이학석사 학위논문

DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer

DRPreter: 지식 기반 그래프 신경망과 트랜스포머를 활용한 해석 가능한 항암 약물 반응 예측

2023 년 8 월

서울대학교 대학원

협동과정 생물정보학전공

신 지 혜

DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer

DRPreter: 지식 기반 그래프 신경망과 트랜스포머를 활용한 해석 가능한 항암 약물 반응 예측

지도교수 김 선

이 논문을 이학석사 학위논문으로 제출함

2023 년 6 월

서울대학교 대학원

협동과정 생물정보학전공

신 지 혜

신지혜의 이학석사 학위논문을 인준함

2023 년 6 월

위 원 장	황대희
부위원장	김선
위 원	임상수

Abstract

DRPreter: Interpretable Anticancer Drug Response Prediction Using Knowledge-Guided Graph Neural Networks and Transformer

Jihye Shin
Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University

Some of the recent studies on drug sensitivity prediction have applied graph neural networks to leverage prior knowledge on the drug structure or gene network, while other studies focus on the interpretability of the model to delineate the mechanism governing the drug response. However, it is crucial to make a prediction model that is both knowledge-guided and interpretable, so that the prediction accuracy is improved and also practical use of the model can be enhanced. I propose an interpretable model called DRPreter (Drug Response PREdictor and interpreTER) that predicts anticancer drug response.

DRPreter learns cell line and drug information with graph neural networks where the cell line graph is further divided into multiple subgraphs with domain knowledge on biological pathways. Transformer encoder-based structure in DRPreter helps detect relationships between pathways and a drug, highlighting important pathways that are involved in the drug response. Extensive experiments on the GDSC (Genomics of Drug Sensitivity and Cancer) dataset demonstrate that the proposed method outperforms state-of-the-art graph-based models for drug response prediction. In addition, DRPreter detected putative key genes and pathways for specific drug-cell line pairs with supporting evidence in the literature, implying that the model can help interpret the mechanism of action of the drug.

Keywords: transcriptomics, Explainable AI, pharmacogenomics, cancer drug sensitivity, Graph Neural Networks, precision medicine, drug discovery

Student Number: 2021-24949

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Drug Response Prediction

The advances in technology and scientific capability enable the acquisition of large amounts of personal omics data at reduced cost (Kellogg *et al.*, 2018). Consequently, there is a growing interest in using individualized health data for precision medicine, leading to a number of data-driven healthcare models (Ahmed, 2020). Pharmacogenomics, one of the branches of precision medicine, is the study of how a person's genetic profile influences their response to medications (Kalamara *et al.*, 2018; Singh, 2019). Prediction of drug response or efficacy using the omics data of patients before the actual treatment is crucial because it can help increase clinical success and minimize adverse drug effects by modifying dosages or selecting alternative medications based on predicted value for personalized chemotherapy. However, obtaining patients' tumor tissues by surgical procedure or biopsy involves safety issues (Cho, 2020), and performing animal experiments for clinical trials to infer human drug efficacy leads to ethical and financial concerns (Singh *et al.*, 2016). In addition, even though correlating drug response and omics data can help improve understanding of drug mechanisms of action (Rees *et al.*, 2016), many candidate drugs still fail to enter clinical trials during the drug discovery process due to an incomplete understanding of the mechanisms (Seyhan, 2019; Kuenzi *et al.*, 2020). In this respect, an interpretable *in silico* model for drug response prediction would be useful for numerous healthcare purposes, especially for precision medicine and drug discovery (Savage, 2021).

1.1.2 Related works and limitations

Molecular profiles of cancer cell lines and high throughput drug sensitivity screening databases are publicly available (Shoemaker, 2006; Barretina *et al.*, 2012; Yang *et al.*, 2012; Basu *et al.*, 2013; Seashore-Ludlow *et al.*, 2015; Iorio *et al.*, 2016) including CCLE (Cancer Cell Line Encyclopedia) (Barretina *et al.*, 2012) and GDSC (Genomics of Drug Sensitivity in Cancer) (Yang *et al.*, 2012; Iorio *et al.*, 2016). Public databases and improved computing technologies such as machine learning and deep learning have contributed to the rapid development of models for predicting anticancer drug sensitivity from cancer cell lines based on their genetic profiles.

The early studies in drug sensitivity prediction utilized machine learning techniques (Güvenç Paltun *et al.*, 2021; Adam *et al.*, 2020; Firoozbakht *et al.*, 2022) such as a random forest (Riddick *et al.*, 2011), support vector machine (Dong *et al.*, 2015), and matrix factorization (Wang *et al.*, 2017; Guan *et al.*, 2019). However, the traditional machine learning-based models can still be improved in terms of predictive performance and generalizability (Kalamara *et al.*, 2018; Baptista *et al.*, 2021). Matrix factorization-based models leave nonlinear relationships unaddressed because they attempt to identify interactions between the drug and cell line using linear combinations of latent features. With the capability of learning complex nonlinear functions and high dimensional representations from raw data, various deep learning techniques have been utilized for predicting drug response (Baptista *et al.*, 2021) and the overall predictive power of drug sensitivity has been improved (Sakellaropoulos *et al.*, 2019). DeepDR (Chiu *et al.*, 2019) and MOLI (Sharifi-Noghabi *et al.*, 2019) are drug-specific models that only use cell profiles such as somatic mutation, gene expression, or copy number variation to predict IC_{50} values of each sample. tCNNs (Liu *et al.*, 2019) introduced a model to predict drug sensitivity for drug-cell pairs using SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) sequences as drug features in addition to the genomic profiles of cells. The models described above used vector representations in common for describing cell or drug features.

Graph-based approaches have been introduced in drug response prediction models to take advantage of the structural information of a drug or a gene network. A drug can be represented as a molecular graph consisting of a set of atoms (nodes) and a set of bonds (edges) and the graph is transformed into a high-level representation by a neural network (Liu *et al.*, 2020; Nguyen *et al.*, 2021). For example, GraphDRP (Nguyen *et al.*, 2021) obtained drug embeddings using graph convolutional networks, while cell line embeddings were derived from binary vectors of genomic aberrations. The state of a cell line can also be characterized as a gene-gene interaction network where genes (nodes) have node features from omics data such as gene expression values (Kim *et al.*, 2021; Zhu *et al.*, 2022; Feng *et al.*, 2021). Zhu *et al.* proposed an end-to-end drug response prediction model TGDRP with cell line graph embedding consisting of genes that hold cancer-related mutations and drug graph embedding obtained by a graph neural network. They also proposed TGSA which updates embeddings from TGDRP with similarity information between cell lines and drugs and predicts drug response again.

While recent studies described above have introduced graphs into the deep learning models to leverage structural information and improve prediction accuracy, the models lack interpretability of the predicted results. Several methods tried to delineate the mechanism governing the drug response, highlighting the important genes or high-level subsystems such as biological pathways that can cause changes in cellular phenotype. SWnet (Zuo *et al.*, 2021) explored the interactions between genetic profile and the chemical structure of drugs using self-attention and identified genes with the strongest predictive power. Deng *et al.* proposed a multilayer perceptron model called pathDNN which incorporates a layer of pathway nodes and quantified the activity of each pathway to explain their effect on drug response. DrugCell (Kuenzi *et al.*, 2020) obtained binary encodings of mutational status via a visible neural network guided by a hierarchy of cell subsystems and measured the predictive performance of the subsystems. Although pathDNN and DrugCell attempted to construct an explainable model with a hierarchical structure, biological pathways were implemented as gene sets rather than gene networks, indicating that domain knowledge in gene-gene interactions was not fully reflected in the models.

1.2 Problem Definition

In this study, a regression model was developed to predict the half maximal inhibitory concentration (IC_{50}), normalized to natural logarithms, in cell line-drug pairs. IC_{50} serves as a representative indicator of drug sensitivity, quantifying the effectiveness of a drug in inhibiting a particular biological or biochemical process. It represents the concentration at which a drug can achieve 50% inhibition of the target activity.

1.3 Motivation and Contributions

According to the existing studies that suggest deep learning models for drug response prediction, it is helpful to incorporate the graph representation for both drug and cell line profiles that enables a detailed description of compound structure and gene network. Moreover, the gene network can be dissected as a set of biological pathways that include gene-gene interactions for each specific biological mechanism, which can help enhance both prediction accuracy and interpretability. However, current interpretable models for drug response prediction simply describe the network as gene-pathway layers, leaving the interaction information inside the biological pathways unused. Here, I propose a novel anticancer drug response prediction model named DRPreter (Drug Response PREdictor and interpreTER) with key features as follows:

1. **Knowledge-guided cell representation with graphs.** DRPreter constructs a cell line network as a set of subgraphs that correspond to cancer-related pathways for the detailed representation of the biological mechanism.

2. **Interpretability of drug mechanisms of action.** Using Transformer's encoder, the interactions between drugs and pathways are derived from the model and putative key pathways for the drug mechanism can be highlighted.

3. **Enhanced performance.** The proposed method DRPreter outperforms state-of-the-art drug response prediction models demonstrated by comparative experiments on the GDSC drug sensitivity dataset.

The following is a description of the graph configuration for cell lines and drugs and the graphical abstract of DRPreter (Figure 1.1).

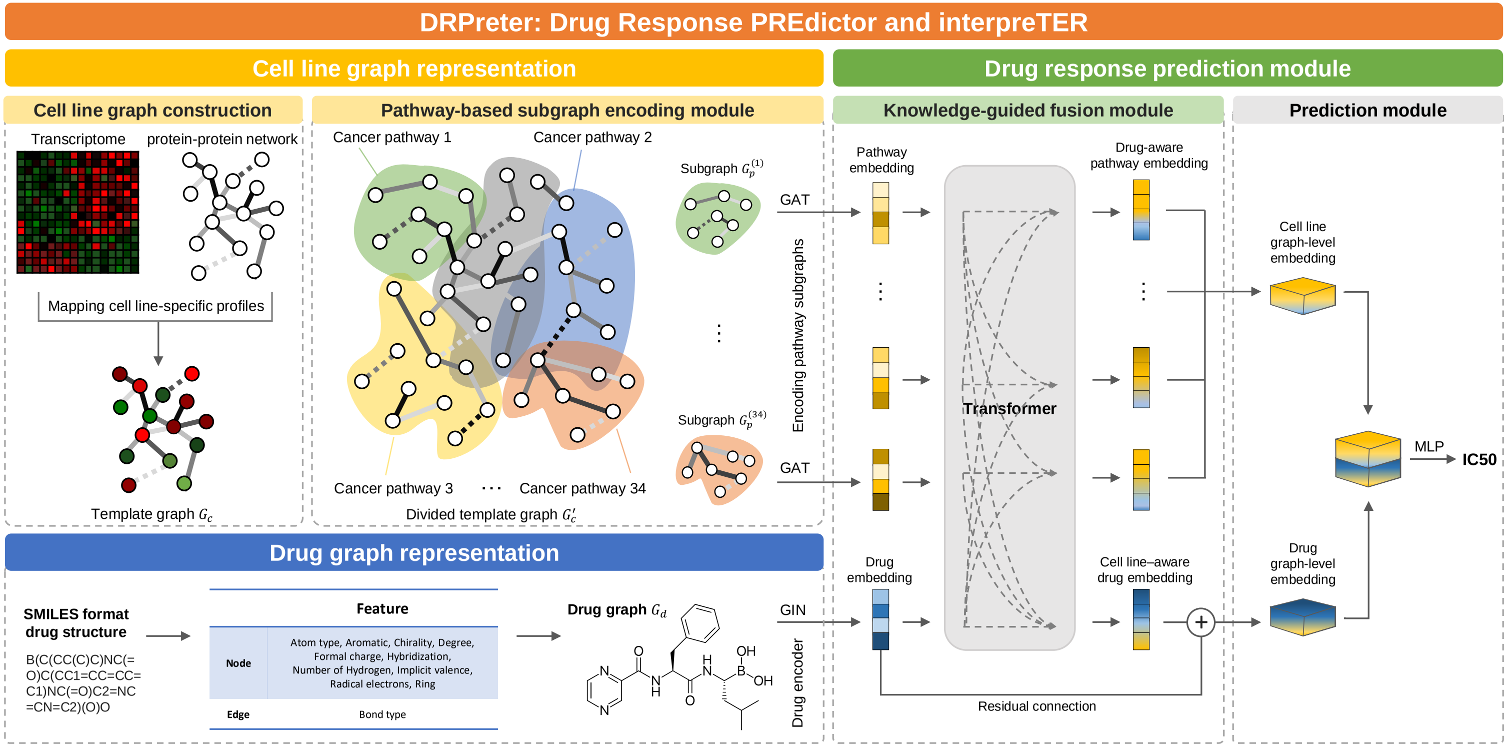
DRPreter: Drug Response PREdictor and interpreTER
Cell line graph representation
Drug response prediction module
Cell line graph construction
Pathway-based subgraph encoding module
Knowledge-guided fusion module
Prediction module
Transcriptome
protein-protein network
Mapping cell line-specific profiles
Template graph G_c
Cancer pathway 1
Cancer pathway 2
Cancer pathway 3
Cancer pathway 34
Divided template graph G'_c
Subgraph $G_p^{(1)}$
Subgraph $G_p^{(34)}$
GAT
Encoding pathway subgraphs
Pathway embedding
Transformer
Drug-aware pathway embedding
Cell line graph-level embedding
MLP
IC50
Drug graph representation
SMILES format drug structure
B(C(CC(C)C)NC(=O)C(CC1=CC=CC=C1)NC(=O)C2=NC=CN=C2)(O)O
Feature
Node
Atom type, Aromatic, Chirality, Degree, Formal charge, Hybridization, Number of Hydrogen, Implicit valence, Radical electrons, Ring
Edge
Bond type
Drug graph G_d
GIN
Drug encoder
Drug embedding
Cell line–aware drug embedding
Residual connection
Drug graph-level embedding

Figure 1.1: An overview of DRPreter. In the graph representation sections, embeddings of pathway subgraphs and a drug molecule were obtained using GNN. With the obtained pathway embeddings and drug embeddings as inputs to the transformer-based cell line and drug fusion module, the embeddings were updated by reflecting inter-pathway relationships and pathway-drug relationships in the model learning process.

Chapter 2

Materials and Methods

2.1 Graph Neural Networks

A graph neural network (GNN) is a type of neural network that operates on graph-structured data. GNN uses the topology of the graph to learn the relationships between the input features. It can perform more effectively than other representation learning methods on input data with topological information. In this study, I represent a graph as $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ is the set of n nodes and $E \subseteq V \times V$ is the set of edges. The node v_i has node feature $x_i \in \mathbb{R}^d$, where d is a dimension of the feature. The node feature matrix of the graph can be represented as $X \in \mathbb{R}^{n \times d}$, where n is the number of nodes in the graph. Adjacency matrix $A \in \mathbb{R}^{n \times n}$ indicates the total connectivity of nodes in the graph, where $A_{i,j} = 1$ means nodes, v_i and v_j are linked, and $W^{(l)}$ represents the parameters of the l-th layer of the graph (Table 2.1).

In each GNN layer, a key mechanism, called message passing, updates the node representation by using the node features of the previous layer and the topology of the graph (Gilmer *et al.*, 2017). The message-passing mechanism

Table 2.1: Notation of graph neural networks used in this paper.

Notation	Description
G	A graph.
V	Set of nodes of a graph.
v	A node included in V.
i, j	Indexes of the nodes.
l	Index of the layer of a graph.
v_i	i-th node in V.
x_i	Node feature of node v_i
$N(i)$	Set of neighbor nodes of a node v_i
E	Set of edges of a graph.
A	Adjacency matrix between nodes.
$W^{(l)}$	Trainable parameter matrix of l-th layer.
$X^{(l)}$	Node feature matrix of l-th layer.
σ	Nonlinear activation function softmax.
ϵ	Learnable parameter.

involves aggregating the information of neighboring nodes and updating the hidden state of each node by combining the node representation from the previous layer and the aggregated messages. For every node in each layer, a transformed feature vector is generated capturing the structural information of the k-hop neighbor nodes. The GNN can update the i-th node representation in the l-th layer as in the following Equation (Li *et al.*, 2021; Dai *et al.*, 2022), where $N(i)$ is the set of neighbor nodes linked to the target node i. For a given node, the AGGREGATE step applies a permutation invariant function to its neighboring nodes to produce the aggregated node feature of neighbors, and the COMBINE step delivers the aggregated node feature to the learnable layer to produce updated node embedding by integrating the existing embedding and the aggregated neighbor embedding.

$$x_i^{(l)} = COMBINE^{(l)}\left(x_i^{(l-1)}, AGGREGATE^{(l-1)}\left(x_j^{(l-1)} : j \in N(i)\right)\right) \quad (2.1)$$

2.2 Cell line Graph Representation

2.2.1 Cell line Graph Construction

I used a biological template network to represent cell lines to simulate gene–gene interactions in actual cells. In the cell line graph $\mathbf{G}_c$, genes are represented by nodes and edges represent the relationships between genes. This template graph contained 2,369 genes selected using the pathway selection method described in the dataset section.

It is known that drugs do not have a universal effect throughout all cellular components, but tend to have distinct effects on specific genes or pathway targets. In this way, cancer cells undergo phenotypic changes as a result of drug molecules inhibiting or activating their target pathways. Motivated by this point, instead of representing the cell line as a homogeneous large-scale graph that contains the entire genes, I divided the template network $\mathbf{G}_c$ into pathway subgraphs $\mathbf{G}_p$ according to the biological domain knowledge inspired by Lee *et al.* and learned graph embeddings from the selected subgraph units. Finally, the divided cell line graph $\mathbf{G}_c^{'}$ was represented as a heterogeneous graph containing multiple subgraphs. I selected pathways that can be targeted by drugs, as they are associated with cancer from the KEGG pathway database (Kanehisa and Goto, 2000), and used these pathways as pre-defined subgraphs of the cell line template.

In the case of template graph $\mathbf{G}_c$, the i-th selected pathway subgraph can be described as $\mathbf{G}_p^{(i)} = (\mathbf{V}_p^{(i)}, \mathbf{E}_p^{(i)})$, where $\mathbf{V}_p^{(i)}$ refers to a set of nodes and $\mathbf{E}_p^{(i)}$ refers to a set of edges of the pathway. Thus, the template graph $\mathbf{G}_c$ is extended as a union of pathway subgraphs, with overlaps between the pathways in the form of $\mathbf{G}_c^{'} = \{\mathbf{G}_p^{(1)}, ..., \mathbf{G}_p^{(34)}\}$. In the template cell line graph $\mathbf{G}_c$, gene sets included in 34 pathways were represented by 2369 nodes and 7954 edges. A divided template graph $\mathbf{G}_c^{'}$ with pathways as subgraphs had

4646 nodes and 12,004 edges after combining the data from all pathways. The types of constituting genes remained the same, but the numbers of nodes and edges increased when the template network was divided into subnetworks due to the overlap of functions.

2.2.2 Cell line Graph Encoder on Pathway Subgraphs

Transcriptomic features of nodes and biological network topology were captured within each subgraph using Graph Attention Network (GAT) (Veličković *et al.*, 2017). Using the self-attention mechanism, GAT calculates a normalized attention score α_{ij} indicating the importance of the features of the neighbor nodes for a target node i, where $j \in N(i)$. A subsequent step in the message-passing process is for each node to reflect the importance of the neighboring nodes' information in accordance with the previously obtained attention scores.

$$X^{(l)} = \sigma\left(\Sigma_{j \in N(i)} \alpha_{ij}^{(l-1)} W^{(l-1)} X^{(l-1)} \right) \tag{2.2}$$

If template graph $\mathbf{G}_c$ is used as it is, edges connected to one gene include interactions from multi pathways which can be noise. Node representations were updated through GAT on the cell line graph constructed in the previous subsection. The cell line graph consists of pathway-based subgraphs, thus the updated node representation can reflect the intra-pathway gene-gene interaction information. To pool the cell line graph-level embedding, I initially used a simple hierarchical permutation-invariant graph pooling strategies (Zhang *et al.*, 2018; Gao and Ji, 2019; Lee *et al.*, 2019). However, the graph pooling strategies I employed resulted in slight performance degradation. I assumed that this may be due to the relatively large size of the cell line graph, and simply pooling the nodes into a vector of the same dimension may lose the information of the nodes in the cell line. As a result, the embeddings of each

Table 2.2: Atomic and bond features of drug graph.

	Feature	Size	Description
Node	Atom type	43	[B, C, N, O, F, ...] (one-hot)
	Aromatic	1	Whether the atom is in the aromatic system (binary)
	Chirality	2	[R, S] (one-hot or null)
	Degree	11	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10] (one-hot)
	Formal charge	1	electric charge (integer)
	Hybridization	5	[sp, sp^2, sp^3, sp^3d, sp^3d^2] (one-hot or null)
	Number of Hydrogens	5	[0, 1, 2, 3, 4] (one-hot)
	Implicit valence	7	[0, 1, 2, 3, 4, 5, 6] (one-hot)
	Radical electrons	1	Number of radical electrons (integer)
	Ring	1	Whether the atom is in ring (binary)
Edge	Bond type	4	[single, double, triple, aromatic] (one-hot)

node learned through GAT were concatenated to form graph-level embedding for each pathway.

2.3 Drug Graph Representation

2.3.1 Drug Graph Construction

I used a graph neural network to learn the drug representation by reflecting the relationships between atoms connected by bonds and the overall molecular structural information. A drug can be represented as a graph, in which atoms are nodes, and bonds are edges. I used RDKit (Landrum *et al.*, 2013) to transform SMILES (Weininger, 1988), a one-dimensional string format drug structure, into graph format that can reflect structural information of an actual drug. The ten initial features of atomic nodes were imported from previous research (Liu *et al.*, 2020; Zhu *et al.*, 2022), which predicted drug sensitivity from GNN-based embeddings of drug structures. The details of atomic and bond features can be found in Table 2.2.

2.3.2 Drug Graph Encoder

A graph isomorphism network (GIN) (Xu *et al.*, 2018) was used to learn the features of the atomic nodes within the drug graph. GIN applies a neighborhood aggregation method similar to the Iisfeiler-Lehman test (Weisfeiler and Leman, 1968) and updates the i-th node feature at the l-th layer as follows.

$$x_i^{(l)} = MLP^{(l)}\left(\left(1+\epsilon^{(l)}\right)\cdot x_i^{(l-1)} + \Sigma_{j\in N(i)} x_j^{(l-1)}\right) \tag{2.3}$$

The graph encoder was chosen following the results of GraphDRP, which involves a comparison of different types of graph neural networks, GIN, GAT, and GCN+GAT in order to analyze the effectiveness of each graph encoder in predicting drug response. In addition, GIN is widely used for the embeddings of drug graphs in various drug response prediction models (Feng *et al.*, 2021; Nguyen *et al.*, 2021; Zhu *et al.*, 2022; Zheng *et al.*, 2022).

2.4 Knowledge-guided Cell line-Drug Fusion Module Using Transformer

2.4.1 Knowledge-guided Fusion Module

The transformer model tracks relationships in sequential data, like the words in the sentence, to discover context and meaning between the components (Vaswani *et al.*, 2017). I used a Transformer-based module to reflect not only inter-pathway interactions but also the interactions between the pathways and each drug, which allows for exploring the pharmacological mechanisms of action at the pathway level during a therapeutic process (Figure 2.1).

The model structure has a single encoder-based layer taking pathway embeddings $X_p^{(l)}$ $(l = 1, ..., 34)$ and a drug embedding X_d derived from knowledge-guided GNNs as input values. Inputs in a typical Transformer's encoder are constructed by adding positional encoding to embeddings of source sequences.

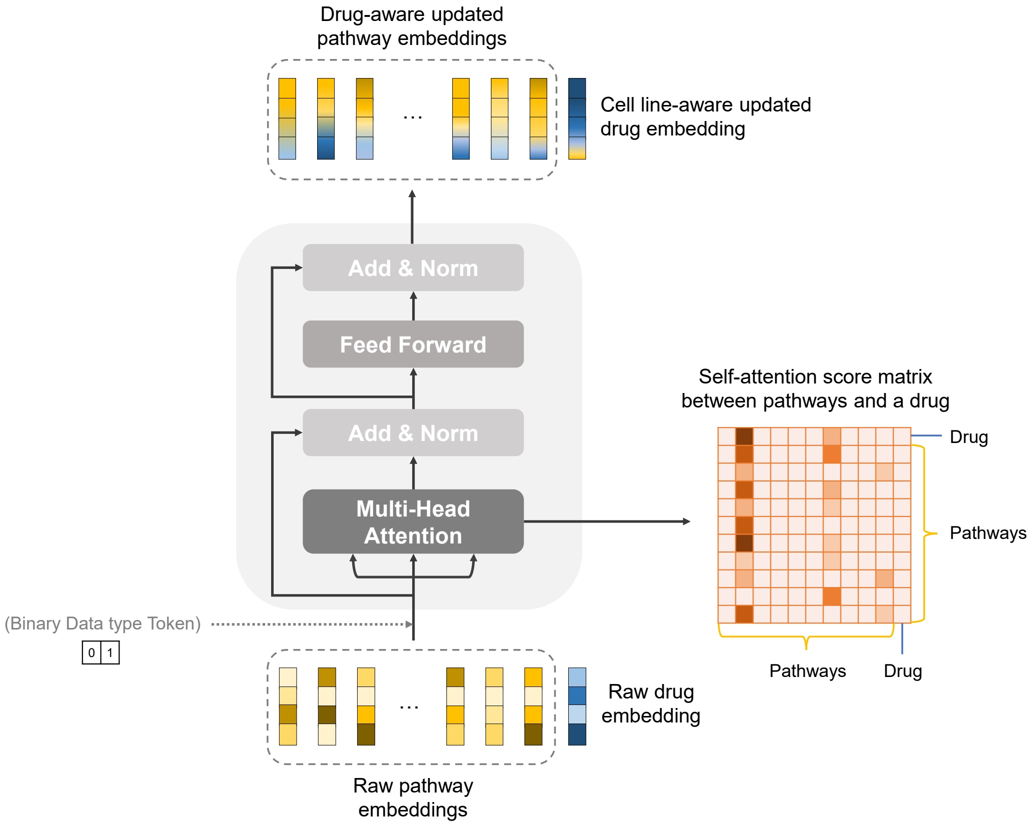

Figure 2.1: A detailed structure of type-aware Transformer encoder reflecting interactions and relationships between pathways and a drug. I extracted drug-pathway interaction information from the modified encoder of the Transformer module and identified putative key pathways for the drug's mechanism of action using a matrix of self-attention scores between pathways and the drug.

Unlike translation, where an order of words in a sentence is important, the pathway embeddings entering the encoder are not affected by the order in which they are encoded, so a Transformer's encoder structure other than positional encoding was used for this study. As an alternative, I added a type encoded token that indicates whether the embedding is a drug or a pathway. In an element-wise manner, type encoded binary tokens are added to the input feature matrix of the same dimension before input embeddings are fed into the module.

To fusion pathway and drug embeddings, self-attention was performed several times through multi-head attention, and the average of each trial was used as the final attention score. After the encoder has completed its execution, the encoder produces drug-aware updated pathway embeddings $X_p'^{(l)}$ and pathways' transcriptome-aware updated drug embedding $X_d^{'}$ reflecting interaction information. These drug-aware pathway embeddings facilitate interpreting the medication's mechanism of action since it can reflect both the drug-pathway interaction information as well as the interaction between the pathways. Drugs have a large structural variation when compared to cell line graphs which are composed of the same genes and structurally equal but have different node feature values. Therefore, it is possible that the variation of the drug embedding may be blurred because the new drug embedding updated as a result of the Transformer is affected by the cell line embedding. Hence, I connect the raw drug embedding obtained through GNN prior to the Transformer structure with the updated drug embedding obtained after the Transformer using residual connection (He *et al.*, 2016). By residual connection, it is possible to preserve the original drug structure information and utilize the cell line-drug interaction information using the updated drug embedding which recognizes the transcriptomic information of each pathway. I concatenated the resulting 34 subgraph embeddings in order to prevent information loss, thereby embedding the entire cell line.

2.5 Improving Predictive Performance using Similarity Graph

Based on the idea that similar drugs and similar cell lines exhibit interchangeable drug response behaviors, some drug response prediction models use prior knowledge of drug and cell line similarity to minimize differences between drugs and cell lines in the latent space. Wang *et al.* applied regularization terms based on chemical structural similarities between drugs and similarities between cell lines based on gene expression profiles to improve prediction accuracy and prevent overfitting.

I followed the similarity-based embedding updating strategy of Zhu *et al.*. From the completed end-to-end model up to Section 2.3, embeddings of all 580 cell lines and 170 drugs can be made. Then I constructed two homogeneous graphs each consisting of cell lines and drug nodes, with the initial feature of each node set as the resulting embeddings of the previous step. Using GraphSAGE (Hamilton *et al.*, 2017), I updated the embeddings of each homogeneous cell line and drug graph. After that, updated embeddings of cell line-drug pair which is aimed to obtain response were derived respectively from two homogeneous graphs. I concatenated two embeddings into the one-dimensional vector and used a multi-layer perceptron to predict final IC_{50} values.

Chapter 3

Results

3.1 Performance Comparison

3.1.1 Dataset

Among the publicly available databases that offer insights into drug response, certain databases provide gene expression data both before and after drug administration, such as LINCS (Library of Integrated Network-based Cellular Signatures) (Subramanian *et al.*, 2017). However, LINCS is limited in terms of gene expression information, as it only covers approximately 1,000 genes that correspond to landmark genes. Consequently, it becomes unfeasible to obtain gene expression values for all the pathway-related genes necessary for the intended analysis. In order to address this limitation, CCLE data was employed as an alternative. Although CCLE does not provide gene expression data after drug administration, it offers gene expression values for a larger number of genes, enabling a broader coverage for the analysis. In the cell line template graph, the initial feature of each gene node was derived from transcriptomic data of each cell line obtained from the CCLE database version of 21Q4 (DepMap, 2021) (`https://portals.broadinstitute.org/ccle`, accessed on 3 December 2021). The gene expression data were TPM values of the protein-coding genes for DepMap cell lines, which were inferred from RNA-seq data using the RSEM tool and were provided after log2 transformation, using a pseudo-count of 1; log_2(TPM+1) (DepMap, 2021). I assigned edges of the graph as only those interactions with high reliability scores and a combined score of at least 990 among the STRING (v11.5) (Szklarczyk *et al.*, 2019) protein–protein interactions. The edges of the template graph and each subgraph were all STRING protein–protein interactions. Only the genes corresponding to each cancer-related pathway were obtained in KEGG, and the genes corresponding to each pathway were used as nodes in the subgraph. The STRING interactions were used as the edges connecting them. Pathways for constructing subgraphs were selected in the following manner. The non-processed pathways listed in categories 6.1 and 6.2. of the KEGG pathway database (`https://www.genome.jp/kegg/pathway.html`, accessed on 16 April 2022) were categorized according to the cancer types. These pathways include common subpathways related to cell signaling, the cell cycle, and apoptosis, which are key in various types of cancer. Consequently, if the cancer pathways provided by KEGG are used as they are, the overlap between the pathways will be excessive, and the meaning of learning for each pathway diminishes. Additionally, KEGG provides information on detailed pathways associated with each cancer type pathway. There were a total of 84 detailed pathways categorized by function. Among these pathways, I eliminated duplicate pathways, metabolic pathways, non-cancer disease pathways, viral infection pathways, and pathways with fewer than 10 genes or gene–gene interaction edges. Furthermore, the focal adhesion pathway (hsa04510) was also eliminated because 91% of the genes constituting this pathway were included in the remaining pathways. To mitigate potential bias in the embedding process arising from variations in pathway sizes, an assessment of each pathway's size was necessary. The selected pathways encompassed a range of 41 to 351 genes, with an average size of 103 genes. Of particular concern was PI3K-Akt signaling pathway (hsa04151), which has the potential to be exhibited a significant bias due to its large size of 351 genes. To ascertain whether this pathway was unconditionally regarded as important in the Transformer-based structure, self-attention score-based pathway interpretations were conducted. The subsequent interpretation results did not highlight PI3K-Akt signaling pathway as the most influential pathway (Figure 3.2, Figure 3.4). Consequently, it can be concluded that the differing scales of the selected pathways in this study did not have a substantial impact on the obtained results. The finally selected 34 cancer-related detailed pathway list can be found in Table 3.1. For drug graph construction, I obtained SMILES strings from PubChem (Wang *et al.*, 2009).

For the performance comparison experiment, the nodes constituting the cell line graph of the existing GNN-based drug response prediction models were configured according to the settings in each comparison paper. I compared performance with state-of-the-art GNN-based drug response prediction models: GraphDRP, TGDRP, and TGSA. As the initial feature for each gene node, the GraphDRP uses mutation (mut) and copy number variation (cnv), and TGDRP and TGSA also use mut and cnv with gene expression (exp). As GraphDRP represents cell lines as one-dimensional binary vectors, one-dimensional CNN is used to get their embeddings. Cell lines and drugs are represented in graph format in TGDRP and TGSA, and the embeddings are obtained by GNN. The cancer driver genes from COSMIC were selected as the genes to represent the cell lines in all baseline models (Sondka *et al.*, 2018). The COSMIC database provides information about mutation-containing genes involved with cancer, as well as how these muta-

Table 3.1: A list of cancer-related pathways used as subgraphs in cell line template graph.

Pathway name	KEGG identifier	Number of genes	Number of edges
Ubiquitin mediated proteolysis	hsa04120	142	534
TGF-β signaling pathway	hsa04350	94	228
Estrogen signaling pathway	hsa04915	137	222
MAPK signaling pathway	hsa04010	294	692
PPAR signaling pathway	hsa03320	74	28
mTOR signaling pathway	hsa04150	155	688
Regulation of actin cytoskeleton	hsa04810	218	552
B cell receptor signaling pathway	hsa04662	79	208
Cell adhesion molecules	hsa04514	146	150
Chemokine signaling pathway	hsa04062	190	514
Apoptosis	hsa04210	136	424
Cytokine-cytokine receptor interaction	hsa04060	293	588
Wnt signaling pathway	hsa04310	167	384
p53 signaling pathway	hsa04115	73	180
Ras signaling pathway	hsa04014	232	600
Notch signaling pathway	hsa04330	59	76
Calcium signaling pathway	hsa04020	239	218
HIF-1 signaling pathway	hsa04066	109	204
T cell receptor signaling pathway	hsa04660	104	336
ErbB signaling pathway	hsa04012	85	326
Cell cycle	hsa04110	126	1076
Melanogenesis	hsa04916	101	110
cAMP signaling pathway	hsa04024	221	222
VEGF signaling pathway	hsa04370	59	102
Hedgehog signaling pathway	hsa04340	56	80
Adherens junction	hsa04520	71	172
Basal transcription factors	hsa03022	44	470
PI3K-Akt signaling pathway	hsa04151	351	1030
JAK-STAT signaling pathway	hsa04630	162	508
Hematopoietic cell lineage	hsa04640	96	102
Toll-like receptor signaling pathway	hsa04620	102	328
Homologous recombination	hsa03440	41	140
ECM-receptor interaction	hsa04512	88	120
NF-κB signaling pathway	hsa04064	102	392

tions can cause cancer. I selected 702 COSMIC Cancer Gene Census (`https://cancer.sanger.ac.uk/cosmic/census?tier=all`, accessed on 3 December 2021) genes that all three omics type data are provided in CCLE, and used the genes equally for the baseline model execution. Moreover, the types of cell lines and drugs used in this study are the same as in the TGDRP and TGSA. The data type used by DRPreter model differs from every baseline model, and those two models used the most numerous omics types among them. To use only cell line-drug pairs with three omics data available, a lot of filtering was done on cell lines and drugs. Since all omics data had to be imported for baseline model execution, the same cell line-drug pair was used as in the most data-intensive models. Consequently, the performance test consists of 580 cancer cell lines that can obtain omics data from CCLE and 170 anticancer drugs provided by GDSC2. The total number of possible cell line-drug pairs is 82,833 with log-normalized IC_{50} values.

3.1.2 Experimental Setups

In the regression experiments for predicting natural log-transformed IC_{50} values based on drug and cancer cell line profiles, I used four standard evaluation metrics to compare the results of different models by computing the statistical correlation and accuracy between predicted and observed IC_{50} values. The metrics include pearson correlation coefficient (PCC), spearman correlation coefficient (SCC), mean absolute error (MAE) and mean squared error (MSE). PCC measures the linear correlation of observed and predicted IC_{50} values while SCC is a non-parametric measure for rank correlation between observed and predicted IC_{50} values. MSE and MAE directly measure the difference between observed and predicted IC_{50} values.

3.1.3 Rediscover Responses of Known Pairs

All possible cell line-drug pairs were randomly divided into train, validation, and test datasets at an 8:1:1 ratio, and the experiments were conducted repeatedly on 10 random seeds. For each model, the test performance is averaged over the seeds and reported as mean ± standard deviation. Comparing the results of different models was based on four common evaluation indicators. The mean squared error and mean absolute error between the predicted IC_{50} and the correct answer IC_{50} and the Pearson correlation coefficient and Spearman correlation coefficient between each IC_{50} distribution were used as evaluation criteria. Compared to the baseline model I selected above, I conducted an ablation study to examine each part's effectiveness of DRPreter (Table 3.2) and showed a performance improvement of about 20% in MSE with my best model (Table 3.3). In the ablation study, a random data experiment was conducted as a control and analyze the contribution of the gene expression as input in The mean squared error and mean absolute error between the predicted IC_{50} and the correct answer IC_{50} and the Pearson correlation coefficient and Spearman correlation coefficient between each IC_{50} distribution were used as evaluation criteria. prediction performance. Instead of using actual gene expression values, random embeddings of the same dimensionality as the transcript data were generated specifically for each cell line based on its index. I replaced the transcript data with the random data and kept the drug features intact. I evaluated the The mean squared error and mean absolute error between the predicted IC_{50} and the correct answer IC_{50} and the Pearson correlation coefficient and Spearman correlation coefficient between each IC_{50} distribution were used as evaluation criteria. prediction performance using the random data and compared it with the performance using the original transcript data to analyze the difference in performance to assess the importance of transcript data in the prediction task. The experimental results revealed a substantial enhancement in performance when utilizing actual gene expression data. Furthermore, the performance was further improved by incorporating several modules that constitute DRPreter.

3.2 Case Study

3.2.1 Interpolation of Unknown Values

The method of missing values prediction has been widely used in drug-response-prediction studies (Liu *et al.*, 2019, 2020; Nguyen *et al.*, 2021; Zhu *et al.*, 2022) to identify whether the model is capable of inductive prediction. For evaluating the inductive predictability DRPreter model, I trained with all the known cell line–drug pairs and predicted values without experimental results of pairs in the GDSC2 database. There were a total of 98,600 pairs using 580 cancer cell lines and 170 drugs, but 15,767 cell lines were not covered by my data due to filtering because of a lack of omics data or due to the absence of drug response experiments in GDSC. The model with the highest performance was used to predict missing drug response values.

I illustrate the distributions of known IC_{50} values in GDSC2 and the predicted values of DRPreter model (Figure 3.1). The box plots are grouped by drugs, and each box represents the distribution of the IC_{50} values within a cell line. I displayed the drugs with the top 10 highest and top 10 lowest median IC_{50} values. After conducting Mann–Whitney Wilcoxon test for each drug distribution, 18 drugs among the 20 selected drugs showed no significant difference between the GDSC2 and predicted unknown IC_{50} value distribution. The result implies the predicted missing IC_{50} values follow the measured value distribution.

Not knowing the actual values for these missing pairs, I conducted literature searches to assess model predictions. Bortezomib had the smallest overall

Table 3.2: Model ablation studies with different settings

Structural settings of DRPreter	Data	MSE (↓)	MAE (↓)	PCC (↑)	SCC (↑)
Random embedding	No mRNA [1]	2.4308 ± 0.0084	1.1832 ± 0.0048	0.8313 ± 0.0014	0.7795 ± 0.0038
Template graph	COSMIC [2]	0.8926 ± 0.0363	0.6909 ± 0.0146	0.9423 ± 0.0027	0.9196 ± 0.0034
Template graph	Pathway[3]	0.8536 ± 0.0420	0.6759 ± 0.0161	0.9449 ± 0.0032	0.9224 ± 0.0035
Pathway	Pathway[3]	0.8645 ± 0.0277	0.6791 ± 0.0113	0.9446 ± 0.0014	0.9233 ± 0.0008
Pathway + Transformer	Pathway[3]	0.8302 ± 0.0156	**0.66760.0051**	0.9465 ± 0.0015	0.9242 ± 0.0015
Pathway + Transformer + Similarity	Pathway[3]	**0.82510.0122**	0.6682 ± 0.0047	**0.94670.0013**	**0.92480.0014**

[1] No mRNA: Instead of using gene expression level data, random embeddings were created based on the cell line index and used as the cell line data. [2] COSMIC: 702 COSMIC genes. [3] Pathway: 2,369 genes of 34 cancer-related pathways.

Table 3.3: Performance comparison with baseline models

Model	Cell Encoder	Data	MSE (↓)	MAE (↓)	PCC (↑)	SCC (↑)
GraphDRP	1D CNN	COSMIC	1.0110 ± 0.0157	0.7618 ± 0.0083	0.9386 ± 0.0018	0.9151 ± 0.0021
TGDRP	GNN	COSMIC	0.9004 ± 0.0341	0.6933 ± 0.0148	0.9417 ± 0.0026	0.9188 ± 0.0040
TGSA	GNN	COSMIC	0.8955 ± 0.0536	0.6913 ± 0.0238	0.9425 ± 0.0043	0.9201 ± 0.0051
DRPreter	Knowledge-guided GNN	Pathway	**0.82510.0122**	**0.66820.0047**	**0.94670.0013**	**0.92480.0014**

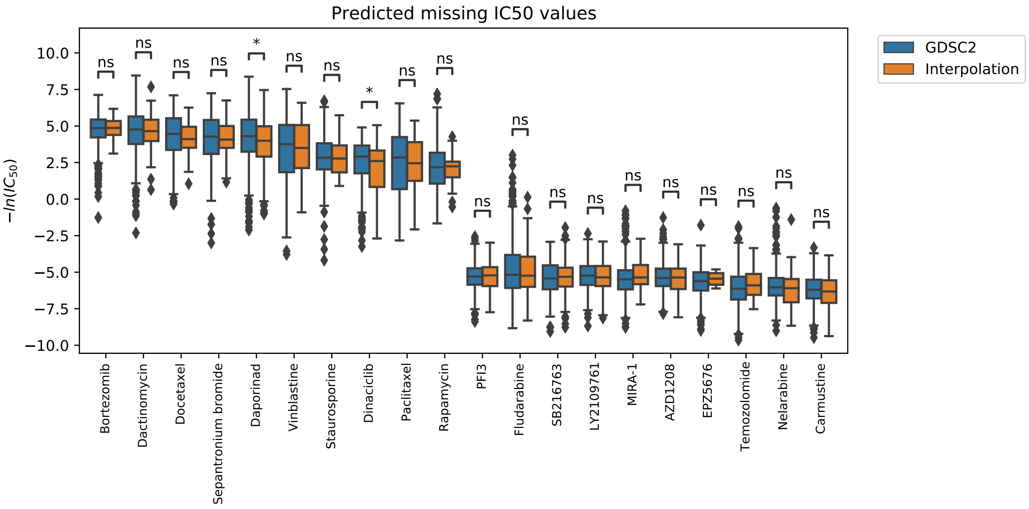

Figure 3.1: Box plot of drug-specific IC_{50} distributions of cell lines. The distribution of GDSC2 data (blue) compared with predicted missing IC_{50} values (orange). The 10 drugs with the highest median IC_{50} values and the 10 drugs with the lowest median were selected. Among the 20 drugs, IC_{50} value distributions of 18 drugs showed no significant differences through the Mann–Whitney Wilcoxon Test. ns: not significant, *: 0.01 ¡ p-value ¡ 0.05.

IC_{50} distribution, and the most sensitive cell line pair was LP-1 in my model. LP-1 is a cell line derived from the peripheral blood of a multiple myeloma patient. Bortezomib is a proteasome inhibitor that is widely used in patients with multiple myeloma (Field-Smith *et al.*, 2006; Kouroukis *et al.*, 2014). Rapamycin was not included among the top 10 sensitive drugs in the known GDSC data but in model predicted values, so I analyzed it further. In this study, rapamycin was most sensitive to the MV-4-11 cell line. The MV-4-11 cells are macrophages that were isolated from the blast cells of a biphenotypic B myelomonocytic leukemia patient. Rapamycin can inhibit leukemic activity in acute myeloid leukemia by mTOR inhibition through the blockade in G0/G1 phase of the cell cycle (Récher *et al.*, 2005).

Based on the biological processes at the cellular and molecular level of cancer cells and drugs, DRPreter can make inductive predictions for cell lines and drugs when there are no known responses and seems to have the potential to select candidates for drug treatment.

3.2.2 Gradient-Weighted Gene Nodes Interpretation

It is essential for drug-response prediction methods to capture significant biological implications and to make accurate predictions. A gene-level analysis was performed first to determine whether the model was taking into account genes that are known as drug targets, involved in target pathways, or biomarkers of disease. I prioritized genes from an input drug and cancer–cell line pair by scoring each gene with a gradient-weighted extent to check whether it is drug-target-related. The importance score of each gene node was determined by GradCAM, which is a widely utilized technique to produce explanations of model decisions (Selvaraju *et al.*, 2017), and I considered the score as the extent of its contribution. In this model, GradCAM determined the influence of input gene nodes on the label by tracing back the gradient backpropagation process of the model for predicting IC_{50} value. Table 3.4 shows the top five most significant genes of each cell line–drug pair in the test dataset.

As verified by literature searches, the bolded genes in Table 3.4 are the target genes or genes associated with the target pathway for each drug-cell line pair. The target was obtained from DrugBank (Wishart *et al.*, 2008) and GDSC, and the gene corresponding to the target pathway was obtained from GeneCards (Safran *et al.*, 2010) and Harmonizome (Rouillard *et al.*, 2016). Afatinib is an irreversible ErbB family blocker (Ioannou *et al.*, 2011) that targets *EGFR* and *ERBB2*, and its target pathway is EGFR signaling pathway. The model found *ERBB2* as a significant gene of the afatinib pair. As with the majority of cancers, *TP53* is the most common mutated gene showing a predominant clonal expression in Non-Small-Cell Lung Cancer (NSCLC) (Canale *et al.*, 2022). It is known to be possible to use *CLDN18* as an early-stage indicator of pancreatic ductal carcinogenesis and to study *CLDN18*'s regulatory mechanisms for uncovering key pathways like the PKC pathway of pancreatic cancer (Tanaka *et al.*, 2011). *CDK2* corresponds to the mTOR signaling pathway, which is the target pathway of Rapamycin. The use of Bortezomib and Paclitaxel suggests the potential for rationally designed treatments for solid tumors with MAPK pathway activation (Mehnert *et al.*, 2011).

3.2.3 Pathway-level Interpretation using DRPreter

I examined which pathways were stimulated in different cancer types that are sensitive to drugs and those that are not, as well as whether this model could capture such meaningful pathways. Self-attention score from the Transformer-based structure (Figure 2.1) was investigated for a drug that is sensitive only to specific cell lines. All the GDSC data with known IC_{50} values were shown in the same way as Figure 3.1(a), and Dasatinib was selected as having the widest IC_{50} distributions. The wide distribution of IC_{50} means that the drug exhibits

Table 3.4: Gradient-based gene importance analysis

Drug	**Cell line**	**Disease**	**Top 5 significant genes**	$\ln(\mathbf{IC}_{50})$	
				True	Predicted
Afatinib	GMS-10	Glioblastoma	ACTR3B, PRR5, PRKCZ, **ERBB2**, LTBR	0.5372	0.5324
Vinblastine	NCI-H1792	NSCLC	CYP7A1, GTF2H2, DVL2, RAB5B, **TP53**	-5.9258	-5.27633
Docetaxel	PANC0327	Pancreatic cancer	**CLDN18**, SOX17, FGF19, WNT7A, CDH5	-3.7668	-3.8204
Rapamycin	IGR1	Melanoma	TYRP1, DCT, TYR, FRZB, **CDK2**	-1.6747	-1.7651
Bortezomib	EBC-1	Lung squamous cell carcinoma Derived from metastatic site: Skin	SHC4, TNR, IL17RA, **MAPK12**, SMURF1	-5.7714	-6.0714

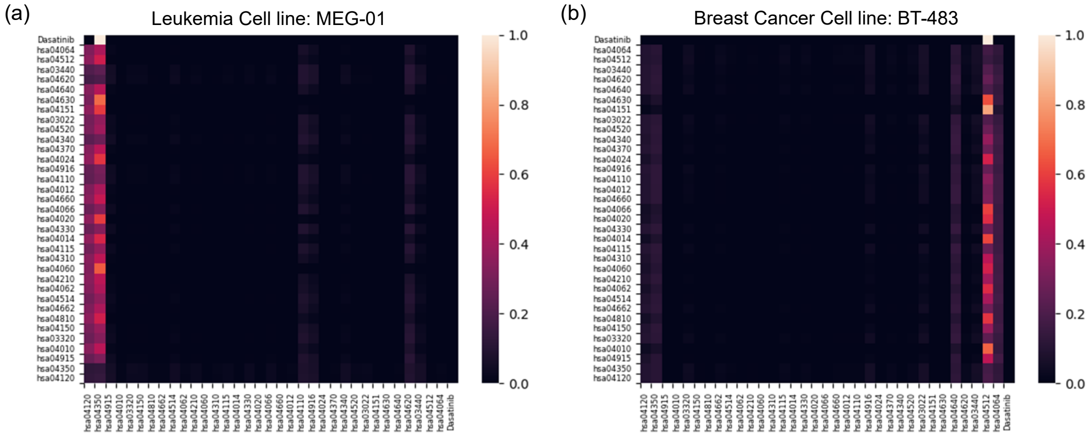

Figure 3.2: Visualization of self-attention score from Transformer. (a) Dasatinib and Leukemia cell line MEG-01 pair (b) Dasatinib and Breast cancer cell line BT-483. The figures show the y-axis as the query of the transformer, and the x-axis as the key. On each axis, there is a drug and 34 pathways which start with "hsa," indicating KEGG pathway identifiers.

the greatest differences in efficacy based on the type of cell line. I compared the self-attention score Transformer on MEG-01, the cell line judged to be sensitive with the smallest IC_{50} value, and BT-483, the most insensitive cell line with the largest IC_{50} value, among the 548 cell line pairs with Dasatinib (Figure 3.2). The MEG-01 cell line was derived from the hematopoietic and lymphoid tissue of a leukemia patient, and the BT-483 cell line was derived from the breast tissue of a breast cancer patient.

TGF-β signaling pathway (hsa04350) is the pathway with the highest attention score in MEG-01 cell line which is the most sensitive to dasatinib. The second most important pathway, ubiquitin-mediated proteolysis (hsa04120), involves the covalent binding of ubiquitin to the target protein and its degradation. It is known that ubiquitin-mediated degradation can regulate the TGF-β signaling pathway (Izzi and Attisano, 2004). The TGF-β signaling pathway suppresses tumors in normal and premalignant cells yet promotes oncogenesis in advanced cancer cells, and its components are regulated by ubiquitin modifying enzymes that abnormalities of the enzymes can cause malfunctioning of the pathway which can cause cancer, tissue fibrosis, and metastasis (Huang and Chen, 2012; Iyengar, 2017; Seoane and Gomis, 2017). In this regard, the ubiquitin modifying enzymes in the pathway and their counterparts are increasingly being explored as potential drug targets (Iyengar, 2017). Dasatinib is the tyrosine kinase inhibitor that can be the treatment of chronic myeloid leukemia (Keskin *et al.*, 2016). Dasatinib functions by binding to the ATP site of the active conformation of BCR-Abl (Sun *et al.*, 2011). As a signal transduction inhibitor, dasatinib inhibits the proliferation of tumor cells by inhibiting tyrosine kinase action, especially blocking transcriptional and promigratory responses to TGF-β through inhibition of Smad signaling (Bartscht *et al.*, 2015). The ubiquitin pathway can regulate the basal level of Smads, and altered Smad proteins can cause a malfunction in responding to the incoming signals due to their importance in transducing TGF-β signals (Izzi and Attisano, 2004). From the ubiquitin to the TGF-β pathway, this model captures the drug's mechanism of action.

Moreover, the ECM-receptor interaction pathway (hsa04512) was found to be most important in breast cancer-oriented cell line which is the most insensitive to Dasatinib. The ECM-receptor interaction pathway has been shown to be possibly useful as a biomarker for breast cancer (Bao *et al.*, 2019), but it does not relate to dasatinib's mechanism of action. Hence, DRPreter model identifies the pathways related to the drug mechanism of action for drug-sensitive carcinoma and focuses on the biomarker for carcinoma without drug efficacy.

3.2.4 Improve CCLE cell line-based model into TCGA patient-based model

I employed breast cancer patient data from The Cancer Genome Atlas (TCGA) (53 and 68, 2013) as an external validation for the study. Due to the inherent differences between TCGA data derived from *in vivo* patient samples and the *in vitro* cell line data on which the DRPreter model was trained, it posed challenges to directly apply the DRPreter for TCGA data. To address this, I sought to enhance the CCLE cell line-based model by adapting it to the TCGA patient-based model for classification purposes. I achieved this by incorporating an additional MLP layer at the front of the DRPreter model, which served to convert the TCGA gene expression space into the CCLE gene expression space. Subsequently, while keeping the model parameters of the DRPreter frozen, I exclusively trained the MLP layer for space transformation using the TCGA data. In this paper, I will consistently refer to the revised model with this setup as 'TCGAtoCCLE'.

I utilized TCGA gene expression data obtained from UCSC Xena (Goldman *et al.*, 2020) (`https://xenabrowser.net/datapages/`, accessed on 13 April 2023), specifically log_2(FPKM) values inferred from RNA-seq data. Furthermore, the binary drug response values indicating responder or non-responder were obtained from Ding *et al.*. The dataset consisted of a total of 360 drug-patient pairs, encompassing 26 drugs and 126 patients.

Binary drug response classification

I conducted experiments using a 5-fold cross-validation approach with three different seeds. The experiments were performed under two settings: random split and unseen patient stratified by subtype. Initially, the complete dataset was randomly split and subjected to 5-fold cross-validation for testing. Subsequently, to evaluate the prediction capability for unseen patients, I conducted an additional experiment where predictions were made for new patients who were not included in the training phase. It was ensured that the same patient was not included in the training, validation, and test sets. Furthermore, each divided set was stratified based on the patient's subtype. An ablation study was conducted in the two aforementioned settings to assess the impact of adding the TCGA to CCLE conversion layer and showed a performance improvement for each metric with the conversion layer (Table 3.5).

Four metrics were used to evaluate the performance of the models: Area Under the ROC Curve (AUC), Balanced Accuracy (BACC), Precision (PREC), and Cohen's Kappa (KAPPA). AUC represents the area under the Receiver Operating Characteristic (ROC) curve. It measures the ability of the model to distinguish between positive and negative samples across various classification thresholds. A higher AUC value indicates better overall model performance.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(f)\, \mathrm{dFPR}(f) \tag{3.1}$$

BACC calculates the average of sensitivity and specificity. It takes into account both the true positive rate and true negative rate, providing a more reliable measure of overall classification performance, especially in imbalanced datasets.

$$\mathrm{BACC} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2} \tag{3.2}$$

Precision is the proportion of correctly predicted positive samples over the total predicted positive samples. By focusing on the correctness of positive predictions, indicates the model's ability to avoid false positives.

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.3}$$

Cohen's Kappa measures the agreement between the predicted and actual labels while considering the agreement that could occur by chance. It takes into account the imbalance in class distribution and provides a normalized measure of agreement.

$$\text{KAPPA} = \frac{\text{Pr(a)} - \text{Pr(e)}}{1 - \text{Pr(e)}} \tag{3.4}$$

In these formulas, TPR represents the True Positive Rate, FPR represents the False Positive Rate, TNR represents the True Negative Rate, TP represents the number of True Positive instances, FP represents the number of False Positive instances, Pr(a) represents the observed agreement, and Pr(e) represents the chance agreement.

Subtype-specific visualization using t-SNE

I conducted a comparative analysis to assess the capability of DRPreter and TCGAtoCCLE models in generating embeddings for distinguishing patients based on their subtypes. Breast cancer subtype data was downloaded from Berger *et al.*, which provides subtypes for 1,050 TCGA patients. The specific counts for each subtype are as follows: 'Luminal A (LumA)' (563), 'Her2' (82), 'Luminal B (LumB)' (208), 'Basal' (192), and 'Normal' (5).

A t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008) plot was constructed using patient embeddings obtained from the models. To quantitatively evaluate the embedding quality of DRPreter and TCGAtoCCLE, mutual information (Steuer *et al.*, 2002) was calculated between the results of agglomerative clustering (Zepeda-Mendoza and Resendis-Antonio, 2013) using embeddings derived from each model and the actual subtypes of patients. Mutual information is a measure of the degree of dependency between two variables and serves as a metric to evaluate how well the clustering results align with the actual patient subtypes. A higher mutual information score indicates better performance of the clustering in correctly grouping patient subtypes using the embeddings from the model. Therefore, according to the mutual information score on the t-SNE plot, it can be concluded that the embedding quality of TCGAtoCCLE is better (Figure 3.3).

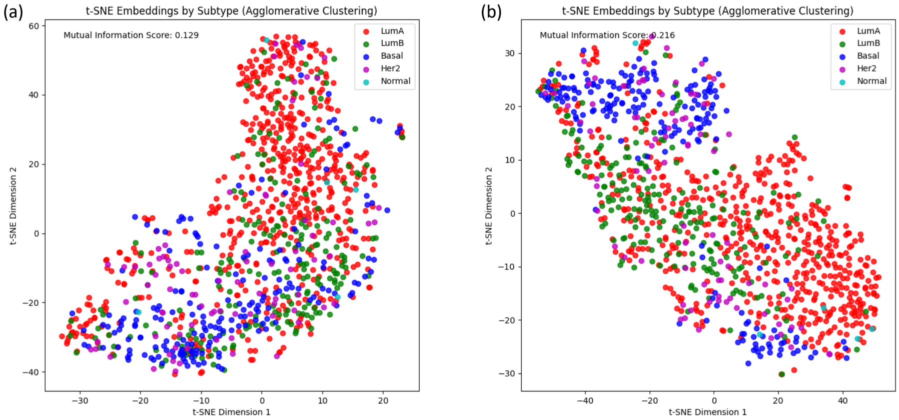

Figure 3.3: Tumor subtype classification potential revealed by t-SNE. Each point represents a patient sample. (a) The t-SNE plot of the 1,050 patients' embeddings using DRPreter. (b) The t-SNE plot of the 1,050 patients' embeddings using TCGAtoCCLE. In each plot, the mutual information value calculated between the results of agglomeration clustering and the actual subtypes of patients was displayed.

Pathway-based interpretation for Capecitabine

Among the 360 drug response data, only Capecitabine showed a clearly different drug response depending on the subtype. Within the data, all LumA patients were responders to Capecitabine, and basal patients were non-responders (Asleh *et al.*, 2023). Consequently, I investigated the divergences in emphasized pathways within the model based on the subtypes (Figure 3.4).

The p53 signaling pathway (hsa04115) emerges as a commonly implicated pathway among LumA breast cancer patients, as supported by the self-attention score analysis of the Transformer model. LumA is characterized by estrogen receptor-positive (ER+) and progesterone receptor-positive, HER2-negative, and also exhibits low levels of the Ki-67 protein that regulates the growth rate of cancer cells. Coutant *et al.* studied p53 gene signatures to predict prognosis and response to chemotherapy in ER-positive and ER-negative breast cancers. They revealed that ER+ breast cancers with p53 dysfunction, measured by a transcriptional signature, are more sensitive to chemotherapy than p53 normal cases. Hence, it seems appropriate for the Transformer to focus on the p53 signaling pathway to assess p53 dysfunction in this transcriptome-based model. In addition, no common pathways were identified in basal patients who were non-responders to Capecitabine. Instead, the ubiquitin-mediated proteolysis pathway (hsa04120) and the Toll-like receptor signaling pathway (hsa04620) were notably important in each respective sample. Previous studies have linked that the ubiquitin ligases of the ubiquitin-proteasome system, which encompasses the ubiquitin-mediated proteolysis pathway, to basal-like breast cancer (Versari *et al.*, 2006; Qi and Ze'ev, 2015; Saucedo-Cuevas *et al.*, 2014; Chan *et al.*, 2011). Toll-like receptors (TLRs) are highly expressed in breast cancer cells (Kidd *et al.*, 2013), and their expression levels vary across different breast cancer subtypes and stages (Shi *et al.*, 2020). Therefore, in line with the findings from analysis in cell lines, the focus for responders is on pathways associated with drug responsiveness, while non-responders receive information related to cancer-type-specific biomarkers.

Also, I conducted a further experiment to investigate the dissimilarities in embeddings among structurally similar drugs. Through a comparative analysis of Capecitabine and the other drug sharing the most similar structural characteristics, I aimed to elucidate the specific interaction of the drug with LumA. This experimental study aimed to determine whether these dissimilarities can shed light on the targeted action of Capecitabine on LumA using DRPreter. Dice similarity with Capecitabine was calculated for all 26 drugs with TCGA drug response data. Subsequently, a comparison was performed between the embeddings of Gemcitabine and Capecitabine, yielding the highest similarity value of 0.3656. Although the similarity is the highest among drugs, it is only 0.3656. However, when I looked at what pathway the model is most interested in between Gemcitabine and the tumor sample pair, it came out almost the same as Capecitabine (Figure 3.5). This finding indicates that the model's transformer-based architecture demonstrates sensitivity towards variations in gene expression, but does not strongly capture the structural characteristics of the drug.

Table 3.5: Binary classification performance comparison in different settings

Setting	Model	AUC (↑)	BACC (↑)	PREC (↑)	KAPPA (↑)
Unseen patient	DRPreter [1]	0.7583 ± 0.0772	0.6035 ± 0.05	0.8927 ± 0.0367	0.2261 ± 0.0924
	TCGAtoCCLE [2]	**0.8086 ± 0.0516**	**0.6314 ± 0.0562**	**0.8958 ± 0.0269**	**0.3128 ± 0.101**
Random split	DRPreter [1]	0.8183 ± 0.159	0.7142 ± 0.158	0.9166 ± 0.0487	0.4425 ± 0.3095
	TCGAtoCCLE [2]	**0.8619 ± 0.1288**	**0.7549 ± 0.1947**	**0.9258 ± 0.0618**	**0.5395 ± 0.3429**

[1] DRPreter: The configuration in the original model involved utilizing a classification layer instead of a regression layer.

[2] TCGAtoCCLE: Transforming TCGA gene expression space into CCLE space by using an additional MLP layer at the front of DRPreter.

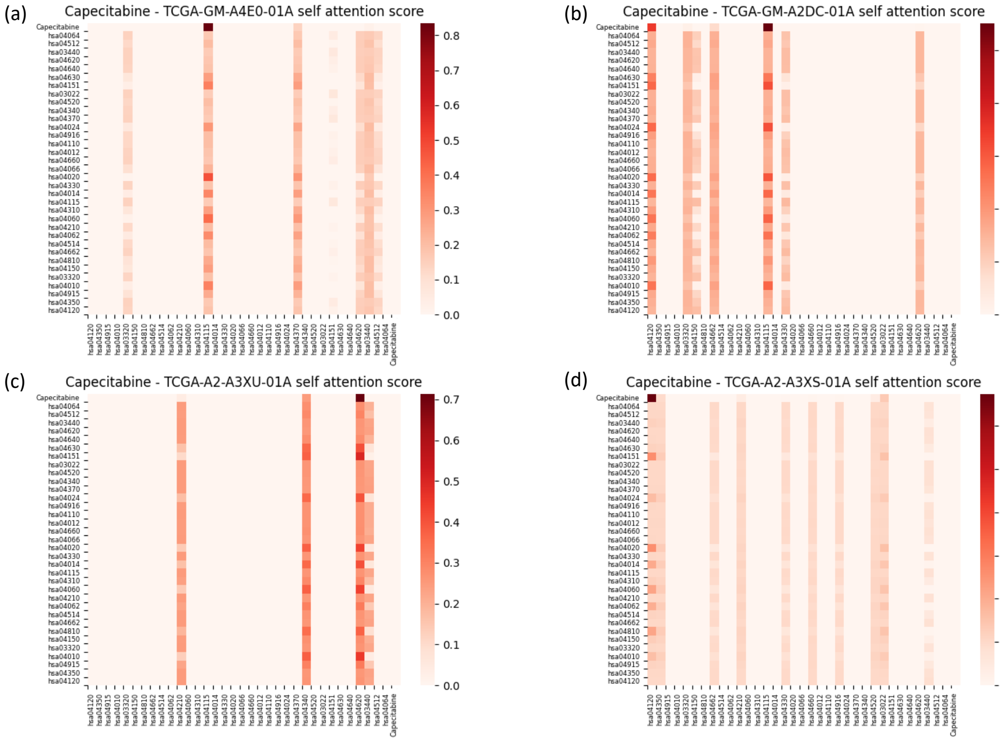

Figure 3.4: Visualization of self-attention score from Transformer. (a) and (b) depict the results for LumA subtype patients who responded to Capecitabine. (c) and (d) represent the results for basal subtype patients who did not respond to the treatment. The figures show the y-axis as the query of the transformer, and the x-axis as the key. On each axis, there is a drug and 34 pathways which start with "hsa," indicating KEGG pathway identifiers.

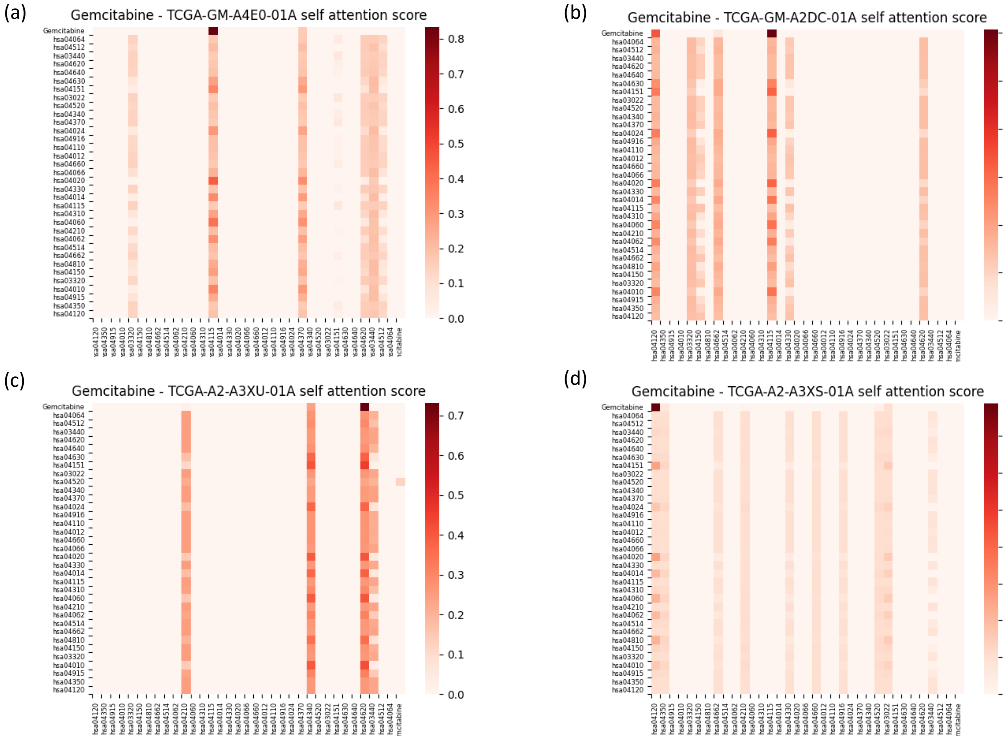

Figure 3.5: Visualization of self-attention score from Transformer. (a) and (b) depict the results of Gemcitabine for LumA subtype patients who responded to Capecitabine. (c) and (d) represent the results of Gemcitabine for basal subtype patients who did not respond to the treatment. The figures show the y-axis as the query of the transformer, and the x-axis as the key. On each axis, there is a drug and 34 pathways which start with "hsa," indicating KEGG pathway identifiers.

Chapter 4

Conclusion

In this section, I will summarize the works in this paper and set forth future works for further improvement.

1. An interpretable drug response prediction model called DRPreter integrates biological and chemical domain knowledge with cutting-edge deep learning technologies to deliver outstanding predictive performance and interpretability.

2. I introduced cancer-related pathways and constructed the cell line network as a set of subgraphs to represent and interpret biological mechanisms in detail.

3. I extracted drug-pathway interaction information from the modified encoder of the Transformer module and obtained putative key pathways for the drug mechanism.

4. Ablation studies verified the effectiveness of each component of the model and performance comparison experiments showed DRPreter has

enhanced predictive power than the state-of-the-art drug response prediction models.

5. Through external validation using TCGA data, the feasibility of extending the model from cell line data to patient data was illustrated.

To properly apply the drug response predicted by the model for clinical use or drug discovery, it is essential to understand the process and mechanism from which it was derived due to safety and reliability issues. Accordingly, I implemented gene and pathway-level analysis via DRPreter, and it has been shown that DRPreter predicts drug sensitivity based on known drug mechanisms of action and target-related factors. I also identified the cell line that would act most sensitively for each drug in the absence of experimental data through a case study and confirmed that it is widely used for each drug currently in the clinical situation. By doing so, patients who have shown resistance to a specific drug may be able to select a drug candidate group that would replace the ineffective drug. However, this study utilized genes within the pre-selected 34 pathways, and the Transformer model was trained to learn relationships exclusively within these pathways. As a result, there is a limitation in the interpretation regarding the significance of pathways outside the selected ones. Therefore, there is room for further study by incorporating a larger number of genes and pathways to explore beyond the current limitations and enhance the understanding of pathway interactions.

Also, the Transformer architecture, commonly used in natural language processing tasks such as machine translation and text generation, may not be the most appropriate choice if data has a different structure or characteristics. Therefore, the suitability of the Transformer architecture for the given input configuration, comprising multiple pathways and a single drug, should be carefully evaluated. This consideration was further substantiated through pathway-based interpretation using the self-attention scores of the Transformer-based structure. Notably, when different cell lines were employed for the same drug, the model's attention was observed to vary according to the gene expression profiles specific to each cell line type. This finding signifies that the Transformer-based structure employed in this study has the capability to extract meaningful information from the gene expression data. However, when experiments with different drugs were performed on the same tumor samples, there were cases where the results were exactly the same (Figure 3.4, Figure 3.5), so meaningless results were obtained when the number of pathways and drugs differs. This suggests that there is potential for further model modifications to enhance pathway-based interpretation performance for data imbalance.

Bibliography

53, D. C. C. B. R. . J. M. A. . K. A. . P. T. . P. D. . W. Y. and 68, T. S. S. L. D. A. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**(10), 1113–1120.

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, **4**(1), 1–10.

Ahmed, Z. (2020). Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. *Human genomics*, **14**(1), 1–5.

Asleh, K., Lluch, A., Goytain, A., Barrios, C., Wang, X. Q., Torrecillas, L., Gao, D., Ruiz-Borrego, M., Leung, S., Bines, J., *et al.* (2023). Triple-negative pam50 non-basal breast cancer subtype predicts benefit from extended adjuvant capecitabine. *Clinical Cancer Research*, pages OF1–OF12.

Bao, Y., Wang, L., Shi, L., Yun, F., Liu, X., Chen, Y., Chen, C., Ren, Y., and Jia, Y. (2019). Transcriptome profiling revealed multiple genes and ecm-receptor interaction pathways that may be associated with breast cancer. *Cellular & molecular biology letters*, **24**(1), 1–20.

Baptista, D., Ferreira, P. G., and Rocha, M. (2021). Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*, **22**(1), 360–379.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., *et al.* (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391), 603–607.

Bartscht, T., Rosien, B., Rades, D., Kaufmann, R., Biersack, H., Lehnert, H., Gieseler, F., and Ungefroren, H. (2015). Dasatinib blocks transcriptional and promigratory responses to transforming growth factor-beta in pancreatic adenocarcinoma cells through inhibition of smad signalling: implications for in vivo mode of action. *Molecular cancer*, **14**(1), 1–12.

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., Ebright, R. Y., Stewart, M. L., Ito, D., Wang, S., *et al.* (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, **154**(5), 1151–1161.

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., *et al.* (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, **33**(4), 690–705.

Canale, M., Andrikou, K., Priano, I., Cravero, P., Pasini, L., Urbini, M., Delmonte, A., Crinò, L., Bronte, G., and Ulivi, P. (2022). The role of tp53 mutations in egfr-mutated non-small-cell lung cancer: Clinical significance and implications for therapy. *Cancers*, **14**(5), 1143.

Chan, P., Möller, A., Liu, M. C., Sceneay, J. E., Wong, C. S., Waddell, N., Huang, K. T., Dobrovic, A., Millar, E. K., O'Toole, S. A., *et al.* (2011). The expression of the ubiquitin ligase siah2 (seven in absentia homolog 2) is mediated through gene copy number in breast cancer and is associated with a basal-like phenotype and p53 expression. *Breast Cancer Research*, **13**, 1–10.

Chiu, Y.-C., Chen, H.-I. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J., Huang, Y., and Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC medical genomics*, **12**(1), 143–155.

Cho, S.-Y. (2020). Patient-derived xenografts as compatible models for precision oncology. *Laboratory Animal Research*, **36**(1), 1–11.

Coutant, C., Rouzier, R., Qi, Y., Lehmann-Che, J., Bianchini, G., Iwamoto, T., Hortobagyi, G. N., Symmans, W. F., Uzan, S., Andre, F., *et al.* (2011). Distinct p53 gene signatures are needed to predict prognosis and response to chemotherapy in er-positive and er-negative breast cancers. *Clinical Cancer Research*, **17**(8), 2591–2601.

Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., and Wang, S. (2022). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570*.

Deng, L., Cai, Y., Zhang, W., Yang, W., Gao, B., and Liu, H. (2020). Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. *Journal of Chemical Information and Modeling*, **60**(10), 4497–4505.

DepMap, B. (2021). Depmap 21q4 public. figshare. dataset.

Ding, Z., Zu, S., and Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, **32**(19), 2891–2895.

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., and Zheng, X. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer*, **15**(1), 1–12.

Feng, R., Xie, Y., Lai, M., Chen, D. Z., Cao, J., and Wu, J. (2021). Agmi: Attention-guided multi-omics integration for drug response prediction with graph neural networks. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1295–1298. IEEE.

Field-Smith, A., Morgan, G. J., and Davies, F. E. (2006). Bortezomib (velcade™) in the treatment of multiple myeloma. *Therapeutics and clinical risk management*, **2**(3), 271.

Firoozbakht, F., Yousefi, B., and Schwikowski, B. (2022). An overview of machine learning methods for monotherapy drug response prediction. *Briefings in Bioinformatics*, **23**(1), bbab408.

Gao, H. and Ji, S. (2019). Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., *et al.* (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, **38**(6), 675–678.

Guan, N.-N., Zhao, Y., Wang, C.-C., Li, J.-Q., Chen, X., and Piao, X. (2019).

Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization. *Molecular therapy-nucleic acids*, **17**, 164–174.

Güvenç Paltun, B., Mamitsuka, H., and Kaski, S. (2021). Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Briefings in bioinformatics*, **22**(1), 346–359.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, **30**.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, F. and Chen, Y.-G. (2012). Regulation of tgf-β receptor activity. *Cell & bioscience*, **2**(1), 1–10.

Ioannou, N., Dalgleish, A., Seddon, A., Mackintosh, D., Guertler, U., Solca, F., and Modjtahedi, H. (2011). Anti-tumour activity of afatinib, an irreversible erbb family blocker, in human pancreatic tumour cells. *British journal of cancer*, **105**(10), 1554–1562.

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.* (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**(3), 740–754.

Iyengar, P. V. (2017). Regulation of ubiquitin enzymes in the tgf-β pathway. *International journal of molecular sciences*, **18**(4), 877.

Izzi, L. and Attisano, L. (2004). Regulation of the tgfβ signalling pathway by ubiquitin-mediated degradation. *Oncogene*, **23**(11), 2071–2078.

Kalamara, A., Tobalina, L., and Saez-Rodriguez, J. (2018). How to find the right drug for each patient? advances and challenges in pharmacogenomics. *Current opinion in systems biology*, **10**, 53–62.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kellogg, R. A., Dunn, J., and Snyder, M. P. (2018). Personal omics for precision health. *Circulation Research*, **122**(9), 1169–1171.

Keskin, D., Sadri, S., and Eskazan, A. E. (2016). Dasatinib for the treatment of chronic myeloid leukemia: patient selection and special considerations. *Drug design, development and therapy*, **10**, 3355.

Kidd, L. C. R., Rogers, E. N., Yeyeodu, S. T., Jones, D. Z., and Kimbro, K. S. (2013). Contribution of toll-like receptor signaling pathways to breast tumorigenesis and treatment. *Breast Cancer: Targets and Therapy*, pages 43–51.

Kim, S., Bae, S., Piao, Y., and Jo, K. (2021). Graph convolutional network for drug response prediction using gene expression data. *Mathematics*, **9**(7), 772.

Kouroukis, T., Baldassarre, F., Haynes, A., Imrie, K., Reece, D., and Cheung, M. (2014). Bortezomib in multiple myeloma: systematic review and clinical considerations. *Current oncology*, **21**(4), 573–603.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., and Ideker, T. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer cell*, **38**(5), 672–684.

Landrum, G. *et al.* (2013). Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*.

Lee, J., Lee, I., and Kang, J. (2019). Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR.

Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics*, **36**(12), 3818–3824.

Li, M. M., Huang, K., and Zitnik, M. (2021). Graph representation learning in biomedicine. *arXiv preprint arXiv:2104.04883*.

Liu, P., Li, H., Li, S., and Leung, K.-S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC bioinformatics*, **20**(1), 1–14.

Liu, Q., Hu, Z., Jiang, R., and Zhou, M. (2020). Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, **36**(Supplement_2), i911–i918.

Mehnert, J. M., Tan, A. R., Moss, R., Poplin, E., Stein, M. N., Sovak, M., Levinson, K., Lin, H., Kane, M., Gounder, M., *et al.* (2011). Rationally designed treatment for solid tumors with mapk pathway activation: A phase i study of paclitaxel and bortezomib using an adaptive dose-finding approach-paclitaxel and bortezomib for tumors with mapk activation. *Molecular cancer therapeutics*, **10**(8), 1509–1519.

Nguyen, T., Nguyen, G. T., Nguyen, T., and Le, D.-H. (2021). Graph convolutional networks for drug response prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, **19**(1), 146–154.

Qi, J. and Ze'ev, A. R. (2015). Dysregulation of ubiquitin ligases in cancer. *Drug Resistance Updates*, **23**, 1–11.

Récher, C., Beyne-Rauzy, O., Demur, C., Chicanne, G., Dos Santos, C., Mas, V. M.-D., Benzaquen, D., Laurent, G., Huguet, F., and Payrastre, B. (2005). Antileukemic activity of rapamycin in acute myeloid leukemia. *Blood*, **105**(6), 2527–2534.

Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., Adams, D. J., Price, E. V., Gill, S., Javaid, S., Coletti, M. E., Jones, V. L., Bodycombe, N. E., *et al.* (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, **12**(2), 109–116.

Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., and Fine, H. A. (2011). Predicting in vitro drug sensitivity using random forests. *Bioinformatics*, **27**(2), 220–224.

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., *et al.* (2010). Genecards version 3: the human gene integrator. *Database*, **2010**.

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., *et al.* (2019). A deep learning framework for predicting response to therapy in cancer. *Cell reports*, **29**(11), 3367–3373.

Saucedo-Cuevas, L. P., Ruppen, I., Ximénez-Embún, P., Domingo, S., Gayarre, J., Muñoz, J., Silva, J. M., García, M. J., and Benítez, J. (2014). Cul4a contributes to the biology of basal-like breast tumors through modulation of cell growth and antitumor immune response. *Oncotarget*, **5**(8), 2330.

Savage, N. (2021). Tapping into the drug discovery potential of ai. *Biopharma Deal*.

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., *et al.* (2015). Harnessing connectivity in a large-scale small-molecule sensitivity datasetharnessing connectivity in a sensitivity dataset. *Cancer discovery*, **5**(11), 1210–1223.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Seoane, J. and Gomis, R. R. (2017). Tgf-β family signaling in tumor suppression and cancer progression. *Cold Spring Harbor perspectives in biology*, **9**(12), a022277.

Seyhan, A. A. (2019). Lost in translation: the valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. *Translational Medicine Communications*, **4**(1), 1–19.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, **35**(14), i501–i509.

Shi, S., Xu, C., Fang, X., Zhang, Y., Li, H., Wen, W., and Yang, G. (2020).

Expression profile of toll-like receptors in human breast cancer. *Molecular medicine reports*, **21**(2), 786–794.

Shoemaker, R. H. (2006). The nci60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, **6**(10), 813–823.

Singh, D. B. (2019). The impact of pharmacogenomics in personalized medicine. *Current Applications of Pharmaceutical Biotechnology*, pages 369–394.

Singh, V. P., Pratap, K., Sinha, J., Desiraju, K., Bahal, D., and Kukreti, R. (2016). Critical evaluation of challenges and future use of animals in experimentation for biomedical research. *International Journal of Immunopathology and Pharmacology*, **29**(4), 551–561.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, **18**(11), 696–705.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**(suppl_2), S231–S240.

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.* (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**(6), 1437–1452.

Sun, H., Kapuria, V., Peterson, L. F., Fang, D., Bornmann, W. G., Bartholomeusz, G., Talpaz, M., and Donato, N. J. (2011). Bcr-abl ubiquitination and usp9x inhibition block kinase signaling and promote cml cell apoptosis. *Blood, The Journal of the American Society of Hematology*, **117**(11), 3151–3162.

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.* (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**(D1), D607–D613.

Tanaka, M., Shibahara, J., Fukushima, N., Shinozaki, A., Umeda, M., Ishikawa, S., Kokudo, N., and Fukayama, M. (2011). Claudin-18 is an early-stage marker of pancreatic carcinogenesis. *Journal of Histochemistry & Cytochemistry*, **59**(10), 942–952.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Versari, D., Herrmann, J., Gossl, M., Mannheim, D., Sattler, K., Meyer, F. B., Lerman, L. O., and Lerman, A. (2006). Dysregulation of the ubiquitin-proteasome system in human carotid atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology*, **26**(9), 2132–2139.

Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, **17**(1), 1–12.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, **37**(suppl_2), W623–W633.

Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, **28**(1), 31–36.

Weisfeiler, B. and Leman, A. (1968). The reduction of a graph to canonical form and the algebra which appears therein. *NTI, Series*, **2**(9), 12–16.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, **36**(suppl_1), D901–D906.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., *et al.* (2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, **41**(D1), D955–D961.

Zepeda-Mendoza, M. L. and Resendis-Antonio, O. (2013). Hierarchical agglomerative clustering. *Encyclopedia of systems biology*, **43**(1), 886–887.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018). An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zheng, K., Zhao, H., Zhao, Q., Wang, B., Gao, X., and Wang, J. (2022). Nasmdr: a framework for mirna-drug resistance prediction using efficient neural architecture search and graph isomorphism networks. *Briefings in Bioinformatics*.

Zhu, Y., Ouyang, Z., Chen, W., Feng, R., Chen, D. Z., Cao, J., and Wu, J. (2022). Tgsa: protein–protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics*, **38**(2), 461–468.

Zuo, Z., Wang, P., Chen, X., Tian, L., Ge, H., and Qian, D. (2021). Swnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC bioinformatics*, **22**(1), 1–16.

국문초록

약물 반응성 예측에 대한 최근 연구 중 일부는 그래프 신경망을 적용하여 약물 구조 또는 유전자 네트워크에 대한 사전 지식을 활용하는 반면, 다른 연구는 약물 반응을 지배하는 메커니즘을 설명하기 위한 모델의 해석 가능성에 초점을 맞추고 있다. 그러나 예측 정확도가 향상되고 모델의 실용성이 향상될 수 있도록 사전 지식에 기반하면서도 해석 가능한 예측 모델을 만드는 것이 중요하다. 따라서 DRPreter(Drug Response PREdictor and interpreTER)라는 해석 가능한 모델을 제안한다. DRPreter는 도메인 지식을 바탕으로 생물학적 패스웨이를 서브그래프로 하여 세포주 그래프를 분할하고, 그래프 신경망을 통해 세포주 및 약물 정보를 학습한다. DRPreter에서 사용한 트랜스포머의 인코더 기반 구조는 약물 반응과 관련된 중요한 패스웨이를 강조하고, 패스웨이들과 약물 사이의 관계를 탐지하는 역할을 한다. GDSC(Genomics of Drug Sensitivity and Cancer) 데이터에 대한 성능 평가 결과는 본 모델이 항암제 반응 예측을 위한 그래프 기반 최신 모델들의 성능을 능가한다는 것을 보여준다. 또한 특정 항암제-세포주 쌍에 대해 핵심 유전자와 패스웨이를 추정하고 문헌에서 이를 뒷받침하는 증거를 찾았고, 이는 본 모델이 약물의 작용 매커니즘을 해석하는 데 도움을 줄 수 있음을 시사한다.

주요어: 전사체학, 설명가능 인공지능, 약물유전체학, 항암제 반응성, 그래프 신경망, 맞춤의학, 신약 발굴

학번: 2021-24949