *p*-values of the one-sided Wilcoxon signed-rank test using the Var1000-AUC dataset comparing the ranks of DrugGCN and the other five methods.

Measure	BR	KRR	KRL	MLP	RWEN
RMSE	0.0957	1.12×10^{-32}	-	5.76×10^{-36}	1.00×10^{-36}
PCC	0.0001	3.56×10^{-11}	1.08×10^{-10}	2.93×10^{-33}	9.47×10^{-27}
SCC	4.52×10^{-6}	8.06×10^{-16}	0.0002	5.81×10^{-35}	4.88×10^{-33}

References

1. Liang, P.; Pardee, A.B. Analysing differential gene expression in cancer. *Nat. Rev. Cancer* **2003**, *3*, 869–876. [CrossRef] [PubMed]
2. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
3. Bao, R.; Connolly, D.C.; Murphy, M.; Green, J.; Weinstein, J.K.; Pisarcik, D.A.; Hamilton, T.C. Activation of cancer-specific gene expression by the survivin promoter. *J. Natl. Cancer Inst.* **2002**, *94*, 522–528. [CrossRef] [PubMed]
4. Hutchinson, L.; DeVita, V.T. The era of personalized medicine: Back to basics. *Nat. Clin. Pract. Oncol.* **2008**, *5*, 623. [CrossRef]
5. Castiblanco, J.; Anaya, J.M. Genetics and vaccines in the era of personalized medicine. *Curr. Genom.* **2015**, *16*, 47–59. [CrossRef] [PubMed]
6. Marquart, J.; Chen, E.Y.; Prasad, V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. *JAMA Oncol.* **2018**, *4*, 1093–1098. [CrossRef]
7. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]

8. Zhang, J.; Baran, J.; Cros, A.; Guberman, J.M.; Haider, S.; Hsu, J.; Liang, Y.; Rivkin, E.; Wang, J.; Whitty, B.; et al. International Cancer Genome Consortium Data Portal—A one-stop shop for cancer genomics data. *Database* **2011**, *2011*, bar026. [CrossRef]
9. Garnett, M.J.; Edelman, E.J.; Heidorn, S.J.; Greenman, C.D.; Dastur, A.; Lau, K.W.; Greninger, P.; Thompson, I.R.; Luo, X.; Soares, J.; et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **2012**, *483*, 570–575. [CrossRef]
10. Chen, J.; Zhang, L. A survey and systematic assessment of computational methods for drug response prediction. *Briefings Bioinform.* **2021**, *22*, 232–246. [CrossRef]
11. Baptista, D.; Ferreira, P.G.; Rocha, M. Deep learning for drug response prediction in cancer. *Briefings Bioinform.* **2021**, *22*, 360–379. [CrossRef] [PubMed]
12. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
13. He, X.; Folkman, L.; Borgwardt, K. Kernelized rank learning for personalized drug recommendation. *Bioinformatics* **2018**, *34*, 2808–2816. [CrossRef] [PubMed]
14. Basu, A.; Mitra, R.; Liu, H.; Schreiber, S.L.; Clemons, P.A. RWEN: Response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics* **2018**, *34*, 3332–3339. [CrossRef] [PubMed]
15. Sharma, H.K.; Kumari, K.; Kar, S. A rough set approach for forecasting models. *Decis. Mak. Appl. Manag. Eng.* **2020**, *3*, 1–21. [CrossRef]
16. Ghosh, I.; Chaudhuri, T.D. FEB-Stacking and FEB-DNN Models for Stock Trend Prediction: A Performance Analysis for Pre and Post Covid-19 Periods. *Decis. Mak. Appl. Manag. Eng.* **2021**, *4*, 51–84. [CrossRef]
17. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef]
18. Zhang, T.; Leng, J.; Liu, Y. Deep learning for drug–drug interaction extraction from the literature: A review. *Briefings Bioinform.* **2020**, *21*, 1609–1627. [CrossRef]
19. Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: Interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329–3338. [CrossRef]
20. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829. [CrossRef]
21. Lee, C.Y.; Chen, Y.P.P. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Briefings Bioinform.* **2020**, *22*, 1884–1901. [CrossRef] [PubMed]
22. Rifaioglu, A.S.; Atas, H.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. *Briefings Bioinform.* **2019**, *20*, 1878–1912. [CrossRef] [PubMed]
23. Sharifi-Noghabi, H.; Zolotareva, O.; Collins, C.C.; Ester, M. MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **2019**, *35*, i501–i509. [CrossRef] [PubMed]
24. Ding, M.Q.; Chen, L.; Cooper, G.F.; Young, J.D.; Lu, X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol. Cancer Res.* **2018**, *16*, 269–278. [CrossRef]
25. Aloysius, N.; Geetha, M. A review on deep convolutional neural networks. In Proceedings of the 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 6–8 April 2017; pp. 588–592.
26. Liu, P.; Li, H.; Li, S.; Leung, K.S. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinform.* **2019**, *20*, 408. [CrossRef] [PubMed]
27. Chang, Y.; Park, H.; Yang, H.J.; Lee, S.; Lee, K.Y.; Kim, T.S.; Jung, J.; Shin, J.M. Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.* **2018**, *8*, 1–11. [CrossRef]
28. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
29. Cortés-Ciriano, I.; Bender, A. KekuleScope: Prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. ChemInform.* **2019**, *11*, 1–16. [CrossRef]
30. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning convolutional neural networks for graphs. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2014–2023.
31. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3844–3852.
32. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
33. Nguyen, T.T.; Nguyen, G.T.T.; Nguyen, T.; Le, D.H. Graph convolutional networks for drug response prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**. [CrossRef]
34. Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph convolutional networks for computational drug development and discovery. *Briefings Bioinform.* **2020**, *21*, 919–935. [CrossRef] [PubMed]
35. Costello, J.C.; Heiser, L.M.; Georgii, E.; Gönen, M.; Menden, M.P.; Wang, N.J.; Bansal, M.; Hintsanen, P.; Khan, S.A.; Mpindi, J.P.; et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **2014**, *32*, 1202–1212. [CrossRef]

36. Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **2016**, *166*, 740–754. [CrossRef]
37. Rhee, S.; Seo, S.; Kim, S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 3527–3534.
38. Lee, S.; Lim, S.; Lee, T.; Sung, I.; Kim, S. Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics* **2020**, *36*, 3818–3824. [CrossRef]
39. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **2017**, *171*, 1437–1452. [CrossRef] [PubMed]
40. Shuman, D.I.; Narang, S.K.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83–98. [CrossRef]
41. Hammond, D.K.; Vandergheynst, P.; Gribonval, R. Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* **2011**, *30*, 129–150. [CrossRef]
42. Dhillon, I.S.; Guan, Y.; Kulis, B. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1944–1957. [CrossRef]
43. Järvelin, K.; Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst. (TOIS)* **2002**, *20*, 422–446. [CrossRef]
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
45. Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*; Springer: Berlin, Germany, 1992; pp. 196–202.
46. Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44.
47. Khatri, P.; Sirota, M.; Butte, A.J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **2012**, *8*, e1002375. [CrossRef] [PubMed]
48. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]