이화여자대학교
EWHA WOMANS UNIVERSITY

Drug Response Prediction via Graph-based Deep Learning

박서연 Soyon Park [1]

[1] Department of Molecular Life & Chemical Sciences, Ewha Womans University, Seoul 03760, Korea; soyon0304@ewhain.net

Abstract: Accurate drug response prediction is vital for personalized medicine. This study compares two approaches for drug representation: fingerprint (using a 1D CNN encoder) and molecular graph (using GCN, GIN models). Cell line data is represented as gene expression data

and encoded using a 1D CNN. Results show the effectiveness of both approaches in predicting drug responses, with slight variations in accuracy. Future work involves exploring network structures, optimizing hyperparameters, pretraining with external databases, incorporating multiomics data, and leveraging domain knowledge to enhance prediction models. These advancements will improve drug response prediction for personalized medicine studies.

Keywords: Transcriptomics; Drug Response; Fingerprint; SMILES; Deep Learning; Graph; Graph Convolution Network; Graph Attention Network; Graph Isomorphism Network

1. Introduction

Cancer is a leading cause of death nationally and worldwide in the 21st century. Cancer Research UK reported 17 million new cases of cancer and 9.6 million deaths from cancer worldwide in 2018 [1]. In addition to high mortality rate, debilitating symptoms and potential toxicity associated with cancer chemotherapies contribute to the profound apprehension surrounding this disease. Traditionally, generalized approach was used to treat cancer, where patients with similar types of cancer receive standardized treatments. However, recent research has revealed numerous molecular lesions that underlie the development of various cancers, showing distinct genomic signatures specific to each tumor type. While the accumulation of genomic damage over an individual's lifespan contributes to the cause of cancer, hereditary genetic variations can also impact cancer susceptibility. Therefore, recent studies of oncogenic mechanisms have started to shape risk assessment, diagnostic classifications, and therapeutic approaches, with an increasing focus to design drugs and antibodies counteract the effects of specific molecular drivers. A substantial number of targeted therapies have been developed and continue to undergo investigation [2].

Precision medicine, also known as personalized medicine, aims to tailor diagnostics and therapeutics to individual patients based on their unique genetic, biomarker, phenotypic, or psychosocial characteristics. One critical aspect of precision medicine is the need for accurate drug response prediction. Estimating how an individual will respond to a particular drug is a challenging task due to the complex and diverse nature of cancer. By incorporating drug response prediction into precision medicine, clinicians can utilize advanced computational methods to analyze diverse biological data from patients including genomics, transcriptomics, epigenetics, proteomics, metabolomics, and clinical information [3-9].

Recently, deep learning, a powerful subset of machine learning, has emerged as a prominent approach in predicting drug responses with high accuracy and precision. The accumulation of extensive chemical and biological data over several decades, coupled with advancements in computational power such as massively parallel computing architectures and the utilization of graphical processing units (GPUs), has led to the integration of artificial intelligence (AI) in the field of drug development [10]. Deep learning leverages neural network architectures to learn intricate

patterns and relationships within large datasets. These methods use non-linear models which enable to discover multiple levels of increasingly complex representations from raw data [11]. In the context of drug response prediction, deep learning models can analyze extensive biological and clinical data to identify meaningful features associated with drug sensitivity or resistance. This enables clinicians to make more informed decisions regarding treatment selection, dosage optimization, and combination therapies for individual patients.

The majority of drug response prediction via deep learning workflows typically involve general steps (Figure 1).

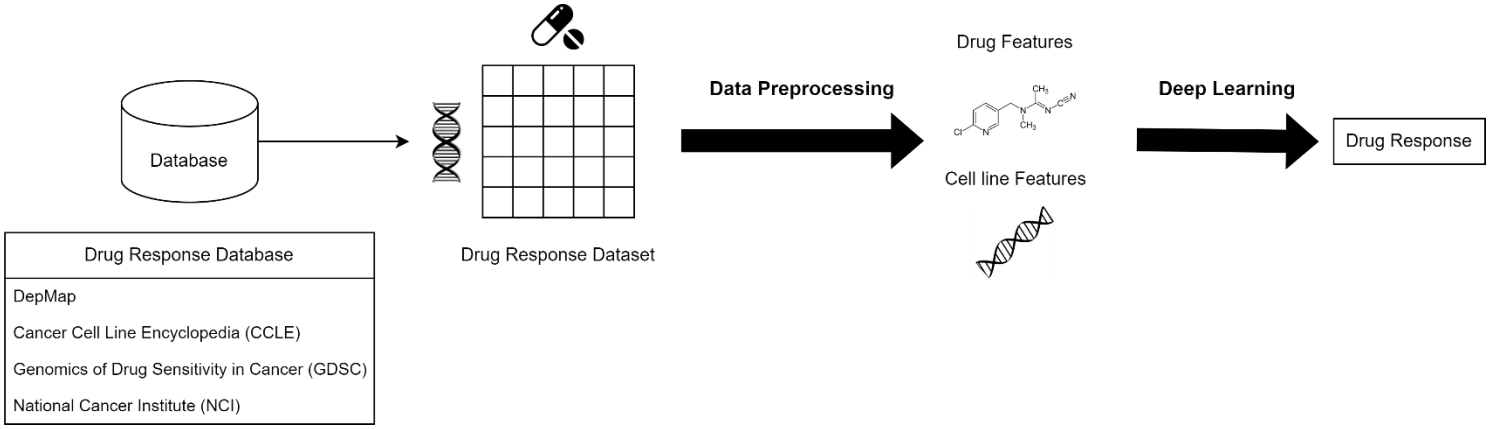

Figure 1. General steps for drug response prediction via deep learning.

First, a drug screening dataset for cancer cell line is collected from a drug response database such as DepMap [12], CCLE [13], GDSC [14], NCI-60 [15], or more. These large-scale drug screening projects include cell line omics, drug, and cell line-drug response data. After selecting the database, input data types are defined. For cell line omics data, the drug screening database provides genomic data such as mutation and CNV, transcriptomic data such as mRNA expression, and epigenomic data including methylation (Figure 2). For drug data, the given drug name and Compound ID number (CID) can be used to select a representation for the drug. Drugs are characterized by several types of molecular representations from simple sequences of molecular entities to molecular graphs [16, 17] (Figure 3). Drug response data shows sensitivity profiles corresponding to various cancer cell lines such as dose-response curve (AUC) and half maximal inhibitory concentration (IC50). AUC represents the Area Under the concentration response Curve, where the AUC value indicates the effectiveness of the drug. IC50 refers to the quantity of a drug required to inhibit cellular activity by 50% [18]. The next step involves preprocessing and feature engineering, where relevant features are extracted from the input data. Once the data is prepared, deep learning models are defined based on the selected data types and representations and trained using algorithms such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). The trained models are then validated using independent datasets to assess their performance and generalizability. When evaluating the model performance, appropriate scoring metrics should be selected. Finally, the validated models can be deployed to predict drug responses for new data [19].

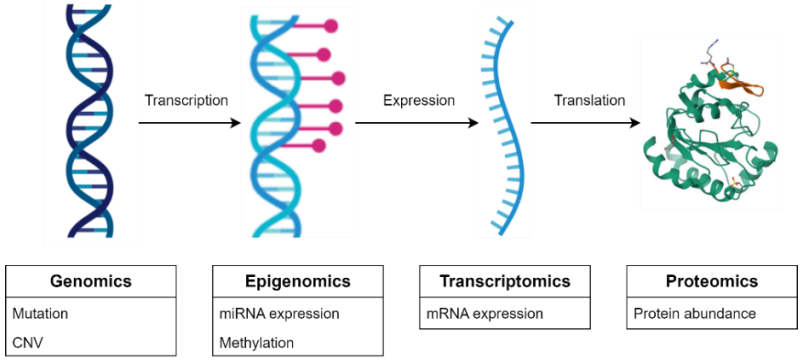

Figure 2. Types of cell line omics data.

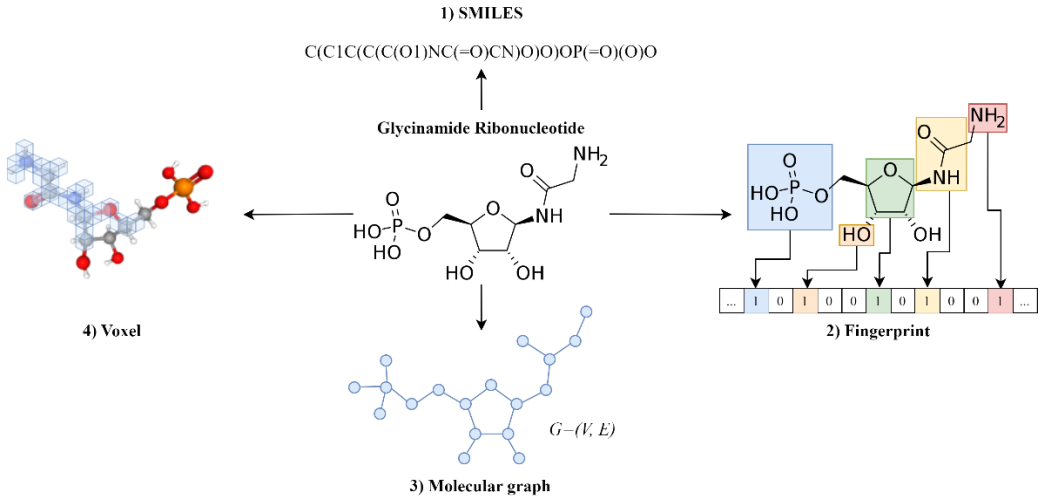

Figure 3. Types of drug representations. For drug response prediction via deep learning, drug can be characterized by different types of representations. In this figure, Glycinamide Ribonucleotide is depicted using four commonly employed drug representations. 1) SMILES: Simplified Molecular-Input Line-Entry System (SMILES) string. A linear notation which encodes connectivity of atoms within a molecule. 2) Fingerprint: a binary or numeric vector that encodes the presence or absence of specific chemical substructures or molecular properties. 3) Molecular graph: drug molecule described as a graph, where atoms are represented as nodes and chemical bonds as edges. 4) Voxel: Three-dimensional space surrounding a drug molecule discretized into a grid of small cubic cells called voxels.

While previous studies in drug response prediction through deep learning have followed the general workflow mentioned above, there is a notable variation in the specific models employed depending on the input data types. For instance, DeepDR [20] and MOLI [21] used separate feature-encoding deep neural networks for each omics data type to predict the IC50 values. These models included various omics data of the patient to predict the drug response but lacked information specific to the drug such as the molecular formula or structure. Therefore, studies such as tCNNS [22] included the molecular formula of the drug as SMILES string and used 1D CNN to encode the features for both the drug and genomic data. To include detailed molecular features such as structural and pharmacological feature of a ligand, drugs were represented as fingerprints in DeepDSC [23] and CDRScan [24]. DeepDSC model used a pretrained stacked autoencoder to encode gene expression data of cancer cell lines and merged the cell line features with molecular fingerprints. The combined features can predict the drug response after training through the deep neural network. Cancer Drug Response profile scan (CDRScan) proposed an ensemble of five convolutional neural network (CNN)-based models, each of which predicting the IC50 value from the genomic signature and the drug signature. These drug representations used for deep learning models lost the structural information of drugs, therefore studies started to represent drugs as molecular graphs. Graph convolutional network (GCN) enabled to learn representations of compound structures from molecular graphs. GraphDRP [25] used a molecular graph converted from the SMILES notation to represent the drug structure and adopted GCN to encode the drug features. Moreover, DRPreter [26] learns both cell line and drug representation with graph neural networks, where cell line graph is a combination of multiple subgraphs with domain knowledge on biological pathways. [16, 18, 19]

In this study, we investigated the application of gene expression data for drug response prediction using deep learning techniques. Specifically, we focused on comparing the effectiveness of different drug representations, namely fingerprints and molecular graphs. We aimed to evaluate how well each representation captures the structural information of drugs and their impact on prediction performance. We explored the utilization of various graph neural network architectures when employing molecular graphs as the representation for drugs. By comparing the performance of different models and representations, we aimed to gain insights into the optimal approach for drug response prediction using gene expression data and deep learning.

2. Materials and Methods

2.1. Data

This section provides the details about the data used in the experiment.

2.1.1. Dataset

Dependency Map (DepMap) portal is a research community which provides open access to key cancer dependencies analytical and visualization tools [12]. DepMap provides the gene expression data of cancer cell lines from the CCLE database, a Cancer Cell Line Encyclopedia which gives access to genomic data and computational analyses for human cancer cell lines. For drug response data, Genomics of Drug Sensitivity in Cancer (GDSC) provides drug sensitivity data as IC50 value for each pair of drug-cell line.

The downloaded files from DepMap and GDSC are:

1. Expression_Public_22Q4.csv, which is a list of gene expression values of 1,408 cancer cell lines for 16,383 genes. Values are inferred from RNA-seq data and log2 transformed ($log_2(TPM + 1)$).
2. GDSC1_fitted_dose_response_24Jul22.csv, which is a GDSC1 IC50 drug screening dataset of 970 cancer cell lines and 378 drugs. The dataset provides a total of 333,161 cell line-drug response pairs.
3. GDSC2_fitted_dose_response_24Jul22.csv, which is a GDSC2 IC50 drug screening dataset of 969 cancer cell lines and 286 drugs. The dataset provides a total of 242,036 cell line-drug response pairs.

2.1.2. Gene Expression Data

Cancer cell lines are characterized by gene expression data, which gives the RNA-seq gene expression value of various genes for each cell line (Table 1). In this study, we implemented gene expression as the primary data type for characterizing the features of the cell line, considering its significance in predicting drug response [27, 28].

Table 1. DepMap database gene expression data.

		Genes			
Cell lines	TSPAN6	TNMD	DPM1	…	MAGEA6
ACH-001113	4.331992	0	7.36466	…	0.028569
…	…	…	…	…	…
ACH-000052	4.249445	0	6.175724	…	0.137504

2.1.3. Drug Representation

For drug representation, we compared two different methods to characterize the drug.

First, we used fingerprint to represent each drug. Among various types of fingerprints, the Morgan fingerprint [29] was selected as the representative fingerprint for the drug (Figure 4). Morgan fingerprint, also known as the circular fingerprint, is a widely used method for representing the structural features of a drug molecule. It works by considering the neighborhood of each atom in the molecule and encoding the presence or absence of specific chemical substructures within a defined radius around each atom. These substructures are represented as unique circular patterns, allowing the fingerprint to capture information about the molecular bonds, functional groups, and overall structural characteristics of the drug. The Morgan fingerprint is particularly advantageous as it can effectively handle molecules of varying sizes and complexities, making it a popular choice for drug discovery and prediction tasks [30]. Therefore, drugs in the dataset were converted into binary vectors representing fingerprints using RDKit [31].

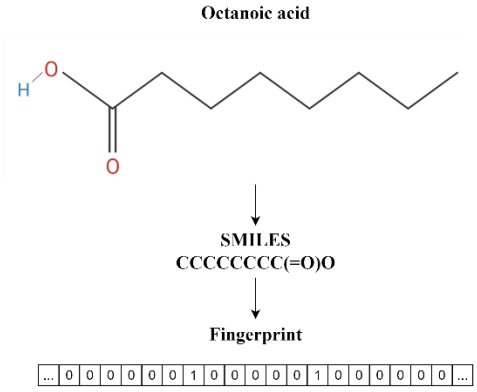

Figure 4. Fingerprint representation of Octanoic acid. Each CID corresponding to the drugs was first transformed into SMILES notation and subsequently converted into binary vectors representing the Morgan fingerprints.

Second, drugs were represented as molecular graphs. Molecular graphs represent atoms as nodes and bonds as edges, allowing for the inclusion of valuable details such as atom types, bond types, and connectivity patterns. In this study, MolGraphConvFeaturizer [32] was employed to convert the SMILES notation of drugs into molecular graphs using DeepChem [33]. The node features in the MolGraphConvFeaturizer consist of various atom properties, resulting in a feature length of 30. The atom type feature represents the type of atom using a one-hot vector, with categories including "C" (carbon), "N" (nitrogen), "O" (oxygen), "F" (fluorine), "P" (phosphorus), "S" (sulfur), "Cl" (chlorine), "Br" (bromine), "I" (iodine), and "other atoms". The formal charge feature indicates the integer electronic charge of the atom, and the hybridization feature is a one-hot vector representing the hybridization state of the atom, with options including "sp", "sp2", and "sp3". Also, it includes the hydrogen bonding feature, aromatic feature, degree feature, and the number of hydrogens. The edges in the molecular graph are represented by pairs of indices indicating the source and destination atoms connected by a bond. The molecular graph incorporates various node features to capture atom properties, while the edge indices represent the connectivity between atoms in the molecule.

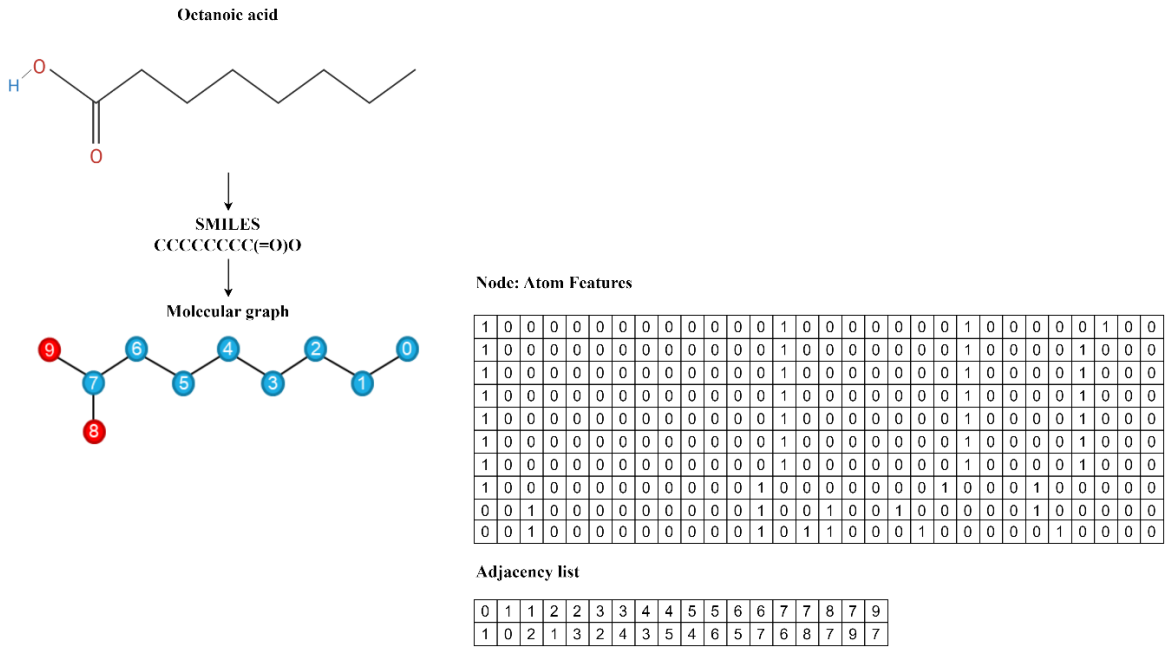

Figure 5. Molecular graph representation of Octanoic acid. Each CID corresponding to the drugs was first transformed into SMILES notation and subsequently converted into molecular graphs. Molecular graphs include node matrix, which provides the node (atom) features of the molecular graph, and adjacency matrix, which describes the bond between atoms.

2.2. Data preprocessing

2.2.1. Normalization

For gene expression data, normalization was applied as a preprocessing method (Table 2).

Table 2. DepMap database gene expression data after preprocessing. Standard Normalization and gene filtering by variance based on each gene.

		Genes		
Cell lines	TSPAN6	DPM1	…	MAGEA6
ACH-001113	0.584224	1.337422	…	-0.736650
…	…	…	…	…
ACH-000052	0.534013	-0.481543	…	-0.696717

In order to preprocess the gene expression data, two normalization methods were utilized: min-max normalization and standard normalization.

Min-max normalization, also known as feature scaling, rescales the data to a specific range, typically between 0 and 1. This normalization technique ensures that the minimum value of the data is mapped to 0, the maximum value is mapped to 1, and the values in between are linearly scaled accordingly. Min-max normalization is straightforward to implement and preserves the original distribution of the data. However, it is sensitive to outliers and can result in a loss of information if the range of the data is not representative.

On the other hand, standard normalization, also known as z-score normalization, transforms the data to have a mean of 0 and a standard deviation of 1. This method calculates the z-score for each data point by subtracting the mean and dividing by the standard deviation of the dataset. Standard normalization is advantageous as it allows for comparison and interpretation of data in terms of standard deviations from the mean. It is less affected by outliers and can handle data with different ranges. However, it assumes that the data follows a Gaussian distribution and may not be suitable for non-normal distributions.

Therefore, both min-max and standard normalization methods were employed in this study to explore the impact of different normalization techniques on the gene expression data analysis, considering their respective advantages and limitations.

2.2.2. Feature Selection

Feature selection was conducted to select genes that are effective features and potentially contribute to drug response prediction. We applied filter method for feature selection [34-37].

The filter method involves selecting features based on their statistical properties or relevance to the target variable, independent of any specific machine learning algorithm. In this study, the variance threshold approach was utilized as a filter method. It eliminates features with low variance, assuming that they may not significantly contribute to the prediction of drug response. Variance threshold was used to choose the top 10% (1639) genes in variance. By employing variance threshold, the study aimed to identify a subset of genes that are most relevant to drug response prediction.

2.3. Model

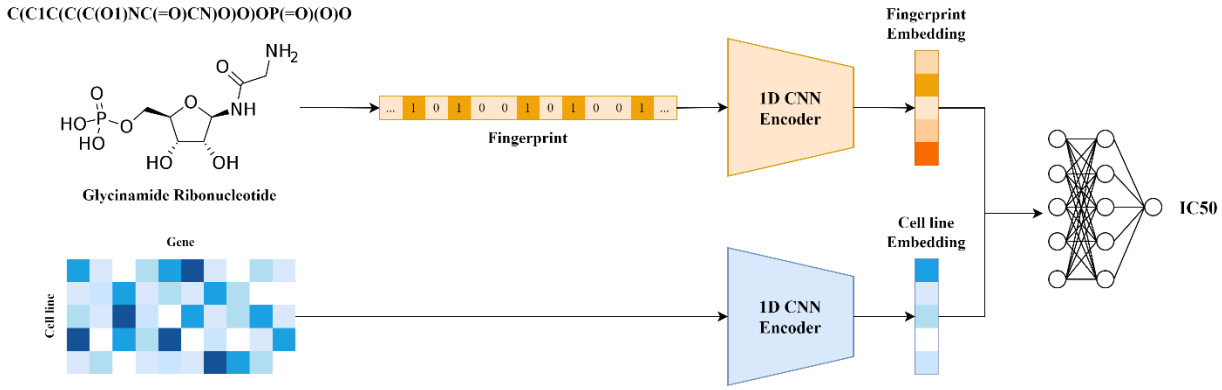

Figure 6. Overview of fingerprint-based drug response prediction model framework. The drug, represented as SMILES notation, undergoes a conversion process to obtain a binary vector fingerprint. The fingerprint then passes a 1D CNN encoder to learn the distinctive features of the drug representation. Similarly, the gene expression matrix undergoes encoding using a 1D CNN encoder to capture the embeddings of the cell lines. Subsequently, the fingerprint and cell line embedding are concatenated, and fully connected layers are employed to predict the IC50 value.

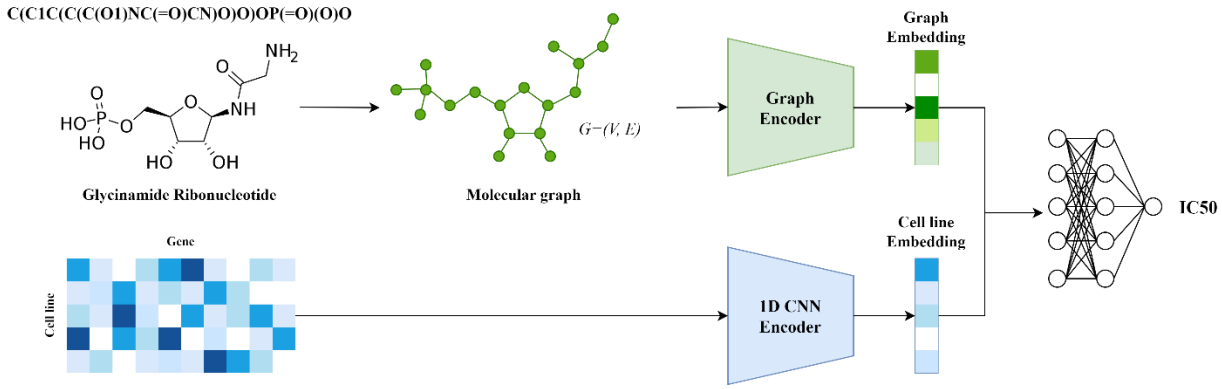

Figure 7. Overview of molecular graph-based drug response prediction model framework. The drugs are initially represented as SMILES notation, and then converted into a molecular graph. This graph representation is then fed into a graph encoder. The gene expression matrix undergoes encoding to capture the relevant features using a 1D CNN encoder. Finally, both embeddings are concatenated and passed through fully connected layers for IC50 value prediction.

2.3.1. Convolutional Neural Network (CNN)

In recent years, a modified variant of conventional 2D Convolutional Neural Networks (CNNs) known as 1D Convolutional Neural Networks (1D CNNs) has emerged. While conventional CNNs are primarily designed for processing 2D data, such as images, 1D CNNs have been developed to effectively analyze 1D signals such as ECG signals [37]. The applications of 1D CNNs extend beyond text analysis to include various domains, including the field of biology. In the context of biological data analysis, 1D CNNs have proven to be valuable tools for processing and extracting meaningful information from biological sequences, such as DNA, RNA, protein sequences, and gene expression profiles. In this study, 1D CNN was employed to reduce the dimensionality of both gene expression data and drug fingerprints, while simultaneously capturing important latent variables. By utilizing 1D CNNs, the study aimed to extract meaningful representations from the high-dimensional input data, facilitating the identification of critical features and patterns relevant to the relationship between gene expression and drug response.

2.3.2. Graph Convolutional Network (GCN)

When representing drugs as molecular graphs, a graph for a specific drug $G = (V, E)$ was stored using two matrices: feature matrix X and adjacency matrix A. Feature matrix $X \in \mathbb{R}^{N \times C}$ consists of N nodes within the graph, where each node is represented by a C-dimensional vector capturing the relevant atom properties. The adjacency matrix $A \in \mathbb{R}^{N \times N}$ illustrates the connectivity between nodes. The original graph convolutional layer takes these two matrices as input and aims to produce a node-level output with F features for each node. The original graph convolutional layer is:

$$AXW \tag{1}$$

where $W \in \mathbb{R}^{C \times F}$is a trainable weight matrix. However, two main drawbacks exist for the original graph convolutional layer. Firstly, it only sums up the feature vectors of neighboring nodes for each node, neglecting the contribution of the node itself. Secondly, the absence of matrix A normalization results in a scale change when performing the multiplication with A. To address these limitations, Graph Convolutional Network (GCN) model [38] was introduced. GCN model incorporates an identity matrix into A and applies a normalization scheme. Thus, the GCN layer is defined as:

$$Z = \widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}XW \tag{2}$$

Here, $\tilde{A}$ represents the graph adjacency matrix with added self-loop, and $\widetilde{D}$ denotes the graph diagonal degree matrix. Therefore, the overall multi-layer Graph Convolutional Network (GCN) follows the layer-wise propagation rule:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{3}$$

where $\sigma(\cdot)$ denotes an activation function, such as the $\mathrm{ReLU}(\cdot) = \max(0,\cdot)$, and $H(l)$ is the matrix of activations in the l th layer. In this study, two GCN layers were employed. Following each GCN layer, ReLU activation function was applied for non-linearity. Subsequently, a global max pooling layer was added right after the final GCN layer to learn the representation of the entire graph [25].

2.3.3. Graph Isomorphism Network (GIN)

Graph Isomorphism Network (GIN) [40] recently appeared as a solution for the graph isomorphism problem. Graph isomorphism refers to determining whether two graphs are structurally identical or can be transformed into each other through a permutation of nodes. Unlike traditional graph neural networks, GIN uses a learnable set function to aggregate information from neighboring nodes, enabling it to capture and propagate structural similarities across different graphs. Therefore, GIN has demonstrated maximum discriminative power among GNNs. GIN updates the node representations by multi-layer perceptron (MLP) as:

$$h_v^{(k)} = MLP^{(k)}((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum\nolimits_{u \in \mathcal{N}(v)} h_u^{(k-1)}) \tag{7}$$

where ϵ is either a learnable parameter or a fixed scalar, h is the node feature vector, and $\mathcal{N}(v)$ is the set of nodes neighbor to v. In this study, five GIN layers with 32 features were stacked to build GIN architecture. Batch normalization layer followed by ReLU activation function was applied after each GIN layers. Unlike previous models, GIN uses a different global pooling layer:

$$h_G = \sum_{i=0}^{N} h_i^0 \parallel \cdots \parallel \sum_{i=0}^{N} h_i^k \tag{8}$$

For each layer, the node embeddings are summed, and the resulting embeddings are concatenated. This approach combines the expressive power of the sum operator, which captures the information from neighboring nodes, with the memory of previous iterations achieved through concatenation. By summing the embeddings, the model retains the local information and captures the collective influence of neighboring nodes. The concatenation operation allows the model to incorporate the memory of previous layers, enabling the integration of both local and global information.

2.4. Evaluation Criteria

In the subsequent section, the performance of the proposed model was assessed by two widely adopted metrics: Mean Squared Error (MSE) and Pearson Correlation Coefficient (PCC). These metrics were employed to quantify the difference between the observed drug responses, denoted as $y = (y_i)_{i=1}^{m}$ (IC50), and the predicted drug responses, represented as $\hat{y} = (\hat{y}_i)_{i=1}^{m}$, where m corresponds to the total number of cell lines [18].

MSE provides an overall measure of the average squared differences between the observed and predicted drug responses:

$$MSE(y, \hat{y}) = \frac{\sum_i (y_i - \hat{y}_i)^2}{m} \tag{9}$$

PCC offers an evaluation of the linear correlation between the two response sets. PCC between the observed drug responses (y) and the predicted drug responses $(\hat{y})$ is computed using the following formula, where $\mu(X)$ represents the mean of a random variable X:

$$PCC(y, \hat{y}) = \frac{\sum_i (\hat{y}_i - \mu(\hat{y}))(y_i - \mu(y))}{\sqrt{\sum_i (\hat{y}_i - \mu(\hat{y}))^2} \sqrt{\sum_i (y_i - \mu(y))^2}} \tag{10}$$

3. Results

To evaluate the performance of the drug response prediction model, the dataset described in "Dataset" section was split into 80% as training set, 10% as validation set, and 10% as testing set. Experiments were implemented using PyTorch, and the models were trained for 100 epochs with a batch size of 512. The models were trained with 1e-4 learning rate and Adam Optimizer was used for optimization with a weight decay of 1e-5.

Table 3 compares the performance between different type of normalizations applied to gene expression data. Normalization methods described in Section "Normalization" were conducted to gene expression data, and drugs were represented as fingerprint to compare the drug response prediction performance. The aim was to assess and compare the predictive performance of drug response prediction based on the different normalization approaches. Table 3 demonstrates that the performance of the different normalization methods yielded similar results, indicating that the choice of normalization technique may have minimal impact on the overall performance of the model. Despite the slightly better result for MSE in minmax normalization, standard normalization is commonly used because it helps to account for the inherent variability in gene expression levels across different genes and allows for a more effective comparison and integration of gene expression data from multiple samples.

Table 3. Performance comparison of different normalization methods for gene expression data and using fingerprint to represent the drug data.

Drug		Gene expression	MSE	PCC
Representation	**Model**	**Normalization**		
Fingerprint	1D CNN	Minmax	1.2778	0.9086
		Standard	1.2811	0.9086

The overall performance of the models employed in this study, as described in the "Model" section, is summarized in Table 4. The drug representations were based on binary vector fingerprints and molecular graphs, leading to the adoption of distinct models for representation learning. Specifically, 1D CNN was utilized to learn representations from the bit string fingerprints, while three different graph-based learning models were employed for the molecular graphs. The results presented in Table 4 provide a comprehensive evaluation of the predictive performance achieved by these models in terms of their ability to capture essential features and accurately predict drug

responses. Fingerprint encoded with 1D CNN shows the best performance for both MSE and PCC, and GCN resulted the best MSE while GIN showed the best PCC.

Table 4. Overall performance comparison of different drug representations and models.

Drug		MSE	PCC
Representation	**Model**		
Fingerprint	1D CNN	**1.2811**	**0.9086**
Molecular graph	GCN	**1.3543**	0.9051
	GIN	1.4012	**0.9061**

4. Discussion

In Table 5, fingerprint showed better performance than molecular graph for drug representation. Despite the initial expectation that molecular graphs would outperform fingerprint-based representations, the observed results did not align with this hypothesis, which could be attributed to several factors. Firstly, the limited number of epochs used for training the models might have influenced the results. With a larger number of epochs, the graph-based models could have had more opportunities to learn and capture the complex structural features of the molecular graphs, potentially improving their predictive accuracy. Secondly, the use of only gene expression data in conjunction with drug representations might have influenced the results. While gene expression data provides valuable insights into the cellular response, it represents a single omics data type. Incorporating additional omics data, such as genomic or proteomic data, could provide a more comprehensive view of the underlying mechanisms and potentially enhance the predictive performance of the graph-based models. Furthermore, the nature of the drug molecules themselves could also play a role. Fingerprint-based representations encode the chemical properties and structural features of the drugs in a binary vector format, which may have facilitated the extraction of relevant information for drug response prediction. On the other hand, the molecular graph-based models rely on capturing the intricate connectivity patterns and interactions within the molecular graphs. If the molecular graphs in the dataset exhibit less prominent structural characteristics or if the available graph-based models were not optimized for the specific dataset, it could contribute to the relatively lower performance compared to fingerprint-based representations.

In addition to the findings presented in this study, there are several potential avenues for future research that could further enhance the experiment conducted in this paper. These areas of exploration include changing network structures, hyperparameter tuning, adopting state-of-the-art models, utilization of pretrained models, integration of multiomics data, and incorporation of existing domain knowledge.

Firstly, further investigation into different network structures, such as varying the number of layers or exploring alternative architectural designs, could provide valuable insights. This exploration could help determine the optimal network structure for drug response prediction, potentially improving the model's performance and interpretability.

Hyperparameter tuning is another aspect that warrants further attention. Fine-tuning hyperparameters like batch size, optimization algorithms, and learning rate could lead to enhanced convergence, generalization, and overall model performance. Conducting more comprehensive hyperparameter searches and adopting advanced optimization techniques could result in even better model performance.

Adopting other state-of-the-art models, including transformer-based models, would contribute to achieve a higher performance for drug response prediction. Also, assessing the performance across a wider range of models and datasets could provide valuable insights into the model's versatility and robustness.

Utilizing pretrained models that have been trained on diverse datasets, such as pretraining through data from databases like The Cancer Genome Atlas (TCGA), could offer significant benefits. For example, DeepDR [20] used two pretrained autoencoders for mutation and expression to learn

representations from cell lines. Leveraging the knowledge and patterns learned from large-scale datasets may enhance the model's ability to extract meaningful features and improve prediction accuracy.

Incorporating multiomics data, encompassing various molecular layers such as genomics, proteomics, and metabolomics, could further enrich the model's predictive capabilities. Integrating multiple data modalities and capturing their interactions could provide a more comprehensive understanding of the underlying biological mechanisms and lead to more accurate drug response predictions.

The incorporation of existing domain knowledge, such as protein-protein interaction (PPI) networks and known drug-target interactions, holds promise for improving drug response prediction. Integrating such knowledge into the model's architecture and leveraging it as constraints or priors can enhance the interpretability and reliability of the predictions.

By addressing these areas of future research, we can further advance the field of drug response prediction. Exploring different network structures, fine-tuning hyperparameters, comparing with other models, utilizing pretrained models, integrating multiomics data, and incorporating existing domain knowledge have the potential to significantly enhance the accuracy, interpretability, and applicability of drug response prediction models. These avenues of exploration pave the way for future studies and contribute to the ongoing efforts in personalized medicine and therapeutics.

5. Conclusions

In conclusion, this study focused on drug response prediction using gene expression data and compared the performance of two different approaches for drug representation: fingerprint and molecular graph. For the fingerprint representation, a 1D CNN encoder was employed, while for the molecular graph representation, various graph-based models, including GCN and GIN, were utilized. Additionally, gene expression data was encoded using a 1D CNN to capture cell line representations.

The results of our experiments demonstrated the effectiveness of both the fingerprint and molecular graph representations in predicting drug responses. The 1D CNN encoder applied to the fingerprint representation effectively learned informative features from the binary vector, enabling accurate predictions. On the other hand, the graph-based models, namely GCN and GIN, demonstrated their capability to capture the structural characteristics of molecular graphs and extract relevant features for drug response prediction.

Comparing the two approaches, we observed that both the fingerprint and molecular graph representations exhibited promising performance, with slight variations in their predictive accuracy. The choice between the two representations may depend on factors such as the nature of the dataset, the availability of additional domain knowledge, and the interpretability of the learned features.

Furthermore, the use of gene expression data in conjunction with the drug representations proved to be valuable for improving the prediction accuracy. The 1D CNN encoder applied to gene expression data successfully captured the distinctive features of the cell lines, enabling a comprehensive understanding of the interaction between drugs and cell lines.

In conclusion, this study highlighted the significance of different drug representation approaches for drug response prediction. Both the fingerprint and molecular graph representations, in combination with gene expression data, provided valuable insights into personalized drug responses. Future research can focus on further refining the models, exploring different network architectures, optimizing hyperparameters, incorporating multiomics data, and integrating existing domain knowledge to enhance the accuracy and interpretability of drug response prediction models. These advancements will contribute to the development of personalized medicine and improve treatment outcomes for patients.

References

1. "Worldwide Cancer Statistics." Cancer Research UK, 22 Aug. 2019, www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer#heading-Zero.
2. Collins, F. S. & Varmus, H. A new initiative on precision medicine. N. Engl. J. Med. 372, 793–795 (2015).
3. Jameson, J. L. & Longo, D. L. Precision medicine--personalized, problematic, and promising. N. Engl. J. Med. 372, 2229–2234 (2015).
4. Adam, G., Rampášek, L., Safikhani, Z. et al. Machine learning approaches to drug response prediction: challenges and recent progress. npj Precis. Onc. 4, 19 (2020). https://doi.org/10.1038/s41698-020-0122-1
5. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision oncology: an overview. J. Clin. Oncol. 31, 1803–1805 (2013).
6. Doherty, M., Metcalfe, T., Guardino, E., Peters, E. & Ramage, L. Precision medicine and oncology: an overview of the opportunities presented by next-generation sequencing and big data and the challenges posed to conventional drug development and regulatory approval pathways. Ann. Oncol. 27, 1644–1646 (2016).
7. Heymach, J. et al. Clinical Cancer Advances 2018: annual report on progress against cancer from the American Society of Clinical Oncology. J. Clin. Oncol. 36, 1020–1044 (2018).
8. Twomey, J. D., Brahme, N. N. & Zhang, B. Drug-biomarker co-development in oncology—20 years and counting. Drug Resist. Updat 30, 48–62 (2017).
9. Johnson, A. et al. The right drugs at the right time for the right patient: the MD Anderson precision oncology decision support platform. Drug Discov. Today 20, 1433–1438 (2015).
10. Rifaioglu, A. S. et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief. Bioinform. https://doi.org/10.1093/bib/bby061 (2018).
11. Lipinski, C.F. Maltarollo, V.G. Oliveira, P.R. Silva, A.B.F. da Honorio, K.M. Advances and Perspectives in Applying Deep Learning for Drug Design and Discovery. Front. Robot. AI. 6, 108 (2019).
12. "Explore the Cancer Dependency Map." DepMap, depmap.org/portal/. Accessed 1 June 2023.
13. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607 (2012).
14. Yang, W. et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 41, D955–D961 (2013).
15. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer 6, 813–823 (2006).
16. Kim, J., Park, S., Min, D. & Kim, W. Comprehensive survey of recent drug discovery using deep learning. Int. J. Mol. Sci. 22 (18), 9983 (2021).
17. An, X., Chen, X., Yi, D., et al. Representation of molecules for drug response prediction. Brief. Bioinform. 23 (1), 1-11 (2021).
18. Kim, S., Bae, S., Piao, Y., Jo, K. Graph convolutional network for drug response prediction using gene expression data. Mathematics. 9, 772 (2021).
19. Baptista, D., Ferreira, P. G. & Rocha, M. Deep learning for drug response prediction in cancer. Brief. Bioinform. https://doi.org/10.1093/bib/bbz171 (2020).
20. Chiu, Y. C. et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med. Genom. 12, 18 (2019).
21. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 35, i501–i509 (2019).
22. Liu, P., Li, H., Li, S. et al. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinformatics. 20 (1), 408 (2019).
23. Li, M., Wang, Y., Zheng. R. et al. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform. 1–1 (2019).
24. Chang, Y. et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci. Rep. 8, 8857 (2018).
25. Nguyen, TT., Nguyen, GTT., Nguyen, T. et al. Graph convolutional networks for drug response prediction. . IEEE/ACM Trans Comput Biol Bioinform. 19 (1), 146-154 (2021).
26. Shin, J., Piao, Y., Bang, D., Kim, S., and Jo, K. DRPreter: Interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. International Journal of Molecular Sciences, 23, 13919 (2022).

27. Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat. Biotechnol. 32, 1202–1212 (2014).
28. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. Cell 166, 740–754 (2016).
29. Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. J. Chem. Doc. 5, 107–113 (1965).
30. Kuenzi, B. M., et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell 38, 672–684.e6 (2020).
31. Landrum G. RDKit: Open-source cheminformatics; 2006. Available from: https://www.rdkit.org/.
32. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. J. Comput. Aided Mol. Des. 30, 595–608 (2016).
33. "deepchem/deepchem: Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology". DeepChem, https://github.com/deepchem/deepchem Accessed 7 June 2023.
34. P. Langley. Selection of Relevant Features in Machine Learning. Proc. AAAI Fall Symp Relevence. 140-144 (1994).
35. Blum, A. L. & Langley, P. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271 (1997).
36. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003).
37. Kiranyaz, S. et al. 1D convolutional neural networks and applications: A survey. Mech. Syst. Signal Process. 151, 107398 (2021).
38. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. Prepint at https://arxiv.org/abs/1609.02907 (2016).
39. Bahdanau, D., Cho, K. & Bengio. Y. Neural machine translation by jointly learning to align and translate. Preprint at https://arxiv.org/pdf/1409.0473.pdf (2014).
40. Xu, K. et al. How powerful are graph neural networks? In Proceedings of the 7th International Conference on Learning Representations. (OpenReview, New Orleans, LA, 2018).