DrugCellGNN: Graph Convolutional Networks for Integrating Omics and Drug Similarities in Cancer Therapy Prediction

Gehad Awad Aly[1], Rania Ahmed Abdel Azeem Abul Seoud[2], Dina Ahmed Salem[3]
Department of Electrical Engineering -Faculty of Engineering, Fayoum University, Fayoum, Egypt[1,2]
Faculty of Engineering-Department of Computer, Misr University for Science and Technology, 6 of October, Egypt[3]

***Abstract*—Predicting drug response in cancer cell lines is a critical step toward precision oncology, enabling more efficient therapeutic discovery and personalized treatment strategies. However, the complexity of drug–cell interactions, driven by diverse omics profiles and structural variability among drugs, poses significant challenges for conventional machine learning approaches. In this study, we propose an end-to-end pipeline that integrates multi-omics data (gene expression, copy number variation, and mutations) with chemical structure representations of drugs to predict binary drug response. Our method employs principal component analysis (PCA) for dimensionality reduction of high-dimensional omics data, followed by the computation of drug–drug and cell–cell similarity matrices. These are used to construct a heterogeneous graph combining intra-class similarities with drug–cell interactions. A customized graph neural network model, DrugCellGNN, is then applied to learn context-aware embeddings of drugs and cells. The fused representations are passed to a downstream multi-layer perceptron for classification. To address class imbalance, we introduce a dynamic focal loss function that adaptively emphasizes hard-to-classify examples. Evaluation on the GDSC dataset with an 80/20 train–test split demonstrates strong performance: Accuracy = 0.8935, F1 = 0.9201, AUC = 0.9510. This work highlights the utility of graph-based integration of multi-omics and drug features for drug sensitivity prediction. By leveraging both molecular and relational information, the proposed framework offers a robust and extensible foundation for advancing computational approaches in precision oncology.**

***Keywords*—Precision oncology; drug sensitivity prediction; graph neural networks (GNNs); multi-omics integration; focal loss; PCA**

I. Introduction

Cancer is a disease of the genome that causes mortality worldwide, characterized by high complexity and inter-patient heterogeneity that complicate effective treatment [1]. Traditional therapies, such as chemotherapy and radiotherapy, often exhibit limited specificity, leading to substantial toxicity in healthy tissues and suboptimal efficacy [2]. However, not all patients react to pharmacological therapy in the same manner, and biological data, such as gene mutations or gene expression, can help predict which patients are most likely to benefit from a certain medicine [3]. Thus, establishing strategies for predicting drug sensitivity is critical for individualized therapy. According to the National Research Council, precision medicine can be used to divide patients into subgroups based on their unique reactions to medical treatment [4]. Customizing effective treatments based on individual characteristics can improve therapy quality, reduce wasteful expenses, and minimize negative side effects. Precision medicine aims to optimize cancer therapy by selecting drugs tailored to each patient's unique genetic profile [4]. This process presents a complex challenge in cancer treatment. These challenges have led to large-scale experiments on human cancer cell lines and a wide range of anticancer drugs, which have led to the availability of free datasets for cancer cell drug sensitivity information and molecular markers of drug responses. The Genomics of Cancer Drug Sensitivity (GDSC) [5] is one of the most widely used projects, providing molecular profiles and drug response data for hundreds of cancer cell lines treated with various anticancer drugs. These datasets provide genetic features of cancer cell line panels, such as gene expression profiles, copy number alterations, and single-nucleotide mutations. This enormous accumulation of data paves the way for the advancement of data analysis and computational techniques, including machine learning and artificial intelligence approaches.

Despite notable progress, existing computational approaches face several limitations. Traditional machine learning models often struggle to capture complex interactions between drugs and cancer cell lines and are limited by the high dimensionality and heterogeneity of multi-omics data. Moreover, many methods treat drugs and cell lines independently, overlooking their intrinsic relational structure, which is crucial for accurate prediction and generalization across unseen samples.

To address these challenges, this study proposes DrugCellGNN, a graph neural network–based framework for binary drug sensitivity prediction. The proposed method integrates multi-omics features, drug structural information, and functional drug response data to construct biologically meaningful drug–cell interaction graphs. Dimensionality reduction using principal component analysis (PCA) is employed to mitigate data sparsity, while a dynamic focal loss is introduced to handle class imbalance. Through comprehensive evaluation, DrugCellGNN demonstrates improved predictive performance compared to existing approaches, highlighting its potential utility in precision oncology.

The remainder of this study is organized as follows: Section II briefly reviews some of the recent work published in the area of prediction of the drug response by using genomics

data. Section III explains the methodology that has been used to create the model. Evaluation criteria is explained in Section IV. The results of the proposed model are presented in Section V. Finally, Section VI concludes the study.

II. Related Work

Several computational approaches have been developed to predict medication response by analyzing gene expression profiles or other molecular information of cell lines. This Computational method for predicting drug response can be grouped into two main approaches: 1) classification (predicting drug responses as sensitive vs. resistant), 2) regression (Estimating a quantitative measure to evaluate a cell line's response to a drug). Various computational approaches have been developed to solve classification problems for predicting the sensitivity of anticancer drugs. Choi et al. [6] created RefDNN, a computational model that utilizes a deep neural network and ElasticNet repressors. They established a collection of reference medications and a baseline for classifying drugs to evaluate others. Drug sensitivity probabilities for a given cell line-drug pair were predicted using the drug's similarity to the reference set. RefDNN has the potential to be utilized for repositioning anticancer medications. Similarly, Logistic matrix factorization with regularization terms was used in DSPLMF [7] to classify drug response. This approach incorporates drug similarity derived from chemical structures and cell line similarity computed from gene expression, copy number variation, mutation profiles, and drug response data. While these methods demonstrate promising performance, they primarily rely on pairwise similarities and do not explicitly model higher-order interactions between drugs and cell lines.

Furthermore, various computational regression approaches have been developed to predict the half-maximal inhibitory concentration (IC50) of cell lines in response to drugs. Wang et al. [4] proposed the Similarity Regularized Matrix Factorization (SRMF) method, which incorporates the gene expression similarity of cell lines and the chemical similarity of drugs to enhance the prediction of anticancer drug responses. Ahmadi Moughari et al. have introduced ADRML, a framework for predicting how well cancer cells will respond to drugs using manifold learning. ADRML maps drug response values into a low-dimensional latent space and predicts the drug response for new cell line-drug pairs based on this latent space. The framework incorporates multiple types of cell line similarities and drug similarities into the manifold learning process [8].

Selecting the optimal set of genetic features for predicting drug response is crucial for developing effective classification models. Consequently, numerous feature selection algorithms employing diverse strategies have been developed. AutoBorutaRF [9] is a method which based on feature selection and designed for predicting drug response. It combines an autoencoder network with the Boruta algorithm [10]. The autoencoder is used initially, and then Boruta is applied to select the most relevant features. These selected features are then used to train a Random Forest classifier for predicting drug response. Dong et al. [11] developed a drug response prediction model called Support Vector Machine Recursive Feature Elimination (SVM-RFE), which employs a wrapper approach, combining recursive feature selection with an SVM classifier. Emdadi et al. [12] introduced a feature selection method designed for drug response prediction. It uses Hidden Markov Models (HMM) and Autoencoders (AE). While these techniques reduce dimensionality and noise, they typically treat samples independently and fail to exploit relational structures between drugs and cell lines.

Considering the complexity and noise inherent in biological data, deep learning methods often outperform traditional machine learning algorithms in predicting drug sensitivity [13]. A hybrid graph convolutional network (GCN) has been proposed in [14], This approach, known as DeepCDR, utilizes a variational autoencoder (VAE) framework to effectively model the complex relationships between multi-omics profiles of cancer cells and the intrinsic chemical structures of drugs, achieving strong results; nevertheless, it performs late-stage omics fusion, employs limited similarity types (primarily structural), and does not explicitly mitigate severe class imbalance common in pharmacogenomic data.

Other studies suggest several statistical learning strategies, such as regularized linear regression models, ridge regression lasso [15], and Support vector machines [16].

Although prior studies have progressively improved prediction accuracy, they share common limitations: over-reliance on pairwise or late-fused similarities, inadequate handling of class imbalance, limited integration of hybrid (structural + functional) drug/cell similarities, and insufficient relational modeling of multi-omics interactions. To overcome these gaps, the proposed DrugCellGNN framework integrates multi-omics features, functional and structural drug similarities, and graph neural networks to explicitly model drug–cell relationships, enabling robust and biologically informed binary drug sensitivity prediction.

III. Methodology

The primary objective of this study is to develop a predictive framework for binary drug response classification, as it may be more important whether a drug is effective for a given cancer cell line or not. There are two classifications of drug Responses: "sensitivity" and "non-sensitivity". We focused on integrating multi-omics profiles of cancer cell lines with structural and functional representations of drugs. The first step begins with extensive data preprocessing, including the removal of samples and features with excessive missingness, followed by a hybrid imputation strategy. Then, for improving the prediction performance, the cleaned datasets are dimensionally reduced to retain the most informative features, and similarity matrices are computed for both drugs and cell lines. Finally, a model is used for predicting the probability that a cell line would become drug-sensitive. Every pair was then categorized into a sensitive or resistant class by applying the threshold to the predicted probability of the cell line-drug pairs. The main scheme of our model is represented in Fig. 1.

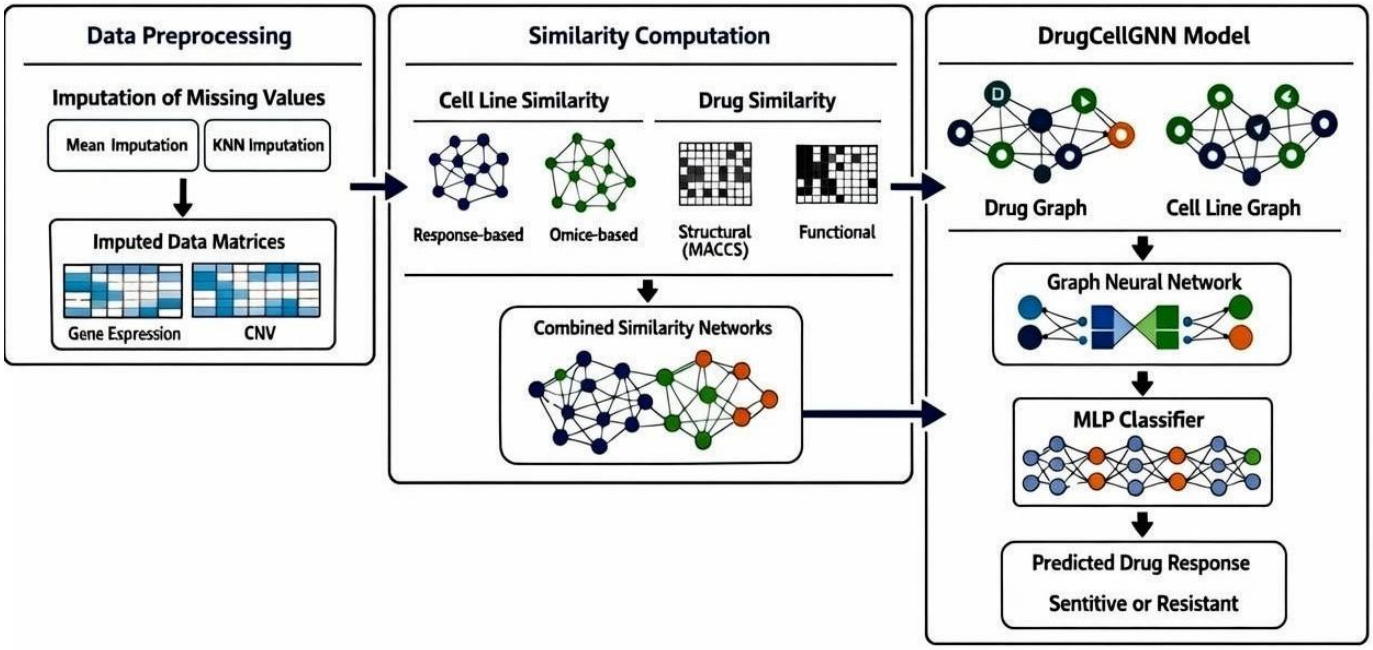

Fig. 1. Overview of the DrugCellGNN framework.

A. Dataset

Genomics of Drug Sensitivity in Cancer (GDSC) [17] is used in this study. This dataset includes drug response measurements, such as the half-maximal inhibitory concentration (IC50 values), Cell line omics information (including gene expression profiles, copy number variation (CNV), and mutation data), and maximum concentration values for drugs. The GDSC has 1,084 cell lines and 344 drugs; however, some values were missing in the dataset, so we filtered the dataset and applied preprocessing steps to address this issue. After the preprocessing, we had 551 cell lines and 91 drugs. In addition to the mentioned cell line information, chemical compound information is used and represented by Simplified Molecular Input Line Entry System (SMILES) [18], which was downloaded from PubChem [19]. And molecular fingerprints of drugs were extracted using the rdkit package in Python [20].

B. Preprocessing

We organize the collected data into three main types of matrices. The first is the response matrix, which captures the drug response data; rows represent cell lines, and columns represent different drugs. Each entry in this matrix holds either an IC50 value or a sensitivity label. The second type is the cell line feature matrices, where each row refers to a specific cell line and the columns represent its associated features. Lastly, drug feature matrices, with drugs as rows and their corresponding features arranged across the columns. Once these matrices were prepared, we carried out a series of preprocessing steps to ensure they were ready for downstream analysis.

The preprocessing steps are as follows:

- Missing value analysis and filtering: Because some of the information above has a high ratio of missing values, we first needed to remove samples with substantial missing data before performing imputation. For the response matrix, according to the distribution of missing values in Fig. 2(A), we initially removed cell lines with more than 95% missing values. Then, we eliminated cell lines and drugs with more than 30% missing data. Fig. 2(B) shows the distribution after filtering. In the gene expression matrix, we identified 72 cell lines that contained only missing values (i.e., no data across all features); therefore, the dataset became fully complete after excluding these cell lines. In the CNV matrix in Fig. 3(A) and Fig. 3(B), we removed features and cell lines with more than 30% missing values. For the mutation matrix, we eliminated features with over 35% missing data, resulting in a fully complete dataset.
- Imputing missing values: After removing samples and features with a high percentage of missing values, some missing entries still remained in the response and CNV matrices. To fill these gaps, we used a hybrid imputation method that combines mean imputation with K-nearest neighbors (KNN) imputation.

Let the drug response matrix be denoted by:

$$R \in R^{n \times m}$$

where, n and m represent the number of cell lines and drugs, respectively.

Mean Imputation is used for Low-Missing Drugs, where each drug j with a missing rate less than a threshold τ (in our study, τ =0.1), for drugs missing <10% values, fills gaps with the average response [Eq. (1)]:

$$\mu_j = \frac{1}{|I_j|} \sum_{i \in I} R_{ij} \quad (1)$$

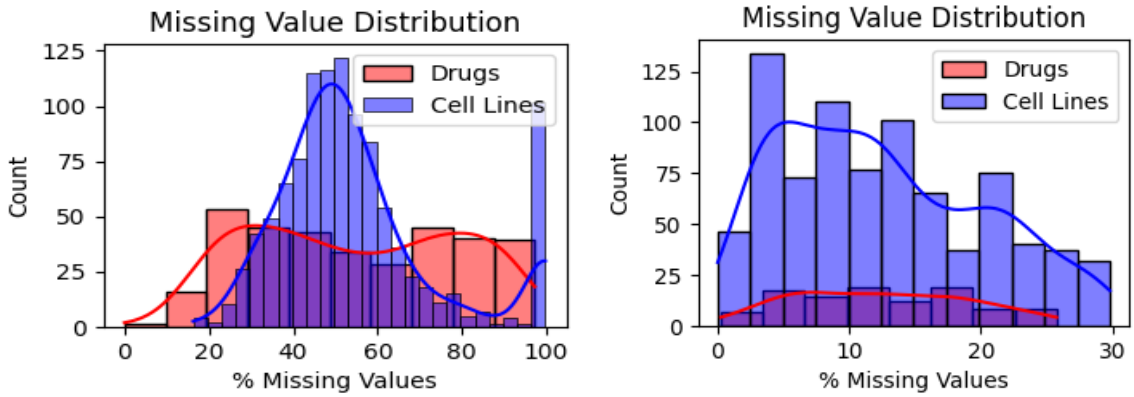

Fig. 2. (A) Missing value distribution for the response matrix before filtering. (B) Missing value distribution for the response matrix after filtering.

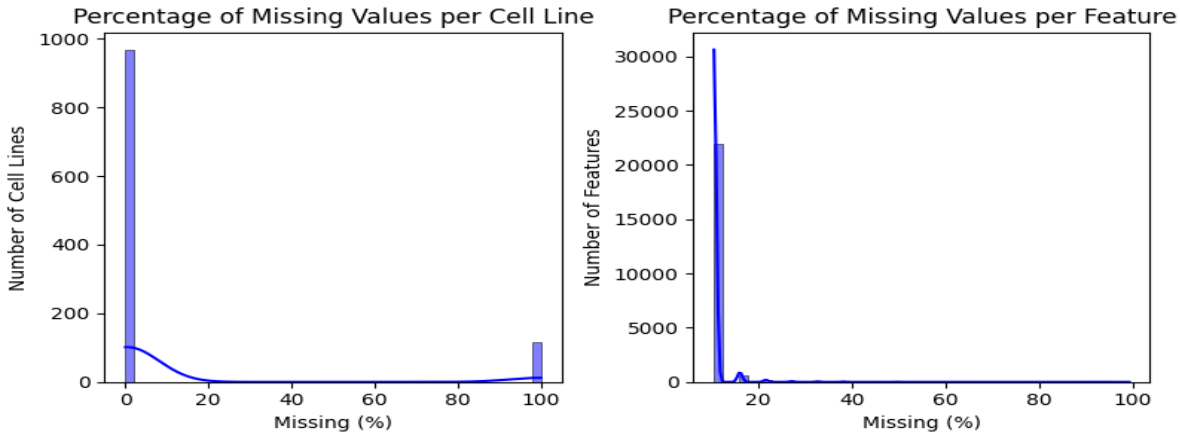

Fig. 3. (A) Missing values distribution for cell lines in the CNV matrix before filtering. (B) Missing values distribution for CNV features in the CNV matrix before filtering.

where, R_{ij} represents the response of cell line i to drug j, and may be missing (i.e., R_{ij}=NaN), $I_j=\{i|R_{ij}$ is observed$\}$, then for all missing entries R_{ij} where j ∈ low-missing drugs, set: $\hat{R}_{ij}=\mu_j$

For drugs with missing values greater than τ, we used KNN Imputation. To prepare for KNN imputation, the matrix is scaled using Eq. (2):

$$\grave{R}_{ij}=\frac{R_{ij}-\text{Median}(R_j)}{\text{IQR}(R_j)} \quad (2)$$

where, $\text{IQR}(R_j)$ is the interquartile range of drug j 's responses.

We then applied KNN imputation on the scaled data. For each missing value $\grave{R}_{ij}$, the imputed value is computed as the average of the same feature from the k nearest samples (rows) based on Euclidean distance [see Eq. (3)]:

$$\grave{R}_{ij}=\frac{1}{k}\sum_{r\in N_i}\grave{R}_{rj} \quad (3)$$

where, N_i is the set of k nearest neighbors of the sample i with observed values in feature j. After imputation, the scaled data were transformed back to their original scale using the inverse of the robust scaling operation [Eq. (4)]:

$$R_{ij}^{final}=\grave{R}_{ij}.\text{IQR}(R_j)+\text{Median}(R_j) \quad (4)$$

This hybrid approach ensures that low-missing features are imputed efficiently using simple statistics, while high-missing features benefit from structure-aware imputation via KNN.

- Binary conversion using IC50: This research aims to develop classification-based modeling, so we converted the continuous drug response data into binary labels (sensitivity and non-sensitivity). Several research studies have adopted different threshold values of IC50 for binarization. Such as median values [7], [9] or a specific deviation from the normalized mean [11]. While others have relied on more biologically grounded pharmacokinetic thresholds, such as the maximum drug concentration (Cmax) [6]. Cmax makes more sense among the different thresholds used to assign labels to drug response data because it is predicated on the medication's pharmacokinetic characteristics. Cmax is the highest concentration of a medication that can be found in plasma. Thus, it is clear that a cell line is resistant to a medicine if it needs a molar concentration greater than the drug's Cmax for 50% inhibition. For this reason, we adopted the Cmax in our study for classifying drug response. And based on it, we labeled a cell line as "sensitive" if the $IC50_{i,j} \leq$ Cmax, which is represented by 1; otherwise, it's labeled as "non-sensitive".

C. Feature Engineering

Omics Integration and Dimensionality Reduction: To construct a comprehensive representation of each cell line, we integrated multiple omics data types, including gene expression, copy number variation (CNV), and mutation profiles. Only cell lines present in all three sources were retained to ensure consistency. The aligned datasets were then concatenated to form a unified omics feature matrix. To reduce scale disparities and improve model stability, the omics matrix was standardized using z-score normalization and subsequently, Principal Component Analysis (PCA) [21]. In our study, PCA was applied to reduce dimensionality while preserving 80% of the variance. This step helped eliminate redundancy and emphasized the most informative biological features.

The cumulative explained variance curve (Fig. 4) demonstrates a smooth and steady increase, indicating that a relatively moderate number of components (~140) is sufficient to retain 80% of the variance. This gradual curve suggests that the variance is well distributed across components, meaning no single feature dominates the representation. Such a distribution supports the stability and robustness of the compressed feature space, as it captures diverse biological signals rather than being overly dependent on a few dominant factors.

To explore the potential benefit of non-linear dimensionality reduction, we compared PCA with a shallow autoencoder trained on the standardized omics feature matrix. Both approaches yielded comparable performance in downstream tasks, suggesting that the underlying structure of the data is predominantly linear. Given PCA's simplicity, interpretability, and computational efficiency, we opted to use PCA for feature compression in the final pipeline.

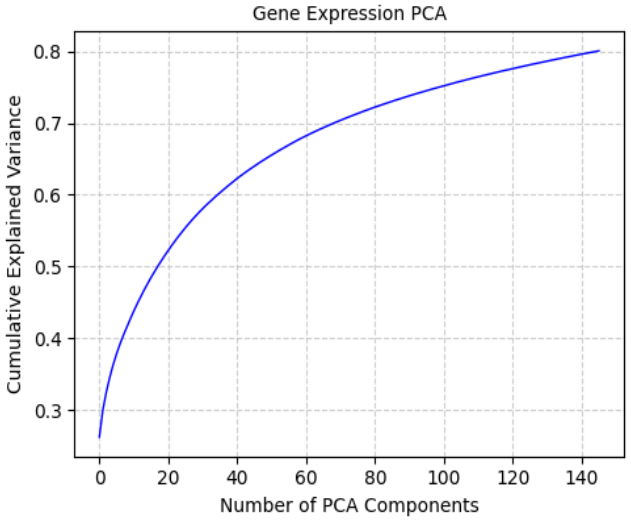

Fig. 4. PCA explained variance curve.

- Similarity Computation: To enhance the representation of both cell lines and drugs, we constructed similarity networks using a weighted combination of functional and structural measures. This approach captures complementary biological and chemical relationships, enabling the model to generalize better across diverse drug-cell line pairs.

For cell lines, similarity was computed using two modalities:

- Response-Based Similarity: We calculated cosine similarities from the drug response matrix $R \in R^{n \times m}$, where R_i and R_j are the response vectors of cell lines i and j. Cosine similarity is defined in Eq. (5):

$$S_{ij}^{response} = \frac{R_i . R_j}{\|R_i\|_2 \|R_j\|_2} \quad (5)$$

- Omics-Based Similarity: derived from the omics feature matrix $O \in \mathbb{R}^{n \times m}$ where O_i and O_j are the feature vectors of cell lines i and j. Using the same cosine similarity equation [Eq. (6)]:

$$S_{ij}^{omics} = \frac{O_i . O_j}{\|O_i\|_2 \|O_j\|_2} \quad (6)$$

The final cell line similarity matrix was computed as a weighted combination [see Eq. (7)]:

$$S^{cell} = \lambda_c . S^{response} + (1-\lambda_c) . S^{omics} \quad (7)$$

λ_c is a parameter that represents the weight of the matrix; in our study, λ_c=0.3 was selected empirically to balance phenotypic response patterns and intrinsic molecular.

This fusion strategy balances phenotypic behavior with intrinsic biological or chemical characteristics.

For drug similarity, the similarity between two drugs i and j is computed using MACCS fingerprints [22] and the response matrix; we computed both similarity matrices using Jaccard similarity.

Structural similarity: computed using the Jaccard similarity over MACCS fingerprints [see Eq. (8)].

$$S_{ij}^{struc} = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} \quad (8)$$

where, F_i, F_j are Sets of active bits in the MACCS fingerprints of drugs i and j, respectively.

Functional similarity: using cosine similarity across binary response profiles S_{ij}^{func}.

The final drug similarity matrix was Eq. (9):

$$S^{drug} = \lambda_d . S^{func} + (1-\lambda_d) . S^{struc} \quad (9)$$

λ_d is a parameter that represents the weight of the matrix; in our study, λ_d=0.3.

D. Model Architecture

To capture the complex interplay between drugs and cancer cell lines, we propose a novel graph-based deep learning framework, DrugCellGNN, that integrates structural drug information, functional cell similarities, and multi-omics features into a unified predictive model.

Graph Construction and Node Embedding: To include the structural relationships between drugs and cell lines, we built two separate graphs: a drug similarity graph and a cell line similarity graph. Each graph G= (V, E) is composed of nodes V (drugs or cell lines) and edges E, which are defined based on a pairwise similarity threshold. Specifically, an edge (i,j) ∈ E is

formed if the similarity score $S_{ij} > 0.7$, a threshold selected to retain only strong associations.

Each node is initialized with a trainable embedding vector of size e from $E_d \in R^{|v_d| \times e}$ for drug and $E_c \in R^{|v_c| \times e}$ for cell lines, where $|v_d|$ and $|v_c|$ denote the number of drug and cell nodes, respectively. These embeddings were then updated using a graph convolution operation defined as Eq. (10):

$$H^{l+1} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{l}W^{l}\right) \quad (10)$$

Where $\hat{A}=A+I$ is the adjacency matrix with self-loops, $\hat{D}$ is the corresponding degree matrix, H^l is the node feature matrix at layer l, W^l is the trainable weight matrix, and σ is a non-linear activation function (ReLU in our case).

Alternatively, this operation can be interpreted from a node-wise perspective as Eq. (11):

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{|N(i)|.|N(j)|}} W^{(l)} h_j^{(l)}\right) \quad (11)$$

This emphasizes how each node i updates its feature by aggregating normalized information from its neighbors N(i).

Using this, the final graph-enhanced embedding is Eq. (12):

$$d_i = GCN_{drug}(E_d, G_d),\ c_j = GCN_{cell}(E_c, G_c) \quad (12)$$

The output embedding for drugs and cell lines — denoted $d_i = GCN_{drug}(E_d, G_d)$ and $c_j = GCN_{cell}(E_c, G_c)$, respectively — are then concatenated with the corresponding omics feature vector $o_j \in R^O$ of the cell line, resulting in a fused representation [see Eq. (13)]:

$$z_{ij} = \left[d_i \| c_j \| o_j\right] \in R^{2e+o} \quad (13)$$

This vector z_{ij} is passed into a multi-layer perceptron (MLP) — a deep learning model for classification. Where the first hidden layer can be represented by Eq. (14):

$$h_1 = RELU(BN(W_1 z_{ij} + b_1)) \quad (14)$$

The second layer [see Eq. (15)]:

$$h_2 = RELU(W_2 h_1 + b_2) \quad (15)$$

To prevent overfitting, a dropout layer with a probability of p =3 is applied after the second fully connected layer. Dropout randomly deactivates a fraction p of neurons during training, encouraging the network to learn distributed representations rather than memorizing specific patterns [Eq. (16)].

$$\tilde{h}_2 = m \odot h_2,\ m \sim Bernoulli(1-p) \quad (16)$$

Finally, the transformed features are passed to a sigmoid output layer to produce the predicted probability of drug sensitivity, as in Eq. (17):

$$\hat{y}_{ij} = \sigma(W_3 \tilde{h}_2 + b_3) \quad (17)$$

where, $\hat{y}_{ij} \in [0.1]$ indicates the likelihood that cell line j responds to drug i.

E. Dynamic Focal Loss Function for Imbalanced Data

To address class imbalance inherent in drug response data, we implemented a custom focal loss function with dynamic weighting, where the per-sample weight α_d is given in Eq. (18):

$$\alpha_d = \frac{\#\text{ sensitive cell lines for drug d}}{\text{Total cell lines tested for drug d}} \quad (18)$$

This formulation increases the loss contribution from drugs with few positive (sensitive) samples, preventing the model from being biased toward majority-class drugs.

The overall loss for a sample (i.j) is in Eq. (19):

$$L_{focal} = \alpha_d.(1-p_t)^{\gamma}.BCE\left(\hat{y}_{ij}, y_{ij}\right) \quad (19)$$

where, $\hat{y}_{ij}$, y_{ij} are the predicted probability and ground truth label, respectively, $p_t = \hat{y}_{ij}$ if $y_{ij}=1$, otherwise $1-\hat{y}_{ij}$. And γ is the focusing parameter.

IV. Evaluation Criteria

The dataset was split into 80% training and 20% testing using stratified sampling, which is widely used in drug response prediction benchmarks to ensure sufficient training data while maintaining an unbiased test set. This setting evaluates the model's ability to generalize across unseen drug–cell line pairs under standard experimental conditions. Training was conducted for 100 epochs with early stopping based on validation AUC. All computations were accelerated via GPU. The model parameters were optimized using the Adam optimizer (lr = 0.001).

The model is evaluated on the test set using a comprehensive set of metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{Tp}{Tp + Fp}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = 2.\frac{\text{precision.Recall}}{\text{precision+Recall}}$$

where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. In addition to the metrics mentioned above, AUPR, AUC, and Specificity were also used.

V. Results

A. Data Description

Data filtering helped to reduce the missing values. For the IC50 response matrix, missing values reduced from 54.39% to 12.50%, and the Omics: RNA missing values reduced from 6.64% to 0.00%, CNV missing values reduced from 11.20% to 2.40%.

Since the GDSC dataset still includes missing entries, we compensated for them by applying a hybrid imputation approach, as outlined in the Materials and Methods section

(Section III). The total number of features and samples after this step is in Table I.

TABLE I. GDSC DATASET INFORMATION BEFORE AND AFTER PREPROCESSING

Dataset	State	Drugs	Cell lines	RNA	CNV	Mutation
GDSC	Raw	343	1084	17612	22903	68
	preprocessed	103	643	17611	22773	55

So we had omics data with more than 40 thousand features, which were integrated with the response matrix to calculate cell line similarities. This analysis produced a 551 × 551 cell line similarity matrix, reflecting the 551 common cell lines shared across the datasets. For the drug data, compounds lacking SMILES representations were excluded. MACCS fingerprints were then computed and combined with the response matrix to derive drug similarities, resulting in a 91 × 91 similarity matrix corresponding to the 91 drugs common across the datasets. Then we reduced the dimensionality of the omics data by applying PCA, ensuring that 80% of the total variance is preserved, reducing the number to 146 features.

For labeling, we used Cmax to convert IC50 values to labeled values. The GDSC dataset exhibits a moderately balanced class distribution, with 62.93% of drug-cell pairs classified as sensitive and 37.07% as resistant, yielding a 1.7:1 ratio. However, per-drug sensitivity rates vary significantly, with some drugs showing as low as 13% or as high as 76% sensitive cell lines, creating drug-specific imbalances (see Fig. 5). To ensure robust performance on both classes, particularly the clinically critical resistant cases, we employed focal loss with dynamic per-drug alpha (α=1-sensative fraction,γ=3). This approach weights the loss inversely to each drug's sensitivity rate, prioritizing underrepresented classes (e.g., resistant cases for highly effective drugs). As a result, the DrugCellGNN achieved balanced performance, with an AUPR of 0.9698, F1-score of 0.9228, and specificity of 0.7853, demonstrating effective handling of per-drug class variations and enhancing its utility for precision oncology.

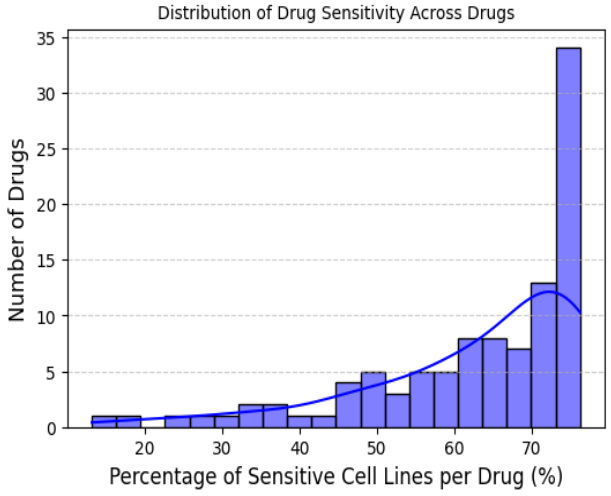

Fig. 5. Distribution of drug sensitivity across drugs.

B. Predictive Performance of the Model

To validate the effectiveness of our approach, we compared its predictive performance with that of the state-of-the-art models, including DSPLMF [7], Auto-HMM-LMF[12], CDSML [23], and MOICVAE [24]. All the methods mentioned above are classification models. All baseline models were evaluated using the same preprocessed dataset and identical train–test splits to ensure a fair comparison. The results in Table II show that in the GDSC cell line data set, the five indicators of our model framework were the highest.

TABLE II. PREDICTION PERFORMANCE OF DIFFERENT ALGORITHMS BASED ON SIX CRITERIA ON THE (GDSC) DATASET

Method	ACC	Rec	SPC	F1	AUC	Precision
DrugCellGNN	**0.895**	**0.941**	**0.807**	**0.921**	**0.951**	**0.902**
AutoBorutaRF	0.653	0.652	0.654	0.650	0.711	0.646
DSPLMF	0.682	0.750	0.615	0.702	0.760	0.671
MOICVAE	0.772	0.787	-	0.775	0.856	0.764
CDSML	0.838	0.9031	-	0.8715	0.9157	0.842
Auto-HMM-LMF	0.70	0.78	0.63	0.73	0.78	0.77

VI. CONCLUSION AND DISCUSSION

In this study, we introduced DrugCellGNN, a pioneering Graph Neural Network framework that revolutionizes cancer drug sensitivity prediction by seamlessly integrating multi-omics data—encompassing gene expression, copy number variations, and mutations—with structural and functional similarities between drugs and cell lines. Leveraging the GDSC dataset, our pipeline addresses key challenges in pharmacogenomics, including data sparsity through hybrid imputation, high dimensionality via PCA reduction, and class imbalance with dynamic focal loss. Evaluated on 50141drug-cell pairs (80/20 train-test split), DrugCellGNN achieved an AUC of 0.9516, F1-score of 0.9228, and AUPR of 0.9698, outperforming traditional baselines Random Forest and SVM by leveraging relational dependencies. These results highlight DrugCellGNN's ability to model complex interactions, enhancing prediction accuracy for rare sensitive cases critical in precision oncology. A key factor contributing to the effectiveness of DrugCellGNN is the integration of functional and structural similarity information for both drugs and cell lines. By combining response-based similarities with molecular and chemical features, the constructed graphs provide a biologically meaningful representation of drug–cell interactions. This fusion strategy allows the model to leverage complementary information, balancing phenotypic drug response patterns with intrinsic biological characteristics of cancer cell lines and chemical properties of drugs.

Future work should aim to expand this framework to other datasets, such as CCLE or CTRP, to evaluate how well it performs across different data sources. Using more advanced molecular representations (like transformer-based encoders or molecular graph embeddings) could boost predictive accuracy. Adding more layers of omics data, such as epigenomic or proteomic information, might also deepen the biological understanding. Lastly, integrating explainability features into

the GNN could reveal mechanistic insights by pinpointing which molecular features or graph edges most influence the predictions.

REFERENCES

[1] C. Tomasetti, L. Li, and B. Vogelstein, "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention," Science (80-.)., vol. 355, no. 6331, pp. 1330–1334, 2017.

[2] A. C. Begg, F. A. Stewart, and C. Vens, "Strategies to improve radiotherapy with targeted drugs," Nat. Rev. Cancer, vol. 11, no. 4, pp. 239–253, 2011.

[3] D. M. Roden and A. L. George Jr, "The genetic basis of variability in drug responses," Nat. Rev. Drug Discov., vol. 1, no. 1, pp. 37–44, 2002.

[4] L. Wang, X. Li, L. Zhang, and Q. Gao, "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization," BMC Cancer, vol. 17, pp. 1–12, 2017.

[5] W. Yang et al., "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," Nucleic Acids Res., vol. 41, no. D1, pp. D955–D961, Jan. 2013, doi: 10.1093/nar/gks1111.

[6] J. Choi, S. Park, and J. Ahn, "RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance," Sci. Rep., vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-58821-x.

[7] A. Emdadi and C. Eslahchi, "DSPLMF: A Method for Cancer Drug Sensitivity Prediction Using a Novel Regularization Approach in Logistic Matrix Factorization," Front. Genet., vol. 11, no. February, pp. 1–14, 2020, doi: 10.3389/fgene.2020.00075.

[8] F. Ahmadi Moughari and C. Eslahchi, "ADRML: anticancer drug response prediction using manifold learning," Sci. Rep., vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-71257-7.

[9] X. Xu, H. Gu, Y. Wang, J. Wang, and P. Qin, "Autoencoder based feature selection method for classification of anticancer drug response," Front. Genet., vol. 10, no. MAR, 2019, doi: 10.3389/fgene.2019.00233.

[10] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," J. Stat. Softw., vol. 36, pp. 1–13, 2010.

[11] Z. Dong et al., "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," BMC Cancer, vol. 15, no. 1, Jun. 2015, doi: 10.1186/s12885-015-1492-6.

[12] A. Emdadi and C. Eslahchi, "Auto - HMM - LMF : feature selection based method for prediction of drug response via autoencoder and hidden Markov model," BMC Bioinformatics, pp. 1–22, 2021, doi: 10.1186/s12859-021-03974-3.

[13] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," Briefings in Bioinformatics, vol. 22, no. 1. Oxford University Press, pp. 360–379, Jan. 01, 2021. doi: 10.1093/bib/bbz171.

[14] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, "DeepCDR: a hybrid graph convolutional network for predicting cancer drug response," Bioinformatics, vol. 36, no. Supplement_2, pp. i911–i918, 2020.

[15] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," Genome Biol., vol. 15, pp. 1–12, 2014.

[16] C. Huang, R. Mezencev, J. F. McDonald, and F. Vannberg, "Open source machine-learning algorithms for the prediction of optimal cancer drug therapies," PLoS One, vol. 12, no. 10, p. e0186906, 2017.

[17] B. Haibe-Kains, "PSet_GDSC2020." Zenodo, Apr. 2023. doi: 10.5281/zenodo.7829915.

[18] E. Anderson, G. D. Veith, and D. Weininger, SMILES, a line notation and computerized interpreter for chemical structures. US Environmental Protection Agency, Environmental Research Laboratory, 1987.

[19] S. Kim et al., "PubChem in 2021: new data content and improved web interfaces," Nucleic Acids Res., vol. 49, no. D1, pp. D1388–D1395, 2021.

[20] G. Landrum, "Rdkit documentation," Release, vol. 1, no. 1–79, p. 4, 2013.

[21] I. Jolliffe, "Principal component analysis," in International encyclopedia of statistical science, Springer, 2011, pp. 1094–1096.

[22] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," J. Chem. Inf. Comput. Sci., vol. 42, no. 6, pp. 1273–1280, 2002.

[23] F. A. Moughari and C. Eslahchi, "A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information," PLoS One, vol. 16, no. 4 April 2021, Apr. 2021, doi: 10.1371/journal.pone.0250620.

[24] C. Wang, M. Zhang, J. Zhao, B. Li, X. Xiao, and Y. Zhang, "The prediction of drug sensitivity by multi-omics fusion reveals the heterogeneity of drug response in pan-cancer," Comput. Biol. Med., vol. 163, no. June, p. 107220, 2023, doi: 10.1016/j.compbiomed.2023.107220.