Applications of Deep Learning and Graph Representation Learning in Precision Cancer Medicine

David Earl D. Hostallero

Department of Electrical & Computer Engineering
McGill University
Montreal, Quebec, Canada

August 10, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

Computationally designing personalized treatment plans to increase a cancer patient's chances of recovery using their molecular profiles has been one of the major objectives of precision cancer medicine. Despite the advancement of high-throughput sequencing and artificial intelligence, drug response prediction has remained a challenging task. This thesis presents novel methodologies for predicting responses to drug treatments, addressing challenges such as limited clinical data and drug-specific biases. Leveraging available datasets, I explored the utility of different information modalities in predictive models.

First, I focused on clinical drug response prediction using only preclinical data. This stemmed from the current situation of cancer drug response datasets, wherein drug responses for preclinical cancer cell line (CCL) samples treated with hundreds of drugs are widely available, while clinical drug responses of tumors are only available in small patient cohorts for a handful of drugs. I developed a deep learning pipeline that leverages tissue information to bridge discrepancies between CCL and tumor samples, enabling models to distinguish between sensitive and resistant patients. I then ventured towards improving drug representation using knowledge graphs composed of CCLs, drugs, and genes. Unlike previous methods that solely rely on the structural properties of drug molecules, I integrated additional response-relevant information, such as molecular profiles of extremely sensitive/resistant CCLs, CRISPR gene effects, and drug targets. My analyses demonstrated superior performance compared to existing methods and baseline approaches.

Beyond drug response prediction, I also identified potential biomarkers of drug response for each model that I presented. This not only enhances model interpretability, but also produces data-driven hypotheses. Many implicated genes and pathways were supported by literature, and in some cases, experimentally validated. I introduced a graph-based

interpretation method to provide further insights and visualize the prediction process at a high level.

The contents of this thesis not only improve drug response prediction but also shed light on potential therapeutic targets, contributing to the advancement of precision cancer medicine.

Abrégé

L'un des principaux objectifs de la médecine de précision en cancérologie est de concevoir par ordinateur des plans de traitement personnalisés afin d'augmenter les chances de guérison des patients atteints d'un cancer en se basant sur leur profil moléculaire. Malgré les progrès du séquençage à haut débit et de l'intelligence artificielle, la prédiction de la réponse aux médicaments reste une tâche difficile. Cette thèse présente de nouvelles méthodologies pour prédire les réponses aux traitements médicamenteux, en relevant des défis tels que les données cliniques limitées et les biais spécifiques aux médicaments. En exploitant les ensembles de données disponibles, nous avons exploré l'utilité de différentes modalités d'information dans les modèles prédictifs.

Tout d'abord, nous nous sommes concentrés sur la prédiction de la réponse clinique aux médicaments en utilisant uniquement des données précliniques. Cela s'explique par la situation actuelle des ensembles de données sur la réponse aux médicaments anticancéreux, où les réponses aux médicaments pour les échantillons de lignée cellulaire de cancer (LCC) précliniques traitées avec des centaines de médicaments sont largement disponibles, tandis que les réponses cliniques aux médicaments des tumeurs ne sont disponibles que dans de petites cohortes de patients pour quelques médicaments. Nous avons développé un pipeline d'apprentissage profond qui exploite les informations sur les tissus pour combler les écarts entre les échantillons de LCC et de tumeurs, ce qui permet aux modèles de faire la distinction entre les patients sensibles et résistants. Nous avons ensuite tenté d'améliorer la représentation des médicaments en utilisant des graphes de connaissances composés de LCC, de médicaments et de gènes. Contrairement aux méthodes précédentes qui s'appuient uniquement sur les propriétés structurelles des molécules de médicaments, nous avons intégré des informations supplémentaires relatives à la réponse aux médicaments, telles que les profils

moléculaires des LCC extrêmement sensibles/résistants, les effets des gènes CRISPR et les cibles des médicaments. Nos analyses ont démontré des performances supérieures à celles des méthodes existantes et des approches de référence.

Au-delà de la prédiction de la réponse aux médicaments, nous avons également identifié des biomarqueurs potentiels de la réponse aux médicaments pour chaque modèle présenté. Cela permet non seulement d'améliorer l'interprétabilité des modèles, mais aussi de produire des hypothèses basées sur des données. De nombreux gènes et voies impliqués ont été étayés par la littérature et, dans certains cas, validés expérimentalement. Nous avons introduit une méthode d'interprétation basée sur les graphes afin de fournir des informations supplémentaires et de visualiser le processus de prédiction à un niveau élevé.

Le contenu de cette thèse permet non seulement d'améliorer la prédiction de la réponse aux médicaments, mais aussi de mettre en lumière des cibles thérapeutiques potentielles, contribuant ainsi à l'avancement de la médecine de précision contre le cancer.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Amin Emad, for their support, guidance, and insightful feedback throughout my entire PhD adventure. Their expertise, encouragement, and patience have been instrumental in shaping this work. It has been my great honor and privilege to have worked under his guidance.

Additionally, I extend my appreciation to the members of my supervisory committee, Dr. Tal Arbel, Dr. Hamed Najafabadi, and Dr. William Hamilton, for their valuable feedback and scholarly input, which have enriched my understanding of the different fields relating to my research, especially during the early years of this degree.

I would like to thank McGill University and Mila - Quebec AI Institute for fostering an exceptional research environment that has nurtured my academic journey. The financial assistance provided through MEDA and FRQNT has been instrumental in supporting my studies. I am also grateful to Calcul Québec, Digital Research Alliance of Canada for the access to state-of-the-art computational resources.

I have been fortunate to work with many talented researchers in the COMBINE lab, especially Jessica, Safyan, Joseph, Chen, Abdulrahman, Yazdan, Antoine, and Reda. Additionally, I would like to thank our collaborators, Dr. Marc Hafner and Dr. Scott Martin. Working with them has been an educational experience.

I would also like to dedicate this thesis to my former adviser during my master's studies, Dr. Yung Yi. While he is no longer with us, his mentorship and dedication to research continue to inspire me.

Finally, I would like to extend my heartfelt thanks to my family and friends for their unwavering love, encouragement, and understanding. Their constant support and belief in me have been a source of strength and motivation.

Contents

List of Figures

List of Tables

List of Acronyms

ADDA	Adversarial discriminative domain adaptation
AUC	Area under the curve
AUROC	Area under the receiver operating characteristic curve
BiG-DRP	Bipartite graph-represented drug response predictor
CCL	Cancer cell line
CCLE	Cancer Cell Line Encyclopedia
CDR	Clinical drug response
CNN	Convolutional neural network
CRISPR	Clustered regularly interspaced short palindromic repeats
CTRP	Cancer Therapeutics Response Portal
CV	Cross-validation
DANN	Domain adaptive neural network
DL	Deep learning
DNA	Deoxyriboneucleic acid
FC	Fully connected
FDR	False discovery rate
FPKM	Fragments per kilobase transcript per million mapped reads
GAT	Graph attention network
GCN	Graph convolutional network
GDSC	Genomics of Drug Sensitivity in Cancer
GEx	Gene expression
GNN	Graph neural network

GSC	Gene set characterization
H-GCN	Heterogeneous graph convolutional network
IC50	Half-maximal inhibitory concentration
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCO	Leave-cell lines-out
LPO	Leave-pairs-out
LSTM	Long short term memory
ML	Machine learning
MLP	Multi-layer perceptron
mRNA	Messenger RNA
MSE	Mean squared error
NECTARE	Knowledge embedding of compounds through targets, response, and essentiality
NSMB	Nested stochastic block model
P2C	Preclinical-to-clinical
PCA	Principal component analysis
PCC	Pearson correlation coefficient
RBF	Radial basis function
RECIST	Response evaluation criteria in solid tumors
ReLU	Rectified linear unit
RI	Rand index
RMSE	Root mean squared error
RNA	Ribonucleic acid
SCC	Spearman's rank correlation coefficient
SD	Standard deviation
SMILES	Simplified molecular input line entry system
STRING	Search tool for the retrieval of interacting genes/proteins
SVM	Support vector machines
SVR	Support vector regression
TCGA	The Cancer Genome Atlas

THCA	Thyroid carcinoma
TINDL	Deep learning pipeline with tissue informed normalization
TPM	Transcripts per kilobase million
tRNA	Transfer RNA
UMAP	Uniform manifold approximation and projection

Chapter 1

Introduction

Precision medicine is a promising concept that tailors the prescribed treatments to the patient's clinical and molecular profiles. Generally, patients would have different responses to different kinds of drugs, even for patients with similar cancer types. As such, it is important to adapt an individual's treatment plan to increase their chances of survival. Unlike standard treatment plans, which are typically prescribed using a handful of criteria (e.g., cancer type), precision medicine aims to cater the treatment strategy to the unique molecular and clinical properties of each patient. However, due to the complexity and size of factors to consider, computational approaches are needed to achieve this goal.

Machine learning (ML) is one such category of computational approaches that has shown great success in modeling complex relationships between observed variables and outcomes. In the past, researchers aiming to utilize ML models typically handcrafted their input features due to limitations in both modeling and hardware capacity. Although this is still applicable in the present day, the re-emergence of neural networks and the invention of powerful processors reduced the burden of feature engineering. Deep learning (DL), a class of ML approaches that is based on artificial neural networks, has been in the spotlight due to its predictive power. DL enabled high-dimensional data to be a feasible form of input, allowing researchers to focus on other aspects of their applications, such as developing appropriate experiments and analyzing their results. However powerful of a technology it may be, the architecture design and careful consideration of various factors, such as quantity and quality of the data, heavily influence the success of a DL model.

In an ideal scenario, a successful model in this application should be able to determine drugs that are suitable to the patient's genetic makeup. However, identifying the most appropriate drug for a patient does not come easily. One of the major challenges in pharmacogenomic applications of ML is the limited amount of clinical drug response (CDR) data available due to technical and ethical constraints. The scarcity of CDR data hinders the development of more complex models trained on clinical data due to the possibility of overfitting.

Instead of identifying suitable drugs, one way of simplifying the task is to predict the patient's response to a specific drug. Multiple studies (e.g., [1–10]) have demonstrated success in predicting *in vitro* (in a controlled environment like a petri dish outside the living organism) drug responses using different *omics* data (genomics, transcriptomics, proteomics, etc.) under the hopes that this is a step closer to predicting CDR. However, the reported performance metrics in many of these approaches are only reliable in specific datasets. Most of the time, models with decent accuracy for a test/validation set within the same dataset perform much worse when tested on a separate dataset of the same type (cross-dataset prediction). In addition, the biological differences between in vitro preclinical data and in vivo clinical patient data give rise to even more challenges in the preclinical-to-clinical (P2C) setting [11].

Additionally, there are many factors that may influence the drug response. Some of these are well-defined and can be conveniently translated into features. However, some factors are much less obvious but can be extracted through careful considerations in the model architecture. I follow this line of thought throughout this thesis, acknowledging similarities, differences, and relationships between different samples, drugs, and factors.

Looking at this prediction problem from the opposite perspective, it is natural to ask whether it is possible to identify and characterize properties that could induce the observed responses. Realistically, answers derived computationally would be difficult to assess for correctness. However, computational approaches can generate data-driven hypotheses that can be experimentally tested, thereby reducing the search space for further studies.

In this thesis, I study different applications of DL in drug response prediction. Considering various limitations, I present appropriate pipelines to model and analyze the data. In

each study, I propose algorithms and model architectures that seek to address gaps identified in prior approaches, as well as generate biological and computational insights, and suggest avenues for improvement. The rest of this chapter discusses the organization of the remaining chapters of this thesis and their contributions.

1.1 Thesis Organization and Contributions

This section describes the organization of the thesis and the main technical contributions of each chapter.

- **Chapter 2**: This chapter provides the background material and a literature review of different approaches in drug response prediction. I first introduce relevant concepts in cancer medicine and molecular biology. Next, I give a small primer on DL, specifically on graph neural networks. Then I present literature on drug response prediction using ML techniques. For each section, I provide a background of the problem and clarify the main task. I also foreshadow some of the contributions of this thesis as I delve into gaps and issues in the prior studies that are being discussed. Additionally, I include studies that have been published after the completion of some of my projects in order to provide an updated insight into the topic.

- **Chapter 3**: I introduce an approach in clinical drug response prediction. I describe traditional and contemporary approaches that were typically used for this task, considering the scarcity of available labeled data. I compare the proposed approach to these methodologies in terms of their ability to segregate responders from non-responders for certain drugs. Detailed analyses of the model and features are also discussed in this chapter. The main contribution of this chapter is a method for preclinical-to-clinical drug response prediction and biomarker identification called *TINDL*, which uses a novel tissue-informed normalization method to allow prediction for clinical samples despite the model only being trained using preclinical samples.

- **Chapter 4**: In this chapter, I focus on *in vitro* drug response prediction. This chapter proposes a novel approach in encoding drug representations by incorporating a bipartite

graph of extreme cell line-drug responses. I also discuss different approaches in data-splitting and performance evaluation depending on the intended use of the model. I demonstrate the superior performance of my proposed method empirically and justify this by analyzing different aspects of the model. The main contribution of this chapter is the drug response prediction model called *BiG-DRP*. The defining characteristic of this model is the way information is propagated through the bipartite graph, a paradigm that has not been thoroughly explored.

- **Chapter 5**: I explore the idea of improving the predictive power of the previous model (Chapter 4) by including extra information to form a knowledge graph. I examine the usefulness of CRISPR gene effects and drug targets in the drug response prediction problem via a graph-based drug representation component called *NECTARE*. Additionally, this chapter analyzes the trained model in terms of the input features and the knowledge graph. In this chapter, I contribute to the field by characterizing the effects of aforementioned auxiliary information, and a heuristic in visualization of relevant entities in the knowledge graph.

- **Chapter 6**: Here, I re-iterate the main contributions of this thesis and discuss the results at a higher level in order to connect the previous three chapters. Additionally, I enumerate different possible directions of this research, including applications to different tasks and strategies to improve the approaches that I have developed.

1.2 Publications and Author Credits

The contents of this thesis are composed of published and unpublished materials. I have only included my own contributions from these publications, which are not part of any other theses/dissertations. In this section, I list the chapters, the relevant articles to the chapter, and the contributions of the authors.

- **Chapter 3**: This chapter is based on our published article [12] under the Creative Commons CC-BY license. Some figures, tables, and supplementary materials were directly lifted from this article.

D.E. Hostallero, L. Wei, L. Wang, J. Cairns, and A. Emad, "Preclinical-to-clinical anticancer drug response prediction and biomarker identification using TINDL," *Genomics, Proteomics, & Bioinformatics*, vol. 21, no. 3, pp. 535-550, 2023.

A. Emad and J. Cairns conceived the study and designed the project. A. Emad led the computational aspects of the study. D.E. Hostallero designed the algorithms, implemented the pipeline and baselines, and performed the statistical analyses of the results. J. Cairns led the experimental validation of the results. L. Wei and L. Wang performed the gene knockdown experiments.

- **Chapter 4**: This chapter is based on our published article [13]. Some figures, tables, and supplementary materials were directly lifted from this article. License to reproduce was provided by Oxford University Press and Copyright Clearance Center.

  D.E. Hostallero, Y. Li, and A. Emad, "Looking at the BiG picture: Incorporating bipartite graphs in drug response prediction," *Bioinformatics*, vol. 38, no. 14, pp. 3609-3620, 2022.

  A. Emad and D.E. Hostallero conceived the study and designed the project. D.E. Hostallero designed the model and implemented the pipeline. D.E. Hostallero and Y. Li ran the baseline methods and performed the statistical analyses of the results.

- **Chapter 5**: This chapter is based on an unpublished project under the Genentech-Mila-McGill collaboration. Permission from the concerned parties has been obtained.

  A. Emad and D.E. Hostallero conceived the study and designed the project. D.E. Hostallero designed the model, implemented the pipeline, and performed the statistical analyses. Guidance on the project was given by S. Martin and M. Hafner.

Chapter 2

Background Material and Literature Review

2.1 Relevant Concepts in Cancer Medicine

Our body is composed of trillions of cells, and the normal growth and division of these cells are regulated by the body's control mechanisms [14]. Cells typically die after a period of time or when they get damaged. However, there are instances when this process takes a different turn, allowing the transformation of normal cells into abnormal cells due to hereditary or environmental factors (subject to natural selection [15]). Abnormal cells may form masses, known as tumors, which can either be malignant or benign. Benign tumors may exhibit growth but lack the capacity to spread. In contrast, malignant tumors possess the ability to both grow and spread to various regions of the body. As such, malignant tumors are "cancerous". Cancer refers to a group of diseases whose main distinguishing factor is the uncontrolled growth and proliferation of abnormal cells, which can invade and destroy other surrounding tissues. Note that there are also types of cancer that do not form tumors, such as hematologic (blood) cancers [16, 17].

2.1.1 Cancer Treatment

Cancer is usually treated using one or a combination of the following procedures: surgery, immunotherapy, radiation therapy, chemotherapy, and targeted therapy [18]. *Surgery* in-

volves physically removing cancer, typically using a scalpel to cut the tumor from the body. However, there are other ways to perform surgeries without cuts, such as cryosurgery, lasers, hyperthermia, and photodynamic surgery. *Immunotherapy* refers to cancer treatment done by helping the patient's immune system act better against cancer. Some examples are drugs that block immune checkpoints. Immune checkpoints prevent the immune response from being too strict; thus blocking these checkpoints enables vigorous response from immune cells in order to fight off cancer. *Radiation therapy* kills cancer cells and shrinks tumors by applying high doses of radiation. *Chemotherapy* utilizes cytotoxic drugs to kill cancer cells. However, a side effect of using chemotherapy drugs is that healthy cells can also be affected, slowing down the normal cells' growth or possibly killing normal cells. *Targeted therapies* are treatments that are designed to focus on specific molecular targets (e.g., proteins) that control the cancer cells' growth and proliferation. This causes less harm to normal cells, unlike chemotherapy drugs. Only chemotherapy and targeted drugs are in the scope of this thesis.

2.1.2 Drug Response

Drug response is quantified in various ways depending on the domain of the study. Preclinical drug responses are commonly measured using the area under the dose-response curve (AUC) and the half-maximal inhibitory concentration (IC50). Clinical drug responses are typically binned into a category using the response evaluation criteria in solid tumors (RECIST) [19, 20].

For preclinical studies, samples (e.g., cancer cell lines, organoids) are treated with a given compound at different doses (concentrations), usually in replicates. Cell viability is then measured after a fixed period of time and compared to control (untreated) samples to calculate the percent viability [21]. After gathering all the measurements, a sigmoidal curve is fitted to the dose-response data, with the percent viability being a function of the concentration. The AUC is the area under this fitted sigmoidal curve, calculated by integrating from the lowest to the highest dose. Usually, the AUC is normalized by the dose range, resulting in AUC values within zero to one. However, some datasets like the Cancer Therapeutics Response Portal (CTRP) [22] allow for values greater than one since they are

not normalized. The IC50 corresponds to the concentration point in the fitted curve, which corresponds to 50% viability. An advantage of IC50 is that it is a physical measurement, which is easier to interpret. However, it is common for the dose-response data not to cross the 50% viability within the given concentration range. In these cases, the IC50 technically does not exist, but they are typically interpolated through the fitted curve, albeit with lower confidence.

For clinical studies, drug response is measured by the change in size of the lesion post-treatment after a period of time. RECIST is a set of criteria that allow practitioners to stratify patients into response classes [19, 20]. First, a baseline measurement is performed on the target lesions before the start of the treatment. Subsequent measurements will be evaluated depending on the protocol. Measurements of the lesions are summarized as the *sum of diameters* and are done either clinically or through imaging. The criterion is divided into four ordinal classes: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). CR corresponds to the elimination of all target lesions. PR is declared when there is at least a 30% decrease in the sum of diameters, while PD indicates that there has been at least a 20% increase in the sum of diameters of the target lesions. When the change is insufficient (between -20% to 30%, exclusive), then the response is SD. Note that other considerations may alter the classification, which can be found in their published guidelines [19, 20]. In some drug response prediction studies [11, 23, 24], these are grouped as responders/sensitive (CR and PR) and non-responders/resistant (SD, PD).

2.2 Relevant Concepts in Molecular Biology

This section briefly summarizes some relevant biological concepts and terminologies that are relevant to this thesis.

The complete set of genetic information in an organism is called the genome. Although the genome is referred to as the blueprint of life, it is merely a constraint, and an organism's *genotype*, alone, does not dictate its observable traits (called *phenotype*). This biological information is carried in the molecule called the deoxyribonucleic acid (DNA), which is composed of two strands winding around each other. Each strand contains a sequence of

nucleotides with one of the four bases: adenine (A), thymine (T), cytosine (C), and guanine (G). Hydrogen bonds are formed between complementary base pairs, A-T and C-G, forming the twisted ladder structure called a double helix. Biological information, such as instructions for the development of an organism and its responses to the environment is encoded in this sequence of bases.

The majority of an organism's DNA is found in the cell nucleus, although there is a small amount of DNA located in the mitochondria. DNA-protein complexes, called chromatins, are formed to help package the DNA into the nucleus. Chromatins fold into a characteristic formation called chromosomes, each of which contains a single molecule of DNA and the packaging proteins (called histones). Additionally, chromatin is also the mechanism that controls how the genome is read across different cells, as each cell contains the same blueprint but is executed differently. The human genome is composed of 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Each pair contains one chromosome from each parent.

Genes are arguably the most-studied regions of the chromosome and are often referred to as the basic unit of heredity. A gene is a segment of the DNA that contains instructions for the construction of specific proteins or ribonucleic acid (RNA) molecules. Proteins are the building blocks of the cell and they carry out most of the functions within organisms. They are also crucial to the structure and regulation of the tissues and organs. Only a tiny percentage of the DNA comprises protein-coding genes, and the majority of the DNA is non-coding. Although initially thought of as "junk," non-coding DNA actually has some purpose [25]. One example is the regulation of other genes, as they determine when the genes are turned on and off [26].

2.2.1 The Central Dogma of Molecular Biology

The central dogma of molecular biology describes the general flow of genetic information in a biological system. This dogma is often simplified as "DNA makes RNA, and RNA makes protein". Related to this is the *process* of gene expression[1], which describes how information encoded within genes is transformed into gene products (e.g., proteins) that

[1]not to be confused with the related data type describing the abundance of mRNA transcripts

affect the organism's phenotype. Gene expression involves two essential processes, namely transcription and translation.

The first part of this process is called *transcription*, where an RNA molecule is created using a portion of the DNA as a template. First, an enzyme called RNA polymerase binds to the DNA template strand, with the help of transcription factors, which determine the DNA sequences that should be transcribed. The RNA polymerase then begins assembling a sequence of nucleotides that is complementary to the DNA template strand. However, thymine (T) is replaced by uracil (U) as the complement of adenine (A) during the synthesis of precursor messenger RNA (pre-mRNA). The pre-mRNA is processed by adding a $5'$ cap and a poly-A tail to the chain. From this, the non-coding regions (introns) are removed, joining the coding regions (exons) together in a process called splicing. The outcome of this process is the mature messenger RNA (mRNA), which is then transported outside the nucleus to the cytoplasm [27, 28].

The second part of this process is called *translation*, where the mRNA is used to create proteins. During translation, the mRNA is read in consecutive triplets of nucleotides called codons. Each codon specifies one amino acid. Translation is initiated as the ribosome attaches itself to the mRNA and finds the start codon (AUG). The ribosome is surrounded by molecules called transfer RNA (tRNA), which consists of two ends: the amino acid attachment site and the anticodon. The tRNA with the anticodon that complements the codon at the current position of the ribosome then binds to the mRNA. The ribosome shifts to the next codon and tRNA molecules attach accordingly, creating a polypeptide chain. When the ribosome encounters a stop codon (UAA, UAG, UGA), the ribosome/mRNA complex is disassembled. The amino acid chain is released and then it folds into an active protein that will perform its function. Since there are four different bases (A, U, C, G), there are 64 possible combinations of triplets. However, there are only 20 different amino acids, allowing for multiple codons to refer to the same amino acid. This is known as redundancy, a mechanism that mitigates possible damages caused by unexpected changes in the sequence [27, 28].

2.2.2 Mutation

Mutation is defined as a change in the DNA sequence of an organism. This divergence may occur as a result of an error during DNA replication, exposure to mutagens (substances that cause mutation), or viral infection. Mutations can be inherited from either parent. These mutations, called germline mutations, occur in the parents' reproductive cells. Alternatively, changes to an organism's DNA that occur in any cell (except egg or sperm) after conception are called somatic mutations.

Mutations happen all the time and rarely have any serious effect on an organism's health. However, mutations can cause a gene to stop working properly and may result in genetic diseases such as cancer. When a mutation happens in a protein-coding region, the triplet (codon) that encodes the amino acid changes, which could possibly change the protein sequence. A *nonsense mutation* occurs when a nucleotide substitution results in the production of a stop codon, therefore prematurely terminating the sequence. When the substitution causes the triplet to code for a different amino acid, it is called a *missense mutation*. If the amino acid is not changed (amino acids can be mapped to multiple codons), then it is called a *silent mutation*. Insertion and deletion of a nucleotide in a protein-coding sequence cause *frameshift mutation*, which alters the groupings of the codons, usually leading to a different protein. Predictive models commonly use a binary representation to indicate whether a sample harbored a mutation for a gene, usually only taking into account nonsense, missense, and frameshift mutations.

2.2.3 Proteomic and Transcriptomic Data

Proteins can be interpreted as workers that execute the instructions in the genome. As such, the measurement of protein abundance (*proteomic data*) is a valuable resource in understanding cellular processes and disease mechanisms. However, technology and cost-related challenges have previously veered researchers away from gathering proteomic data. Nevertheless, proteomic data are recently becoming more and more available [29].

Following the logic of the central dogma, the quantification of mRNA has been extensively used as a proxy for proteomic data. However, this substitution has its caveats due to the imperfect correlation of protein and mRNA [30, 31]. In the context of *transcriptomic*

data, the term "gene expression" or GEx refers to the abundance of mRNA transcripts for a particular gene. Thus, a sample is often represented as a vector where each element corresponds to a gene. GEx is typically measured using RNA sequencing (RNA-seq) and microarrays, although the latter is falling out of favor due to limitations in its range and sensitivity.

2.2.4 CRISPR Gene Knockouts

The genome-wide CRISPR (clustered regularly interspaced short palindromic repeats) - Cas9 knockout screening, or simply *CRISPR knockout screening*, is an approach to uncovering relationships between genotype and phenotype through ablation of genes and analysis of the outcomes. In a simplified manner of explanation, CRISPR-Cas9 is a gene editing technology that utilizes a guide RNA (gRNA) designed to target a gene and Cas9 (CRISPR-associated protein 9). The gRNA guides the CRISPR-Cas9 system to a precise location in the genome for which the Cas9 will create a double-stranded break, ultimately leading to knockout [32, 33]. Data gathered from CRISPR knockout screening is usually presented as *gene effect scores*. This thesis used scores from the Chronos pipeline [34], where negative scores indicate cell death or inhibition of cell growth following the gene knockout.

2.3 Pharmacogenomic Datasets

Data used in this thesis were gathered from publicly available databases. This section describes high-level information about the obtained datasets. These data were filtered and processed differently depending on each chapter's objectives, resulting in non-uniform numbers across the different chapters.

The clinical dataset used in this thesis is sourced from The Cancer Genome Atlas (TCGA)[2] [35]. TCGA comprises patient and molecular data from 33 different projects, each focusing on a specific cancer type. For this thesis, data related to gene expression (GEx), mutations, RECIST drug response, and metadata were obtained from TCGA. The data from TCGA are categorized as *clinical* as they were collected from patient tumors and their

[2] `portal.gdc.cancer.gov`

responses to drugs. The atlas ensures homogenized data across the various projects. Only the molecular profiles of primary tumors at baseline (i.e., before treatment) were included in my studies. Although the dataset contains over 20,000 samples, only a small percentage includes drug response information. Additionally, some patients received multiple drugs either simultaneously or sequentially.

For preclinical drug responses, two of the most popular datasets are (Genomics for Drug Sensitivity in Cancer (GDSC)[3] [36] and Cancer Therapeutics Response Portal (CTRP) [4] [22]. Both datasets provide drug responses of hundreds of cancer cell lines (CCLs) to hundreds of drugs. For GDSC, drug responses are given as IC50, AUC, and binary (sensitive/resistant) labels. In this thesis, the two versions of GDSC (v1 and v2) are combined since there is only a small overlap between the two versions. Redundant data points (e.g., drug synonyms, and replicates with different dose ranges) were addressed in the following chapters. For CTRP, drug responses are in the form of unnormalized AUC. Note that these databases are continually being updated and there is a substantial overlap between the GDSC and CTRP datasets (100+ drugs and 500+ CCLs as of 2024).

Molecular profiles of the CCLs can be obtained from GDSC's website although a more updated version can be obtained from Cell Model Passports[5] [37]. CTRP uses CCLs that are profiled in the Cancer Cell Line Encyclopedia (CCLE) Project [38], in which the data can be accessed through the DepMap Portal[6]. The DepMap Portal also provides CRISPR knockout screening data [34] corresponding to the same CCLs.

2.4 Relevant Concepts in Deep Learning

Deep learning (DL) is a general term used to denote machine learning approaches based on artificial neural networks. Deep neural networks (DNNs) refer to the class of models trained using DL, and the word "deep" pertains to the idea that typical DNNs are composed of multiple layers. In most contexts, a layer is an abstraction of a set of operations that typically involves trainable parameters (weights and biases) and an activation function.

[3] cancerrxgene.org
[4] portals.broadinstitute.org/ctrp
[5] cellmodelpassports.sanger.ac.uk
[6] depmap.org

2.4.1 Multi-layer Perceptron

The most widely-used form of DL and arguably the most basic DNN is the multilayer perceptron (MLP), which is composed of multiple *fully connected layers* or *dense layers*[7] stacked on top of each other. The term MLP is also often interchanged with "fully connected feedforward neural networks" or simply "feedforward networks". In this context, feedforward alludes to the fact that there is no recurrence (i.e., no dependence on previous states), unlike recurrent neural networks.

Given a sample $\mathbf{x} \in \mathbb{R}^d$, where d is the number of features, a forward pass for a fully connected layer $f(\cdot)$ is given by:

$$f(\mathbf{x}) = \sigma(\mathbf{W}^\top \mathbf{x} + \mathbf{b}). \tag{2.1}$$

Here, $\mathbf{W} \in \mathbb{R}^{d \times s}$ is the weights of the layer, and $\mathbf{b} \in \mathbb{R}^s$ is the bias. The activation function is denoted as σ. The output of the layer is a vector of size s. This is the building block of many neural network architectures, and these layers can be stacked on top of each other to increase the representation power of the model.

2.4.2 Graph Neural Networks

Here, I introduce the basics of graph neural networks (GNN). GNN is a class of neural networks that cater to graph-structured data through a form of neural message passing where "messages" are exchanged among nodes. This is called the *message-passing framework* [40].

A graph comprises a set of nodes V connected through a set of edges E. Each node u has some information which is called the *hidden embedding* $\mathbf{h}_u^{(k)}$ at the kth message-passing iteration (or the layer). Here, $\mathbf{h}_u^{(0)}$ would be a transformation (including identity) of the initial features (also called attributes) of the node. Using Hamilton's [40] notation, the message-passing framework in the perspective of node u is given by:

[7] Not to be confused with DenseNet [39], a type of convolutional network.

$$\mathbf{h}_u^{(k+1)} = \text{UPDATE}^{(k)}\left(\mathbf{h}_u^{(k)}, \text{AGGREGATE}^{(k)}(\{\mathbf{h}_v^{(k+1)}, \forall v \in N(u)\})\right) \tag{2.2}$$

$$= \text{UPDATE}^{(k)}\left(\mathbf{h}_u^{(k)}, \mathbf{m}_{N(u)}^{(k)}\right) \tag{2.3}$$

where $N(u)$ is the neighborhood of u. The $\text{AGGREGATE}^{(k)}$ function is an arbitrary differentiable function that combines the embeddings of the neighborhood $N(u)$ to generate an aggregated *message* $\mathbf{m}_{N(u)}^{(k)}$. The $\text{UPDATE}^{(k)}$ function is another arbitrary differentiable function that merges the message $\mathbf{m}_{N(u)}^{(k)}$ from the neighborhood and the previous embedding $\mathbf{h}_u^{(k)}$. To summarize, in a single iteration (or layer) of the message-passing framework, each node receives a message from its neighborhood and updates its own embedding using the received message and its previous embedding. After K iterations of message-passing, the output would be the final embedding of the node, which would contain information propagated from the K-hop neighbors of u.

A primary manifestation of this framework is the basic GNN. For a layer k, the basic GNN is given by

$$\mathbf{h}_u^{(k+1)} = \sigma\left(\mathbf{W}_{\text{self}}^{(k+1)}\mathbf{h}_u^{(k)} + \mathbf{W}_{\text{neigh}}^{(k+1)}\sum_{v \in N(u)}\mathbf{h}_v^{(k)} + \mathbf{b}^{(k+1)}\right) \tag{2.4}$$

where $\mathbf{W}_{\text{self}}^{(k+1)}$ and $\mathbf{W}_{\text{neigh}}^{(k+1)}$ are trainable parameters of size $d^{(k+1)} \times d^{(k)}$, $\mathbf{b}^{(k+1)} \in \mathbb{R}^{d^{(k+1)}}$ is the bias term, and σ is the activation function. One could see the resemblance of this equation to a feedforward layer in Equation 2.1, as it looks similar except for the extra term concerning the neighborhood. This could also be rewritten in the message-passing form.

$$\text{UPDATE}^{(k)}(\mathbf{h}_u^{(k)}, \mathbf{m}_{N(u)}^{(k)}) = \sigma(\mathbf{W}_{\text{self}}^{(k+1)}\mathbf{h}_u^{(k)} + \mathbf{W}_{\text{neigh}}^{(k+1)}\mathbf{m}_{N(u)}^{(k)}) \tag{2.5}$$

$$\mathbf{m}_{N(u)}^{(k)} = \text{AGGREGATE}^{(k)}(\{\mathbf{h}_v^{(k)}, \forall v \in N(u)\}) \tag{2.6}$$

$$= \sum_{v \in N(u)}\mathbf{h}_v^{(k)} \tag{2.7}$$

Note that many variations of the GNN can be defined using different UPDATE and AGGREGATE functions. For example, GraphSAGE [41] uses concatenation instead of addition for UPDATE. Some GNNs such as GIN [42] also use multiple layers of feedforward networks instead of simply using a parameter matrix $\mathbf{W}_{\text{self}}^{(k)}$.

2.5 Approaches in Preclinical Drug Response Prediction

In the previous decade, high-throughput sequencing has provided a large amount of molecular "omics" profiles (e.g., genomic, transcriptomic, proteomic, epigenomic, etc.) for hundreds of different CCLs. Multiple studies, such as GDSC [36], CCLE [38], and CTRP [22], have released these omics profiles along with the CCLs' responses to hundreds of drugs. These datasets have expedited the development of sophisticated models of preclinical drug response prediction. Drug response for CCLs is usually measured using continuous values such as IC50 and AUC, which naturally translates to a regression task. However, classification (discretized labels) and ranking tasks are also being used in the literature.

2.5.1 Methods based on Traditional Machine Learning

The National Cancer Institute and the Dialogue on Reverse Engineering Assessment and Methods (NCI-DREAM) challenge for drug sensitivity prediction [1] has shown the popularity and effectiveness of traditional machine learning in the drug response prediction task. The participants of this challenge were asked to rank a set of CCLs according to their sensitivity to a specific drug. Support vector machines (SVM), regularized regression (lasso, elastic net, ridge), and random forests were some of the most used models in the challenge. The winning team proposed Bayesian multitask multiple kernel learning [1], which allowed simultaneous training for all drugs using multiple kernels that focus on different "views" of the sample (i.e., different omics data types). The Bayesian part comes from their assumption that the parameters of the model are random variables under a specific distribution. Other methods also had the same idea of leveraging the different omics data. In [43], they separately trained random forests for each data type. The outputs of these random forests

were then aggregated using least squares regression. Although the top-performing models in the NCI-DREAM challenge favor multi-omic integration, the evaluation of all submitted models to this challenge concluded that GEx profiles were the most informative features [1, 44]. As such, many subsequent studies, especially in deep learning, have focused on GEx as the representation of CCLs [3–9].

The high dimensionality of GEx has prompted researchers to find a way to reduce the feature size. Principal component analysis (PCA) and correlation-based filtering were two of the most utilized dimensionality reduction techniques in the NCI-DREAM challenge [1]. Knowing that genes and their protein products interact with each other in a cell, ProGENI [45] leveraged protein-protein interaction (PPI) networks to create a ranking of genes using random walks with restart. This ranking can then be used as a method to eliminate genes that are deemed less critical. MDREAM [46] focused on predicting the response of acute myeloid leukemia samples, which allowed them to select some of their features manually. Their model used a stacking technique, where they trained individual SVM models per drug and then used the outputs of these individual models as input for another set of SVM models. The rationale is that stacking their models allows information to be shared across different drugs.

2.5.2 Methods for Imputation of Drug Response Matrices

Labels for in vitro drug response are usually presented as a response matrix consisting of CCLs as columns and drugs as rows. Given that, some methods are only interested in imputing missing data from the response matrix. In these methods, the test set is composed of the same set of CCLs and drugs from the training set, although the CCL-drug as a pair was not encountered during training. MCDRP [47] used a matrix completion algorithm called soft-impute [48] to fill in the missing data. NRL2DRP [2] took a different approach by incorporating a graph with CCLs, drugs, and genes as nodes. These nodes are connected through sensitivity (CCL-drug), mutation (CCL-gene), and protein-protein interaction (gene-gene). They then generated node embeddings using LINE [49] and used these LINE-based embeddings as features for an SVM. Note that LINE is a transductive method that relies on the fixed topology of a network, which means that predicting for drugs and CCLs that are not

in the graph during training cannot be conveniently done. The advantage of both MCDRP and NRL2DRP, however, is that they both eliminated the need for high dimensional omics features to represent the CCLs.

SRMF [50] calculated a drug similarity network and a CCL similarity network using the drugs' fingerprints and CCLs' GEx profiles, respectively. They then proposed to impute the response matrix by using a matrix factorization method that is regularized by the constructed similarity networks. Unlike SRMF, which utilized a fully connected similarity network (i.e. a similarity matrix), WGRMF [51] used a sparse similarity network, opting to use local neighborhoods of highly similar drugs/CCLs rather than global similarities. Liu et al. [3] proposed to use a combination of matrix factorization and ridge regression (using GEx). These methods lose the "featureless" advantage of matrix factorization but are conceptually better in terms of sample representation since they enable CCLs/drugs to be independent of the response matrix. The issue with these methods is that their predictive ability is limited to the CCLs and drugs that they used to train (i.e., cannot predict on new CCLs/drugs) due to the limitations of their matrix factorization. Simply put, a new row/column (drug/CCL) without any prior data (labels) will not give the model anything to base their predictions on, essentially treating all unknown drugs/CCLs the same. Additionally, the factored matrix cannot be used if the dimensions of the training and test sets are different. Another matrix factorization method is CaDRReS [52], which decomposed the response matrix into drug/CCL-specific biases, drug latent features, and CCL latent features. Although CCL latent features are calculated as a linear transformation of the CCLs' GEx profiles in CaDRReS, the CCL-specific bias in their formulation implies that test CCLs must also exist during training, or alternatively, be given as a "known" bias if predicting for a new CCL.

2.5.3 Methods based on Deep Learning

Due to its predictive power, DL has been a popular approach in the past decade. However, since DL models require a large number of samples, many DL methods in this field frame the problem as a *paired prediction task* to increase the sample size. In this paradigm, the input consists of the CCL-drug pairs [53] as opposed to the one-model-per-drug formula-

tion. In addition to the CCL features, drug features are required to represent the drug. Extended-connectivity fingerprints (also known as Morgan fingerprints) [54] and drug descriptors (physical and chemical properties of the molecule, such as the number of hydrogen bond donors/acceptors) are commonly used features in this task [4, 55–57]. PathDNN [5] and ConsDeepSignaling [58] used drug targets as features. Although drug targets are intuitive, they are not unique to a drug, and therefore, multiple drugs may be identically represented in the model. Furthermore, drug target information can be unavailable in rare cases. For kinase inhibitors, DEERS [59] introduced kinase inhibition profiles as drug features, which correspond to their strength of inhibition for a panel of protein kinases. Another uncommon drug representation is the differential gene expression profiles of a specific CCL post-treatment [60]. These alternative drug representations establish relationships to the CCL component of the input. However, these representations are more difficult to acquire, and particular considerations must be taken into account (e.g., choices of CCLs and drugs).

Since drug molecules are commonly represented as a string using the SMILES notation, Liu et al. [6] proposed to translate the SMILES notation, which has 72 different symbols, into a binary matrix where each row is a distinct symbol and each column corresponds to a character in the SMILES sequence. They then used a one-dimensional convolutional neural network (1D-CNN) to encode this matrix and the CCL's genetic features, namely copy number alterations (CNA) and somatic mutations. CDRScan [61] used a similar convolutional architecture on somatic mutations but opted for drug descriptors generated by PaDEL [62]. However, since 1D-CNNs work in interval windows within an ordered sequence, the use of this technique on non-sequential data, such as mutations and drug descriptors, is not ideal.

DeepCDR [63], GraphDRP [64], GraphCDR [65], and DRPreter [66] used graph neural networks to encode the graph structure of the drugs, represented by their molecular graphs (atoms as nodes, bonds as edges). DGSDRP [67] utilized both the SMILES sequence and the molecular graph of the drug, which they have observed to have superior performance over using only one of the two. Over the years, different methods of drug fingerprinting have also been proposed [62, 68, 69]. Zagidullin et al. [70] compared some of these fingerprinting methods and have shown that deep graph infomax [71], which was not specifically created for drug representation, performed best among other fingerprints.

Uniquely, GraphCDR [65] formulated drug response prediction as a *link prediction problem* by constructing a bipartite graph (apart from the molecular graph of the drug) composed of CCL nodes and drug nodes. An edge is present between a CCL node and a drug node if the CCL is sensitive to the drug, according to some IC50 threshold. GraphCDR was trained using a combination of contrastive and supervised learning to predict whether an edge exists between two nodes.

Other DL studies focused on CCL representation. DeepDSC [4] utilized the CCLs' GEx and drugs' Morgan fingerprints [54] for prediction. They proposed to extract the GEx's latent features using a stacked autoencoder before training for the drug response prediction task due to the high dimensionality of the GEx. Similarly, VAEN [72] used a variational autoencoder for dimensionality reduction and then trained elastic net models using the compressed representations to predict drug responses. Ding et al. [73] also trained an autoencoder to extract features from the GEx but utilized the encodings from multiple hidden layers of the autoencoder instead of the typical "bottleneck" layer.

2.5.4 Methods for Interpretability

Model interpretability is a crucial requirement for computational models, enabling the identification of biomarkers and mechanistic insights in drugs' mechanisms of action. However, most DL models, as well as SVMs and principal component regressions, lack interpretability. Jang et al. [44] suggested the use of ridge regression or elastic net in drug sensitivity modeling. The significance of this approach lies in the direct interpretability of the magnitude of the learned coefficients, which is deemed proportional to the feature's importance. However, it is important to note that these models' predictive capacity is often inferior to more sophisticated yet "black box" models.

NCFGER [74], HIWCF [75], and Dual-layer CSN/DSN [7] have proposed collaborative filtering methods by looking at responses of similar CCLs to similar drugs, a reminiscent of the nearest neighbors algorithm. These approaches offer a reassuring level of transparency since predictions can be readily mapped back to the neighboring CCLs/drugs that bear high similarity to the query (test) CCL/drug. However, this transparency does not directly lead to biological insights and statistical analyses must be performed on the mapped CCLs/drugs

to reveal relevant properties.

For deep learning models, some methods attempted to incorporate "prior knowledge" into their models to alleviate the black box nature of neural networks. PathDNN [5], PASNet [8], and ConsDeepSignaling [58] used pathway information to constrain neural network connectivity. These approaches utilize a so-called pathway layer, where the trainable weight matrix is masked with a binary matrix representing pathway membership of the genes. In this case, the input to the layer corresponds to gene-level information (e.g., GEx, mutation), and the output of this layer corresponds to the pathway-level summary based on the pathway's members. DrugCell [9] expands on this idea by defining a hierarchy of biological processes, therefore nesting various gene sets to represent cellular subsystems at different scales.

Along with pathway membership constraints, attention mechanisms were used in HiDRA [57], enabling straightforward attribution of gene/pathway relevance in the form of attention scores. PathDSP [76] took a more direct approach by performing pathway enrichment analysis on the gene-level information of the samples before feeding it to an MLP. DRPreter [66] superimposed a template graph on the CCLs using STRING PPI networks [77]. Graph neural networks were used to create subgraph embeddings, in which a pathway defines each subgraph. They then used a transformer architecture below their final drug response predictor to provide attention scores for interpretability. Although these models claim to have improved the level of interpretability, Li et al. [78] have noted that prior information can sometimes function unexpectedly (e.g., no actual biological meaning are being embedded). Additionally, Bertin et al. [79] have observed that the curated graphs, like the STRING PPI, seem to have limited benefits in incorporating prior knowledge.

2.6 Approaches in Clinical Drug Response Prediction

Although many methods have been reported to be successful in predicting in vitro drug responses, many of these methods do not directly translate to clinical drug response (CDR) prediction due to the distributional discrepancies of the omics features and biological differences of the samples. Notably, CCLs are more homogeneous than tumors, and their growth is

limited in a controlled 2D environment, whereas tumors have microenvironments. However, it has been shown that preclinical datasets still hold value, albeit less ideal [11, 24, 80–84]. Given the scarcity of available CDR data, augmenting the data or fully training the models using preclinical datasets is a viable (and in many cases a necessary) option. Throughout this thesis, I will loosely use the term "domain" as a generalization of data types or data sources, typically characterized by their feature distribution or overall context.

2.6.1 Methods based on Preclinical Data

Geeleher et al. [24, 80] and Huang et al. (TG-LASSO) [11] utilized linear models trained using GEx and drug responses of preclinical (CCL) data and then predicted the CDR from the GEx of the patient tumors, hence the term preclinical-to-clinical (P2C) drug response prediction. Both methods attempted to solve the domain discrepancy of tumors and CCLs using a batch-effect removal method called ComBat [85]. In these methods, the GEx values of both CCLs (training set) and patient tumors (testing set) were corrected for "batch effects" prior to training, which implies that their models were trained specifically to predict on the pre-determined test samples. Unique to TG-LASSO, their method employed a tissue-based underfitting approach to take into account the tissue-specific distributions of the tumor samples.

PRECISE [81] and TRANSACT [82] addressed the domain discrepancies using subspace alignment [86]. In subspace alignment, they first extracted the factors (i.e., principal components) of the GEx matrices independently per domain, matched the factors across domains, and then selected the factors with the highest similarity (implying commonality across domains). The factors were used to project the tumors and CCLs into the same subspace, for which a regression model was trained using only CCL data and then tested with tumor data. TRANSACT represents a broader framework than PRECISE, where it enhances the similarity function for factor selection by incorporating kernel functions.

Some studies have also applied the same trend of CCL-tumor homogenization of GEx via ComBat prior to training but utilized DL models [83, 84]. MOLI [84] is a DL model that incorporates multiple omics data (GEx, somatic mutation, CNA) for binary drug response classification. They applied a late-integration approach, where each data type has its own

feature encoders, and the outputs of the encoders were concatenated before inputting to the drug response predictor. Additionally, MOLI used a triplet loss [87] to create more robust latent representations. However, instead of training classifiers for individual drugs, they proposed to train on drugs with the same targets, which they demonstrated to have significant performance improvement.

2.6.2 Methods based on Integrated Data

A more recent approach, called PACE [88], proposed to minimize the maximum mean discrepancy (MMD) [89] to align the latent embeddings of the CCL and tumor features. As the name implies, MMD measures the difference between the distribution of the two domains by comparing the means of their representations. Therefore, a common latent space between the two domains is born by minimizing such metric.

TUGDA [90] addressed the domain discrepancy using domain adaptation via adversarial training. Adversarial methods are characterized by the existence of a *discriminator* whose task is to classify the domain of a given embedding. The goal is to adapt the feature extractors so the discriminator cannot identify the samples' original domains from their latent space representation while keeping the information relevant to the original task. However, TUGDA was designed as a transductive model. The test tumor samples are also part of the training set in their formulation, although they did not use the tumor labels for training. Therefore, their models were only able to adapt to specific samples in the target domain.

AITL [91] utilized transfer learning via adversarial training, which assumes the existence of some labeled training data from the tumor samples. Unlike P2C methods, this further reduces the test set (i.e., labeled tumor data), which could hinder statistical evaluations from being performed.

DeepDR [92] trained autoencoders using the TCGA mutation and GEx data. Once trained, they used the encoder part as feature extractors for a subsequent drug response predictor trained on preclinical data. This is reminiscent of the classic transfer learning, where the models were pretrained on a different task/domain with more data and then finetuned for the target task. However, in their case, they pretrained on tumors, which arguably have more unlabeled data, hence the use of autoencoders. Finally, they predicted

the response of tumor samples using their full model.

2.6.3 Methods based only on Clinical Data

Ding et al. [23] attempted to predict CDR (sensitive or resistant) by training ensemble classifiers based on logistic regression. Since using a high dimensional feature with only a few training samples is impractical, their input features were selected from various molecular data (mRNA and miRNA expression, CNA, methylation) using a combination of univariate logistic regression and elastic net. However, they observed an overall poor performance with some exceptions and discussed that the complexity of the task and the limited dataset contributed significantly to this observation.

Some methods focused on patient survivability instead of drug sensitivity. GPBDN [93] and TransSurv [94] are two examples of DL models that utilized a combination of pathological images and molecular profiles to predict the patients' survival. Although these methods are based on patient data, these methods did not take into account the effects of drug intervention.

2.7 Gaps

As shown in the previous sections, many paradigms have been proposed to tackle the DRP problem in clinical and preclinical domains. Despite this, there is much more room for improvement in both performance and methodological perspectives. Most methodologies opt for task-agnostic drug features (e.g., structural properties and targets) for drug representation. Although some studies have shown the merit in using such molecular graphs, it is still unclear whether they are the most appropriate for the drug response prediction task. When using drug targets, there is a lack of consideration for unknown drug targets and non-unique information, such as when multiple drugs have the exact set of targets. Incorporation of biological priors (e.g., pathway membership, transcription factors) was also mostly focused on cancer cell representation, but not much was done for the drug components of the input. There is also a gap in integrating high-level information that can potentially be confounding factors, such as tissues of origin and cancer types. Many existing methods also have a form

of data leakage caused by their methodological framing (e.g., using test samples for pretraining/adaptation) or evaluation oversight (e.g., testing on random splitting of CCL-drug pairs, where CCLs and drugs can independently be part of the training set). Finally, there are still mechanisms of drug response that are yet to be characterized by interpreting more accurate models with appropriate consideration of the nuances of the biological data presented. These gaps motivate the different chapters of this thesis, for which finer details will be discussed.

Chapter 3

Preclinical-to-clinical Drug Response Prediction

One of the primary objectives of personalized medicine is to predict how patients will respond to various treatments and to pinpoint biomarkers that facilitate such predictions. High-throughput sequencing technologies, coupled with significant initiatives like The Cancer Genome Atlas (TCGA) [35], have created an opportunity for machine learning (ML) algorithms to address these challenges. Nevertheless, ML models, particularly those utilizing deep learning (DL) approaches, necessitate a substantial number of samples with documented drug responses to train generalizable models. However, clinical drug response (CDR) data for cancer patients, even in extensive databases like TCGA, is typically limited for most drugs and is not conducive to training DL models.

In the previous decade, there has been an effort to document drug responses of in vitro cancer cell lines (CCLs) to hundreds of compounds along with their molecular profiles [22, 36, 38]. Enabled by these large databases, various ML algorithms have been developed for the prediction of drug response in an effort to harness the power of artificial intelligence in pharmacogenomics [1, 2, 95]. Although these models have shown promising results in predicting the drug response of held-out CCLs, they lack the ability to sufficiently generalize when presented with tumor data from cancer patients. Due to biological differences between CCLs and tumors, as well as the statistical nuances of the data, most of these methods have been shown to exhibit significant performance deterioration [11].

Studies have attempted to address this issue by using tumor samples with known CDRs in training their models. Some methods fully shifted their datasets such that only tumor samples are utilized by the model [23, 93, 96]. Others have employed more sophisticated approaches such as transfer learning and incorporated the tumor samples in addition to the CCLs [91]. A caveat of this strategy is that these studies were only able to develop models for a handful of drugs since many drugs do not have an adequate number of samples with known CDRs. Alternatively, one could train ML models solely on CCLs but address the statistical differences between CCLs and tumors using other computational approaches. For instance, a batch effect removal method called ComBat [85] has been used by multiple approaches as an attempt to cut down the negative effects of the CCL-tumor disparities. In these methods, the gene expression (GEx) profile for CCLs (training data) and tumors (testing data) are used as inputs to ComBat, which adjusts the feature values of both datasets before training the model. However, in practice, CDR prediction for new cancer patients in real-time (i.e., originally not a part of the given testing data) would entail retraining of the model because the feature adjustment with respect to the new patients has to be performed prior to training.

In this chapter, I developed a DL-based computational pipeline to (1) predict the CDR of cancer patients and (2) identify biomarkers of drug response of various anti-cancer drugs, with the additional requirement of solely training on GEx profiles and preclinical drug responses of CCLs. This additional constraint is a realistic consideration given the scarcity of currently available clinical data. Inspired by the work of Huang et al. [11], which demonstrated that integrating tissue (or cancer) type information of test samples enhances the prediction performance of computational models, I developed a DL pipeline with tissue-informed normalization (TINDL) to pursue my objectives. Unlike the previously presented techniques, TINDL's preprocessing of the training data is not intertwined with the test data. Therefore, there will be no need to retrain the model if a new test sample is added during inference.

There are two phases in the TINDL pipeline. The first phase of the pipeline is concerned with the prediction of CDR of cancer patients using the GEx of their tumor samples. The second phase focuses on making these predictions interpretable by selecting a small number of genes that have substantially contributed to the model's predictive capability. In this chapter, I focus on drugs that are common between the Genomics of Drug Sensitivity in

Cancer (GDSC) [36] and TCGA [35]. TINDL employs a technique called tissue-informed normalization, a simple yet effective normalization strategy to reduce the statistical discrepancies between the GEx profiles of CCL and tumor samples. Performance evaluations have shown that TINDL can distinguish between the sensitive and resistant patients for 10 (out of 14) drugs. This is a considerable improvement over other DL-based models that attempt to explicitly remove these domain discrepancies using other approaches such as ComBat or domain adaptation [97, 98]. TINDL also identified important genes that are linked to responses in different drugs, which are also corroborated by previous literature and our in vitro experiments[1].

3.1 Problem Statement

I consider the preclinical-to-clinical (P2C) drug response prediction problem, where the main goal is to predict the CDR of a patient tumor sample (clinical) to a drug. However, the only drug response data available during training are CCL (preclinical) screens.

This problem has two domains: the source domain and the target domain. In P2C drug response prediction, the source domain comprises a set of CCL samples D_S and their responses to drugs, quantified by the natural logarithm of their half-maximal inhibitory concentration or log IC50 (continuous). The target domain is the set of tumor samples D_T and their resistance/sensitivity (binary) to drugs. For both preclinical and clinical samples, their GEx vectors are given as $\mathbf{x} \in \mathbb{R}^m$. However, the two domains come from different distributions ($p(D_S) \neq p(D_T)$) due to the technical and biological differences between CCLs and tumors. Let the subscripts u and v be identifiers for arbitrary samples. For a given drug d, the goal is to train a model $f_d(\mathbf{x})$ using D_S that will accurately predict for D_T. However, the labels are given as $y_u \in \mathbb{R}$ if $u \in D_S$, and $y_u \in \{0, 1\}$ if $u \in D_T$. The evaluation of the models is re-calibrated such that a successful model would give $f_d(\mathbf{x}_u) > f_d(\mathbf{x}_v) | \forall y_u = 1, y_v = 0$ in the target dataset.

[1]The related publication [12] contains more details regarding the in vitro experiments as this portion was performed by my collaborators.

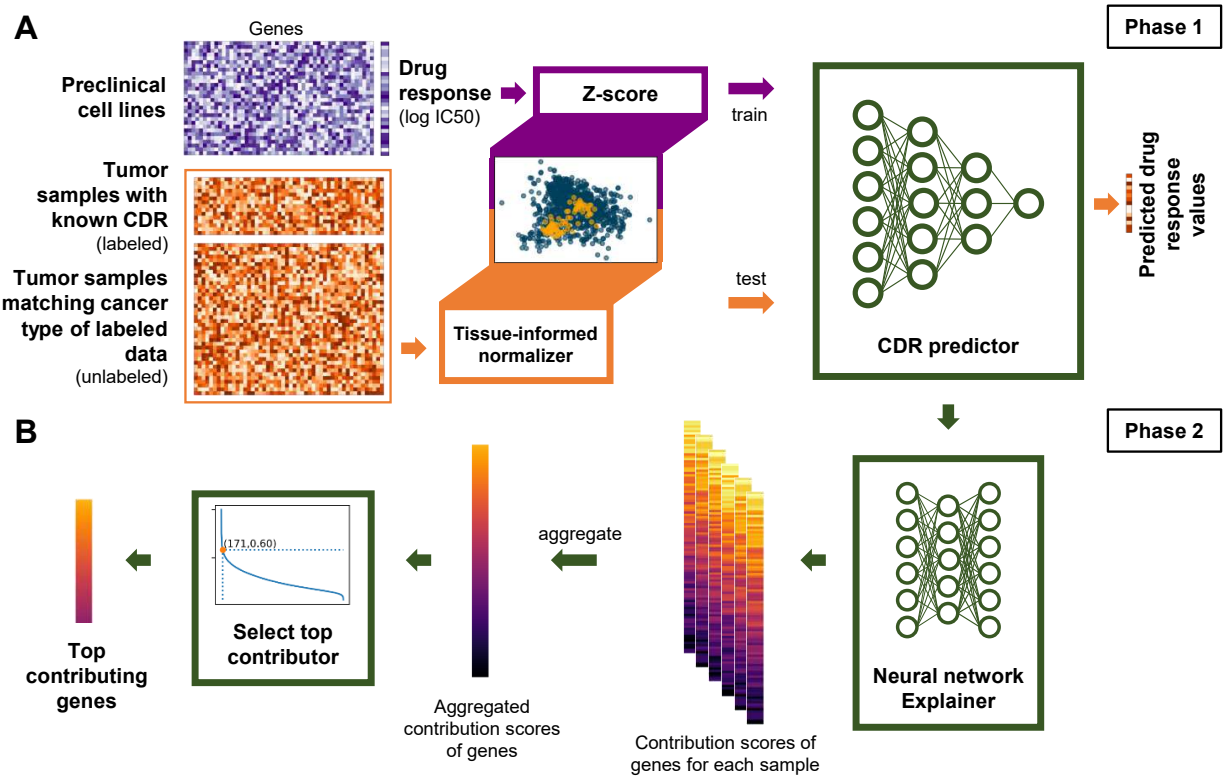

Figure 3.1: The TINDL pipeline. A) During phase 1, the gene expression profiles of the CCLs (training features) and the log IC50 (training labels) were both z-score normalized, while the GEx profiles of the tumor samples (testing features) were normalized using the tissue-informed normalizer. Subsequently, I trained a predictor for CDR using the CCL data. Following training, the model outputs response predictions for the tumor samples. B) In phase 2, a neural network explainer was trained to gain insights into the CDR prediction using the same training data. The trained explainer was utilized to assign gene contribution scores for each gene in each test sample. These scores were aggregated across samples, and the top genes were selected by estimating the point of maximum curvature.

3.2 Methods

3.2.1 TINDL Pipeline Overview

In this chapter, I present a pipeline called TINDL [12] to (1) predict the CDR of cancer patients and (2) identify predictive biomarkers of drug response. The pipeline has two phases: the drug response modeling phase and the gene scoring phase. Figure 3.1 illustrates an overview of the pipeline.

In the response modeling phase (Figure 3.1A), I trained a neural network using the GEx profiles of CCLs as the features and their drug responses (log IC50) as the labels. After

training, I used the model to predict the drug response of cancer patients, represented by the GEx profiles of their primary tumors. The tumor GEx profiles were normalized using the *tissue-informed normalizer* prior to inference.

The second phase, called the gene scoring phase (Figure 3.1B), aims to allocate a contribution score to each gene based on its role in the trained predictive model as a means of interpretability. First, I used CXPlain [99] to calculate the contribution scores of each gene for each individual sample. These individual scores were then averaged across all samples for each gene and subsequently normalized to produce a final contribution score. I then used the distribution of these scores to estimate the threshold at which gene contributions decline. This allowed us to refine the list of top-ranked genes for further investigation, such as pathway enrichment analysis or gene knockdown experiments.

3.2.2 Dataset Acquisition and Preprocessing

For the preclinical (training) data, I used the publicly accessible GEx data of 958 CCLs from GDSC, which were provided as RMA-normalized GEx. As for the clinical (test) data, I utilized RNA sequencing data (in FPKM) from primary tumors in TCGA. In both datasets, I filtered out genes with missing values, eliminated genes not expressed (FPKM < 1) in at least 90% of all TCGA samples, and transformed the remaining genes using $log_2(\text{FPKM} + 0.1)$. Only genes present in both datasets, totaling 15,650 genes, were included. Drug responses of CCLs were quantified as log IC50. I normalized GDSC GEx data (gene-wise) and log IC50 values (drug-wise) using z-score transforms. The CDRs of cancer patients were obtained from the supplementary file of Ding et al. [23].

Given the relatively small number of samples with known CDR in TCGA, my analysis included samples that received multiple drugs during their treatment. I focused exclusively on drugs common to both datasets with at least 20 samples with documented CDR in TCGA, quantified using the Response Evaluation Criteria in Solid Tumors (RECIST). TCGA samples that do not have RECIST CDR in the selected drugs but have GEx profiles are denoted as *unlabeled tumor samples* in the remainder of this chapter. Tissue-informed normalization, detailed below, was employed. Additionally, I reclassified CDRs into sensitive (comprising complete and partial responses) and resistant (comprising stable disease and clinically

progressive disease) to alleviate the scarcity and imbalance of the labels. Note that this stratification only affects the calculation of the performance metrics since the models were trained using continuous log IC50 values. Thus, the concept of relative sensitivity (i.e., more/less sensitive) persists, and predictions can be mapped to the original four classes if the dataset permits. Details on sample counts and tissue types per drug are in Table 3.1 and Supplementary Table A.1.

3.2.3 Tissue-informed Normalization

The normalization step for GEx profiles of patient tumors was designed to tackle two significant challenges. First, the approach should be able to reduce the effects of the discrepancy in statistical properties between GEx profiles of the CCLs and patient tumors, which stem from both technical dissimilarities in data measurement protocols and biological differences between preclinical CCLs and clinical tumors. Secondly, I sought to integrate information regarding the tissue of origins (or cancer types) of tumors into the prediction task. The prior study by Huang et al. [11] demonstrated the importance of tissue information in this prediction task. However, conventional methods commonly used for drug response prediction lack the ability to appropriately integrate such information. The TINDL pipeline highlights a simple yet effective normalization approach called *tissue-informed normalization*.

Since TINDL trains a separate model for each drug, this normalization was carried out independently for each model. First, I identified the set of tissues/cancer types, denoted as T_d, for which a drug d was administered in the TCGA samples (hereafter referred to as "target tissues"). I then collected all the unlabeled tumor samples from the target tissues, forming the unlabeled dataset. The means ($\boldsymbol{\mu}_{T_d}$) and standard deviations ($\boldsymbol{\sigma}_{T_d}$) of the GEx in the unlabeled dataset were calculated gene-wise and were used to normalize the (labeled) test samples corresponding to drug d. For a gene at position i of an arbitrary sample in the test set, $\mathbf{x} = [x_1, ..., x_i, ..., x_m]$, the normalized value x_i would be:

$$x_i = \frac{\tilde{x}_i - \mu_{i,T_d}}{\sigma_{i,T_d}} \tag{3.1}$$

where $\tilde{x}_i$ is the expression of gene i of the sample.

The idea is similar to that of z-scoring, where the distribution of each gene/feature is transformed to have unit variance and zero mean. However, since the number of unlabeled tumor samples is considerably larger, this normalization process remains unaffected by the significantly smaller size of the test set. For example, in a hypothetical case where a user is trying to predict the response of three patients, it would not make sense to normalize their GEx using the statistics of only three samples. The alternative is to use the statistics of the training set, but this is equivalent to not addressing the domain discrepancies. Furthermore, because the normalization is conducted independently for both the training set and the test set, there is no necessity to retrain the model each time the drug response of a newly acquired test sample needs to be predicted.

3.2.4 Network Architecture, Hyperparameter Selection, and Training

For the drug response prediction models, the number of epochs, batch sizes, and learning rates were selected using grid-search and 5-fold cross-validation. Only the training data corresponding to CCLs were used to conduct the hyperparameter search. Since the primary goal of the end task is to segregate the responders from the non-responders, I opted to select the set of hyperparameters with the highest average Pearson's correlation coefficient on the validation set across the five folds. The selected hyperparameters for TINDL are in Supplementary Table A.5. I fixed the network architecture for simplicity and to prevent the exponential growth of the hyperparameter search space, considering that different drugs are trained separately. I kept the architecture simple with only three hidden layers, each with 512, 256, and 128 hidden nodes in order. For the activation function, all hidden layers were attached with rectified linear units (ReLU). A dropout layer with a 0.2 dropping probability was added prior to the output layer.

To train the models, I used the normalized log IC50 as the labels (y) and the mean squared error (MSE) as the loss function. The MSE loss was chosen because it assigns greater importance to samples with label magnitude greater than one ($|y| > 1$), thereby focusing on the differences between samples with higher sensitivity/resistance rather than those within one standard deviation of the mean. During hyperparameter tuning, models were permitted

to train for a maximum of 1000 epochs, with early stopping implemented if the model's validation MSE failed to decrease after 30 epochs. Following hyperparameter tuning, a final model was retrained from scratch using all labeled CCL samples. To ensure the robustness of the results, I used ten distinct random initializations (seeds) and formed an ensemble by averaging their predictions. Note that individual models were trained independently. A similar methodology was applied to ADDA-DL, DANN-DL, ComBat-DL, TrainNorm-DL, and TestNorm-DL.

3.2.5 Calculating Contribution Scores of Genes

The second phase of TINDL (Figure 3.1B) is concerned with biomarker identification. I used CXPlain [99] to determine the contribution score of each gene in each sample with respect to the model's performance. CXPlain is a black box explainer that attempts to provide causal explanations for the trained model's predictions. This is done by training an "explainer" model, a separate model that takes inputs similar to the trained model (called "predictor") and outputs scores corresponding to each input feature's contribution. This approach, inspired by Granger causality [100], aims to assess the impact of features (genes in this case) individually by zeroing out features one by one and measuring the normalized difference between the predictor's initial error and the error incurred when the feature under consideration is zeroed out. Here, zeroing out was used as an approximate feature removal for practical purposes. The error is defined as:

$$\varepsilon_u = (y_u - \hat{y}_u)^2, \tag{3.2}$$

where y_u is the true value and $\hat{y}_u$ is the output of the predictor. Note that here, the subscript u is a sample identifier, and the input feature vector would be denoted as $\boldsymbol{x}_u$. Additionally, since the predictor is an ensemble, $\hat{y}_u$ is the average of the individual models.

As "ground truth" for the explainer, a vector of "real" contributions, denoted as $\boldsymbol{\Omega}_u = [\omega_1(\boldsymbol{x}_u), \ldots, \omega_m(\boldsymbol{x}_u)]$ were also calculated for each training sample prior to training the explainer. The contribution of the feature at position i was calculated as follows:

$$\omega_i(\boldsymbol{x}_u) = \frac{\Delta\varepsilon_{u,i}}{\sum_{j=1}^{p}\Delta\varepsilon_{u,j}}, \tag{3.3}$$

where

$$\Delta\varepsilon_{u,i} = \varepsilon_{u\setminus\{i\}} - \varepsilon_u. \tag{3.4}$$

In Equation 3.4, $\varepsilon_{u\setminus\{i\}}$ denotes the predictor's error when given $\boldsymbol{x}_u$ but with feature i zeroed out (i.e., $[x_{u,1},\ldots,x_{u,i},\ldots,x_{u,m}] \rightarrow [x_{u,1},\ldots,x_{u,i-1},0,x_{u,i+1},\ldots,x_{u,m}]$). The explainer is designed with an architecture where the dimensions of the input vector match those of the output vector. Each output corresponds to the predicted contribution for its respective feature. The explainer is trained by minimizing the KL divergence, $KL(\boldsymbol{\Omega}_u, \hat{\boldsymbol{\Omega}}_u)$, of the real contributions $\boldsymbol{\Omega}_u$, and predicted contributions $\hat{\boldsymbol{\Omega}}_u$ of the training set.

As noted in Schwab and Karlen's work [99], it is possible to use the calculated ground truth, $\boldsymbol{\Omega}_u$, instead of training an explainer model in cases where the explicands' labels are available. However, in this case, the explicands' (test set) labels cannot be applied to our error function (Equation 3.2). I modified the code linked in [99] to fit this application, which is also included in the published code.

The explainer was trained as a neural network comprising two layers with 512 hidden units. I utilized the ensemble mode, training ten independent explainers (with different initializations) and reporting their median as the final contribution values. I used the explainer model to obtain the contribution vectors for each of the samples in the test set. Drug-specific gene contribution scores were calculated as the gene-wise average contribution score across all the labeled test samples for that drug. I then normalized the scores such that the gene with the highest contribution score for a drug equals 1.

3.2.6 Identifying Genes with Substantial Contribution Scores

Once the contribution scores of each gene for a drug have been calculated, it is more manageable to analyze a smaller subset of genes that have considerably influenced the model's predictions. First, I sorted the genes according to their scores and plotted a curve, where the x-axis is the rank of the gene i and the y-axis is the drug-specific contribution score $\bar{\omega}_i$ of gene i. I utilized the kneedle algorithm [101] to pinpoint the "knee," representing the

point of maximum curvature, which I adopted as the cutoff for selecting the top genes. The idea of kneedle is that if one forms a line l from $(1, \bar{\omega}_{\max})$ to $(n, \bar{\omega}_{\min})$ and rotate the curve around the point $(n, \bar{\omega}_{\min})$, the "knee" can be approximated by the set of points in the local maxima. The point farthest from the line l is considered to be the knee or the contribution threshold.

3.2.7 Knowledge-guided Pathway Enrichment Analysis

From the top-ranked genes identified for each drug, I associated some pathways using KnowEnG's gene set characterization (GSC) pipeline [102]. This pipeline integrates gene interactions to augment the analysis in its network-guided mode, for which I selected the STRING Experimental protein-protein interaction (PPI) [77]. This PPI graph contains experimentally verified protein-protein interactions as edges. I used the default 50% network smoothing parameter and chose the Enrichr pathway collection [103]. Unlike common GSC pipelines, the network-guided GSC does not provide a P value. Instead, a score called "difference score" is used to implicate the top pathways. Using the random walk with restarts algorithm, the scores corresponding to the relevance of the pathways are calculated with the query nodes (genes) as the restart set. The difference score is the computed relevance subtracted by a baseline score. Pathways above the 0.5 threshold are considered to be associated with the input query gene set.

3.2.8 Precision at kth Percentile

In the analysis, I introduced a performance indicator called *precision at kth percentile*. For each drug, I collected TINDL's predictions for the tumor samples, interpreted as a prediction of log IC50 due to the nature of the training dataset. Next, I identified the kth percentiles of the distribution ($k \leq 50$), which I denoted as t_k. The predictions were then grouped, ensuring that all values below t_k were classified as positives (i.e., sensitive). Subsequently, the precision at kth percentile was computed using the formula:

$$\text{Precision@}k = \frac{TP_k}{TP_k + FP_k} \tag{3.5}$$

where TP_k and FP_k represent the true positives and false positives at the kth percentile, respectively.

3.2.9 Baseline Approaches

This section describes the models used for the comparative evaluations.

Traditional Machine Learning Approaches

I used SVR, random forest, and LASSO regression to represent traditional ML methods that do not take into account the domain differences between the clinical and preclinical samples. These three approaches were implemented using scikit-learn [104].

The comparison also included two traditional ML approaches that regarded the domain differences as "batch effects." The first one is Geeleher's method [24], which I reimplemented using scikit-learn and pyComBat, a Python implementation of the batch effect removal tool called ComBat [85, 105] (see Appendix A.2). The second approach is called TG-LASSO [11], for which I used the implementation provided by the authors. All hyperparameters were tuned as described in the previous subsections except for TG-LASSO, which has a unique built-in hyperparameter tuning technique.

Deep Learning Approaches

All baseline DL models utilized a similar architecture to TINDL to guarantee an equitable comparison. Furthermore, the hyperparameter selection and training procedures closely mirrored those outlined above for TINDL. Note that there is no validation set for the labeled clinical dataset. Therefore, none of the models were tuned to optimize the main performance indicators, and the architecture that performs best in the preclinical dataset was chosen. Consistent with TINDL, both labeled and unlabeled TCGA samples were utilized for this task. The models and their specific considerations are described below.

Basic DL Workflows. TrainNorm-DL and TestNorm-DL represent the two "default" workflows that are usually applied when domain discrepancies are considered to be negligible. In TrainNorm-DL, the feature-wise means and standard deviations were calculated solely from

the training set and then applied to the normalization of both the training and test sets. This approach assumes that the training and test sets belong to the same domain and that it is not possible to peek into the distribution of the test set. On the other hand, TestNorm-DL employs a per-dataset normalization technique, where the test set is normalized using its own mean and standard deviation, while the training set utilizes its own statistics. The TestNorm-DL approach assumes that the empirical distribution of the test set is adequately close to the real distribution such that the test set-based normalization does not yield unexpected values. The same model as TINDL is employed for these baseline scenarios, as the difference in normalization influences only the test set.

Additionally, I created a baseline called ComBat-DL, which I consider the DL analogue of Geeleher's method [24]. As such, I applied ComBat [85] (see Appendix A.2) to address discrepancies between the TCGA and GDSC datasets prior to training and then followed the workflow of TrainNorm-DL, where statistics of the ComBat-processed training (CCL) data were used for normalization of both training and testing sets. However, since ComBat performs adjustments by looking at the distribution of both datasets, the adjusted data is expected to have some form of data leakage unlike TrainNorm-DL.

Deep Domain Adaptation. Two deep domain adaptation techniques were also included in the comparison, namely ADDA-DL and DANN-DL, which, as their names imply, use *adversarial discriminative domain adaptation* (ADDA) [98] or *domain adaptive neural network* (DANN) [97] to remove the domain discrepancies between TCGA and GDSC datasets. Here, the source dataset is GDSC (i.e., the source task is the drug response prediction for the CCLs), while the target dataset is comprised of the TCGA samples.

ADDA is a domain adaptation technique that requires a neural network model that is already trained on the source task. ADDA adapts a copy of the pre-trained model in order to generalize to the target dataset by adjusting how the target data is represented in the latent space (i.e., embeddings from the first few layers). In particular, the target data's embedding distribution is matched to the source data's embedding distribution using a discriminator and adversarial losses. The goal is to confuse the discriminator, whose task is to segregate the embeddings of the source and target data. In the evaluations, I used a copy of the base

model of TINDL as the pre-trained network and applied ADDA to adapt the parameters of the said model. The unlabeled tumor samples from the drugs' target tissues were used as the training dataset associated with the target data.

DANN is another domain adaptation technique that I used as a DL baseline to remove the discrepancy between the TCGA and GDSC datasets. DANN creates a latent feature space that is shared between the two domains. This allows the model to predict for the target dataset even though the model was only trained using the labels from the source dataset. This method also employs a discriminator that aims to segregate embeddings from the two domains. Unlike ADDA, the domain adaptation and learning of the source task happen simultaneously in DANN. In addition to the source task objective (e.g., MSE for regression), DANN incorporates a gradient-reversed discriminative loss function, a negation of the discriminator's objective such that it becomes difficult to distinguish between source data and target data. Similarly, I used the unlabeled tumor samples from the drugs' target tissues as the training dataset associated with the target data.

Graph Neural Networks. Graph convolutional networks (GCN) [106] and graph attention networks (GAT) [107] represent two variations of graph neural networks, both designed for graph-structured data. I used the STRING co-expression graph [77] as the input graph structure for both architectures. Each node in the graph corresponds to a gene and is characterized by the concatenation of a unique trainable embedding vector (gene-specific and shared across samples) and the gene's expression value (sample-specific). These gene-specific vectors allow GCN and GAT to distinguish differences between genes, which GCN and GAT would otherwise overlook due to their permutation invariance properties. I only considered genes present in all STRING, GDSC, and TCGA datasets. The complete model resembles that of TINDL, with the initial two layers substituted by GCN or GAT, facilitating two-hop information propagation within the graph structure.

Recurrent Neural Networks. Long short-term memory (LSTM) is a type of recurrent neural network typically used for sequential data. In this baseline, I utilized the gene indices from the input file to establish an artificial order, dividing the features into ten distinct

Table 3.1: The number of TCGA samples and the performance of TINDL in predicting their CDR for 14 drugs.

Drug	Number of clinical samples	Number of sensitive samples	Number of resistant samples	P value
Cisplatin	303	237	66	6.36E-4
Tamoxifen	20	14	6	1.14E-3
Etoposide	84	73	11	4.00E-3
Doxorubicin	100	68	32	1.42E-2
Paclitaxel	158	111	47	2.29E-2
Vinorelbine	30	23	7	2.41E-2
Oxaliplatin	54	33	21	2.41E-2
Temozolomide	95	11	84	2.94E-2
Bleomycin	52	46	6	3.41E-2
Gemcitabine	157	75	82	4.57E-2
Cyclophosphamide	101	96	5	5.60E-2
Pemetrexed	38	18	20	2.86E-1
Irinotecan	23	6	17	3.04E-1
Docetaxel	102	67	35	7.04E-1

windows. The embedding generated from the final window (the 10th iteration through the LSTM) was used as input for the subsequent fully connected layers. Given that the parameter count of a single LSTM layer aligns more closely with that of two layers in a fully connected network, I opted for a single LSTM layer. The overall architecture mirrors TINDL's, albeit with the initial layer substituted by an LSTM layer.

3.3 Results

3.3.1 Performance of TINDL in P2C Drug Response Prediction

To evaluate TINDL's performance in predicting the CDR of cancer patients, I gathered the GEx profiles of primary cancer tumors from the TCGA database [35]. I utilized the RECIST data curated by Ding et al. [23] that corresponded to the TCGA drug responses and focused on the 14 drugs that have satisfied the filtering conditions laid out in the methods section.

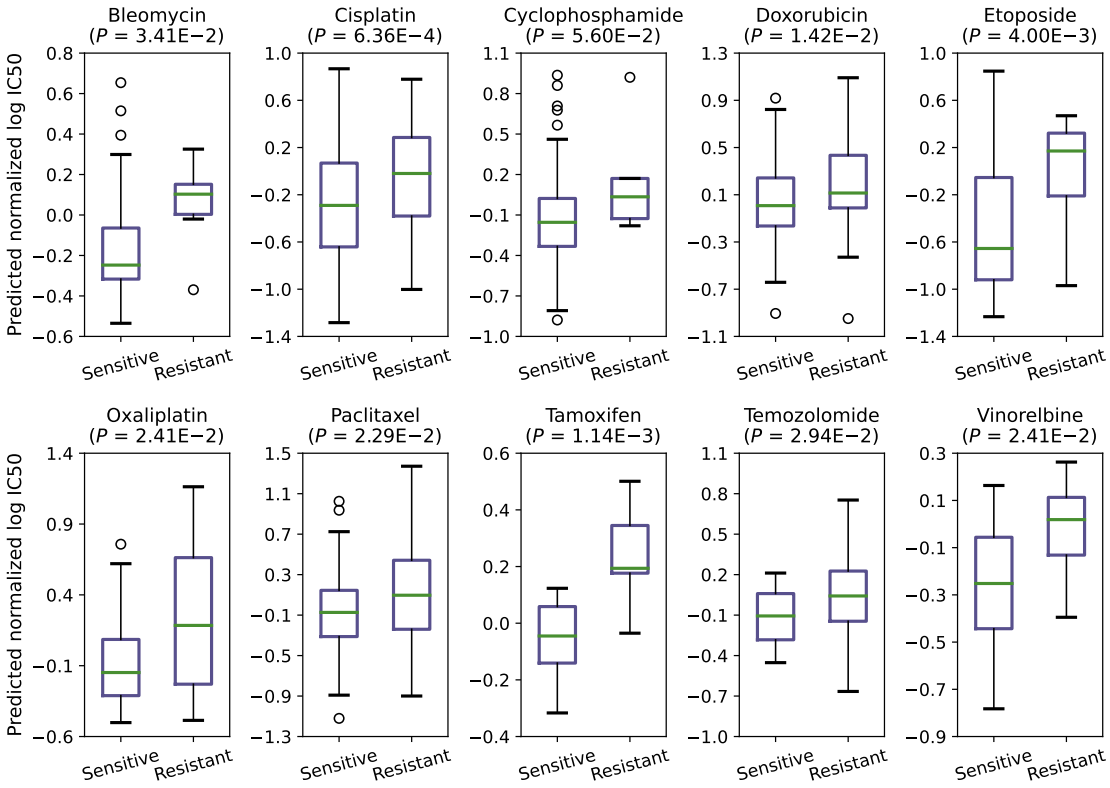

Figure 3.2: Box plots of the predicted drug response for patients, categorized by their true drug response. A one-sided Mann-Whitney U test was used to calculate the P value in each panel. Only drugs with significant P values are shown.

Similar to prior research [11, 24], I utilized a one-sided Mann-Whitney U test to determine whether there is significant evidence to suggest that the predicted values for resistant patients are greater than those of sensitive patients. Rejecting the null hypothesis indicates that the model effectively predicts higher values for resistant patients compared to sensitive patients, thus the model's predictions are consistent with the CDR. The performance of TINDL in the CDR prediction of TCGA samples for various drugs is presented in Table 3.1. My analysis showed no association between the model's predictive performance in terms of AUROC (Table A.3) and the sample size of either the labeled or unlabeled tumors. TINDL effectively discriminates between resistant and sensitive patients for 10 out of 14 drugs ($P < 0.05$, one-sided Mann-Whitney U test), yielding a combined P value of 2.77E-10 (using Fisher's method). The distribution of predicted CDR values for sensitive and resistant patients for these drugs is depicted in Figure 3.2.

Following that, I introduced a performance metric called precision at kth percentile to

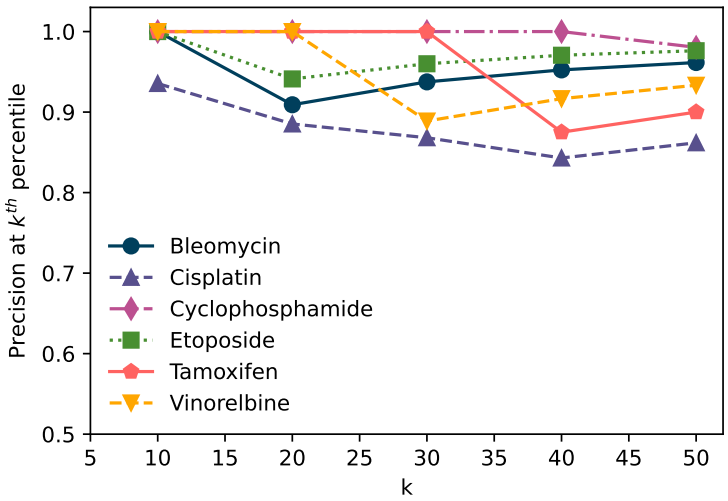

Figure 3.3: Precision at kth percentile for identification of sensitive patients. Only six drugs are shown for visibility (see Supplementary Table A.4 for full details).

assess whether patients with predictions falling within the lower range of the distribution correspond to drug-sensitive individuals. For various values of k, tumors with predicted log IC50 in the bottom k% were classified as sensitive, and their tally was utilized to compute precision. Figure 3.3 and Supplementary Table A.4 present the precision at kth percentile of TINDL across different k values. These findings indicate that for six drugs (tamoxifen, etoposide, vinorelbine, cyclophosphamide, bleomycin, and cisplatin), TINDL can accurately identify responders with precision at kth percentile above 84% regardless of the chosen k.

3.3.2 Comparison of TINDL and Other Methods for Predicting CDR

Next, I evaluated the performance of TINDL in comparison to alternative computational models. To achieve this, various traditional and state-of-the-art ML models [11, 24] for predicting the CDR of cancer patients based on preclinical CCLs were assessed. Detailed performance metrics for each drug and model can be found in Supplementary Tables A.2-A.3, with a summary of results provided in Table 3.2. In this table, the combined P value of 14 drugs was utilized to summarize the performance of different methods using Fisher's method. As illustrated in Table 3.2, TINDL successfully distinguishes between sensitive and resistant patients for 10 out of 14 drugs (with a combined P value of 2.77E-10 for all drugs). In

Table 3.2: The performance TINDL and traditional ML models in predicting CDR of tumor samples using models exclusively trained on CCLs.

Method	Drugs with $P < 0.05$	Drugs	Combined P value (Fisher)
TINDL	10	14	2.77E-10
LASSO	7	14	7.47E-7
TG-LASSO, Huang et al. [11]	6	14	8.32E-7
SVR (RBF kernel)	5	14	1.89E-6
Geeleher, et al. [24]	4	14	5.63E-3
Random forests	4	14	3.12E-3

Table 3.3: The performance of DL baselines and DL-based approaches designed to mitigate discrepancies between preclinical and clinical datasets.

Method	Drugs with $P < 0.05$	Drugs	Combined P value (Fisher)
ComBat-DL	7	14	6.73E-10
ADDA-DL	7	14	2.16E-7
DANN-DL	7	14	1.66E-6
TrainNorm-DL	6	14	4.68E-7
TestNorm-DL	8	14	1.80E-9

contrast, the second-best method in this table only achieves this distinction for seven drugs. Consistent with prior study [11], this analysis also showed that LASSO and its variant, TG-LASSO, demonstrate reasonably good performance across all drugs, whereas support vector regression (SVR) and random forests exhibit comparatively lower performance.

As previously discussed, a major challenge in predicting the CDR of cancer patients using ML models trained on preclinical CCLs lies in the statistical discrepancies between these sample sets. In order to evaluate TINDL's performance against other DL models explicitly designed to mitigate these statistical differences, I examined three alternative methods (ADDA-DL, DANN-DL, and ComBat-DL), along with two baselines (TrainNorm-DL and TestNorm-DL) representing potential "default workflows" had I not anticipated the impact of these domain discrepancies.

I trained models for these methods with an architecture similar to that of TINDL, except for the discriminators, which are specific to ADDA [98] and DANN [97] and are utilized

for domain adaptation. The specifics of these methods, including their architecture and training procedure, are outlined in the Methods section. The performance of these DL-based approaches is presented in Table 3.3 and Supplementary Tables A.2-A.3. The results revealed that across all three instances of explicit discrepancy removal, the response predictions for the sensitive patients were significantly lower than those of resistant patients only for 7 out of 14 drugs. As anticipated, TrainNorm-DL displayed inferior performance (6 out of 14) compared to the other DL approaches. Surprisingly, TestNorm-DL managed to segregate patients in 8 drugs, placing second to TINDL. However, TestNorm-DL may not be well-suited for applications with minimal samples in the test set.

3.3.3 Latent Space Analysis of Adaptive Methods

To evaluate TINDL's superior performance compared to the DL-based models with explicit discrepancy removal mentioned above, I examined their ability to mitigate the discrepancies between preclinical and clinical samples. The default workflows were excluded from this analysis as they overlooked this issue. To accomplish this, I used the pairwise Euclidean distance of samples based on their learned representations by the DL models' first few layers (henceforth referred to as "encoder"). I analyzed the encoder's output since methods utilizing domain adaptation do not alter the input features but instead aim to address domain disparities in the latent space. Furthermore, comparing these latent representations offers greater significance than evaluating input representations since the embeddings are utilized by the subsequent layers in the prediction process. I calculated the distance between the learned representations of preclinical and clinical samples (termed "P2C distances") through Ward's method, a widely adopted approach in hierarchical clustering [108]. This method not only examines Euclidean distances among data points but also incorporates their variance when determining the distance between the two sample groups.

Through a one-sided Wilcoxon signed-rank test, I tested whether the median of differences of TINDL's and a baseline's P2C distances is less than zero. I found that TINDL's learned representations of clinical samples exhibit a significantly smaller average distance to preclinical samples compared to ComBat-DL (P = 6.10E-5), ADDA-DL (P = 4.27E-4), and DANN-DL (P = 6.10E-5), across all drugs (Figure 3.4A). Additionally, the effective-

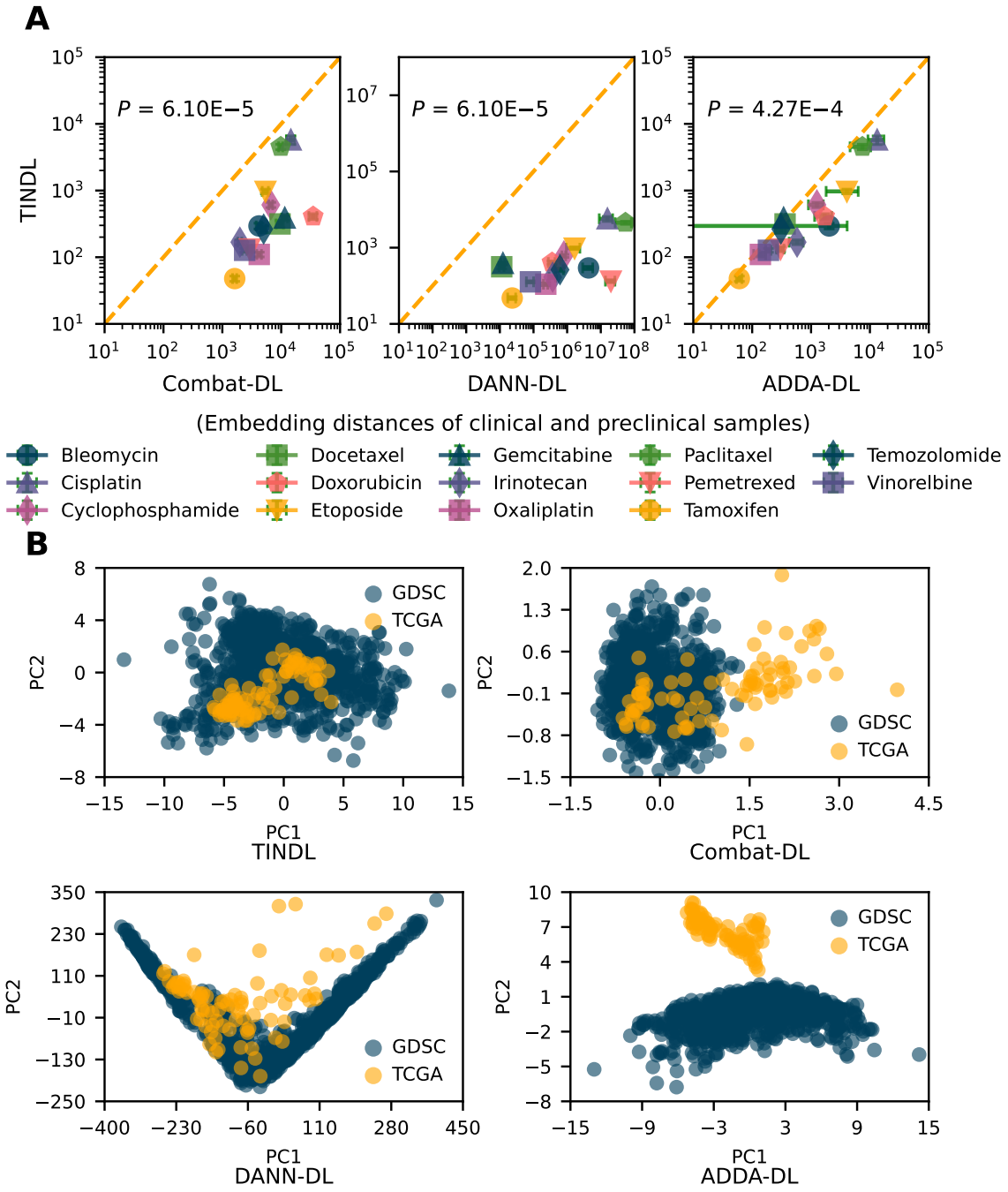

Figure 3.4: Assessment of latent representations generated by deep learning models. A) Scatter plots depict the comparison of distances between preclinical and clinical samples in the embedding space for each drug. Each point represents a drug. P values were computed using a one-sided Wilcoxon signed-rank test. Error bars denote 95% confidence intervals, derived from ten runs of each method with random initializations. B) PCA visualization of embeddings utilized by each method to predict etoposide response. Notably, TCGA samples were visibly more intermixed with GDSC samples in TINDL compared to alternative methods, suggesting reduced separability.

Table 3.4: The performance of alternative neural network architectures utilized as feature extractors.

Method	Drugs with $P < 0.05$	Drugs	Combined P value (Fisher)
GAT	7	14	2.75E-11
GCN	6	14	2.85E-7
LSTM	6	14	1.86E-5

ness of tissue-informed normalization in TINDL for rectifying the statistical discrepancy between preclinical and clinical embeddings can be visually observed using principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) plots of the embeddings learned by each method (Figure 3.4B, Supplementary Figures A.1-A.4).

I then sought to investigate whether the similarity in embeddings has an impact on the predictive performance of TINDL for drug responses across various drugs. My analysis revealed a negative Spearman's rank correlation (r = -0.17, P = 3.93E-2) between the calculated distances and the models' area under the receiver operating characteristic (AUROC) for predicting responses to different drugs. Notably, tamoxifen exhibited the highest AUROC (Supplementary Table A.3, AUROC = 0.92) and also displayed the smallest average distance between clinical and preclinical representations among all drugs in TINDL. These findings further support the conclusion that minimizing the gap between the statistical characteristics of clinical and preclinical samples plays an important role in the effectiveness of TINDL in predicting drug responses.

3.3.4 Comparison of Various Neural Network Architectures

I also evaluated the performance of different neural network architectures as feature extractors instead of the fully connected (FC) networks utilized in the preceding section. Specifically, I employed LSTM, GCN [106], and GAT [107] for the first few layers of the model (see Methods). All models underwent the same protocol and evaluation methodologies as the other DL approaches based on FC networks. A summarized overview of the findings is presented in Table 3.4, with comprehensive evaluation metrics available in Supplementary Tables A.2-A.3. While GCN and GAT theoretically might offer advantages over the FC architecture, given that the input graph (gene co-expression) provides valuable information

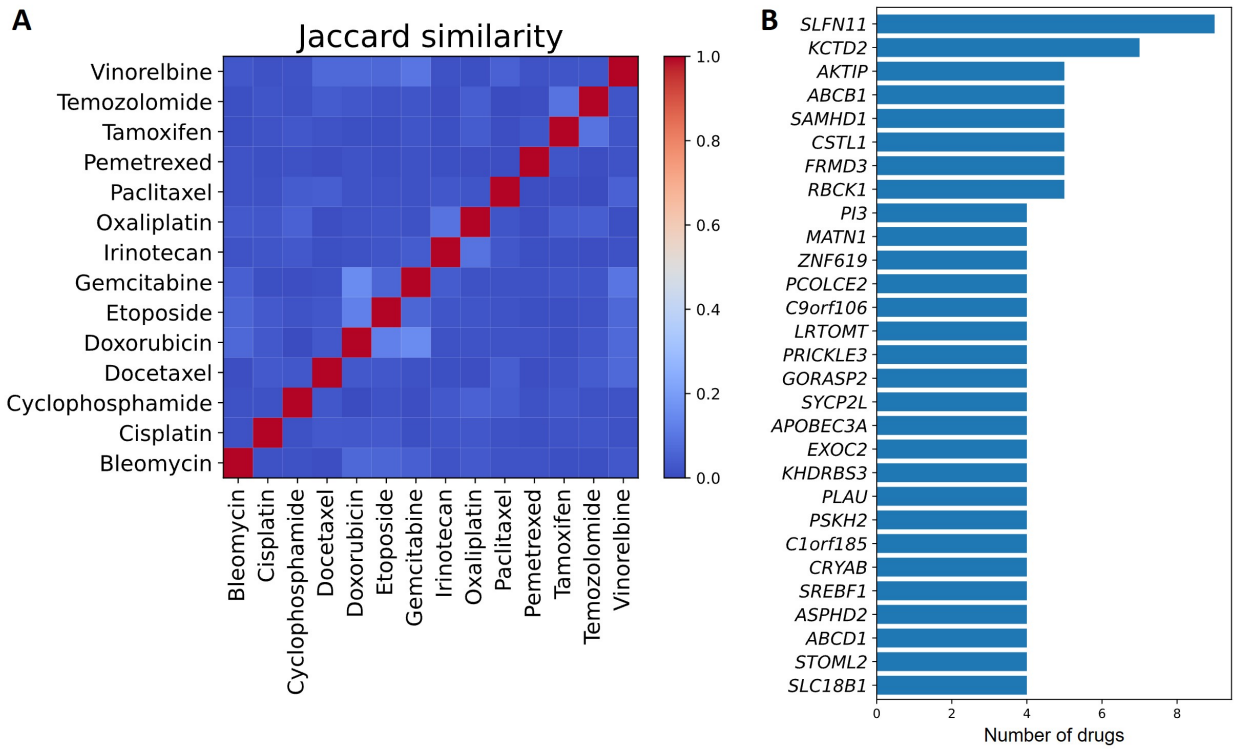

Figure 3.5: Top genes identified by TINDL that are shared across different drugs. A) The heatmap of the Jaccard index of the identified top genes in the 14 drugs. B) The number of implicating drugs for genes commonly identified as top response indicators by at least four drugs.

about the features, GCN and GAT did not exhibit improved performance compared to FC networks. As expected, LSTM did not perform strongly due to the non-sequential nature of the data. Nonetheless, it was noteworthy that LSTM managed to separate the sensitive and resistant patients for some of the drugs.

3.3.5 Identification of Biomarkers of Drug Sensitivity

I utilized TINDL (Figure 3.1B) to assign a score to the contribution of each gene in the trained model. Depending on the drug, this approach identified 64 (for pemetrexed) to 243 (for bleomycin) genes. The ranked list of genes identified by TINDL using this drug-specific threshold is provided in Supplementary File A.1.

Next, I assessed whether the identified genes exhibit drug specificity. For this purpose, I computed the Jaccard similarity coefficient for each possible pair of drugs (Figure 3.5A). I

observed a notable degree of drug specificity, as evidenced by an average Jaccard similarity coefficient of 0.027 across all unique drug pairs. Nevertheless, certain genes were implicated across multiple drugs (Figure 3.5B). Previous research has highlighted the involvement of these genes in various cancers and their association with sensitivity to multiple drugs [109–115].

Multidrug resistance (MDR) is a prominent factor in diminishing the effectiveness of many anti-cancer agents [116]. MDR is characterized by resistance to therapeutic substances not associated with structure or mechanism of action [117]. The classical mechanism of MDR is linked to the overexpression of ATP-binding cassette (ABC) transporter genes (ABCB1, ABCD1, etc.), which leads to a reduced effective drug concentration through the efflux of drugs from the cells [118].

Apart from the classical MDR mechanism associated with ABC gene overexpression, atypical mechanisms also exist [119–121]. Notable examples of these atypical mechanisms include evading adaptive immune responses [119]. Dysregulation of numerous genes, such as APOBEC3A, fosters the evolution and progression of cancers, facilitates evasion of adaptive immune responses, and contributes to the emergence of drug resistance in various cancers [122, 123].

Another atypical mechanism is the dysregulation of genes associated with macrophage infiltration and polarization, such as CRYAB [120], and the dysregulation of genes governing drug-induced apoptosis through the activation of survival pathways like the MEK/ERK signaling and inhibition of the mitochondrial apoptosis pathway in cervical cancer cells [121]. Specifically, Schlafen family member 11 (SLFN11) was associated with nine drugs (Figure 3.5B) and emerged as the top contributor for bleomycin, cisplatin, doxorubicin, etoposide, gemcitabine, and irinotecan, while ranking as the third top contributor for oxaliplatin. SLFN11 is a putative DNA/RNA helicase that is recruited to the stressed replication forks, where it inhibits DNA replication. Dysregulation of DNA replication can induce genome instability [124], a hallmark of cancer that fosters genetic diversity during tumorigenesis [125]. Several studies have demonstrated that the expression of SLFN11 sensitizes cancer cells to various chemotherapeutic agents, including cisplatin, oxaliplatin, irinotecan, gemcitabine, doxorubicin, and etoposide [126–128].

Acquired chemotherapy resistance is promoted by EZH2 through the epigenetic silencing of SLFN11. Thus, the acquisition of chemoresistance may be preventable through the targeting of EZH2 [129]. Several potent and selective inhibitors of EZH2 are currently undergoing clinical development across various stages, including phase II trials by Epizyme and phase I trials by Constellation and GSK, showing promising safety profiles in multiple solid tumor and hematological indications. This data support the idea that combining EZH2 inhibitors to downregulate SLFN11 with conventional chemotherapeutic agents warrants consideration in multiple cancer types [130].

3.3.6 Characterization of TINDL-identified Biomarkers

To gain deeper insights into the functional characteristics of genes identified by TINDL across multiple drugs, I used KnowEnG's GSC pipeline [102]. This tool facilitated the identification of pathways linked to 29 genes identified by TINDL for at least four drugs (Figure 3.5B). Leveraging network-guided analysis, this pipeline allows for the exploration of associated pathways while considering interactions among genes and their protein products. The following five pathways were implicated in the analyses: (1) Regulation of toll-like receptor signaling pathway, (2) Alpha-synuclein signaling, (3) Arf6 trafficking events, (4) Insulin pathway, and (5) RalA downstream regulated genes.

Innate immune receptors such as toll-like receptors (TLRs) are responsible for recognizing molecular patterns linked to pathogens, forming meaningful molecular connections between innate cells and adaptive immune responses. Activation of TLRs on dendritic cells (DCs) facilitates communication between the innate and adaptive immune systems, triggering the maturation and migration of DCs into lymph nodes, which in turn leads to activation, proliferation, and survival of tumor antigen-specific naive CD4+ and CD8+ T cells [131]. Notably, tumor cells lack molecules capable of inducing DC maturation. Therefore, the use of TLR agonists is a vital component of immunotherapy protocols aimed at activating T cells [132]. Furthermore, TLR agonists have been proposed as adjuvants for cancer vaccines [133]. For instance, utilizing a TLR3 agonist as an adjuvant alongside conventional chemotherapy can counteract the tolerogenic or immunosuppressive effects induced by the tumor, promoting T cell responses and tumor rejection [134, 135].

Alpha-synuclein (α-syn) is a neuronal protein that is responsible for regulating synaptic vesicle trafficking. It is commonly expressed in different brain tumors and melanoma [136], with its upregulation associated with the aggressive nature of meningiomas [137]. Additionally, the loss of α-syn leads to dysregulation in iron metabolism and inhibition of melanoma tumor growth [138]. Cancer development is facilitated by the oncogenic activation of synuclein, which promotes tumor cell survival through the activation of the JNK/caspase apoptosis pathway and ERK and confers resistance to certain chemotherapeutic drugs [139]. This underscores synuclein as a promising therapeutic target for future treatments aimed at overcoming resistance to certain chemotherapeutic agents.

Trafficking of bioactive cargos to tumor-derived microvesicles (TMVs) are controlled by ADP-ribosylation factor 6 (ARF6). TMVs comprise a class of extracellular vesicles released from tumor cells that facilitate communication between the tumor and the surrounding microenvironment [140]. Invasive tumor cells release TMVs that contain bioactive cargo and utilize them to degrade the extracellular matrix during cell invasion [141]. As such, multiple studies have pointed out a correlation between ARF6 expression and the invasion and metastasis of various cancers [142, 143], indicating that modulation of ARF6 signaling could control TMV shedding and influence the overall invasion process.

Insulin is a signaling molecule crucial for regulating systemic metabolic balance. It can be perceived as facilitating tumor development by enabling PI3K activation and promoting enhanced glucose uptake [144, 145]. Additionally, insulin influences the response to cytotoxic therapy [146]. RAS-related protein RalA, a member of the Ral family, contributes to anchorage-independent growth, tumorigenicity, migration, and metastasis through the RalA pathway [147, 148].

Overall, the association between genes implicated across multiple drugs and these pathways, which play diverse roles in cancer progression, may indicate shared mechanisms of action among different anti-cancer drugs. I also conducted a similar pathway enrichment analysis for genes implicated in each drug individually, and detailed results are provided in Supplementary File A.2.

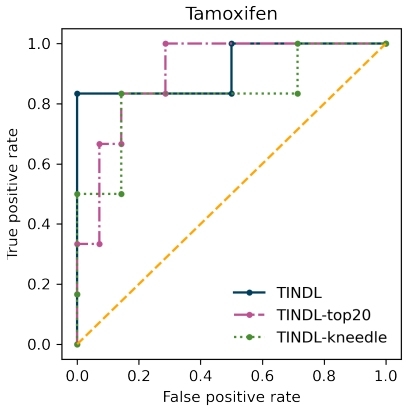

Figure 3.6: Comparison of the ROC curves when using different numbers of genes in CDR prediction of tamoxifen. TINDL represents the default model that utilized the GEx values of all genes (AUROC = 0.92). TINDL-top20 (AUROC = 0.90) and TINDL-kneedle (AUROC = 0.83) utilized the GEx values of the top 20 genes and the top genes identified by kneedle, respectively, while the rest of the genes were assigned a value of zero.

3.3.7 Validation of TINDL-identified Genes for Tamoxifen

I sought to assess the predictive ability of the top identified genes by TINDL for drug response, both computationally and experimentally, focusing on tamoxifen due to TINDL's strong predictive performance for this drug (AUROC = 0.92, P = 1.14E-3 for Mann-Whitney U test). Using only the top implicated genes for tamoxifen (n = 136, based on the threshold identified by kneedle), a high AUROC value and a significant Mann-Whitney U test P value (Figure 3.6A, AUROC = 0.89, P = 2.32E-3) were consistently observed. Subsequently, I reduced the number of genes in the model to only the top twenty and found that the AUROC remained high even with this smaller gene set (Figure 3.6A, AUROC = 0.90, P = 1.65E-3). These results show that even a limited panel of twenty genes can effectively predict the CDR of tamoxifen, indicating potential clinical applications in precision medicine for such small gene panels.

Finally, to ascertain whether the genes pinpointed by TINDL as indicators of tamoxifen response could be associated with substantial shifts in drug sensitivity in vitro, 10 genes were selected to be investigated experimentally. This includes the top nine genes from TINDL's ranking (RPP25, EMP1, EXTL3, EXOC2, NUP37, RPL13, WBP2NL, RPS6, and GBP1) as well as JAK2, which is ranked 19 in the list and is known for its role in the type II interferon

signaling pathway, a crucial pathway in cancer [149]. Breast CCLs that are estrogen receptor-positive, namely MCF7 and T47D, were utilized because tamoxifen is predominantly used in treating breast cancer patients that are estrogen receptor-positive. Additionally, 85% of the tamoxifen-treated patients in the dataset correspond to breast cancer. The knockdown experiments of all ten genes show that they have significantly impacted the sensitivity to tamoxifen in MCF7 cell lines, validating all tested genes in this context. For the T47D cell lines, seven of these genes were also confirmed. For more details, please refer to the publication corresponding to this chapter [12].

3.4 Discussion and Conclusion

Predicting responses to cancer treatments and identifying potential biomarkers of drug response are pivotal objectives in personalized medicine. Given the relative ease of in vitro data generation and collection compared to clinical samples, computational models that can predict clinical drug responses while only utilizing preclinical in vitro data for training can make a significant impact. This is particularly beneficial for newly developed or approved drugs, where clinical sample availability may be limited or nonexistent. However, the inherent biological and statistical differences between CCLs and patient tumors present undeniable challenges. A recent investigation [11] evaluated the capabilities of various ML models trained on preclinical CCLs to predict the CDR of cancer patients, including methods that integrate supplementary data like gene interaction networks. This study validated the complexity of this task and highlighted the necessity of meticulously designing computational methods to address such a challenging problem.

In this chapter, I introduced TINDL and demonstrated substantial improvement compared to state-of-the-art models, utilizing both traditional ML and DL techniques. The results emphasized on the importance of mitigating the statistical discrepancies between preclinical and clinical samples, alongside integrating information regarding tissue/cancer types of the tumor samples. TINDL is not a mere drug response predictor but rather facilitates the identification of the most predictive biomarkers for each drug. The biomarkers identified across multiple drugs (see Figure 3.5B) brought up crucial genes and signaling path-

ways that are potentially important in the mechanism of action of various anti-cancer drugs. Many genes highlighted in my analysis have been previously reported to exhibit altered expression levels in response to specific drugs, notably SLFN11 for multiple chemotherapies [126–128, 150, 151], SALL4 for cisplatin [152], ABCB1 for taxane and doxorubicin [153, 154], PIGB for gemcitabine [155], and BAX for oxaliplatin [156]. These observations suggest that my preclinical-to-clinical model can yield biologically relevant candidate genes and pathways, offering insights into the mechanisms underlying drug resistance and potentially presenting novel combinational therapies to overcome such resistance.

Using the top genes identified by the proposed pipeline, this research also demonstrated that a small panel of 20 genes could maintain the predictive performance of TINDL for tamoxifen (Figure 3.4). This shows a lot of potential in clinical applications since handling fewer genes would relatively be more practical than a large panel. Additionally, reasoning out predictions via a handful of genes may be more informative to experts who already have domain knowledge. In relation to this, the functional validation conducted on MCF7 and T47D cell lines through siRNA knockdown experiments of ten genes identified by TINDL confirmed the direct involvement of these genes in tamoxifen response. Although these genes were only corroborated in CCLs, such confirmation strengthens the claim to the effectiveness of this pipeline.

The interpretation phase of TINDL uses the *black box* version of CXPlain, in which another neural network was trained as an explainer. This can be seen as contradictory since this is an attempt to interpret the trained predictor using a non-interpretable explainer model. While I noted that the differences in the label space (binary vs continuous) necessitated this approach, it can be argued that interpretable models, say linear models, could have been used. However, linear explainers may have difficulty modeling the "decision process" of a nonlinear black box predictor. This also questions the necessity of the nonlinearity in the predictor if linear explainers were proven to be sufficient. Nevertheless, this would be an interesting avenue to explore.

My analysis indicates that TINDL outperforms alternative approaches, particularly in discerning between resistant and sensitive tumors across a greater number of drugs. While its superior performance over traditional ML models can be credited to DL's capability

to better model complex and nonlinear relationships, TINDL's superiority over DL-based domain adaptation techniques showcases its proficiency in mitigating discrepancies between preclinical and clinical samples.

The additional assessments in the latent space also confirm the hypothesis above, both quantitatively and qualitatively. Upon scrutinizing the principal components and the UMAPs of the samples derived from the two datasets in the latent space (Figure 3.4B, Supplementary Figures A.1-A.4), it was clearly visible that the distributions of GDSC and TCGA samples were notably distinct when employing domain adaptation models or ComBat. However, embeddings produced by TINDL suggested a better reduction of domain discrepancies, as demonstrated by the blending of GDSC and TCGA samples. I quantified this observation using the average inter-domain distance among samples in the latent space, where a smaller value is desirable. As illustrated in Figure 3.4A, TINDL exhibited a significantly lower average distance in comparison to the other methods. A potential explanation for this observation is the lack of integration of prior knowledge about the target domain in other approaches. TINDL capitalizes on the distinct patterns of GEx profiles inherent in specific tissues. By simply aligning the mean and standard deviations of the target tissues' GEx, the model is able to view the test distribution as similar to that of the training set. However, the tissue-informed normalization's simplicity can also be its weakness since the adjustments were done independently per gene. Therefore, considerations for gene dependencies should be further explored. Another factor is the difficulty of evaluating the level of adaptation in domain adaptation models, particularly since visual verification of GEx vectors, unlike images, is not feasible. Furthermore, domain adaptation techniques are susceptible to a problem akin to "mode collapse," wherein all samples are mapped to a tiny subspace in the latent space for which the discriminator becomes confused. This can be erroneously equated to having a sufficient level of adaptation.

Although current domain adaptation methodologies have many shortcomings in this application, I remain convinced that novel domain adaptation methods can be developed to enhance outcomes. However, such methodologies must be tailored for GEx data with the consideration of biological factors that influence cancer patients' responses to different drugs. Moreover, incorporating information regarding cancer type or even subtype for each cancer

may be necessary to achieve better results. Another factor to consider is the limitations of CCLs in emulating patient tumors (e.g., growth in a 2D environment, greater homogeneity compared to tumors, and inability to capture the intricacies of the tumor microenvironment), which restrict the predictive capacity of computational models trained on CCLs. Even if these models successfully address statistical discrepancies between training and test sets, their ability to accurately predict the CDR of cancer patients is inherently hindered. Furthermore, tumor-specific characteristics that may be useful in the response prediction task, but are not represented in preclinical samples, can be lost or unintentionally altered during the domain adaptation process. As a result, the availability of large datasets with better models of cancer, such as patient-derived organoids or xenografts, becomes pivotal in improving the predictive power of computational models.

In this study, I focused on models trained solely on GEx profiles of samples, as prior research has demonstrated this data type to be particularly informative regarding drug response [6]. However, adopting a multi-omics approach that incorporates various molecular characteristics of samples could potentially offer a more in-depth understanding of the relationships between cancer and drug response mechanisms. Nonetheless, designing such models entails additional challenges, such as higher chances of overfitting due to the extra features. Another limitation of this study was that all computational models were trained exclusively on CCLs and their responses to individual drugs. As mentioned, some of the patients in the TCGA dataset received multiple drugs, either in sequence or in combination, over the course of treatment. Although not optimal, I included these patients in the analysis due to the limited number of samples with known CDR. In such instances, computational models trained on single drugs only provide a rough approximation. To enhance the predictive capability under these circumstances, computational models must also account for the synergistic and antagonistic effects of drugs. Recent large publicly available datasets like DrugComb [157] and DrugCombDB [158], containing responses of various well-studied CCLs to pairs of drugs, offer an avenue for developing such methodologies.

Chapter 4

Incorporating Response Similarity via Bipartite Graphs

The advancement of machine learning (ML) and statistical analyses within precision medicine has garnered significant attention over the past decade. Predicting drug response based on molecular profiles of samples is a fundamental challenge in this domain, leading to the proposal of various methodologies [1, 11, 12, 80, 84]. Among these, gene expression (GEx) profiles of samples are frequently utilized due to higher predictive ability compared to other molecular profiles [1]. The establishment of public databases containing GEx profiles of hundreds of cancer cell lines (CCLs) and their corresponding responses to a multitude of drugs, such as GDSC [36], has expedited the advancement of novel methodologies in this realm.

Given the molecular and structural similarities among different drugs and their mechanisms of action, there is substantial interest in employing ML methods capable of leveraging these similarities. Rather than training individual ML models for each drug, there is a trend toward framing drug response prediction as a paired prediction problem. In this setup, a model is fed a CCL-drug pair as input, allowing for the training of a unified model across multiple drugs and CCLs [5, 6, 64]. This approach not only expands the sample size but also facilitates information sharing across various drugs and drug families. Chemical structure data (e.g., from PubChem [159] and ChEMBL [160]) prove particularly valuable for representing drugs, and multiple studies have developed methods to extract this information

[54, 68, 161].

Some methods [50–52] have framed drug response prediction as a matrix factorization problem, constructing a matrix composed of drugs and CCLs as columns and rows. One advantage of this formulation lies in its direct usage of the "entities" (i.e., drugs and CCLs) and their responses, thus eliminating the need to map feature representations of these entities to their response labels [50, 51]. Although available features can be incorporated for regularization, this framing is inherently transductive because samples and drugs are presumed to be in the matrix. Consequently, these models cannot be directly employed to predict the response of a new CCL to a drug unless the CCL possesses drug response information in the training set for some other drugs prior to training. Another set of methodologies employs collaborative filtering [74, 75], where predictions are computed using an entity's neighborhood, defined by similarities derived from gene expressions, molecular fingerprints, and drug responses. Given that these approaches require the computation of drug response similarities, an underlying assumption is the existence of some known responses for each unique CCL and drug. This is a less stringent assumption compared to that of matrix factorization methods because it does not demand known responses of the test set prior to training, but a few known responses for unseen samples (i.e., test CCLs/drugs) should still be provided during inference.

Inspired by previous approaches' concept of "entity", while also seeking to address their limitations stemming from their transductive nature, this chapter proposes to leverage the underlying matrix by transforming these entities into drug and CCL nodes to construct a bipartite graph. My hypothesis posits that integrating information from CCLs that are highly sensitive or resistant to a drug could enhance the representation of the drug, which would be beneficial in predicting drug responses. In my proposed method called Bipartite Graph-represented Drug Response Predictor (BiG-DRP and BiG-DRP+), I construct this graph by selecting the most sensitive and resistant CCLs for each drug and then linking the selected CCLs to the drugs via edges. Although drugs are not directly linked through edges amongst themselves, the incorporation of two-hop message passing enables the propagation of drug similarity information. The model takes the descriptors of the drugs and GEx profiles of the CCLs as input and utilizes them as node attributes for the bipartite graph and as

features for the samples. The output is a continuous drug response value corresponding to the predicted normalized log IC50.

To assess the performance of BiG-DRP and BiG-DRP+, I utilized 5-fold cross-validation and compared these outcomes against various baselines and alternative drug response prediction methodologies, specifically NRL2DRP [2], PathDNN [5], and tCNN [6]. I conducted tests utilizing two data-splitting techniques: 5-fold leave-pairs-out and 5-fold leave-CCLs-out, representing two likely scenarios of data availability. This study demonstrated significant enhancements compared to alternative approaches across both scenarios. Furthermore, leveraging a computational pipeline that I developed to identify the features that are most influential to the model, I was able to pinpoint genes that are indicative of biological processes and signaling pathways implicated in the mechanisms of action of the drugs.

4.1 Problem Statement

Given a fixed set of drugs $D = \{\mathbf{d}_1, ..., \mathbf{d}_n\}$, training data composed of CCLs $X = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, and a possibly incomplete training drug response matrix $Y \in \mathbb{R}^{m \times n}$, the goal is to train a model $f(\mathbf{x}, \mathbf{d})$ that can predict the response of any CCL $\mathbf{x}$ for a drug $\mathbf{d} \in D$. This is a multitask problem in the sense that a single model is used to predict for various drugs. Each CCL is represented by a feature vector $\mathbf{x} \in \mathbb{R}^p$, and each drug is represented by a feature vector $\mathbf{d} \in \mathbb{R}^q$.

4.2 Methods

4.2.1 Bipartite Graph-based Drug Response Prediction

I developed a method for the drug response prediction task that uses a novel architecture incorporating a bipartite graph between CCLs and drugs. The method is called Bipartite Graph-represented Drug Response Predictor (BiG-DRP) [13]. An extension of this method that accounts for constantly changing drug representations was also developed, which is called BiG-DRP+. Figure 4.1 shows an overview of the architecture of both models.

The BiG-DRP pipeline works as follows. As "fixed" inputs, the model receives a GEx

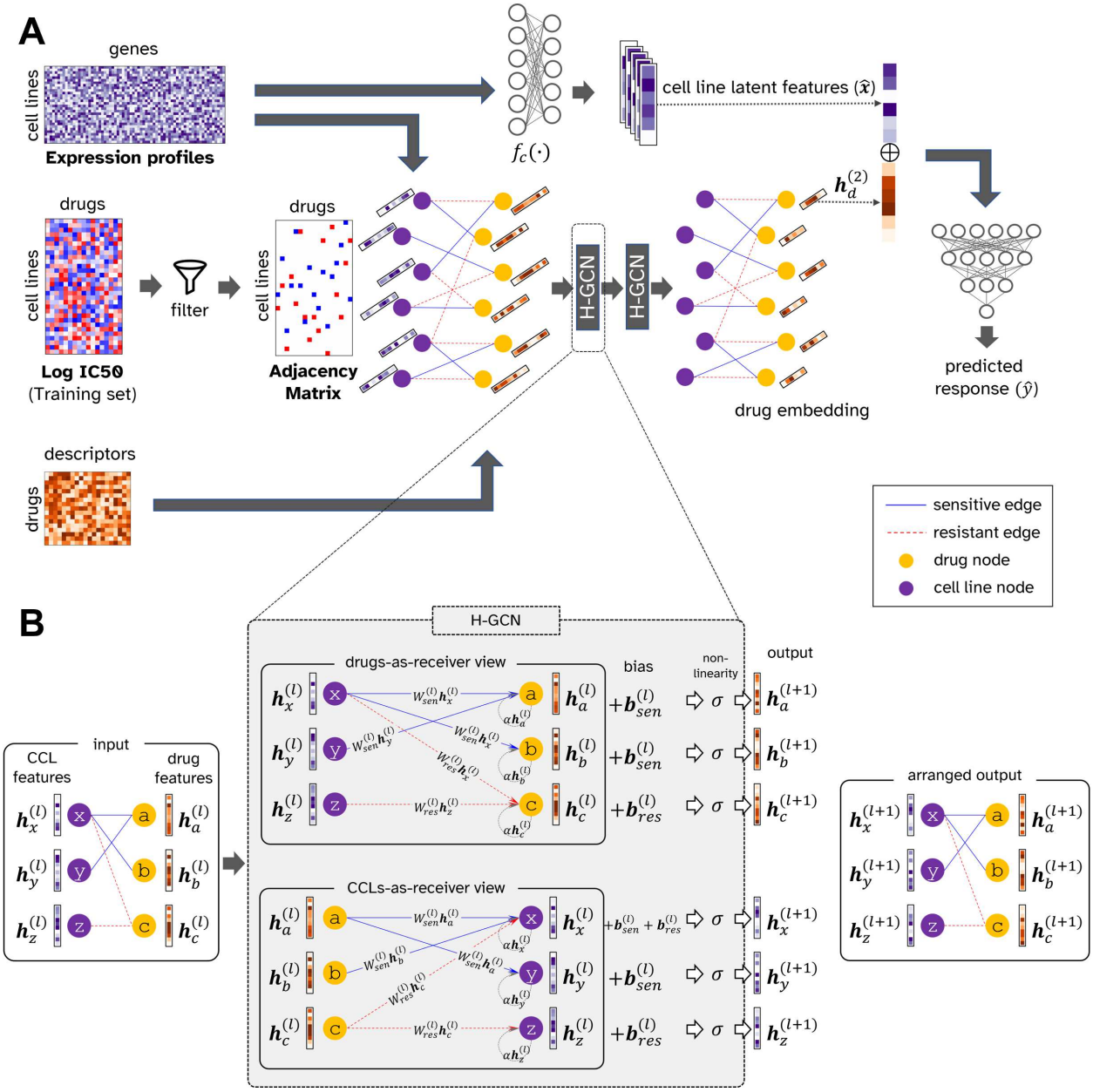

Figure 4.1: The BiG-DRP model. A) Drug embeddings are generated using two layers of heterogeneous graph convolutional network (H-GCN) applied to a CCL-drug bipartite graph with drug descriptors and GEx profiles as node attributes. Simultaneously, CCL embeddings are generated using an encoder based on the CCLs' GEx profiles. The predictor then utilizes these embeddings to predict the drug response. B) An overview of a layer of H-GCN is shown. Node attributes are multiplied to the weight matrices (W_{sen} and W_{res}) to generate "messages". The H-GCN propagates the messages to neighboring nodes through the graph structure. Subsequently, each node aggregates the received messages, biases, and self-information. The output will be the same graph, with each node having updated attributes incorporating information from their respective neighborhoods.

matrix of training CCLs, a matrix of drug features (descriptors in this case), and a possibly incomplete matrix of training drug responses (log IC50). These were termed "fixed" since these inputs do not change regardless of the CCL/drug being predicted. Furthermore, I would like to emphasize that the aforementioned inputs only come from the training set to prevent data leakage. Given these inputs, latent embeddings of all the drugs are obtained using the graph-based encoder. In parallel, the model receives another set of inputs: the GEx profile of a single CCL sample and the drug identifier (say d) for which the response is to be predicted. A separate encoder (f_c) is applied to the GEx to obtain the CCL's latent embedding denoted as $\hat{\boldsymbol{x}}$ in Figure 4.1. The drug identifier d is used to single out the corresponding drug embedding, denoted as $\boldsymbol{h}_d^{(2)}$ in Figure 4.1. These embeddings are then concatenated and sent to a predictor neural network, which uses them for the drug response prediction task.

Drug embeddings are obtained using a drug-specific encoder that utilizes a graph neural network. First, a heterogeneous bipartite graph composed of drug and CCL nodes is formed. Two types of edges are used to connect these nodes, reflecting the relationship between the connected CCL-drug pair: sensitive edges and resistant edges. The edge type is determined based on the log IC50 values of each CCL-drug pair in the training set. When a CCL is highly sensitive to a drug (low log IC50), they are connected using a sensitive edge; on the other hand, when a CCL is highly resistant to a drug (high log IC50), they are connected using a resistant edge. To define the "highly" sensitive/resistant relationship, I connected each drug to the 1% CCLs with the lowest/highest log IC50 values in the training set. For each node in the graph, I also defined attribute vectors: GEx profiles for CCL nodes and drug descriptors for drug nodes. This bipartite graph was then used as input to a heterogeneous graph convolutional network (H-GCN) to obtain drug embeddings ($\boldsymbol{h}_d^{(2)}$ in Figure 4.1). The idea behind using H-GCNs is that the obtained embedding for a drug captures different types of patterns in the data. On one hand, it captures the molecular characteristic of the drug itself (based on its node attributes). On the other hand, it also captures the characteristics of drugs that induce a similar sensitive/resistant pattern in the CCLs to further augment the obtained embedding. Moreover, the use of GEx profiles of CCLs as node attributes in the graph enables the model to have a broader awareness of the drug similarity patterns by

also utilizing CCL similarity patterns.

I should note that while it is possible to utilize the H-GCN (and the bipartite graph) to obtain CCL embeddings, I opted for an independent CCL encoder. The rationale behind such a choice was that H-GCN embeddings would limit the use of the model to only predict drug responses of CCLs that are already present in the training set. More specifically, had I used the H-GCN, I would not be able to obtain embeddings for a testing CCL that is not present in the training set, since it would have been in the form of a single disconnected node in the bipartite graph. In practical applications of precision medicine, it is more logical to seek a model that can predict for newly acquired samples, which would have not been seen by the model during training. Therefore, it was crucial to have an independent CCL encoder.

One of the challenges in training BiG-DRP was the continuous modification (across batches) of the drug embeddings stemming from the underlying message-passing mechanism of the H-GCN. To overcome this issue and stabilize the trained model, I developed an extension of BiG-DRP, called BiG-DRP+, which has the same architecture but uses a slightly different training procedure. After BiG-DRP was completely trained (i.e., after its final training epoch), an extra epoch of training was executed, but with a reduced learning rate and frozen drug embeddings. Lowering the learning rate ensures that the predictor would not overfit while freezing the drug embeddings ensures that the input to the predictor does not constantly change for the same drug across batches.

4.2.2 Heterogeneous Bipartite Graph Construction

Let the heterogeneous bipartite graph be denoted as $G(V_C, V_D, E_r, E_s)$. Here, V_C represents the set of CCL nodes present in the graph, V_D represents the set of drug nodes, E_r represents the set of resistant edges, and E_s represents the set of sensitive edges. Given a predetermined value of k, a drug is connected to CCLs whose log IC50 is among the top (bottom) k% of all CCLs via a resistant (sensitive) edge. The set V_C is the union of all such CCLs, which is potentially a subset of all CCLs in the training set. The edges used in the bipartite graph are unweighted—the log IC50 values are only used to determine the presence or absence of an edge, but not for its weight. In the majority of the results presented in this chapter, I used

$k = 1$. However, my analyses showed that the performance of BiG-DRP and BiG-DRP+ was not too sensitive to the exact choice of k, implying that tuning for this hyperparameter may not be necessary.

4.2.3 Details of the H-GCN encoder for Drug Embeddings

I used a 2-layer H-GCN architecture to obtain drug embeddings using the bipartite graph explained above. H-GCN is a variation of the graph convolutional network architecture [106], which allows the utilization of multiple edge types (here two). The following equation shows the forward pass of an H-GCN:

$$\boldsymbol{h}_v^{(l+1)} = \sigma\left(\sum_{r\in R}\left(\boldsymbol{b}_r^{(l)} + \sum_{u\in N(v,r)}\frac{1}{c_{u,v}}\boldsymbol{h}_u^{(l)}W_r^{(l)}\right) + \alpha\boldsymbol{h}_v^{(l)}\right). \tag{4.1}$$

In this equation, $\boldsymbol{h}_v^{(l)}$ is the embedding of node v at the lth layer. I denote the nonlinearity function as σ, the set of edge types as R, and the neighbors of node v through edge type r as $N(v,r)$. $W_r^{(l)}$ and $\boldsymbol{b}_r^{(l)}$ represent the weights and biases, respectively, for the edge type r at the lth H-GCN layer. The denominator $c_{u,v}$ generalizes the notation to different types of normalization to prevent extremely large values due to the size of the node neighborhood. In this study, I used $c_{u,v} = \sqrt{|N(v,r)|}$.

It is a common practice to add self-loops in the input graphs used by GCNs. This enables the node to retain some self-information from the previous layer, preventing the node's embedding from completely relying on its neighbors' information. However, self-loops increase the complexity of the H-GCNs by adding another set of parameters. To avoid this while achieving a similar final effect, I introduced a residual term ($\alpha\boldsymbol{h}_v^{(l)}$) to the forward pass, simulating the effect of self-loops. I used the hyperparameter α (fixed to α=0.5) to control the amount of information to be retained from the previous layer.

Applying the 2-layer H-GCN to the constructed bipartite graph, the resulting drug embeddings capture relevant information from the CCLs that are highly sensitive/resistant to the drug (its one-hop neighbors), as well as information from drugs to which these CCLs have similar or inverse response patterns (its two-hop neighbors). In other words, the H-GCN component enables the sharing of information among drugs that are connected to similar sets

of CCLs via similar edge types. Conversely, information regarding the contrasting effects of drugs on a set of CCLs is also propagated.

4.2.4 Training Procedure

The CCL encoder receives the CCLs' GEx vector $\boldsymbol{x}$ as input and produces a latent representation denoted as $\hat{\boldsymbol{x}} = f_c(\boldsymbol{x})$. This encoding and the drug d's node embedding (denoted as $\boldsymbol{h}_d^{(2)}$) are then concatenated and used as input to the predictor, which is designed as a 3-layer neural network that outputs the predicted drug response values (denoted as $\hat{y}$).

I trained the model in an end-to-end fashion using the Adam optimizer [162] with the MSE $L = (y - \hat{y})^2$ as the loss function. I kept the loss function, similar to that of Chapter 3, because the distinction between highly sensitive and resistant samples (i.e., extremes of the label distributions) is more relevant in generalizing to unseen CCL data and is expected to give better signals for the subsequent interpretation. I fixed the learning rate to 0.0001 and the batch size to 128. In the Results section, I will discuss and analyze the effect of different hyperparameter choices on the performance of the model. I used Leaky ReLU as the activation function defined as $\sigma(x) = max(0, x) + 0.1 \times min(0, x)$. Based on a validation set formed by randomly selecting samples from the training data, early stopping was used to determine the optimal number of training epochs. Subsequently, the model was re-trained using the entire training set for the determined number of training epochs. Each batch constituted a set of CCL-drug pairs, even though all drug embeddings could be generated simultaneously in each forward pass.

As briefly explained earlier, BiG-DRP+ was designed to stabilize the training of BiG-DRP. Below, I explain the issue that this model tries to resolve. Since the embeddings generated by GCNs rely on node connectivity, a small perturbation of a node's embedding may dramatically affect the embeddings of its neighbors in the next layer of GCN (or H-GCN), even with a relatively small learning rate. This may result in the predictor getting confused, as it may not be able to easily map the "new" perturbed embedding to the "known" ones. The problem is particularly severe if the drug of interest was not a part of the batch during the previous training step. Effectively, this means that the predictor is receiving an infinite number of drugs (instead of the finite set of available drugs), making the learning process

challenging. To overcome this "moving embedding" problem, I developed BiG-DRP+, which modifies the training procedure of BiG-DRP described above. In this modification, the idea is to halt the training of the H-GCN after several epochs but continue the training of the predictor network using the "frozen" drug embeddings and with a lower learning rate (I used a learning rate of 0.00001). In the BiG-DRP+ model, I froze the drug embeddings after training for the set number of epochs (determined by 5-fold CV) but continued the training of other components of the model for one extra epoch. This stabilized the training of the predictor and enabled it to capture the patterns of CCLs treated by the same drug since half of the predictor's input (i.e., the drug's embedding) was fixed during this epoch.

4.2.5 Dataset Acquisition and Preprocessing

I used the Genomics of Drug Sensitivity in Cancer (GDSC) database for the drug response data (log IC50 values) [36]. I only limited the set of drugs to those that had both known log IC50 values and binarized responses. This enables the calculation of several performance metrics, some of which require the knowledge of binarized responses. Additionally, I consolidated data pertaining to drug duplicates and only kept the one for which there were more CCLs with documented responses. Duplicates corresponded to any of the following cases: (1) cases in which the same drug was measured in different batches (with different drug IDs), (2) cases in which synonyms were used to name the drugs, or (3) cases labeled as "rescreens". I performed z-score normalization on log IC50 values of each drug across CCLs (one drug at a time). This normalization was performed to enable comparison of results across different drugs, which may have widely different ranges of log IC50 values.

To obtain drug features, I first acquired string representations of the drug molecules in the form of Simplified Molecular Input Line Entry System (SMILES) encoding [163]. RDKit [161] was utilized to generate drug descriptors (e.g., molecular weight, number of aromatic rings) based on these SMILES strings. Descriptors that contained missing values for the selected drugs were excluded from the analysis. After filtering, the dataset is left with 237 unique drugs, each represented by a 198-long feature vector of drug descriptors. These drug feature vectors were then z-score normalized across all drugs, one descriptor at a time. Morgan fingerprints, which I used as alternative drug features in one of the analyses, were

also generated from the SMILES strings using the RDKit software. I kept the default length of 512 for the Morgan fingerprints and skipped normalization since these are binary vectors.

RNA-seq GEx profiles of 1001 CCLs were obtained from the Cell Model Passports [37] in Fragments Per Kilobase of transcript per Million mapped reads (FPKM). I then processed the GEx using $log_2(\text{FPKM} + 1)$. Genes with missing values in some CCLs or those with a standard deviation less than 0.1 were excluded. Only CCLs with drug responses and GEx were included in the analysis. Upon completion of these steps, 944 unique CCLs and their GEx profiles corresponding to 13,823 unique genes were left in the dataset. Overall, the dataset used in this study contained a total of 181,380 labeled CCL-drug pairs.

4.2.6 Evaluation and Cross-Validation

I used 5-fold cross-validation (CV) for training and evaluation of all models. I ensured that the fold kept aside for evaluation (i.e., test fold) was not used for the training of the parameters or hyperparameters of the models. To form the folds in the CV procedure, I used two data splitting strategies: leave-pairs-out (LPO) and leave-CCLs-out (LCO).

In the LPO-CV strategy, the folds were randomly selected from the set of all CCL-drug pairs. This means that a CCL or a drug in the testing set may have been observed by the model during training, but never at the same time. In the LCO-CV strategy, the folds were randomly selected from the set of all CCLs. In this setup, a CCL in the testing set has never been observed by the model during training. GEx values of CCLs were z-score normalized per gene using the means and standard deviations calculated from the unique CCLs in the training folds to prevent data leakage between the training and test sets. To ensure a fair comparison among the models, identical folds were used for all methods. For each drug, the predictions obtained for the five folds on their respective test sets were gathered and used to calculate performance metrics and evaluate different methods.

In one of the experiments, I assessed the generalizability of the models to independent cancer tumor datasets from The Cancer Genome Atlas (TCGA) [35]. I obtained GEx profiles of tumors (FPKM values) as well as their RECIST-based clinical drug responses. I followed the procedure used in previous studies [11, 24] and grouped the patients into either resistant (stable disease, progressive disease) or sensitive (complete response, partial response).

Table 4.1: List of models evaluated and their inputs.

Model	Drug features/attributes	Other inputs
BiG-DRP+	Descriptors	GEx
BiG-DRP	Descriptors	GEx
Inverted BiG-DRP	Descriptors	GEx
MLP	Descriptors	GEx
NRL2DRP	None	Drug-CCL-gene network
tCNN	SMILES one-hot encoding	Genetic features
PathDNN	Drug targets	GEx, pathway information
SVR-RBF (w/RFE)	Descriptors	GEx
SVR-Linear (w/RFE)	Descriptors	GEx
SVR-RBF	Descriptors	GEx
SVR-Linear	Descriptors	GEx

I obtained data corresponding to four drugs: cisplatin (n = 398), paclitaxel (n = 233), gemcitabine (n = 226), and doxorubicin (n = 208). These drugs were present in the training data, had more than 50 samples (tumors) in each category of resistant/sensitive, and had a large number of samples with known clinical drug responses. Similar to the preprocessing above, the expression of each gene was first processed using $log_2(\text{FPKM}+1)$ and then z-score normalized across the tumors. I used PyCombat [105] to mitigate the effect of statistical discrepancies between samples from GDSC and TCGA.

4.2.7 Alternative Methods for Benchmarking

Several baseline methods were used to benchmark against the proposed models. They were selected to include both deep learning (DL)-based and traditional ML methods. The first baseline method was a multilayer perceptron (MLP) that had architecture and hyperparameters similar to those of BiG-DRP. Moreover, the input to this model also consists of GEx profiles of samples and drug descriptors. The main difference was that instead of an H-GCN encoder, I used a fully connected neural network to obtain drug embeddings. As representatives of traditional ML methods, I included linear support vector regressor (SVR) and SVR with radial basis function (RBF). For the SVRs above, the input was a concatenation of the CCL's GEx profile and the drug's features. I used Nystroem's transformation [164] to approximate the SVR's kernels. This was done to improve the training efficiency, which was

an obstacle due to the large size of the data. A nested CV approach was used to tune the number of Nystroem components, regularization factor, and gamma for RBF, all of which were considered hyperparameters. Additionally, I included variations of the SVR models above that utilized recursive feature elimination (RFE) [165] to identify the most relevant features for the task.

In addition to the models above, I also included several state-of-the-art (SOTA) drug response prediction approaches in my comparative analysis. Descriptions and considerations for these approaches are summarized below.

NRL2DRP [36] is a model based on graph representation learning and uses a graph consisting of gene, drug, and CCL nodes. These nodes are connected by different types of edges, including those capturing sensitivity, mutation, and protein-protein interactions. NRL2DRP does not utilize DL methodologies in its pipeline; instead, it uses LINE [49], which is a topology-based graph embedding algorithm that is typically used in transductive learning settings. I had to slightly modify NRL2DRP since it was originally designed for a binary classification task, and it needed to work with continuous data in a regression task (i.e., using SVR instead of SVM to predict the response from the node embeddings).

PathDNN [5] is a DL-based drug response predictor that tries to add some degree of prior knowledge to the architecture. It achieves this by constraining the connectivity of the neural network using a pathway mask, effectively grouping together various functionally related genes within the model's inner workings. This method additionally uses drug targets (in the form of genes) and GEx profiles, both of which should be a member of one of the pathways present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway collection [166]. I obtained the data available in PathDNN's repository corresponding to drug targets and pathway information. The drug targets present in this repository were represented using their normalized STITCH [167] confidence score. To run PathDNN on the processed dataset, I had to exclude three compounds (foretinib, bx795, and i-bet 151), since they did not have any known targets among the pathways present in KEGG.

tCNN [6] is a DL-based method that utilizes 1-D convolutional neural networks (CNNs)

in its architecture. In this method, the canonical SMILES representation of a compound is encoded into a sequence of one-hot vectors representing each character. To deal with the varying length of SMILES strings, extra zeros are padded at the end of the sequence to obtain equal-sized encodings for all compounds. In the end, this results in an $m \times n$ binary matrix, where n is the length of the encoding and m is the number of unique characters. To represent CCLs, this model utilizes genetic features pertaining to mutations and copy number alterations, which I obtained from GDSC.

When assessing the performance, I took several measures to ensure a fair comparison and a fair chance for the baseline models. First, I fixed the folds in the CV evaluation for all models and used identical folds across the board. Moreover, when a SOTA method was used with additional information in the original study, I provided those data to the model following the descriptions provided in each method's manuscript, even when they were not used by BiG-DRP(+). Table 4.1 provides a summary of the inputs for each approach.

4.2.8 Identification of Biomarkers of Drug Response

Some of the follow-up analyses I performed required the identification of CCL features (i.e., genes) that are predictive of drug response. For this purpose, I used CXPlain [99] and the approach discussed in the previous chapter to aggregate contributions across CCLs and identify top contributors (Section 3.2.5-3.2.6). CXPlain is a neural network explainer that utilizes the concept of Granger's causality [100] to quantify feature attribution. More specifically, it tries to predict the increase in a sample's loss when the feature's value is set to zero. In the analyses, I decided to train separate explainers for each drug. This approach simplifies the analysis since one no longer needs to learn attributions for multiple drugs simultaneously. Moreover, it eliminates the need to learn attribution for drug features. After feature/gene attributions in the test CCLs were obtained, I calculated the mean score of each gene across all CCLs for each drug. Then, I used the kneedle algorithm [101] (with sensitivity S=2) to identify the knee point and the threshold above which the genes were considered influential.

Table 4.2: Results of 5-fold LPO-CV evaluation. Values depicted in boldface reflect the best performance values. Both mean and standard deviations (shown in brackets) are calculated across the drugs. *Three drugs had to be excluded for PathDNN, since this method requires each drug to have targets present in KEGG's signaling pathways.

Model	AUROC mean (±SD)	RMSE mean (±SD)	SCC mean (±SD)	PCC mean (±SD)
BiG-DRP+	**0.878** (±0.068)	**0.843** (±0.241)	**0.748** (±0.100)	**0.758** (±0.102)
BiG-DRP	0.875 (±0.068)	0.855 (±0.244)	0.742 (±0.099)	0.752 (±0.102)
Inverted BiG-DRP	0.862 (±0.075)	0.888 (±0.253)	0.721 (±0.110)	0.730 (±0.110)
MLP	0.835 (±0.083)	0.954 (±0.273)	0.675 (±0.120)	0.681 (±0.119)
NRL2DRP	0.804 (±0.085)	1.153 (±0.345)	0.516 (±0.119)	0.514 (±0.123)
tCNN	0.787 (±0.082)	1.086 (±0.336)	0.587 (±0.119)	0.591 (±0.117)
PathDNN*	0.766 (±0.083)	1.165 (±0.355)	0.516 (±0.115)	0.529 (±0.117)
SVR-RBF (w/RFE)	0.738 (±0.101)	1.182 (±0.384)	0.503 (±0.125)	0.500 (±0.130)
SVR-Linear (w/RFE)	0.738 (±0.101)	1.181 (±0.393)	0.498 (±0.130)	0.497 (±0.135)
SVR-RBF	0.737 (±0.100)	1.182 (±0.383)	0.502 (±0.123)	0.499 (±0.129)
SVR-Linear	0.736 (±0.101)	1.184 (±0.393)	0.494 (±0.129)	0.493 (±0.134)

4.2.9 Pathway Characterization Analysis of Top Genes

To functionally characterize the top genes identified to be most predictive of drug response, I performed pathway enrichment analysis through KnowEnG's cloud platform [102] using the Reactome pathway collection [168]. The pipeline generates P values corresponding to Fisher's exact tests. These P values were corrected for multiple tests (i.e., multiple pathways) using the Benjamini-Hochberg false discovery rate (FDR).

4.3 Results

4.3.1 Performance based on Leave-Pair-Out Cross-Validation

I compared the performance of BiG-DRP and BiG-DRP+ with the aforementioned baseline methods in a five-fold LPO-CV, discussed earlier. Table 4.2 shows a summary of the results. In this table, the area under the receiver operating characteristic curve (AUROC) is calculated based on the binarized drug responses, while root mean squared error (RMSE), Pearson's correlation coefficient (PCC), and Spearman's correlation coefficient (SCC) were

calculated based on the continuous log IC50 values. The values correspond to mean and standard deviations across drugs. As can be seen in this table, BiG-DRP+ and BiG-DRP outperform all other methods. Comparison between BiG-DRP+ and the MLP, which has a similar architecture except for the drug encoder, reveals the importance of the H-GCN encoder: BiG-DRP+ has a ∼5% higher AUROC and ∼11% higher SCC and PCC compared to MLP.

In Table 4.2, I also reported the performance of a variation of the model called "inverted BiG-DRP." In this model, I substituted the role of the H-GCN and the independent encoder: the former was used to obtain CCL embeddings, while the latter was used to obtain drug embeddings. Results showed that this variation outperforms other baselines, except for BiG-DRP and BiG-DRP+. However, one should note that such a model cannot be used to predict the response of unseen CCLs during inference time.

Next, I compared the performance of BiG-DRP+ and other baseline methods for individual drugs. Figure 4.2 shows the SCC of the predicted values and the ground truth for each drug, represented by points. The plot compares the performance of BiG-DRP+ on the y-axis and a baseline method on the x-axis. Although, on average, the SCC of BiG-DRP+ and BiG-DRP are very close (Table 4.2), the first panel in Figure 4.2 shows that the performance improvement achieved by BiG-DRP+ is greater than zero (one-sided Wilcoxon signed-rank test P = 2.26E-36). I attribute this slight but consistent improvement to the stabilizing procedure that was incorporated in BiG-DRP+. Compared to other baselines, BiG-DRP+ shows a better performance in the majority (and in many cases in all) of the drugs. This can be visually observed by the fact that most of the points in each panel are above the diagonal line and have statistically significant P values.

4.3.2 Performance based on Leave-CCLs-Out Cross-Validation

In the next experiment, I compared the performance of BiG-DRP(+) with different baseline models in a five-fold LCO-CV. In this setup, a CCL in the testing set has never been observed by a model during training. This stricter setup is more appropriate for precision medicine applications since this scenario aims to predict the response of a new sample to a drug. The results of this experiment are provided in Table 4.3 based on the four metrics of RMSE,

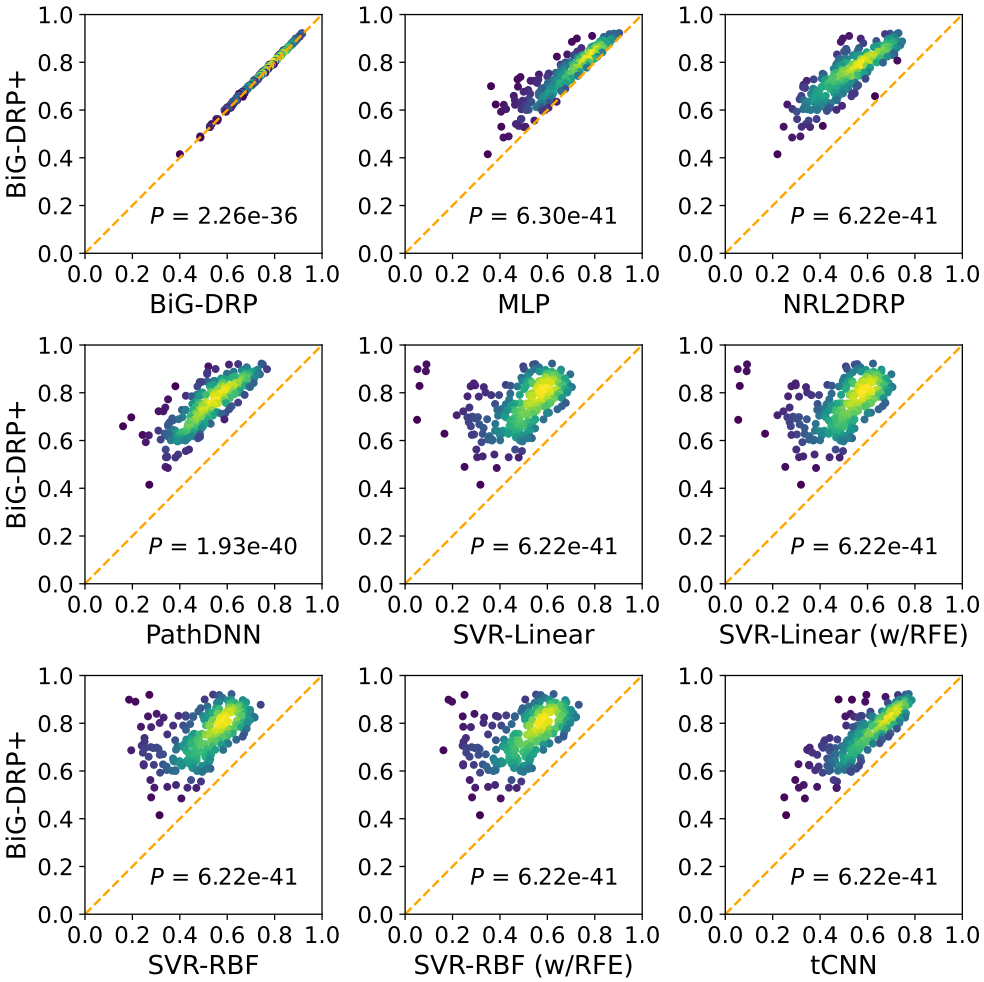

Figure 4.2: The Spearman's rank correlation coefficient (SCC) of BiG-DRP+ versus other methods. Each point reflects a drug. The color scheme represents the density of points in each area to improve visualization. A one-sided Wilcoxon signed-rank test was used to calculate the P value in each panel.

Table 4.3: The performance of BiG-DRP, BiG-DRP+ and baseline methods using 5-fold LCO-CV evaluation. Best performance values are in boldface and underlined. The mean and standard deviations are calculated across the drugs. *Three drugs had to be excluded for PathDNN, since this method requires each drug to have targets present in KEGG's signaling pathways.

Model	AUROC mean (±SD)	RMSE mean (±SD)	SCC mean (±SD)	PCC mean (±SD)
BiG-DRP+	**0.746** (±0.077)	**1.204** (±0.367)	**0.431** (±0.094)	**0.450** (±0.105)
BiG-DRP	0.743 (±0.077)	1.210 (±0.368)	0.426 (±0.095)	0.443 (±0.106)
MLP	0.730 (±0.086)	1.219 (±0.374)	0.413 (±0.100)	0.430 (±0.111)
SVR-RBF (w/RFE)	0.682 (±0.107)	1.276 (±0.404)	0.354 (±0.116)	0.360 (±0.127)
SVR-RBF	0.680 (±0.110)	1.278 (±0.403)	0.348 (±0.120)	0.354 (±0.135)
SVR-Linear	0.666 (±0.102)	1.292 (±0.420)	0.324 (±0.119)	0.331 (±0.126)
SVR-Linear (w/RFE)	0.666 (±0.102)	1.293 (±0.421)	0.322 (±0.118)	0.330 (±0.124)
PathDNN*	0.612 (±0.074)	2.201 (±0.698)	0.193 (±0.061)	0.170 (±0.078)
tCNN	0.586 (±0.060)	1.369 (±0.427)	0.147 (±0.068)	0.147 (±0.072)

AUROC, SCC, and PCC. Note that I was not able to use NRL2DRP in the LCO-CV setup, since a significant component of their model, namely LINE [49], is designed in a transductive manner.

This table shows that BiG-DRP+ and BiG-DRP had the best and second-best performance, respectively, compared to all other baselines. The performance gap between MLP and BiG-DRP+ further emphasizes the importance of the H-GCN encoder using the bipartite graph. The superior performance of BiG-DRP+ was also evident in the drug-wise analysis, in which it showed a statistically significant improvement compared to other methods (one-sided Wilcoxon signed-rank test, Supplementary Table B.1).

As the third experiment, I set out to assess how BiG-DRP+ generalizes to a completely independent dataset. For this task, I focused on the prediction of response to cisplatin, gemcitabine, doxorubicin, and paclitaxel in primary cancer tumors from the TCGA dataset. Since the ground truth drug responses of these tumors were available in the form of RECIST annotations (four ordinal categories), I first binarized them into resistant and sensitive categories (described in Methods). Then, I used a statistical test (one-sided Mann-Whitney U test) to determine whether there is significant evidence to suggest that the predicted values for resistant patients are greater than those for sensitive patients, implicating consistent pre-

Table 4.4: LPO-CV Performance of BiG-DRP and BiG-DRP+ at different values of k.

k	BiG-DRP+		BiG-DRP	
	SCC mean (± SD)	AUROC mean (± SD)	SCC mean (± SD)	AUROC mean (± SD)
0.5	0.748 (±0.100)	0.878 (±0.069)	0.742 (±0.100)	0.874 (±0.069)
1	0.748 (±0.100)	0.878 (±0.068)	0.742 (±0.100)	0.875 (±0.068)
2	0.746 (±0.100)	0.878 (±0.068)	0.741 (±0.100)	0.875 (±0.068)
5	0.745 (±0.100)	0.877 (±0.068)	0.739 (±0.101)	0.874 (±0.069)
10	0.742 (±0.101)	0.875 (±0.070)	0.736 (±0.101)	0.871 (±0.070)

dictions with the ground truth. I used Mann-Whitney U test (instead of a t-test) since one of the drugs did not pass the test of normality. The results were statistically significant for cisplatin ($P = 2.19\text{E-}7$), doxorubicin ($P = 8.80\text{E-}3$), and gemcitabine ($P = 3.40\text{E-}2$). In the aforementioned analyses, I included any tumor that had received these drugs (alone or in combination with other drugs). I repeated this analysis, but removed any sample that had received the drug of interest with another drug, or had received a different drug beforehand. While this restriction reduced the number of samples (and hence the statistical power of the tests), the results were still significant for cisplatin ($P = 1.82\text{E-}2$) and doxorubicin ($P = 4.29\text{E-}2$). Here, I used Welch's t-test since the data for all drugs passed the test of normality. Supplementary Tables B.2-B.3 provide detailed information regarding the samples and the results of different statistical tests.

4.3.3 The Effect of Different Components on the Performance of BiG-DRP+

As described in Methods, the threshold used to determine sensitive and resistant edges was fixed in advance at $k = 1$ percent. Given the importance of this component, as evident from comparisons with an MLP with a similar architecture, I set out to determine the sensitivity of the results to this choice. As a reminder, a drug is connected to a training CCL with a resistant (sensitive) edge if the log IC50 of the CCL is among the top (bottom) $k\%$ of all the CCLs in the training set. I tested different choices of $k \in \{0.5, 1, 2, 5, 10\}$ and constructed the bipartite graphs accordingly. The results, based on LPO-CV (Table 4.4) and LCO-CV

Table 4.5: LCO-CV Performance of BiG-DRP and BiG-DRP+ at different values of k.

k	BiG-DRP+		BiG-DRP	
	SCC mean (± SD)	AUROC mean (± SD)	SCC mean (± SD)	AUROC mean (± SD)
0.5	0.432 (±0.094)	0.747 (±0.077)	0.426 (±0.094)	0.744 (±0.078)
1	0.431 (±0.094)	0.746 (±0.077)	0.426 (±0.095)	0.743 (±0.077)
2	0.430 (±0.095)	0.745 (±0.078)	0.425 (±0.094)	0.742 (±0.077)
5	0.429 (±0.094)	0.743 (±0.079)	0.423 (±0.093)	0.738 (±0.080)
10	0.428 (±0.096)	0.742 (±0.080)	0.423 (±0.097)	0.739 (±0.081)

Table 4.6: The performance of BiG-DRP+ with different drug attributes. The rows show the results of BiG-DRP+ when drug descriptors (vectors of length 198), Morgan fingerprints (vectors of length 512), or the combination of both (vectors of length 710) are used as node attributes.

Drug Attribute	LPO-CV		LCO-CV	
	AUROC mean (± SD)	SCC mean (± SD)	AUROC mean (± SD)	SCC mean (± SD)
Descriptors	0.878 (±0.068)	0.748 (±0.100)	0.746 (±0.077)	0.431 (±0.094)
Morgan FP	0.878 (±0.068)	0.748 (±0.100)	0.743 (±0.080)	0.426 (±0.098)
Both	0.879 (±0.068)	0.748 (±0.099)	0.746 (±0.077)	0.433 (±0.095)

(Table 4.5), showed that the model remains robust to the choice of this hyperparameter. While a minor deterioration was observed as k increased, this was only less than 1% in all evaluations when ranging k from 1 to 10 percent. The reason for this deterioration is that an increase in the value of k also increases the potentially erroneous edges in the bipartite graph, adding low-confidence edges to the model.

Since I used drug descriptors in the main analyses, I asked whether alternative drug features as attributes in the bipartite graph can improve BiG-DRP(+)'s performance. To investigate this, I used Morgan fingerprints [54] as an alternative drug representation, alone or in combination with the drug descriptors. Similar to the drug descriptors, these representations were used as node attributes in the bipartite graph. The results (shown in Table 4.6) revealed that using both representations simultaneously improves the results. However, the difference in performance is relatively small.

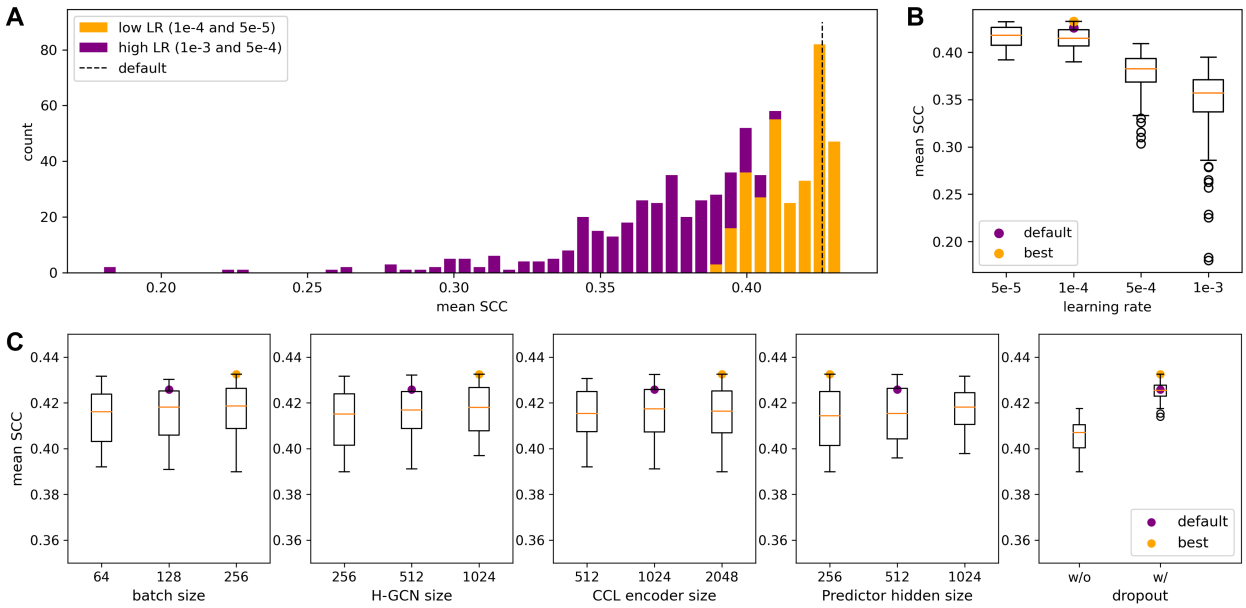

Figure 4.3: The effect of different hyperparameter combinations on the performance of BiG-DRP+. A) The panel shows the distribution of mean SCCs of the models in a 5-fold LCO-CV setup. Two distinct colors are used to distinguish between combinations involving low or high learning rates. B) The panel contains boxplots of the mean SCCs based on the learning rate choice. C) The panel contains boxplots of the mean SCCs for hyperparameter combinations using low learning rates (1E-4 and 5E-5) for different hyperparameter choices. The horizontal line shows the median. In all boxplots, the purple point represents the default hyperparameter combination used in the analyses, while the orange point shows the combination resulting in the best performance.

Finally, I evaluated the effect of different choices of hyperparameters on the performance of BiG-DRP+. I considered 648 different combinations of hyperparameters (as a grid). The following choices were considered: four learning rate options (5E-5, **1E-4**, 5E-4, 1E-3), three batch sizes (64, **128**, 256), three choices for the size of the CCL encoder (512, **1024**, 2048), three choices for the size of the H-GCNs (256, **512**, 1024), three choices for the size of the predictor hidden layer (256, **512**, 1024), and finally the **presence** or absence of dropout. Figure 4.3 shows the effect of these choices on the performance. I observed that there were 82 combinations of hyperparameters that resulted in a performance on par with that of the default parameters (in boldface), and 47 combinations resulted in better performance. These results suggest that while the default hyperparameters give reasonable performance, one can further improve the performance by tuning these hyperparameters, albeit at the expense of increased computational overhead.

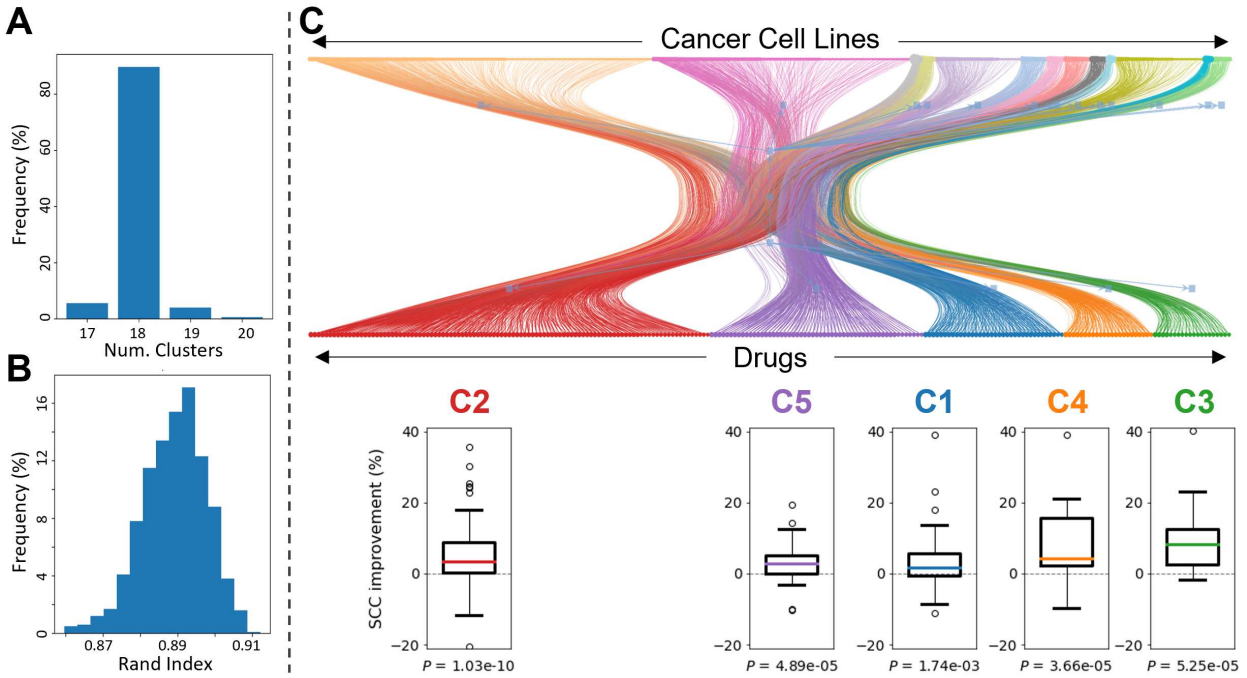

Figure 4.4: The aggregated bipartite graph and its clusters. A) The number of clusters found by NSBM in each run of the algorithm (out of 1000 runs). B) The histogram Rand Index between each clustering and the final cluster assignment. C) The top panel shows the bipartite graph of CCLs (top) and drugs (bottom). The bottom panel shows the boxplots corresponding to the SCC improvements by BiG-DRP+ (compared to MLP) in the LCO evaluation. A one-sided Wilcoxon signed-rank test is used to calculate the P value in each boxplot.

Carefully analyzing each hyperparameter separately, I found that the learning rate has a significant effect on the performance (Figure 4.3B)—the mean SCC deteriorates with a relatively large learning rate, while learning rates of 5E-5 or 1E-4 (the default value) work well with the model. Interestingly, with the lower learning rates above, other hyperparameters have a relatively small effect, with most hyperparameter combinations resulting in good performance (Figures 4.3A and 4.3C). I also note that the inclusion of dropout seems to slightly improve performance.

4.3.4 Detailed Analysis of the Bipartite Graph

Here, I set out to characterize the CCL-drug bipartite graph used in the analysis. Since a slightly different bipartite graph based on the training set was generated in each run of the 5-fold CV, I first formed a single consensus graph by aggregating them. To do this, I extracted

the union of the edges between CCL-drug pairs across the five graphs. To identify clusters in this graph, I used nested stochastic block model (NSBM) [169]. This algorithm finds the modular substructure of the graph, while considering the edge types at the same time. I ran the algorithm 1000 times and selected the number of clusters and the partitioning that was most frequently supported. Figure 4.4A shows the histogram of the number of clusters automatically identified by the algorithm in each run (by maximizing the likelihood of the graph being generated from the partitioning). I found 18 clusters (corresponding to 5 drug clusters and 13 CCL clusters) to be the most frequently observed, which I selected as the final number of clusters. The Rand Index (RI) [170] between each run and the final partitioning showed a high degree of concordance with mean RI = 0.89 $\pm$ 0.01 (distribution shown in Figure 4.4B).

The final bipartite graph and the identified clusters are shown in Figure 4.4C. While drugs in all five identified drug clusters benefited from the bipartite graph (as is clear from the boxplots in Figure 4.4C, comparing SCC-LCO of BiG-DRP+ and an MLP with similar architecture), cluster C3 particularly had the highest median improvement (8.4% SCC improvement, one-sided Wilcoxon signed-rank, P = 5.25E-5). Interestingly, the drugs in this cluster had similar mechanisms of action: 13 out of 20 were protein kinase inhibitors, 8 of which target members of the serine/threonine protein kinase family; another 5 drugs target members of the tyrosine kinase family. From these observations, I conclude that some groups of drugs benefit more from the bipartite graph and its information sharing.

To better understand the characteristics of the CCL clusters, I performed hypergeometric tests (corrected for multiple tests using Benjamini-Hochberg FDR) to evaluate their enrichment in cancer types, tissues, and cancer driver mutations. The majority of clusters (9 out of 13) were enriched in at least one driver mutation. The analysis also revealed that two clusters were enriched in specific cancer types, namely cluster 1 in B-Lymphoblastic Leukemia and cluster 4 in Chronic Myelogenous Leukemia. Characteristics that were deemed significant (FDR $<$ 0.05) are listed in Supplementary Table B.4. These findings indicate that the bipartite graph's connectivity patterns extend beyond tissue or cancer types, capturing molecular-level patterns and possibly other biological intricacies.

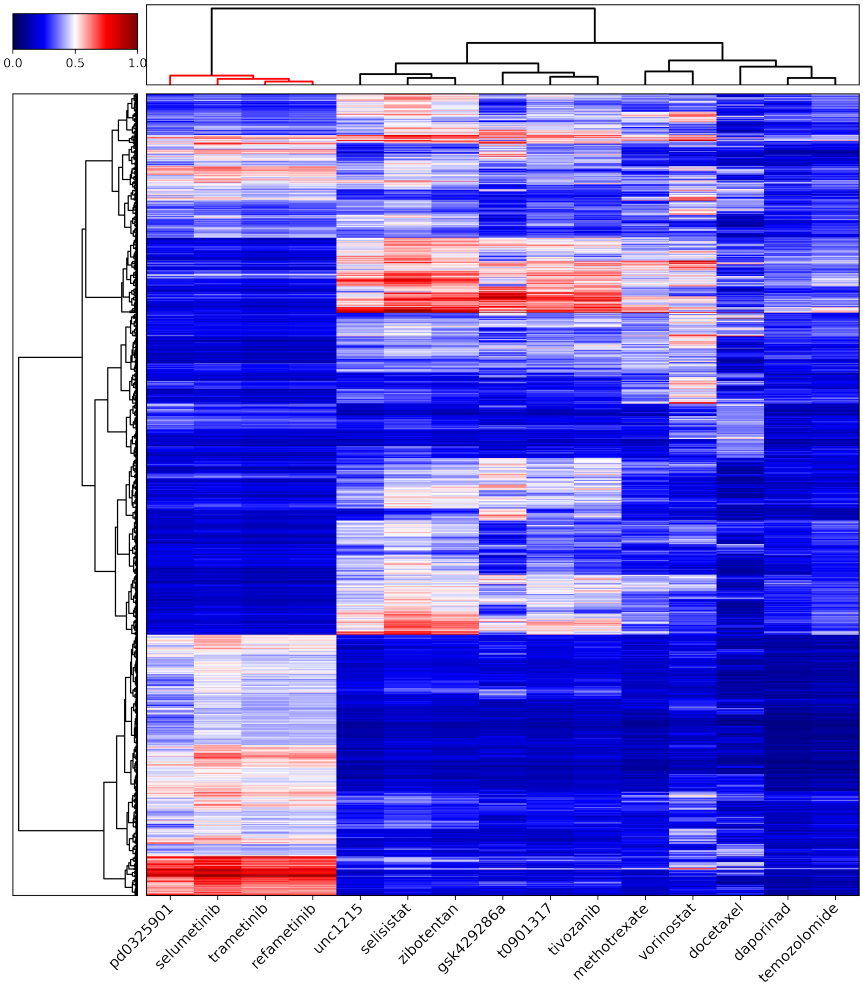

Figure 4.5: The hierarchical clustering of the 15 top-performing drugs for BiG-DRP+ in LCO-CV. The clustering is performed based on the contribution scores of the union of their top genes. The heatmap shows the contribution score of the genes for each drug.

4.3.5 Identification of Genes Associated with Drug Sensitivity

To identify genes whose expression has a considerable contribution to the predictive model, I utilized a similar pipeline as discussed in the previous chapter (Section 3.2.5-3.2.6) [12]. This method yields an aggregated contribution score for each gene used by the models as features, and then uses these scores to systematically identify the top-contributing genes in each drug. I directed my focus to 15 drugs for which BiG-DRP+ yielded the highest SCC values in the LCO-CV evaluation. The ranked list of genes implicated for each of the 15 drugs is provided in the Supplementary File B.1. I conducted clustering of the drugs based on the contribution scores of all implicated genes (Figure 4.5). Intriguingly, four

drugs formed a distinct cluster apart from the others: trametinib, refametinib, selumetinib, and pd0325901. Further examination revealed that these drugs are all MEK inhibitors (i.e., inhibit the mitogen-activated protein kinase kinase enzymes) and share some similar mechanisms of action [159].

I then focused on genes associated with trametinib, a drug for which BiG-DRP+ demonstrated the best performance (SCC in LCO-CV). Among these genes, ETV5 exhibited the highest prediction contribution. ETV5, along with ETV4 (the fourth highest contributor), belongs to the ETS family of oncogenic transcription factors. Upregulation of this family's expression has been observed in solid tumors, and they are known to play roles in tumor progression, metastasis, and chemoresistance [171]. ETV5 was also shown to be regulated by ALK, a receptor tyrosine kinase, in a MEK/ERK-dependent manner in neuroblastoma cell lines [172]. Additionally, it has been observed that trametinib treatment downregulates ETV5 in various CCLs [172–174]. Furthermore, the overexpression of ETV4 and ETV5 has been associated with decreased sensitivity of different CCLs to this drug [174].

I conducted pathway enrichment analysis to better understand the functional characteristics of the genes implicated for trametinib (see Supplementary File B.2 for results of pathway enrichment analysis of all 15 drugs). The results revealed the involvement of several important pathways, including MAPK signaling, EGFR signaling, and IL2 signaling (assessed via Fisher's exact test, FDR < 0.05). In summary, these findings suggest that the genes contributing to the predictive power of BiG-DRP+ for trametinib highlight key genes and signaling pathways involved in its mechanism of action.

4.3.6 Associating the Mutation Status of TCGA Tumors to their Drug Response

I then proceeded to examine the mutation landscape of tumors within the TCGA dataset and their associations with drug response as predicted by BiG-DRP+. For this analysis, I examined primary tumor samples that have both GEx profiles and mutation data from TCGA [35]. This corresponded to 9067 samples with 32 different cancer types. I generated a binary matrix indicating the mutation status of genes for each sample by parsing the Mutation Annotation Format (MAF) file provided in TCGA. Following Chiu et al. [92],

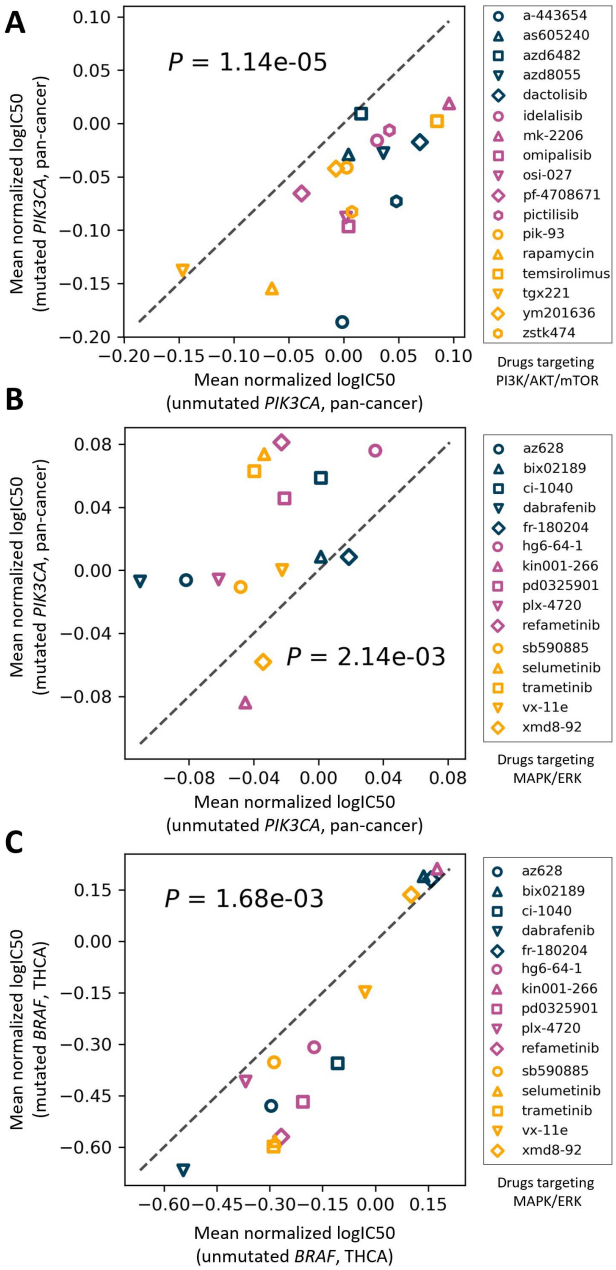

Figure 4.6: The association between mutation status and drug response predictions for TCGA tumors. The scatter plots compare the mean prediction (normalized log IC50) for tumors harboring a certain mutation against those of tumors without such mutation. A one-sided Wilcoxon signed-rank test was used to calculate the P value in each panel. A) Pan-cancer association between drug response and PIK3CA mutation for drugs targeting the PI3K/AKT/mTOR pathway. B) Pan-cancer association between drug response and PIK3CA mutation for drugs targeting the MAPK/ERK pathway. C) THCA-specific association between drug response and BRAF mutation for drugs targeting the MAPK/ERK pathway.

only four types of mutations were considered in this analysis: nonsense, missense, frameshift insertions, and frameshift deletions. Additionally, only mutations that exist in at least 10% of the tumors were selected.

Using the tumors' GEx profiles as input to BiG-DRP+, I predicted the response of the 9067 TCGA tumor samples for 237 drugs in the training set. I then conducted two-sided Mann-Whitney U tests to evaluate the relationship between mutations in the selected genes and drug response (i.e., if there is significant evidence to suggest that the predicted values for patients harboring a specific mutation are different from those without such mutation). The reported FDR values in this section were derived from these tests. Given its significance in determining drug response across various cancers and its potential as a therapeutic target, this section primarily focuses on insights based on the PIK3CA mutation [175] (statistical test results for all genes are available in Supplementary File B.3).

PIK3CA is an oncogene whose mutation triggers hyperactivation of the PI3K/AKT/mTOR pathway, which is linked to cancer progression and unfavorable outcomes across various cancer types [176–179]. A range of targeted therapies has been developed to inhibit this pathway in patients exhibiting deregulation and hyperactivity of PI3K/AKT/mTOR signaling, stemming from PIK3CA mutation or other mechanisms like PTEN loss or inactivation [180]. Furthermore, studies have demonstrated that a positive response to PI3K inhibitors is associated with mutations in this gene, both in vitro and in vivo [180, 181]. Consistent with these findings, the pan-cancer analyses revealed that tumors harboring such mutation exhibit significantly higher sensitivity to drugs targeting the PI3K/AKT/mTOR pathway (Figure 4.6A, one-sided Wilcoxon signed-rank test, P = 1.14E-5), such as the pan-AKT kinase inhibitor GSK690693 (FDR = 2.13E-59) and the pan-class I PI3K inhibitor ZSTK474 (FDR = 1.15E-29).

Conversely, mutation in this gene was associated with higher resistance to drugs targeting the MAPK/ERK signaling pathway. Specifically, drugs targeting this pathway have significantly higher predictions for PIK3CA-mutated tumors compared to tumors lacking this mutation (one-sided Wilcoxon signed-rank test, P = 2.14E-3, Figure 4.6B). Multiple studies, both in vivo and in vitro, have highlighted a regulatory connection between the MAPK/ERK and PI3K/AKT/mTOR pathways, with inhibition of MAPK/ERK signaling

associated with increased activity in the PI3K/AKT/mTOR pathway ([182] and references therein). Previous research has demonstrated that hyperactivity in the PI3K/AKT/mTOR pathway resulting from PIK3CA mutation contributes to drug resistance against dabrafenib and trametinib (drugs targeting the MAPK/ERK pathway), which supports these findings (dabrafenib FDR = 2.93E-18, trametinib FDR = 7.64E-9, Supplementary File B.3). Mutation in PIK3CA has also been demonstrated to bestow resistance to PD0325901 [183], a MEK inhibitor that reduces MAPK/ERK pathway activity. Moreover, genetic ablation of the mutant allele of this gene has been shown to increase sensitivity to PD0325901 in MEK inhibitor-resistant cells [183]. My analysis corroborates these findings, indicating that tumors harboring PIK3CA mutation exhibit increased resistance to this drug (FDR = 1.33E-5). Of the four drugs targeting IGF1R, three exhibited significantly higher predicted log IC50 values in PIK3CA-mutated tumors. Prior studies have established a connection between this protein and PIK3CA-driven ovarian cancer [184], as well as breast cancer tumors carrying mutations in this gene [185], suggesting the potential of dual inhibition of PI3K and IGF1R as a new therapeutic strategy. Cetuximab, an epidermal growth factor receptor inhibitor, is another notable observation from my analyses (FDR = 5.98E-3). Previous research has demonstrated an association between PI3K/AKT/mTOR pathway activity and resistance to this drug [186].

To evaluate the impact of mutations on drug resistance in a cancer-specific context, I focused on thyroid carcinoma (THCA), the most prevalent endocrine malignancy, as an illustrative example [187]. For this cancer type, only BRAF manifested mutations in over 10% of the samples, with a mutation frequency of 57.7% in tumors. The mutation primarily observed in this gene, notably the V600E mutation, activates the MAPK/ERK pathway, leading to sustained cell proliferation and adverse phenotypes [187]. Various studies have suggested this pathway as a therapeutic target and demonstrated that cancer cells (including thyroid cancers) carrying this mutation are substantially more sensitive to BRAF inhibitors (e.g., AZ628 [188]) and multiple MEK inhibitors [189]. Similarly, my analyses showed that THCA tumors with BRAF mutations display significantly higher sensitivity to drugs targeting the MAPK/ERK pathway (Figure 4.6C, one-sided Wilcoxon signed-rank test, P = 1.68E-3), including BRAF inhibitors like AZ628 (FDR = 2.25E-21) and HG6-64-1 (FDR

= 5.62E-12), as well as MEK inhibitors such as trametinib (FDR = 2.83E-26), refametinib (FDR = 1.69E-25), and selumetinib (FDR = 1.75E-25).

Collectively, these findings indicate the effectiveness of the proposed model in offering insights for pharmacogenomic investigations.

4.4 Discussion and Conclusion

In this chapter, I presented two novel deep learning methods based on graph neural networks to integrate information on CCL sensitivity/resistance, GEx profiles, and chemical characteristics of drugs in order to acquire more comprehensive drug representations. Through cross-validation and diverse data partitioning scenarios, I demonstrated significant improvements compared to conventional and state-of-the-art approaches. Leveraging a computational pipeline for neural network interpretability, I identified a subset of genes that substantially contribute to the predictive model. My analyses of these genes implicated important signaling pathways and alluded to both unique and shared mechanisms of action in the drugs. Additionally, I explored the connection between the mutations from TCGA cancer tumors and their predicted drug response, revealing various insights that were corroborated by independent research and thereby highlighting the utility of this approach in pharmacogenomics research.

Furthermore, a thorough assessment of the methods illustrated BiG-DRP(+)'s robustness towards variations in the drug response threshold (k) used for connecting the nodes in the bipartite graph. This further justifies the importance of the various techniques I implemented to ensure the stability of this proposed framework, namely the normalization factor and the injected self-loop in the H-GCN's forward pass. Specifically, the injected self-loops ensure that nodes retain a degree of their own information, thereby promoting a certain level of distinctiveness among the node embeddings. Additionally, the normalization factor prevents the received messages from becoming too large, thus maintaining some balance between messages and self-loop contributions. However, I acknowledge that this robustness may not be universally applicable to all scenarios. For instance, in a disconnected star subgraph where a drug is connected to CCLs that are not connected to other drugs, the second H-

GCN layer does not benefit the drug in terms of information propagation. Similarly, the second H-GCN layer will be obsolete if all the drugs happen to form disconnected stars, rendering the graph-level information sharing across drugs non-existent. Another example is when a new drug is introduced in the graph, resulting in a disconnected node that will fail to integrate CCL information into the drug embedding. This ultimately undermines the purpose of the H-GCN. To this end, I would like to note that the architecture that I proposed in this chapter is simply a framework with countless possible design choices, for which my choices were selected for simplicity of implementation. For example, although it would be interesting to use signed or weighted edges instead of using distinct sensitive/resistant edges, it would also raise more concerns such as the choice of using either edge features or simply using edge weights as scalers/multipliers.

As mentioned, many prior models (e.g., NRL2DRP [2]) can only handle prediction for CCLs and drugs that are already in the training set. Unlike these models, BiG-DRP was engineered to predict the response of unseen cell lines (those absent from the training data). However, since the drug embedding component of the model (the H-GCN) relies on the connectivity of the nodes, it also implies that drugs in the test set must exist in the bipartite graph provided during training. As a consequence, this model is generally unsuitable for predicting how CCLs will respond to newly introduced drugs. While one could hypothetically address this by assuming known connections involving the new drug node and some CCL nodes in the bipartite graph, this solution is impractical and difficult to implement without reducing the size of test set. However, in many practical scenarios, such as predicting drug response in cancer patients [11, 12], the focus should be directed to the model's ability to generalize to unseen samples (CCLs or patients). This is because extensive in vitro studies on CCLs are typically conducted before a new drug progresses to clinical trials or clinical use. Therefore, it is reasonable to anticipate having access to molecular features and drug response data for a drug, even when predicting responses in a newly acquired set of samples.

In this study, instead of directly utilizing the log IC50 values of drugs, I opted to normalize the log IC50 values of each drug separately across the CCLs. This approach served two main purposes: (1) to prevent any artificial inflation of prediction performance results, and (2) to facilitate comparability between the drug response ranges of different drugs, thereby enabling

the model to learn meaningful representations across drugs. However, it is important to note that this normalization procedure means that the predicted values should not be used to compare the potency of different drugs directly. Instead, the predictions should be regarded as a resistance score relative to other CCLs for a specific drug. Consequently, when presenting the performance results, I calculated the performance metrics for each drug individually across the CCLs. Should one wish to rescale to log IC50 values, these predictions can be easily adjusted to reverse the normalization, thereby allowing for the comparison of different drugs for the same CCL.

The pan-cancer and THCA-specific mutation analyses have shown some interesting biological relationships that were not explicitly represented in the bipartite graph. Although the models were trained using samples encompassing different cancer types, it would also be interesting to explore cancer-specific considerations. For example, one could finetune the models using only samples from a certain cancer type. This would allow the model to distinguish cancer-specific nuances, which could further refine not only the performance, but also the insights that can be extracted from analyzing the model. However, the success of such an approach depends on the availability and quality of the samples within the same cancer type.

One of the primary motivations behind this study was to enhance the representations of drugs for drug response prediction. While conventional approaches often rely on direct drug targets or chemical structure information, I posit that these representations can be refined by considering the effects of drugs on CCLs. This can be achieved either by assessing changes in the GEx profiles of CCLs following drug administration (e.g., LINCS dataset [190]) or by employing the bipartite graph formulation proposed in this study. Comprehensive drug representations are particularly important in tackling more complex tasks such as predicting responses to drug combinations, where the vast number of potential drug combinations implies that experimental measurements can only cover a fraction of these possibilities. Therefore, the development of more informative and robust drug representations becomes essential for creating models that can generalize well to drug combinations.

Chapter 5

Integration of Gene Essentiality and Drug Target Information in the Drug Response Prediction Model

Although several studies have shown gene expression (GEx) to be one of the most informative data modalities regarding drug response [1, 44], other data modalities, such as genomic or proteomic data, have also been shown to carry useful information when representing cancer cell lines (CCLs) [1, 29, 84]. However, less is known about the ability of other data modalities to improve drug representations.

Most studies on improving drug representations in the context of drug response prediction (or other tasks) have focused on obtaining drug embeddings irrespective of their relationship to CCLs to which such drugs are administered. While Morgan fingerprints, drug descriptors, and one-hot encoding of drug targets are popular choices [5, 10, 13, 54], some recent studies have obtained embeddings from the molecular structure of drugs using transformer-based models such as ChemBERTa [191] and SELFormer [192]. However, methods that can integrate different sources of information to capture the relationship between drugs and CCLs to which these drugs are administered are needed to provide complementary views and task-specific drug representations. For example, one can hypothesize that a CCL would be more responsive to a drug that targets a gene essential in that CCL. As a result, a model that can systematically incorporate such information, capturing CCL-drug-gene relationships in

the form of heterogeneous knowledge graphs, can be particularly useful, since it can reveal important cancer dependencies and improve model interpretability.

Given the success of BiG-DRP(+) [13] in incorporating CCL-drug information in the form of a heterogeneous bipartite graph, I set out to incorporate other network information in the drug response prediction task. For this purpose, I developed NECTARE (Knowledge Embedding of Compounds through Targets, Response, and Essentiality, pronounced "nectar"), which is an extension of BiG-DRP that integrates a multi-layer heterogeneous graph of genes, CCLs, and drugs to predict drug responses.

5.1 Problem Statement

Given a fixed set of drugs $D = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n\}$, training data composed of CCLs $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, and a possibly incomplete training drug response matrix $Y \in \mathbb{R}^{m \times n}$, the goal is to train a model $f(\boldsymbol{x}, \boldsymbol{d})$ that can predict the response of any CCL $\boldsymbol{x}$ for a drug $\boldsymbol{d} \in D$. This is a multitask problem in the sense that a single model is used to predict for various drugs. Each CCL is represented by a feature vector $\boldsymbol{x} \in \mathbb{R}^p$. Similarly, a drug is represented as a feature vector $\boldsymbol{d} \in \mathbb{R}^q$.

In this chapter, I extend this problem, as defined in Chapter 4, by exploring possible improvements in the model through a knowledge graph. Specifically, I would like to use CRISPR knockout screens (capturing gene essentiality) [34] and drug targets as additional elements to generate the knowledge graph, which would provide a more comprehensive view of the various relationships between CCLs, drugs, and genes. I hypothesize that this auxiliary information will lead to better node representations, which will be used in the drug response prediction problem and particularly in improving model interpretability.

5.2 Methods

In order to incorporate additional graph information in the BiG-DRP model [13], it was necessary to extend this model to support (1) the integration of multi-layer heterogeneous graphs containing more than two node types and (2) the inclusion of directed edges. The first requirement is needed since I am interested in adding gene nodes to the graph, resulting

in three node types, namely CCLs, drugs, and genes. Moreover, CCL-gene relationships (based on the essentiality of the gene in the CCL) and drug-gene relationships (capturing drug target information) needed to be incorporated into the graph. The second requirement was needed to add additional flexibility to the model, since not all relationships between two nodes are bi-directional and symmetric. For example, a gene can be highly essential to a CCL compared to other genes (CCL perspective), but the essentiality score of the CCL for that gene is not as notable as other CCLs (gene perspective), which may only merit an edge in one direction.

To achieve these goals, I developed a drug response prediction model that replaces the drug feature extractor module of BiG-DRP with a new module called NECTARE. The details of this model and NECTARE are provided in the following sections. In addition to this model, another major contribution of this chapter is proposing computational methods to interpret different aspects of the model, including the input features as well as the components of the heterogeneous knowledge graph.

The following terminologies are used throughout this chapter. First, "graph" is used in the context of the knowledge graph, which I used as an input structure to the NECTARE component. The term "network" is only used in the context of a neural network (to avoid confusion with the knowledge graph), which comprises the trainable components of the model. I use the term "embedding" only for node embeddings (i.e., graph-based representation), while "encoding" is only used for latent features extracted by fully connected neural networks. Finally, the term "attribute" is used to refer to features only when they are in the context of nodes of the graph.

5.2.1 Model Overview

The prediction model is composed of three components, which are depicted in Figure 5.1. First is the *CCL encoder*, whose input is simply the feature vector representing each CCL. In this chapter, I used the normalized GEx profiles as the cell line features. The CCL encoder is implemented as a simple 2-layer fully connected neural network that compresses the high-dimensional input (here of length ∼15.8-17.6k, see preprocessing) into a lower-dimensional CCL encoding (length=1024) to be used for prediction.

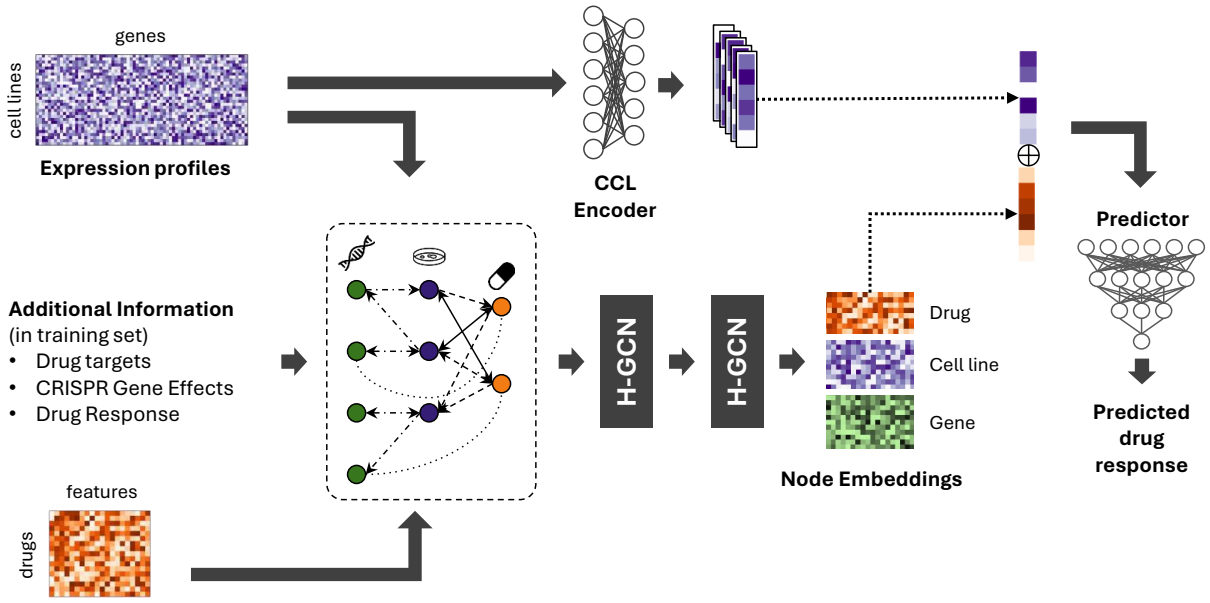

Figure 5.1: An overview of the drug response prediction model with NECTARE. The CCL encoder (top branch) is responsible for generating CCL representations. NECTARE (used in the bottom branch) generates node embeddings using two H-GCN layers and a knowledge graph that is inputted to these layers. This knowledge graph includes three types of nodes: CCLs, drugs, and genes. The directed edges between these nodes capture gene essentiality (using CRISPR knockout screening data), drug target information, and drug response information. The CCL representations and drug embeddings are concatenated and then used as input to a predictor neural network to predict the drug response.

The second component is NECTARE, which consists of two layers of Heterogeneous Graph Convolutional Networks (H-GCN). The H-GCN takes a knowledge graph as an input and outputs node embeddings (length=512) based on the topology and the node attributes in the graph. The knowledge graph constitutes three node types: drugs, CCLs, and genes. Edges between the nodes are categorized into response, essentiality, and target. Response edges exist between CCLs and drugs with extreme responses, indicating resistance or sensitivity to the drug. Essentiality edges exist between CCLs and genes that are highly essential to the cell line. Target edges connect drugs to their known target genes. Specific details about the construction of the knowledge graph are in the subsequent section. Each of the nodes are given unique attributes. Gene nodes use trainable embeddings (length=512) as their attributes, similar to how words are represented in natural language processing. These trainable embeddings are randomly initialized, and are trained alongside other parameters of the prediction model. CCL nodes use their normalized GEx as attributes. Drug nodes use their drug features/descriptors, which are characteristics that are calculated from their molecular structure (through their SMILES representation). The GEx (length $\sim$15.8-17.6k) and drug descriptors (length=197) were passed through a linear layer so that all nodes, regardless of the node type, have initial attributes of length 512. The idea is that through graph convolutions, NECTARE captures the drug's characteristics that are not easily identifiable from their provided molecular features. In addition to inducing similarity/contrast for drugs that exhibit similar/opposite effects on CCLs, the essentiality and target connections may also prove useful in the task. For example, if a gene is known to be a target of a drug, and this gene is essential to some cell lines, associating these types of relationships can be useful in the model's learning process as well as the analysis of the model for follow-up interpretation.

The final component of the model is the *predictor*, which uses the node embedding of a drug node and the CCL encoding to predict the CCL's response to the drug. Specifically, the drug node embedding and the CCL encoding are concatenated and fed to a 3-layer neural network that outputs a scalar drug response. Although CCL embeddings can be obtained from the H-GCN as well, I opted to use a separate CCL encoder instead to allow prediction on CCLs that are not in the training set. This is necessary for such a setup since CCLs

that are not present in the training set would also not exist in the knowledge graph, causing several issues in obtaining their embeddings. For example, although it is possible to connect such "new" CCL nodes to the graph for CCL nodes that have gene essentiality data, this approach would not be applicable to CCLs without gene essentiality information. Moreover, for such new CCL nodes, response edges are not available, which makes their embedding less informative compared to training CCLs. Additionally, the current model does not support a dynamic graph whose nodes and edges can change through time. However, in a matrix imputation application in which all CCLs and drugs that appear in the test set are also present in the training set, NECTARE's embeddings for both drugs and CCLs can be used by the predictor neural network.

5.2.2 Knowledge Graph Construction

I extended my previous model (BiG-DRP) by including more information in the graph. In addition to the response-based edges, I also incorporated CRISPR-based and drug target-based edges. I denote this extended knowledge graph as $G(V, E)$. Here, the set of vertices V is the union of V_C, V_D, and V_P, which are the set of cell line nodes, drug nodes, and gene nodes, respectively. Moreover, the set of edges E is the union of seven directed and undirected edge types, defined below.

- An undirected edge $(d, p) \in E_t$ exists between drug node d and gene/protein node p, if p is a known target of d.
- A directed edge $(c, d) \in E_{s,C\to D}$ exists from CCL node c to drug node d, if the response of c to d belongs to the lowest $k\%$ of the training set's responses to d (i.e., the CCL is among the most sensitive CCLs to the drug).
- A directed edge $(c, d) \in E_{r,C\to D}$ exists from CCL node c to drug node d, if the response of c to d belongs to the highest $k\%$ of the training set's responses to d. (i.e., the CCL is among the most resistant CCLs to the drug).
- A directed edge $(d, c) \in E_{s,D\to C}$ exists from drug node d to CCL node c, if the response of c to d belongs to the lowest $k\%$ of c's known drug responses (i.e., the drug is among the most effective drugs in killing the CCLs).

- A directed edge $(d, c) \in E_{r,D\to C}$ exists from drug node d to CCL node c, if the response of c to d belongs to the highest $k\%$ of c's known drug responses (i.e., the drug is among the least effective drugs in killing the CCLs).

- A directed edge $(c, p) \in E_{C\to P}$ exists from CCL node c to gene node p, if the CRISPR score of gene p in c belongs to the lowest $k\%$ of the CCLs in the training set (i.e., the CCL is among the CCLs for which the gene p is very essential).

- A directed edge $(p, c) \in E_{P\to C}$ exists from gene node p to CCL node c, if the CRISPR score of a *non-globally essential* gene p for c is less than or equal to -1.

In all analyses that will be presented, I set the value of $k = 1$. However, this is a hyperparameter that can be tuned (if desired) using a validation set. Moreover, one can choose to select different values of k for each edge type.

For simplicity, I will use the term "response edges" to refer to the union of edges $E_{s,C\to D} \cup E_{s,D\to C} \cup E_{r,C\to D} \cup E_{r,D\to C}$. I also use the term "essentiality edges" to refer to the union of edges $E_{C\to P} \cup E_{P\to C}$. Whenever a distinction is not needed, I will be using E and V to collectively refer to the set of all edges and nodes, respectively.

One issue with including gene essentiality in constructing the graph using any sort of measurement is that some genes are "globally essential," meaning these genes are essential in most (if not all) cells in the population. Without additional consideration, edges in $E_{C\to P}$ will consist only of edges between CCLs to these globally essential genes. In such a scenario, this type of connection would not contribute any useful information to the node embeddings. This is why in this study, I restricted the set of gene nodes (V_P) to be only genes that are either targets of a drug or are genes that are not globally essential but have at least one edge in $E_{P\to C}$.

5.2.3 Dataset Acquisition and Preprocessing

Drug responses of CCLs were obtained from the Cancer Therapeutics Response Portal V2 (CTRP) [22] in the form of the area under the curve (AUC), and Genomics of Drug Sensitivity in Cancer (GDSC) [36] in the form of log IC50. SMILES encoding of the drugs were either provided by the database (CTRP) or collected systematically (GDSC) using PubChemPy

[193]. I used RDKit [161] to generate drug descriptors (e.g., molecular weights, number of aromatic rings) from the SMILES representation. Note that since SMILES encoding is a simplified string representation of molecules, it sometimes fails to represent slight differences in similar compounds (e.g., isomers). This could result in having identical outputs from the preprocessing tools that I utilized. For example, the drug descriptors and Morgan fingerprints for epirubicin and doxorubicin calculated by RDKit from SMILES are identical. For these instances, I excluded the drug with less labeled data. For the rest of this study, I used 479 drugs for CTRP and 392 drugs for GDSC, represented by vectors of drug descriptors. Drug descriptors with missing values were dropped, while the remaining 197 features were z-score normalized.

For the CCLs in CTRP, I obtained the RNA-seq GEx profiles of the CCLs from DepMap (depmap.org), which were already processed using log_2(TPM+1). I excluded genes that showed small variability across the CCLs (genes with standard deviation < 0.1) as well as genes with missing values in some cell lines. Finally, only genes that were expressed (non-zero) for at least 10 percent of the CCLs were kept. This resulted in 814 CCLs represented by 17,630 genes. For the CCLs in GDSC, RNA-seq GEx profiles of CCLs were obtained from the Sanger Cell Model Passports [37] in FPKM, which was transformed using $log_2(\text{FPKM} + 1)$. I subjected the GDSC data to the same filtering process as previously described for CTRP, resulting to 972 CCLs represented by 15,869 genes.

All in all, this totals 310,729 labeled CCL-drug pairs for CTRP, and 325,705 labeled CCL-drug pairs for GDSC. Similar to the previous chapter's approach, the drug responses were z-score normalized. To reiterate, per-drug normalization of AUCs and log IC50s is necessary since drugs have their unique distributions of drug responses. Utilizing the un-normalized drug responses as labels for model training could negatively affect the model's performance by motivating it to focus on drug-specific biases instead of biologically relevant information to capture CCL-specific variations. I kept the summary statistics (mean and standard deviation) used in the normalization so that it would be possible to inverse the normalization to accurately represent the scale of the predictions in the original drug response space (i.e. AUC and log IC50).

Gene effect scores from CRISPR knockout screens (denoted as CRISPR scores hence-

forth) were also downloaded from DepMap. Only 584 of the labeled CCLs in CTRP have CRISPR scores. For GDSC, I mapped the CCLs using the metadata in DepMap and found that only 580 of the labeled GDSC CCLs have CRISPR scores. A list of genes that were identified to be common dependencies [34] across all CCLs (common essential genes) were also provided in DepMap.

Drug target information was provided in both CTRP and GDSC databases. GDSC genes that existed in the CRISPR screens were manually mapped from synonyms. Only 353 of the 392 drugs were associated with at least one of the 331 targets in GDSC. No further processing was done for CTRP, as the provided targets were already clean. Out of the 479 drugs in CTRP, 360 drugs had known targets spanning 322 genes. In addition to the given targets, I also obtained drug targets from STITCH [167]. A target was only included if the combined score was greater than 700.

5.2.4 Training and Evaluation

The model was trained end-to-end using the mean squared error (MSE) as the loss function. I used the Adam optimizer [162] to train the network parameters with a fixed learning rate of 0.0001. I also fixed the batch size to 128. I used Leaky ReLU for all nonlinearity functions. These values were deemed to be reliable in the previous Chapter, although further hyperparameter tuning can be executed. To choose the number of training epochs, the training set was further divided into training and validation sets. An instance of the model was trained and validated until maximum epochs or the early stopping criterion had been met. The epoch with the optimal loss (lowest MSE) was selected and used to re-train the model using the full training set.

Models were evaluated using 5-fold cross-validation. The dataset was divided using leave-pairs-out (LPO-CV) and leave-CCLs-out (LCO-CV) as described in Section 4.2.6. Note that since the drug response (labels) were z-score normalized per drug, I used the saved summary statistics to scale the predictions back to the original drug response space.

5.2.5 Identification of Predictive Genes from the CCL Encoder

To calculate the contribution score of each gene in a sample, I developed a variation of CXPlain [99] used in Sections 3.2.5 and 4.2.8 to overcome some of its shortcomings. The contribution scores were then aggregated across samples for a specific drug, using only the gene features passed to the CCL encoder. In Sections 3.2.5 and 4.2.8, the output of the model without a certain feature was estimated by replacing the feature value with zero. In the context of the current input data, zero-replacement can be interpreted as using the mean value of the GEx across the training set (post-normalization). However, for some genes, the feature distribution is skewed, and zero-replacement may not be appropriate. To address this issue, I borrowed a technique presented by SHAP [194] where they utilized "background" samples and marginalized across the predictions using these background data.

Let $f(\cdot)$ be the trained predictive model, and $L(\cdot)$ be the loss function. Let sample (CCL) u be the explicand represented by the features $\boldsymbol{x}_u = [x_u[1], \ldots, x_u[m]]$ with a scalar label y_u. I use the notation $\boldsymbol{x}_{u\setminus i} \in \mathbb{R}^{m-1}$ to represent the features of u except at position i. I then introduce an evaluation of the model as

$$f(\boldsymbol{x}_{u\setminus i}, x_u[i] = k) = f([x_u[1], \ldots, x_u[i-1], k, x_u[i+1], \ldots, x_u[m]]). \tag{5.1}$$

This means that f is evaluated on the modified features of u, where the value at position i is replaced with some value k. I then introduce a set of background samples denoted as B, represented as $\boldsymbol{x}_v = [x_v[1], \ldots, x_v[m]]$, $\forall v \in B$. The contribution of the gene at position i for the sample u, denoted by $\Delta\varepsilon_{u,i}$ is calculated by finding the difference:

$$\Delta\varepsilon_{u,i} = \bar{L}_{u,i} - \varepsilon_u. \tag{5.2}$$

Here, ε_u is the original loss for the sample. In this case, this is the squared error $\varepsilon_u = L(y_u, f(\boldsymbol{x}_u)) = (y_u - f(\boldsymbol{x}_u))^2$. The first term $\bar{L}_{u,i}$ is the loss obtained by modifying u. Here, I utilize the background samples to define this loss:

$$\bar{L}_{u,i} = \frac{1}{|B|} \sum_{v \in B} L(y_u, f(\boldsymbol{x}_{u\setminus i}, x_u[i] = x_v[i])). \tag{5.3}$$

This marginalization across background samples is the main distinction from the previous method (aka CXPlain). In fact, if B is only composed of a single sample represented by a vector of all zeros, this approach simplifies to that of CXPlain. The advantage of using background samples is that they reflect the distribution of the features for which the model is supposed to be applied, especially when the background samples are also the training set. However, the computational cost is multiplied by the number of background samples. I do note that this can be considered as the last (or first, depending on the view) step of calculating the Shapley value of each feature [195].

Additionally, I assume that the labels for the explicands are available. With the addition of this assumption, there is no need to train another black box model to function as an explainer. I also do not normalize the contributions as in Equation 3.3, because such normalization downplays the relative importance of the individual samples when aggregating across samples. For instance, scores of samples in the extremes of the label distribution are usually more meaningful than scores of samples in the middle of the pack. Instead, I consider $\Delta\varepsilon_{u,i}$ in Equation 5.2 as the sample-specific gene score for a drug, which I then averaged across samples to provide a drug-wide score.

5.2.6 Calculating the Importance of Nodes and Edges in the Knowledge Graph

Given the trained model, the next objective was to identify the importance of the nodes and edges in the knowledge graph. To simplify the task, I calculated the contribution scores only for a specific drug d and formed an "importance subgraph" for the drug. This subgraph contains nodes and edges of the knowledge graph that substantially contributed to the prediction of responses to drug d.

I do note that the following notations omit the input of the CCL encoder of my proposed model, since these contribution scores only pertain to the knowledge graph. For a graph $G(V, E)$ and drug d, let $f(G, d)$ be the trained predictive model whose output is a vector of predictions $\hat{\boldsymbol{y}} \in \mathbb{R}^N$, N being the number of test samples. Let E and V correspond to the collection of edges and vertices in G, respectively. Each edge is denoted by a triplet (t, u, v), where t is the edge type, while u and v are source and destination nodes, respectively. I

introduce the notation $H_e(G, t, u, v) = G(V, E \setminus \{(t, u, v)\})$ to indicate the exclusion of the edge (t, u, v) from the graph G.

Let $\varepsilon(G)$ be the error corresponding to $f(G, d)$. In this study, I will be using the MSE. I calculate the contribution of an edge (t, u, v) using the following equation:

$$\Delta_e(t, u, v) = \varepsilon(H_e(G, t, u, v)) - \varepsilon(G). \tag{5.4}$$

Similarly, I calculate the contribution of a node u by removing the node and all its associated edges from G. I denote this removal as $H_n(G, u) = G(V \setminus \{u\}, \{(t, a, b) \in E | a \neq u, b \neq u\})$. The contribution of node u is calculated according to

$$\Delta_n(u) = \varepsilon(H_n(G, u)) - \varepsilon(G). \tag{5.5}$$

When removing a node or an edge in the knowledge graph, it is important to consider the effects of the removal on the degree of the node, which in turn may affect the normalization factor in the H-GCN ($c_{u,v}$ in Equation 4.1). In this case, I used $c_{u,v} = \sqrt{|N(u, r)|}\sqrt{|N(v, r)|}$, which depends on the in-degree of u and the out-degree of v given an edge type r. Changes in the prediction caused by decreasing the node degrees are not meaningful in this context because they do not provide any insight regarding the information propagated by/to the node. I addressed this issue using a sink node trick detailed in Appendix C.1.

Only nodes that are within two hops to the corresponding node of drug d are included in the importance subgraph of d. Once the values of Δ_n and Δ_e are calculated for these nodes and edges, the scores are filtered to construct a meaningful importance subgraph. To find a set of nodes with relatively large node scores, I used the kneedle algorithm [101] to find a threshold by estimating the point of maximum curvature given an ordered list of node scores. I recognized that nodes with direct edges to d are more likely to have higher values for Δ_n. I then collected the set of edges whose destination nodes had a node score higher than the previously calculated threshold. Kneedle is then applied once more on the edges in this set to find an edge score threshold. The importance subgraph of the drug is then constructed using the edges that exceeded the edge score threshold, the nodes involved in these edges, as well as the direct edges to d from the nodes that were chosen using the node

Table 5.1: The naming convention used in this chapter.

Name	Response (CCL-drug)	Essentiality (CCL-gene)	Target (drug-gene)
RET (NECTARE)	✓	✓	✓
RE	✓	✓	
RT	✓		✓
R	✓		

score threshold. Therefore, a node is considered implicated for a drug if its node score is high or if it is a node that is connected by a highly scored edge.

5.3 Results

5.3.1 Including Gene Essentiality and Drug Targets Improve Performance

I set out to determine if adding information regarding essentiality of genes in CCLs and targets of drugs to the knowledge graph improves the performance. For this purpose, I trained different variations of the model, each using a different graph containing different edge types. To simplify the naming scheme, each variation is named with the initials of the edge types included: `R` for response, `E` for essentiality, and `T` for targets (Table 5.1). Here, the `R` model is similar to BiG-DRP [13] (excluded from tables to minimize visual clutter), but with asymmetric edges due to the imposed directionality during the graph construction. Note that for this analysis, it was not possible to impartially compare the performance of NECTARE (`RET`) with the variation in which the graph only included information regarding essentiality and targets, since some drug nodes would not have any connections in the latter graph. As the baseline, I included an MLP model that has a comparable architecture to that of the graph-based models, but without the graph convolutions.

Table 5.2 shows the performance of each variation of the model applied to CTRP and GDSC datasets in LCO-CV and LPO-CV framework. The RMSE is calculated for one drug at a time, and the mean and standard deviation across drugs are reported. These

Table 5.2: The test performance of different models on CTRP and GDSC in different cross-validation setups. The values in the table show the per-drug RMSE, with mean and standard deviation calculated across different drugs. The lowest mean RMSE values are shown in boldface.

	LCO-CV		LPO-CV	
	CTRP	GDSC	CTRP	GDSC
RET	**0.9181** (±**0.0767**)	**0.8556** (±**0.0595**)	**0.6935** (±**0.1317**)	0.6197 (±0.1025)
RE	0.9199 (±0.0764)	0.8579 (±0.0644)	0.6958 (±0.1317)	0.6188 (±0.1092)
RT	0.9185 (±0.0782)	0.8556 (±0.0633)	0.6949 (±0.1304)	**0.6176** (±**0.1107**)
R	0.9186 (±0.0764)	0.8569 (±0.0614)	0.6968 (±0.1312)	0.6180 (±0.1098)
MLP	1.5547 (±0.4921)	0.8649 (±0.0630)	1.1690 (±0.4066)	0.6498 (±0.1110)

Table 5.3: CTRP test set performance based on Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC). These metrics were calculated for each fold, when treated as the test set, and mean and standard deviation across folds are reported. These metrics are based on comparison of the unnormalized predictions and ground truth AUC values. The best performance is shown in boldface.

	LCO-CV		LPO-CV	
	PCC	SCC	PCC	SCC
RET	**0.8315** (±**0.0103**)	**0.7885** (±**0.0116**)	**0.9171** (±**0.0008**)	**0.8828** (±**0.0012**)
RE	0.8313 (±0.0107)	0.7881 (±0.0116)	0.9166 (±0.0009)	0.8822 (±0.0012)
RT	0.8313 (±0.0109)	0.7879 (±0.0119)	0.9167 (±0.0019)	0.8825 (±0.0027)
R	0.8311 (±0.0108)	0.7875 (±0.0120)	0.9163 (±0.0011)	0.8818 (±0.0015)
MLP	0.8244 (±0.0131)	0.7824 (±0.0131)	0.9032 (±0.0015)	0.8643 (±0.0020)

results show that using all edge types together has the highest performance in both datasets. An interesting observation is that even though including additional information improves performance, the biggest jump (compared to MLP) occurs when response edges are used in the model (Table 5.2). I would also note that as expected, there is no statistically significant difference between the performance of BiG-DRP and the `R` model (Wilcoxon signed rank test $P > 0.05$).

Since recent studies have shown AUC to be a more robust measure of drug response and better represents the effectiveness of a drug (compared to log IC50), I focused on CTRP, which uses this measurement for the rest of this study [196]. Figure 5.2 shows the MSE for

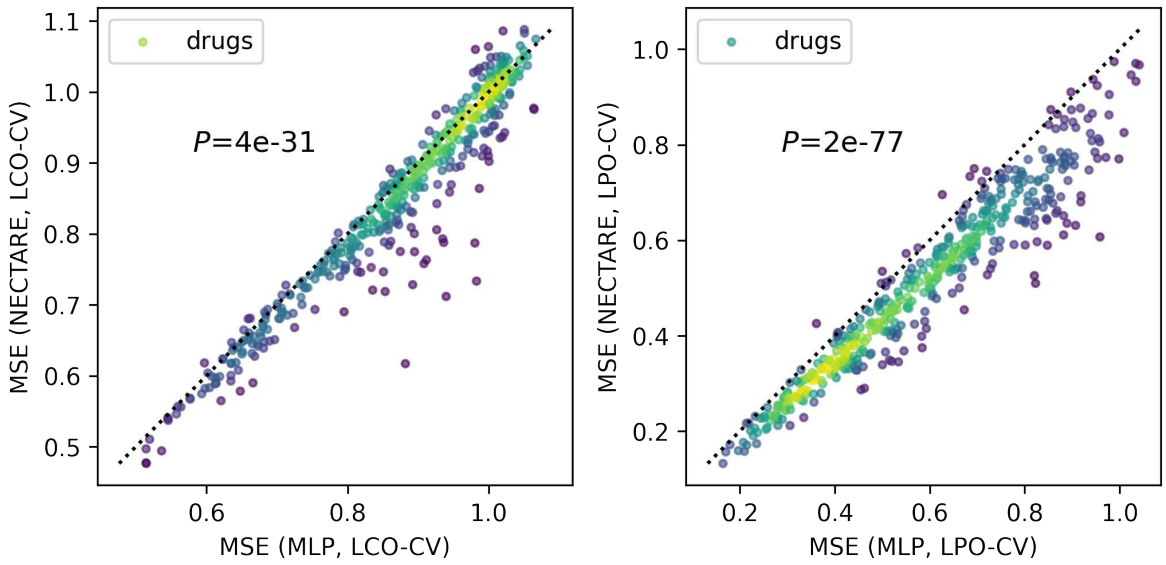

Figure 5.2: Comparison of NECTARE against MLP in CTRP. Each circle in this figure corresponds to a drug in the CTRP dataset. The y-axis corresponds to the test MSE of the drug using the predictive model that uses NECTARE (aka RET), and the x-axis shows the MSE using MLP. The left panel corresponds to LCO-CV, while the right panel corresponds to LPO-CV evaluation. In both cases, the MSE of NECTARE was significantly smaller than that of MLP (one-sided Wilcoxon signed-rank test P = 3.95E-31 for LCO-CV and P = 2.43E-77 for LPO-CV).

each drug in the CTRP dataset, comparing the full model with NECTARE against the MLP baseline. In addition to this visually apparent improvement, there is sufficient statistical evidence that such improvement (difference in MSE) is greater than zero in both LCO-CV (one-sided Wilcoxon signed-rank test P = 3.95E-31) and LPO-CV (one-sided Wilcoxon signed-rank test P = 2.43E-77). Table 5.3 shows the performance of different variations of the model on CTRP using two additional metrics: Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC), confirming that the NECTARE model performs better compared to alternatives.

5.3.2 The Effect of Different Sources of Drug Targets on Performance

In the results discussed in the previous section, I used the list of targets provided by the original database for each drug. Next, I asked how the performance changes if I instead

Table 5.4: CTRP test set performance using NECTARE with different sources of drug target information. These metrics were calculated for each fold, when treated as the test set, and mean and standard deviation across folds are reported. These metrics are based on comparison of the unnormalized predictions and ground truth AUC values. The best performance is shown in boldface.

Source of drug targets	LCO-CV			LPO-CV		
	PCC	SCC	RMSE	PCC	SCC	RMSE
CTRP	**0.8316 (±0.0103)**	0.7885 (±0.0116)	**1.4346 (±0.0395)**	**0.9171 (±0.0008)**	**0.8828 (±0.0012)**	**1.0296 (±0.0069)**
STITCH	0.8313 (±0.0092)	0.7885 (±0.0111)	1.4340 (±0.0353)	0.9168 (±0.0014)	0.8830 (±0.0021)	1.0327 (±0.0067)
Merged (CTRP ∪ STITCH)	**0.8316 (±0.0095)**	**0.7891 (±0.0111)**	**1.4346 (±0.0369)**	0.9170 (±0.0011)	0.8828 (±0.0020)	1.0314 (±0.0071)

use STITCH [167], a large database of interactions between chemicals and proteins, or if I augment the list of targets by combining the two sources. I trained another set of models using the STITCH [167] chemical-protein network information. I only used associations that have STITCH combined scores of at least 700 to ensure high confidence for the associations. The results are provided in Table 5.4. Using the merged drug targets from CTRP and STITCH yielded the best performance in the LCO setup, based on Spearman's rank correlation (SCC). In this setup, the performance based on the merged dataset was comparable with using only CTRP drug targets when evaluating the model using PCC and RMSE. In the LPO setup, however, the performance using the CTRP drug targets alone resulted in the best performance. These results suggest that the CTRP metadata provides a good source of information for the drug-target edge type, and increasing the size of the knowledge graph by incorporating STITCH targets (which also increases the computational cost of the model) does not provide enough benefit to be justified.

5.3.3 GEx-based Identification of Drug Sensitivity Biomarkers

In this section, I highlight some of the genes that the model utilized as indicators of drug response for specific drugs. For a given drug, each test sample receives a vector of scores corresponding to the relative contribution of individual genes, which are then aggregated

across test samples. Note that for this section, gene contributions only correspond to their effects in the CCL encoder. I used the CTRP dataset for this analysis and focused on drugs that had a drug-specific PCC larger than 0.5 in the LCO framework. Here, I describe some of the implicated genes and corresponding literature evidence supporting these observations.

For topotecan and etoposide, Schlafen 11 (SLFN11) was ranked as the top contributing gene. Topotecan is a topoisomerase I (TOP1) inhibitor, and etoposide is a topoisomerase II (TOP2) inhibitor. SLFN11 functions as a suppressor of DNA replication, triggering cell death in reaction to DNA damage. Its role involves eliminating cells with faulty replication, contributing to the maintenance of genomic integrity [197–199]. Zoppoli et al. [197] illustrated that SLFN11 imparts sensitivity to TOP1 and TOP2 inhibitors, which include topotecan and etoposide. SLFN11 knockdown was shown to suppress apoptosis while its overexpression induced programmed cell death for small cell lung cancer cells [200]. Additionally, SLFN13, which is an important paralog of SLFN11 and a novel tRNA/rRNA-targeting RNase with potent anti-HIV activity, was also implicated for topotecan in the analysis [201].

My analysis revealed that BCL2L1 was commonly implicated for Polo-like kinase 1 (PLK1) inhibitors (bi-2536, brd-k70511574, and gsk461364) and was one of the top-ranking genes. PLK1 is a proto-oncogene that is often found highly expressed in tumor cells, and its depletion is associated with inhibiting cell proliferation and inducing apoptosis [202]. On the other hand, BCL2L1 encodes a protein that belongs to the Bcl-2 protein family, which is known for regulating apoptosis at the mitochondrion. Bcl-2 member proteins either promote cell death (pro-apoptotic) or inhibit cell death (anti-apoptotic), which makes the family an important source of information, considering that apoptosis is generally recognized as the prominent mechanism for tumor suppression [203]. Additionally, Dinaciclib, a CDK inhibitor, also ranks the gene BCL2L1 on the top of the list, along with BCL2 Antagonist/Killer 1 (BAK1), which also encodes a protein that belongs to the Bcl-2 protein family. Silencing Bcl-xL, a protein encoded by BCL2L1, has been shown to dramatically increase cell death in low nanomolar concentrations of dinaciclib. Furthermore, dinaciclib was observed to trigger a reduction in mitochondrial membrane potential and induce a conformational change in BAX and BAK1, leading to the initiation of cytochrome c release and caspase

activation, ultimately leading to cell death [204].

Another interesting drug is the Nutlin-3, for which the top-ranked genes include several known associations with drug response. The proposed method has recovered MDM2, which CTRP annotated to be a known target of Nutlin-3. CKDKN1A, RPS27L, DDB2, BAX, SPATA18, EDA2R, ZMAT3, and AEN, all of which are p53 target genes were also ranked among the top indicators [205, 206]. Szwarc et al. [205] observed that Nutlin sensitivity is associated with high basal expression of p53 target genes. This aligns with my observation where the same genes are negatively correlated with Nutlin-3 response.

5.3.4 Knowledge Graph-based Interpretation of the Model

Next, I analyzed the contribution of the individual nodes and edges in the knowledge graph to the final model. The nodes and edges in the knowledge graph are used to obtain drug embeddings, which are inputted into the predictor. As a result, the knowledge graph (specifically, its nodes and edges) affects the response prediction of all CCLs to a drug.

Since I only used two layers of H-GCN, only the nodes whose messages could reach the drug of interest affect the model's prediction. I use the term "importance subgraph" to indicate the subgraph whose nodes and edges were implicated as important by the proposed scoring algorithm for a specific drug. Using the approach described in Methods, I obtained the importance subgraph for each drug of interest.

First, I demonstrate that the graph interpreter does not solely base its scores on the degree of the nodes. Figure 5.3 shows the relationship between the contribution score of a node in various drugs and the degree of the same node. As can be seen in this figure, a high node degree does not immediately translate to a high importance score. However, nodes with lower degrees have higher variance, indicating that the relevance of the connection and the quality of the information being propagated are prioritized by the model. I observed that for drugs with known targets, 338 (93%) of them implicated at least one of their targets in their respective importance subgraphs. Gene nodes that are known targets have a median node rank of 7.5 for their targeting drugs. Among drug nodes, drugs that share target genes tend to be implicated in each other's importance subgraphs. Around 90% of the drugs implicated at least one other drug node that shares its target. One example is the BRAF inhibitor

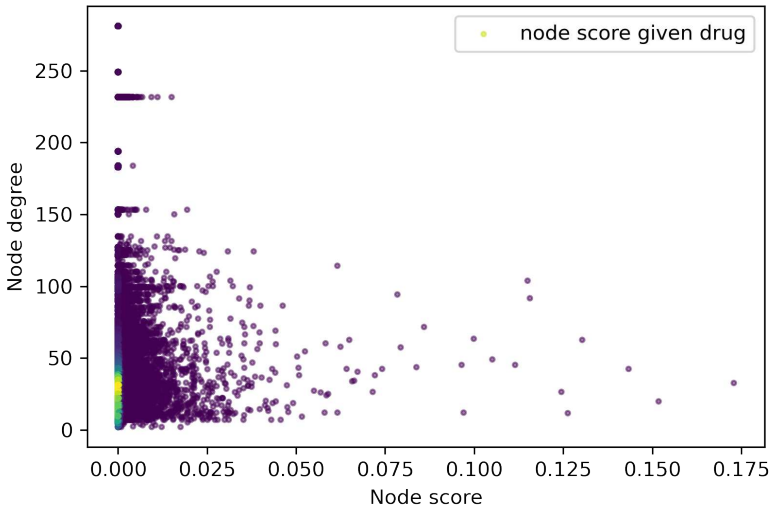

Figure 5.3: Comparison of node scores and node degrees. Each point represents a node's score for a specific drug. Only valid nodes (2-hop of the drug) are included. The color of the points represents the local density within an area, with yellow denoting high density.

dabrafenib (Figure 5.4), where the top five drug nodes implicated are also BRAF inhibitors (raf265, plx-4032, mln2480, plx-4720, regorafenib).

As another example, Figure 5.5 shows the importance subgraph of VX-680 (tozasertib), an Aurora kinase inhibitor that has been shown to be effective in inhibiting tumor growth in melanoma, leukemia, colon, and pancreatic tumors in xenograft models [207, 208]. While it was expected that AURKB would be implicated in the importance subgraph due to its annotation as VX-680's target, the implicated CCLs' cancer types match those of previous studies: leukemia (OCIAML5, SEM, AML193, HUNS1, SUPT1, REH), colon/colorectal cancer (SNUC4), pancreatic cancer (SNU410), and melanoma (HS936T). Other aurora kinase inhibitors, namely alistertib and barasertib, are also implicated. Interestingly, ruxolitinib, which targets JAK1 and JAK2, is also in the subgraph. VX-680 is also known to target JAK2, although this information was not provided in the dataset [209].

BCL2 was implicated in several drugs' importance subgraphs, including nilotinib and JQ-1. Nilotinib targets a fusion protein BCR-ABL, which is a persistently active tyrosine kinase that is responsible for sustaining proliferation, suppressing differentiation, and imparting resistance to apoptosis [210]. This occurs in patients with chronic myeloid leukemia (CML) [210, 211]. CML cells benefit from BCR-ABL through the upregulation of BCL2, MCL1, and

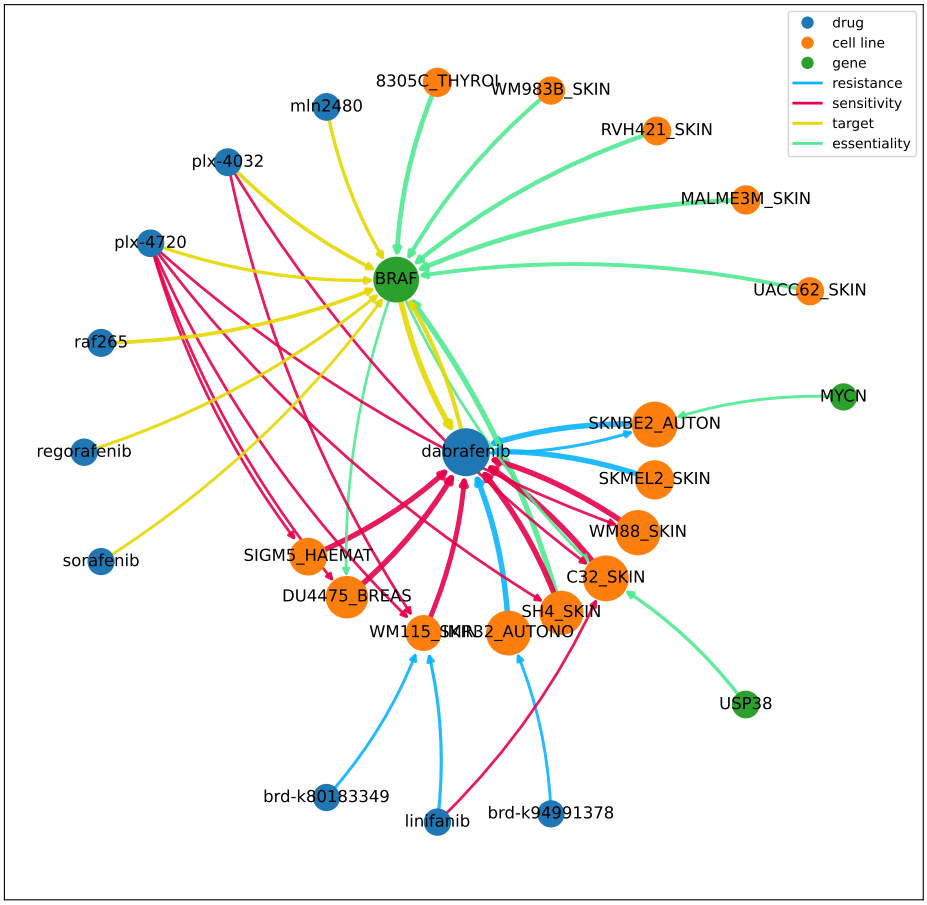

Figure 5.4: Importance subgraph of Dabrafenib. Node size and edge widths are proportional to the node and edge scores, respectively.

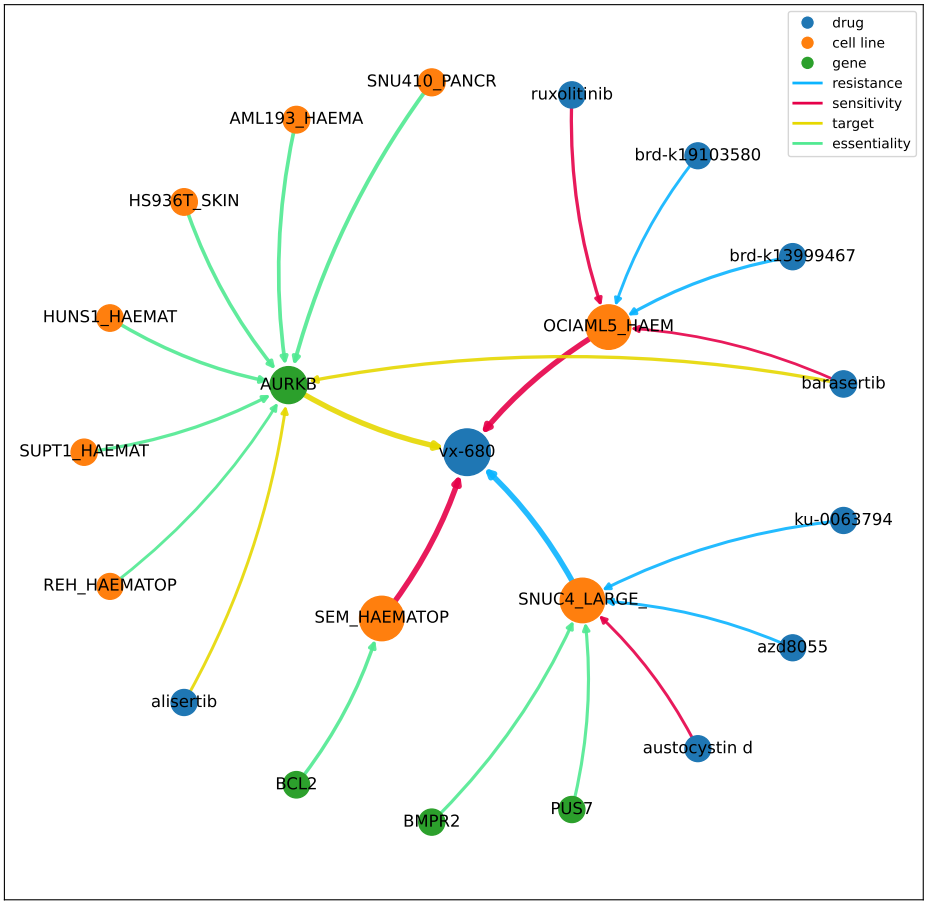

Figure 5.5: Importance subgraph of VX-680. Node size and edge widths are proportional to the node and edge scores, respectively.

BCL-xL (anti-apoptotic protein members of the Bcl-2 family) [212]. Additionally, downregulation of BCL2 was observed in response to tyrosine kinase inhibitors for vascular smooth muscle cells [213]. Parry et al. [212] proposed that inhibiting the anti-apoptotic Bcl-2 family proteins in addition to BCR-Abl inhibition could be a promising treatment for patients with blast phase CML.

Similarly, a combination treatment using JQ-1 and a BCL2 inhibitor (ABT-236) was found to be promising for small cell lung cancer (SCLC) [214]. This was hypothesized due to the common occurrence of BCL2 protein overexpression and MYCN family gene amplification in SCLC. JQ-1 is a BET inhibitor that has been shown to inhibit N-Myc, resulting in the expression of Bim. This then sensitizes MYCN-amplified SCLC cells to ABT-236 [214]. Zhang et al. [215] have also demonstrated that pro-survival genes (BCL2, Cyclin D1, and MYC) are being downregulated by BET inhibition through JQ-1 treatment on malignant transformation induced by 12-O-tetradecanoylphorbol-13-acetate (TPA) in mouse skin epidermal JB6 P+ cells.

5.4 Discussion and Future Work

In this chapter, I proposed an approach called NECTARE to incorporate a knowledge graph in the task of drug response prediction. NECTARE combines drug target information, gene essentiality, and known extreme drug responses to improve drug representation. Using two databases and two data-splitting techniques, I have shown improvement over baselines. Additionally, I interpreted the model on two fronts: (1) the CCL encoder and (2) the knowledge graph. For the CCL encoder, I proposed a modification of the previously utilized method called CXPlain [99] by taking into account the distribution of the features in the dataset. I then proposed to score the nodes and edges in the knowledge graph using an "explaining by removing" technique reminiscent of multiple previously proposed black box explainers [216]. From the node and edge scores, I reconstructed a drug's importance subgraph that visually represents the portion of the knowledge graph that significantly influenced the drug embedding.

Although on average, the improvement of the full NECTARE model is small, this im-

provement is consistent across most drugs in both GDSC and CTRP datasets. This may indicate that some information is gained from the combination of the additional edges. However, it is also possible that most of the explicitly represented biological priors through the target and essentiality edges were already redundant to those of the response edges. Nevertheless, the additional information enables the alignment of the model to the existing body of knowledge, which also increases the model's value in hypothesis generation, as well as its interpretability. I also note that the quality of the knowledge graph plays a significant role both in the performance of the model and in the post-hoc analysis. In this model, the gene nodes were given trainable embeddings as node attributes, which are initialized randomly. This implies that learned embeddings after training may not have any relationship to the gene that the node is supposed to represent, but is only an amalgamation of the information propagated from its neighbors. Careful consideration of alternative node attributes for gene embeddings is an avenue for improvement in the future.

As mentioned earlier, the model has been trained in a static knowledge graph. Although it is computationally feasible to predict for a new or updated graph, the model does not have the capability to generalize in such cases. Allowing updates in the graph enables us to predict for new drugs and possibly utilize the node embeddings of the CCLs, dropping the external CCL encoder. This will open up new avenues in interpretability because the importance subgraph will depend on a pair of nodes (drug and CCL). In the current state of the model, some possibly related genes will not be implicated in a drug's importance subgraph because of their distance on the knowledge graph. Although one could potentially increase the number of H-GCN layers to expand the information propagation range, this comes with the risk of network smoothing, in which repeatedly aggregating neighborhood data leads to less information retained. It would also be interesting to expand the scope of the knowledge graph, which may include gene-gene, gene-disease, and mutation associations [2, 217]. Incorporating more data modalities such as proteomics, epigenomics, gene ontologies, and even text-mining data would help contextualize the nodes in alternative views, potentially enabling the model to consider possible confounding factors that could have been neglected. However, it is important to recognize that as the amount of data modalities increases, additional challenges, such as data imbalance in certain cancers or drugs, will be difficult to avoid and should

be addressed. On a positive note, due to the specific architecture that I proposed (i.e., only drug embeddings of a fixed set of drugs are extracted from the knowledge graph), only the training speed is heavily affected by the size of the graph. During inference, the computational time required for calculating drug representation from larger graphs can be offset by pre-calculating the drug embeddings, making larger graphs still feasible in real-time use. Finally, as more information is being embedded through the knowledge graph, these embeddings may be applied to other tasks such as drug combination recommendation and synergy prediction.

Chapter 6

Discussion, Future Works, and Conclusion

6.1 Discussion

The overarching goal of this thesis is to develop and apply deep learning methodologies to the drug response prediction problem. I have considered problem settings ranging from predicting in vitro drug responses to generalizing to patient tumors. This section provides a high-level discussion of the results and ties together the different chapters of this thesis.

Chapter 3 presented a preclinical-to-clinical (P2C) pipeline for predicting drug responses and identifying biomarkers called TINDL. In this study, I was specifically interested in the P2C paradigm because of the scarcity of available CDR data. In the P2C paradigm, predictive models can only be trained using preclinical labels, although *unlabeled* clinical samples can be utilized. I emphasized that models that were trained using cancer cell line (CCL) data generally do not translate well to predicting drug responses from patient tumor data without special consideration of the differences between CCLs and tumors. My proposed method addressed this issue through a technique called tissue-informed normalization, which utilizes the statistics of unlabeled patient tumor samples to adjust the distribution of the GEx profiles of the test patient samples. The results showed that TINDL differentiates between resistant and sensitive tumors for 10 out of 14 drugs, outperforming other models.

I argued that the P2C setup reasonably mimics practical scenarios because in vitro screen-

ing is relatively easier to collect than clinical samples. However, in the future, it may be possible to include clinical labels in the training set. As seen in the clinical sample counts in Table 3.1, the sample sizes for many drugs are less than 100. Further splitting the data to include labeled clinical samples in the training set will render some of these drugs unfit for statistical tests. Performance metrics that were calculated from an even smaller test set reduce the credibility of such measurements. This is further exacerbated by the imbalance in the number of sensitive and resistant samples.

The analysis in Chapter 3 also included comparisons against methods that used domain adaptation or batch effect removal. Although my assessments have shown the superiority of TINDL, it would also be interesting to know what the other methods "did right" and what could have been done to improve them. This goes back to the discussion about how difficult it is to evaluate whether a sample has been sufficiently adapted. I also want to acknowledge that it might even be impossible to completely remove these domain discrepancies as differences between CCLs and tumors go beyond the statistical properties of their gene expressions. Nevertheless, results suggest that there is merit in including domain knowledge, particularly tissues of origin, in predictive models for drug response.

I then shifted the focus to preclinical drug response prediction in Chapter 4, where I introduced BiG-DRP, a model based on graph representation learning. BiG-DRP is designed somewhat like a multitask model, where a single shared model is trained to predict drug responses across a wide range of drugs. This is a deviation from the single model per drug approach that I applied for TINDL. The advantage of training a shared model is that they implicitly learn drug similarities during training. Additionally, this eliminates the need to train hundreds of models.

BiG-DRP was born out of the idea that highly sensitive and highly resistant samples for a specific drug have some characteristics that could be leveraged to "describe" the drug. An analogy of this is when a person is asked which type of cuisine they like the most and the least; their answers allow us to interpolate their preferences in flavor profiles. In the case of CCLs, it is not trivial to pinpoint the properties of the highly sensitive/resistant CCLs that could be incorporated to improve the representation of the drug. However, pointing to the general direction of such knowledge is already informative, which materialized in the

form of a bipartite graph. BiG-DRP utilizes a heterogeneous graph convolutional network to propagate data in the said graph. This not only extracts characteristics from the connected CCLs, but also integrates information regarding drug similarities.

In my evaluations, I introduced two data-splitting scenarios: leave-CCLs-out (LCO) and leave-pairs-out (LPO). The former evaluates the model's ability to generalize to unseen CCLs, while the latter evaluates the model's ability to fill in unknown values in a drug response matrix (CCLs $\times$ drugs). Results have demonstrated the superiority of BiG-DRP(+) against multiple baselines and other approaches in these two scenarios. Despite this, my model is limited to the drugs that are already in the training set, a limitation that I have already discussed in the corresponding chapter. However, I would like to relate the topic of predicting for new drugs to the analysis regarding drug features. My assessment showed that the model's performance when utilizing drug descriptors is not much different when replaced with the drug's Morgan fingerprints. This raises the question of whether either of them is actually meaningful in the context of drug response prediction. It is quite possible that the model is just utilizing these features as unique identifiers for the drugs. If this is the case, then BiG-DRP's drug representation via information propagation becomes even more relevant. To this end, it would also be worthwhile to explore different data availability scenarios. For example, one could use an auxiliary dataset (say, use CCLE [38] in addition to GDSC [36]) in training the model, where the graph connectivity of the unseen drug can be derived from the auxiliary dataset.

I continued the pursuit for better drug representation in Chapter 5, where I proposed NECTARE, a drug representation component that is based on a knowledge graph. In this chapter, I was interested in exploring different data modalities as sources of information for drug response prediction. I was especially interested in CRISPR gene effects and drug targets as CCL-gene and drug-gene relationships, respectively, in addition to the response-based connectivity from BiG-DRP. This came from the idea that if a drug targets a gene, and that gene is essential to a CCL, then I could hypothesize that the CCL is likely sensitive to the drug.

The ablation studies for the different edge types used by NECTARE have shown that there is merit in using essentiality (CCL-gene) and target (drug-gene) edges. However, the

improvements are only slight increments in performance compared to using the response (CCL-drug) edges alone. This shows that prior knowledge is an important resource in developing predictive models. Nonetheless, knowing how to incorporate prior information into the model constitutes a significant fraction of the design process.

These drug response prediction models are inherently *black boxes*. Special considerations must be made in order to gain insights regarding how the model predicts. As such, I have introduced an interpretability pipeline in Chapter 3, based on CXPlain [99], which assigns contribution scores for each of the input genes and identifies the subset that had a substantial contribution. This has served as an essential tool in generating hypotheses regarding potential biomarkers of drug response. The gene set characterization analyses in Chapters 3 and 4 have implicated important pathways related to various drugs' mechanisms of action. Furthermore, many of the identified genes were corroborated by literature or by experimental validation [12].

In Chapter 5, I improved upon the previous gene scoring methodology by taking into account the GEx distribution instead of approximating feature removal by zero-replacement. Although the argument for using the distribution seems promising, it is quite difficult to assess the amount of improvement in terms of the actual relevance of the features. One idea is to use synthetic datasets in which important features are pre-determined. I have performed such analyses outside of this thesis, but the formulation of such analysis is ridden with so many philosophical questions. For example, in addition, $a + b$, where a and b are features, is the addend with higher magnitude more important, or are they equally important? For practicality, many methods would just set assumptions based on their applications [194, 218]. However, setting appropriate assumptions is already tricky, especially for feature sets with some form of dependency, such as GEx profiles.

Until now, it is still unclear as to how to appropriately portray model interpretations in non-visual applications. There is no doubt that feature importance scoring has been helpful, but this type of explanation is not always intuitive. To this end, I proposed another view of interpretability in the form of importance subgraph, albeit the approach is tied to the method NECTARE. Results have shown some interesting relationships in the knowledge graph, which I have discussed in the corresponding chapter.

However, although importance subgraphs show relationships that are relevant to the drug, they do not directly show *why* some nodes and edges are more relevant. I also relate this to the clustering analysis in Chapter 4, where I clustered the bipartite graph, revealing multiple communities of CCLs and drugs that exhibit shared properties like the drugs' mechanism of action or the CCLs' cancer driver mutations and tissues of origin. This just demonstrates the amount of information that is aggregated by my proposed methods, NECTARE and BiG-DRP(+), simply by "pointing to the general direction". However, it would be interesting to delve into which characteristics of the nodes the H-GCN is actually propagating.

6.2 Future Directions

6.2.1 Utilization of Single-Cell Data

High throughput sequencing technologies combined with extensive drug screening studies have provided an abundance of data to study the response of cancer cells to different therapies. This thesis, for example, has taken advantage of this influx of data to develop predictive models of drug response. However, for years, researchers have focused on utilizing *bulk* gene expression data due to its wide availability and because it is perceived as one of the most informative data modalities for drug response prediction.

Single-cell (SC) sequencing, particularly scRNA-seq, can provide molecular profiles of cancer samples at the SC resolution, offering in-depth biological information and insights. As scRNA-seq is becoming increasingly accessible, one promising future direction would be the prediction of drug response at the single-cell level [219, 220]. Instead of predicting the response of a sample as a whole, the inference of drug response can be performed per cell. However, a significant challenge in training such a model lies in the availability of drug response data at the SC level. Alternatively, the problem can be re-framed as an application of multiple instance learning, where labels are given for a *bag of samples* instead of one-to-one correspondence [221, 222].

An exciting application of SC-level prediction goes back to the CDR prediction problem for patient tumors. Tumors have microenvironments composed of various cells, in addition to the actual cancer cells. Having access to the SC-level information allows researchers more

freedom in addressing this tumor heterogeneity. Additionally, SC-level prediction opens up a possibility for overcoming drug resistance by recommending combinatorial drug treatments that are appropriate for different cancer subpopulations.

6.2.2 Polytherapy Response and Drug Synergy

As mentioned in Chapter 3, the clinical drug response data provided in the TCGA database include patients who were treated with multiple drugs either simultaneously or sequentially. Thus, a model that can predict "overall" responses to polytherapy is of utmost interest.

Since interactions between drugs that were administered simultaneously are expected, it is crucial to predict how well these drugs interact with each other. Drug synergy is the level of interaction between the drugs that contributed to the sensitivity that cannot be attributed as an effect of a single drug in the combination [223]. Databases like DrugComb [157] and DrugCombDB [158] have already curated CCL drug synergy data from multiple studies. However, these databases are far from complete since the search space for discovering synergism is already quadratic for a combination of two drugs.

Revisiting BiG-DRP (Chapter 4) and NECTARE (Chapter 5), it appears that the models that I proposed can be extended for synergy prediction. Furthermore, it would be interesting to incorporate relationships such as synergism and antagonism in the knowledge graph.

6.2.3 Model Interpretability

This thesis has repeatedly utilized explainers to shed some light on the models' prediction process. Although valuable insights were collected from this process, there is still a need for more interpretable approaches. Some approaches have introduced domain knowledge, such as pathway information and gene ontologies, to their model architectures in an attempt to induce some level of interpretability [5, 9, 57, 58, 76]. However, in a previous study that I helped carry out [78], we observed that randomly generated (non-meaningful) gene sets that serve as pseudo-pathways in many of these methods provide comparable performance to using actual biological pathways, implying that the gene sets are acting as random regularizers. This suggests that the incorporated domain knowledge's actual function in the model

does not always reflect the expectations, which compromises the method's interpretability. Therefore, there is still a lot of space for improvement in this area of study.

In Chapter 5, I introduced a different way of interpretation in the form of the importance subgraph. Recent neural network architectures, such as transformers, have the potential to improve graph-level interpretation through the attention mechanism. Additionally, instead of importance scores, model explanations can also be represented in other forms, such as counterfactuals, where the explanation for a prediction would specify the changes to the input that would result in a different output [224, 225].

6.3 Conclusion

In this thesis, I introduced multiple methods for predicting the response of cancer samples to drug treatments. I was primarily interested in developing models that leverage the currently available datasets while addressing various challenges, such as the scarcity of clinical drug response data and drug-specific biases. I looked through different lenses and saw the potential of incorporating high-level information like tissue of origin and knowledge graphs that describe various relationships involving drugs, genes, and samples.

In Chapter 3, I focused on clinical drug response prediction, a task that was rendered more challenging by the scarcity of labeled data. Using TINDL, I was able to distinguish sensitive and resistant patients using a model that was only trained on preclinical samples by leveraging tissue information. Chapter 4 introduced BiG-DRP, a model that utilizes a bipartite graph and heterogeneous graph convolutional networks to incorporate genetic information from highly sensitive and resistant cell lines into the drug embeddings. Finally, in Chapter 5, I presented NECTARE, which extends upon the previous model by also integrating CRISPR gene effects and drug target information, thereby transforming the graph into a more detailed knowledge graph. My analyses demonstrated that the information propagated through the graphs enhanced the predictive performance of the models using various evaluation setups, surpassing multiple prior and baseline methods in drug response prediction.

In addition to the prediction of drug responses, I also presented approaches to identifying

biomarkers of drug response as a way to infuse a degree of interpretability into the models. Many genes and pathways that were implicated in the proposed pipelines were validated experimentally or have shown associations in the existing literature. I also introduced a graph-based approach in model interpretation based on the knowledge graph. The combination of the predictive models and interpretability pipelines not only improves trust in the model, but also generates new hypotheses, which could prove useful in pharmacogenomics research.

This thesis adds to the growing body of literature on precision medicine. As more sophisticated techniques in machine learning and biotechnology emerge, the ultimate goal of personalized medicine comes closer to reality. However, as we navigate towards clinical applications, some issues, such as ethical considerations regarding data privacy and equity in access to personalized treatments, will come into question. With this, there is much space for innovation, some of which may require collaborative efforts not only in technical fields but also in social sciences.

Appendix A

A.1 Supplementary Tables for Chapter 3

Table A.1: Information regarding the unlabeled TCGA samples used as auxiliary data.

Drug	Tissues/Cancer types	Number of unlabeled samples
Bleomycin	Cervix, Testis	321
Cisplatin	Bladder, Cervix, Esophagus, Head and Neck, Liver, Lung, Soft Tissue, Stomach, Testis, Uterus	3170
Cyclophosphamide	Breast	805
Docetaxel	Bladder, Breast, Head and Neck, Lung, Prostate, Soft Tissue, Stomach, Uterus	3824
Doxorubicin	Bladder, Breast, Cervix, Lung, Soft Tissue, Stomach, Thyroid, Uterus	3670
Etoposide	Bladder, Brain, Lung, Stomach, Testis	2116
Gemcitabine	Bladder, Cervix, Esophagus, Liver, Lung, Pancreas, Soft Tissue, Uterus	2607
Irinotecan	Colon, Brain, Pancreas	1173
Oxaliplatin	Colon, Pancreas, Stomach	1018
Paclitaxel	Bladder, Breast, Cervix, Esophagus, Head and Neck, Lung, Ovary, Pancreas, Stomach, Uterus	3984
Pemetrexed	Lung	867
Tamoxifen	Breast, Soft Tissue	1073
Temozolomide	Brain	459
Vinorelbine	Breast, Lung	1754

Table A.2: The P values of the one-sided Mann-Whitney U test comparing the distribution of predictions by various approaches for sensitive and resistant patients.

Drug	TINDL	LASSO	TG-LASSO [11]	SVR	Geeleher et al. [24]	Random forest	ComBat-DL
Bleomycin	3.41E-02	2.97E-02	9.39E-02	4.43E-02	4.43E-02	2.55E-01	1.09E-02
Cisplatin	6.36E-04	2.44E-04	7.92E-04	3.38E-04	7.94E-01	4.25E-02	1.47E-04
Cyclophosphamide	5.60E-02	1.68E-01	1.88E-01	1.13E-01	8.93E-01	4.72E-01	6.15E-02
Docetaxel	7.04E-01	9.91E-01	9.43E-01	6.27E-01	7.77E-01	9.97E-01	7.70E-01
Doxorubicin	1.42E-02	5.22E-01	9.25E-01	5.90E-02	1.35E-01	1.68E-02	6.63E-02
Etoposide	4.00E-03	8.50E-03	9.81E-03	5.66E-03	3.89E-02	3.67E-02	1.92E-03
Gemcitabine	4.57E-02	4.32E-01	1.77E-01	1.37E-02	7.77E-01	2.60E-01	3.84E-02
Irinotecan	3.04E-01	2.81E-01	2.36E-01	2.58E-01	2.16E-01	3.04E-01	3.29E-01
Oxaliplatin	2.41E-02	2.74E-02	1.41E-02	4.43E-02	7.84E-03	6.71E-03	2.98E-02
Paclitaxel	2.29E-02	5.12E-01	4.44E-01	1.35E-01	6.93E-03	6.36E-02	8.20E-02
Pemetrexed	2.86E-01	7.87E-01	7.33E-01	7.70E-01	2.96E-01	4.71E-01	1.96E-01
Tamoxifen	1.14E-03	3.22E-03	1.14E-03	1.37E-01	4.84E-01	2.22E-01	5.86E-03
Temozolomide	2.94E-02	3.71E-02	4.53E-02	8.77E-02	1.44E-01	6.25E-01	1.03E-01
Vinorelbine	2.41E-02	5.84E-03	4.96E-03	2.82E-01	4.43E-01	5.00E-01	7.99E-03

Drug	ADDA-DL	DANN-DL	TrainNorm-DL	TestNorm-DL	GAT	GCN	LSTM
Bleomycin	4.43E-02	5.68E-02	2.74E-01	1.04E-01	2.07E-02	2.07E-02	3.32E-01
Cisplatin	9.79E-05	1.14E-03	1.59E-03	5.62E-04	4.82E-06	1.71E-05	4.71E-03
Cyclophosphamide	4.04E-02	4.18E-02	1.88E-01	5.09E-02	2.66E-02	6.15E-02	6.54E-02
Docetaxel	9.73E-01	9.81E-01	9.13E-01	6.94E-01	5.17E-01	8.79E-01	6.92E-01
Doxorubicin	8.82E-01	2.86E-01	7.60E-01	4.03E-02	1.62E-01	2.85E-02	3.22E-01
Etoposide	2.12E-02	1.43E-02	9.59E-04	1.69E-03	2.44E-04	3.70E-03	2.73E-02
Gemcitabine	8.12E-02	2.29E-02	6.17E-02	4.37E-02	7.19E-03	4.95E-02	1.29E-01
Irinotecan	3.80E-01	4.32E-01	6.20E-01	3.29E-01	1.96E-01	6.96E-01	6.20E-01
Oxaliplatin	2.21E-02	1.55E-02	2.41E-02	2.98E-02	5.15E-02	2.19E-01	3.65E-02
Paclitaxel	1.03E-01	1.66E-01	3.98E-02	2.89E-02	3.07E-02	4.75E-02	8.91E-02
Pemetrexed	3.06E-01	3.92E-01	1.59E-01	2.86E-01	4.03E-01	2.77E-01	2.30E-01
Tamoxifen	7.69E-03	3.11E-02	1.63E-02	1.65E-03	5.86E-03	2.73E-01	1.28E-02
Temozolomide	2.64E-02	1.12E-01	2.00E-01	5.36E-02	1.66E-01	8.05E-02	4.87E-02
Vinorelbine	5.91E-02	4.31E-02	2.46E-03	2.41E-02	8.69E-02	2.82E-01	2.72E-02

Table A.3: The AUROC per drug of TINDL and other approaches.

Drug	TINDL	LASSO	TG-LASSO [11]	SVR	Geeleher et al. [24]	Random forest	ComBat-DL
Bleomycin	0.73	0.74	0.67	0.72	0.72	0.59	0.79
Cisplatin	0.63	0.64	0.63	0.64	0.47	0.57	0.65
Cyclophosphamide	0.71	0.63	0.62	0.66	0.34	0.51	0.71
Docetaxel	0.47	0.36	0.40	0.48	0.45	0.34	0.46
Doxorubicin	0.64	0.50	0.41	0.60	0.57	0.63	0.59
Etoposide	0.75	0.72	0.72	0.74	0.67	0.67	0.77
Gemcitabine	0.58	0.51	0.54	0.60	0.46	0.53	0.58
Irinotecan	0.58	0.59	0.61	0.60	0.62	0.58	0.57
Oxaliplatin	0.66	0.66	0.68	0.64	0.70	0.70	0.65
Paclitaxel	0.60	0.50	0.51	0.56	0.62	0.58	0.57
Pemetrexed	0.56	0.42	0.44	0.43	0.55	0.51	0.58
Tamoxifen	0.92	0.88	0.92	0.67	0.51	0.62	0.86
Temozolomide	0.68	0.67	0.66	0.63	0.60	0.47	0.62
Vinorelbine	0.75	0.81	0.82	0.58	0.52	0.50	0.80

Drug	ADDA-DL	DANN-DL	TrainNorm-DL	TestNorm-DL	GAT	GCN	LSTM
Bleomycin	0.72	0.70	0.58	0.66	0.76	0.76	0.56
Cisplatin	0.65	0.62	0.62	0.63	0.68	0.67	0.60
Cyclophosphamide	0.73	0.73	0.62	0.72	0.76	0.71	0.70
Docetaxel	0.38	0.37	0.42	0.47	0.50	0.43	0.47
Doxorubicin	0.43	0.54	0.46	0.61	0.56	0.62	0.53
Etoposide	0.69	0.71	0.79	0.78	0.83	0.75	0.68
Gemcitabine	0.56	0.59	0.57	0.58	0.61	0.58	0.55
Irinotecan	0.55	0.53	0.46	0.57	0.63	0.43	0.46
Oxaliplatin	0.66	0.68	0.66	0.65	0.63	0.56	0.65
Paclitaxel	0.56	0.55	0.59	0.60	0.59	0.58	0.57
Pemetrexed	0.55	0.53	0.60	0.56	0.52	0.56	0.57
Tamoxifen	0.85	0.77	0.81	0.90	0.86	0.60	0.82
Temozolomide	0.68	0.61	0.58	0.65	0.59	0.63	0.65
Vinorelbine	0.70	0.72	0.84	0.75	0.68	0.58	0.75

Table A.4: Precision at kth percentile of TINDL.

Drug	k = 10	k = 20	k = 30	k = 40	k = 50
Bleomycin	1.000	0.909	0.938	0.952	0.962
Cisplatin	0.935	0.885	0.868	0.843	0.862
Cyclophosphamide	1.000	1.000	1.000	1.000	0.980
Docetaxel	0.455	0.619	0.645	0.659	0.647
Doxorubicin	0.800	0.800	0.833	0.800	0.780
Etoposide	1.000	0.941	0.960	0.971	0.976
Gemcitabine	0.625	0.563	0.553	0.571	0.544
Irinotecan	0.000	0.200	0.286	0.222	0.333
Oxaliplatin	0.667	0.727	0.688	0.727	0.741
Paclitaxel	0.688	0.719	0.771	0.730	0.759
Pemetrexed	0.250	0.625	0.417	0.467	0.526
Tamoxifen	1.000	1.000	1.000	0.875	0.900
Temozolomide	0.200	0.211	0.172	0.184	0.167
Vinorelbine	1.000	1.000	0.889	0.917	0.933

Table A.5: Hyperparameters selected from the 5-fold CV for the TINDL models.

Drug	Learning rate	Batch size	Number of epochs
Bleomycin	1E-5	128	38
Cisplatin	5E-4	128	24
Cyclophosphamide	1E-4	128	6
Docetaxel	5E-5	64	10
Doxorubicin	1E-4	64	23
Etoposide	1E-4	64	33
Gemcitabine	5E-5	64	28
Irinotecan	1E-4	64	21
Oxaliplatin	1E-5	64	31
Paclitaxel	5E-4	64	38
Pemetrexed	1E-5	128	50
Tamoxifen	1E-5	64	21
Temozolomide	1E-5	128	39
Vinorelbine	5E-5	128	8

A.2 Batch-effect Removal using ComBat

ComBat [85] is a method that allows researchers to combine datasets from different batches by reducing their inter-batch variations caused by methodological and environmental inconsistencies between batches. Although this method was originally proposed to remove non-biological differences of samples with the same nature, some approaches studies have repurposed ComBat to homogenize CCL and tumor gene expression data [11, 24]. Similarly, one of the DL baselines in this Chapter was ComBat-DL, which uses a ComBat-based preprocessing before training.

Given gene g of sample j batch i, ComBat assumes that the expression is modeled by the following equation:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \tag{A.1}$$

where α_g is the overall mean expression, X is a design matrix of covariates, and β_g is the regression coefficients. The additive batch effect is denoted by γ_{ig}. The noise is given by ε_{ijg}, and this is scaled by the multiplicative batch effect δ_{ig}.

First, the model parameters α_g, β_g, γ_{ig} are estimated as $\hat{\alpha}_g$, $\hat{\beta}_g$, $\hat{\gamma}_{ig}$ using ordinary least-squares and with the constraint $\sum_i n_i\hat{\gamma}_{ig} = 0$, where n_i is the number of samples in batch i. Next, it calculates $\hat{\sigma}_g^2 = \frac{1}{N}\sum_i j(Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig})^2$, where N is the number of all samples across batches. The data is then standardized as Z_{ijg} using the following:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g}. \tag{A.2}$$

The additive and multiplicative batch effects are assumed to hail from prior distributions $\gamma_{ig} \sim N(Y_i, \tau_i^2)$ and $\delta_{ig}^2 \sim \text{InverseGamma}(\lambda_i, \theta_i)$. This means that the additive batch effect for a given batch across all genes are assumed to be from the same normal distribution. Similarly, the multiplicative batch effects are assumed to come from the same inverse gamma distribution. The parameters γ_i, τ_i^2, λ_i, and θ_i are all estimated empirically from the standardized data. From the assumed distributions, the batch effect parameters γ_{ig}^* and δ_{ig}^* are given by their conditional posterior means (see original publication [85] for full details). Since these two parameters are dependent on each other, these were estimated iteratively as

$\hat{\gamma}^*_{ig}$ and $\hat{\delta}^*_{ig}$. The adjusted data is then calculated by

$$\gamma^*_{ijg} = \frac{\hat{\sigma}_g}{\hat{\delta}^*_{ig}}(Z_{ijg} - \hat{\gamma}^*_{ig}) + \hat{\alpha}_g + X\hat{\beta}_g. \tag{A.3}$$

A.3 Supplementary Figures for Chapter 3

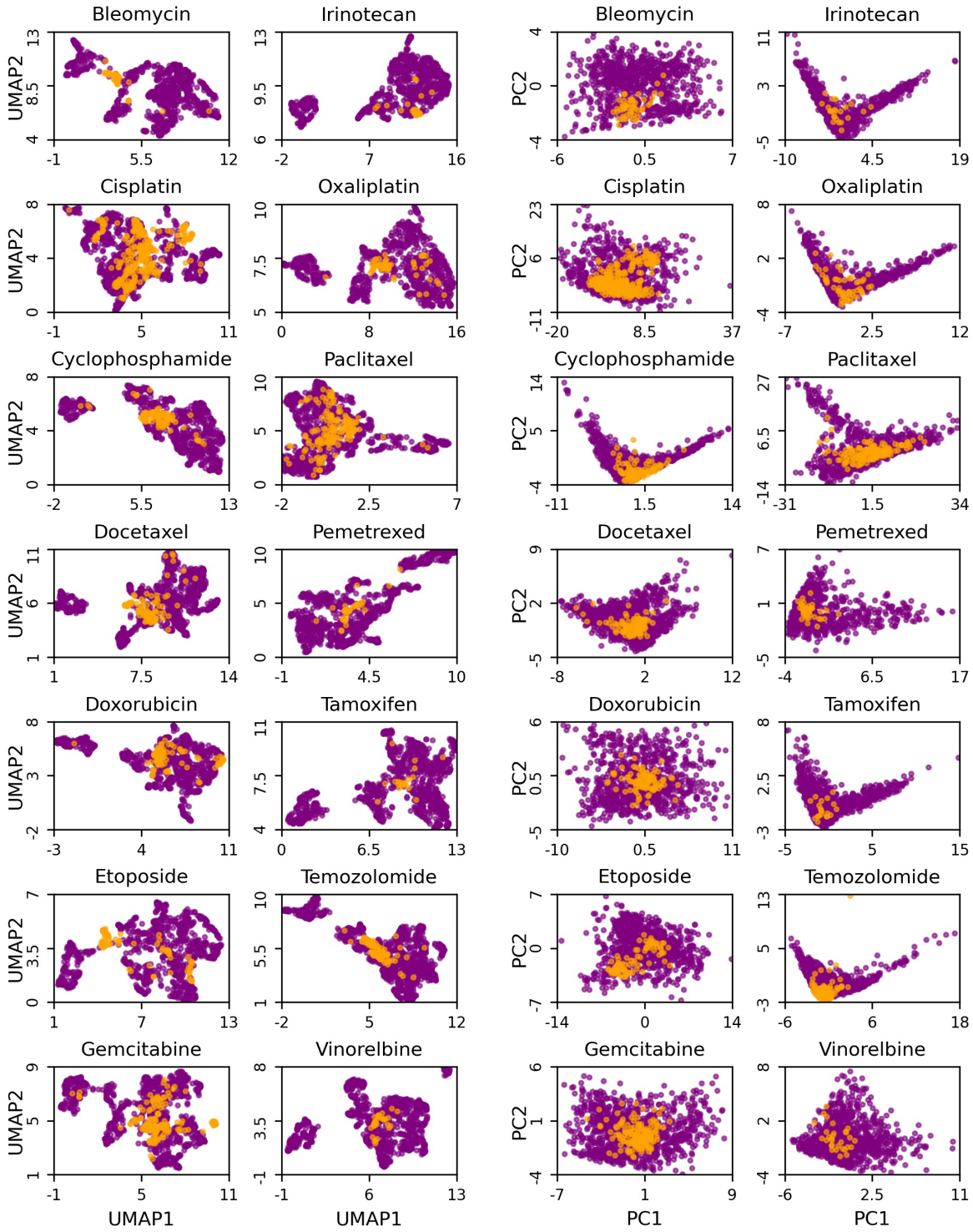

Figure A.1: PCA and UMAP plots of the CCLs (purple) and tumors (orange) for different drugs learned by TINDL.

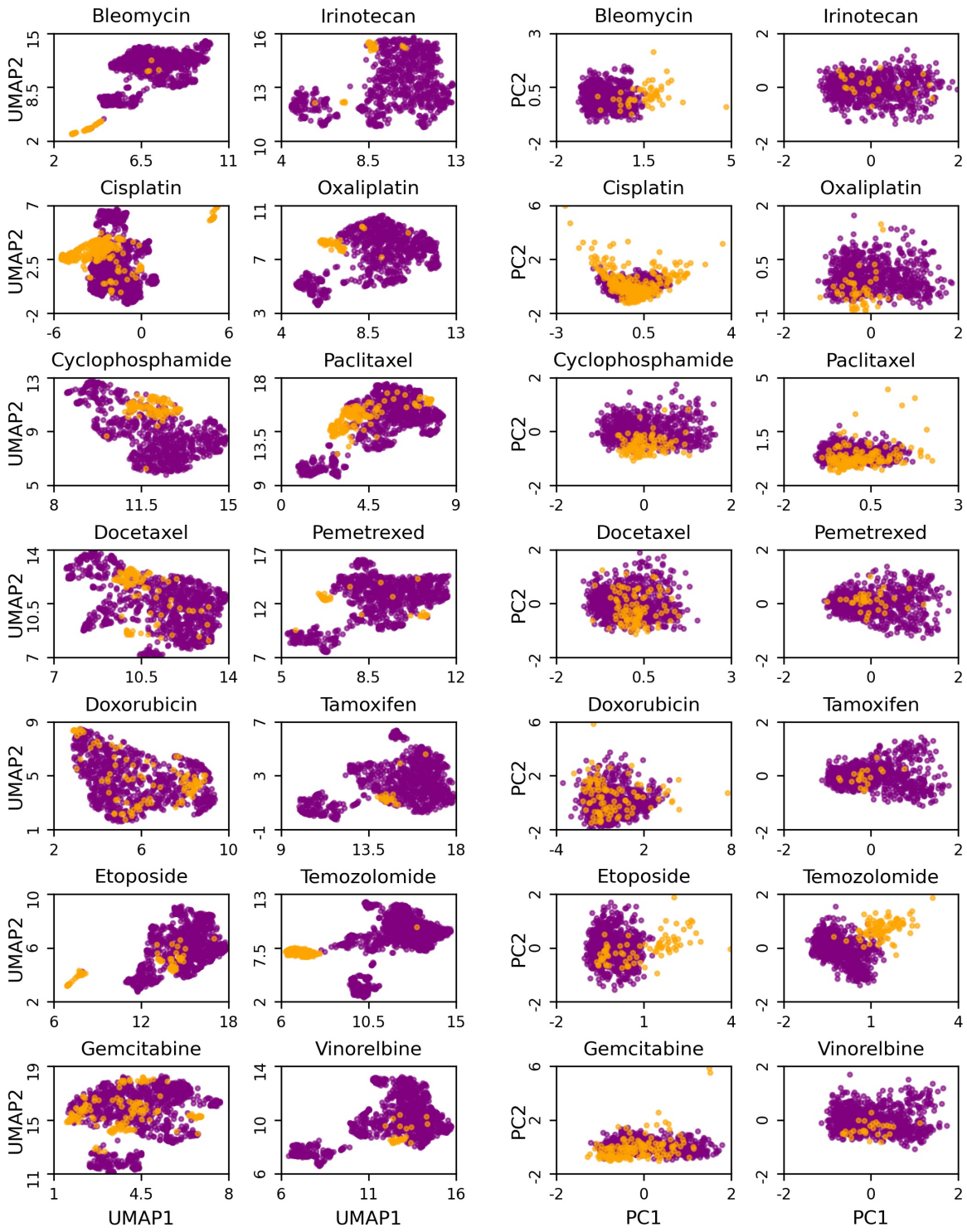

Figure A.2: PCA and UMAP plots of the CCLs (purple) and tumors (orange) for different drugs learned by ComBat-DL.

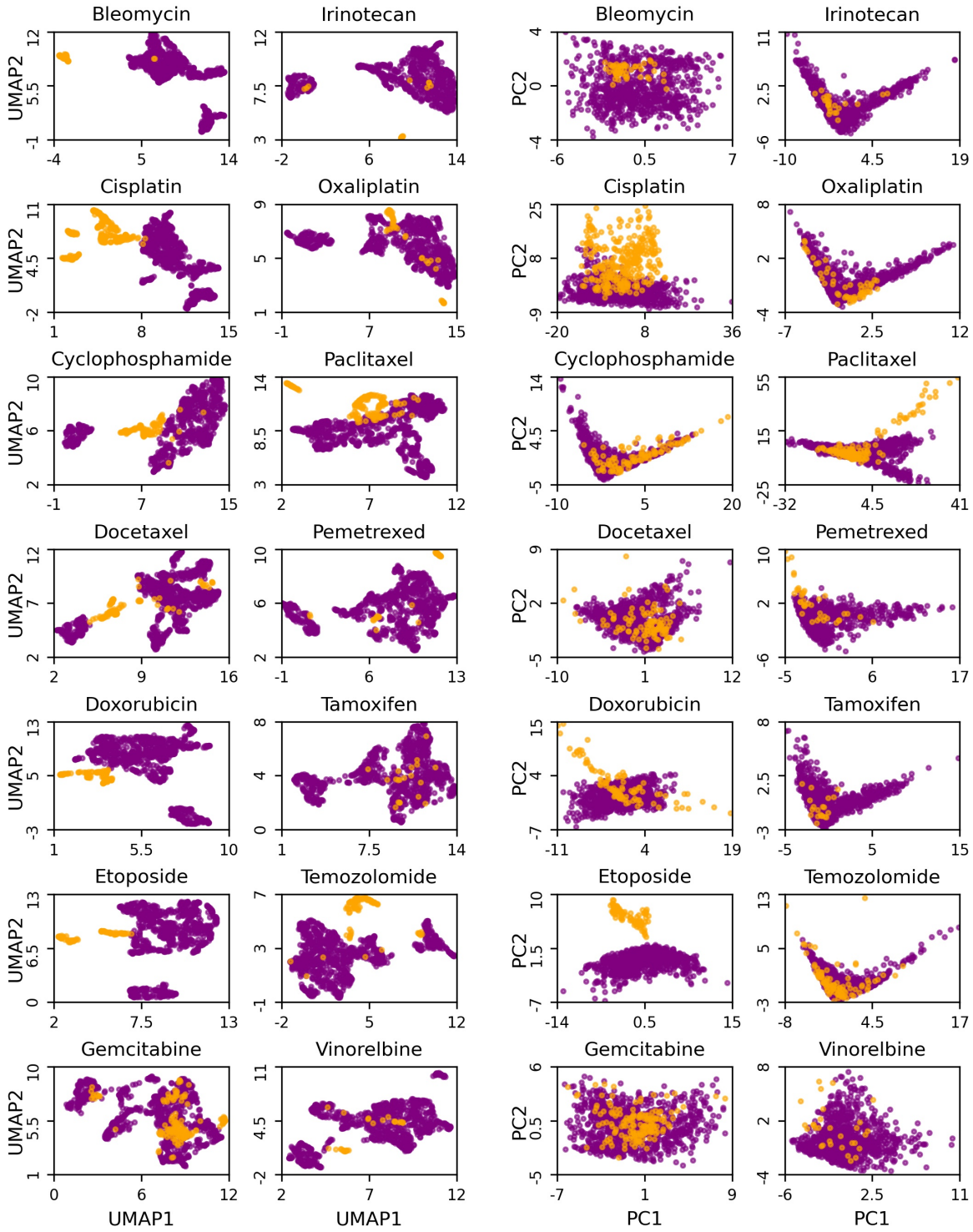

Figure A.3: PCA and UMAP plots of the CCLs (purple) and tumors (orange) for different drugs learned by ADDA-DL.

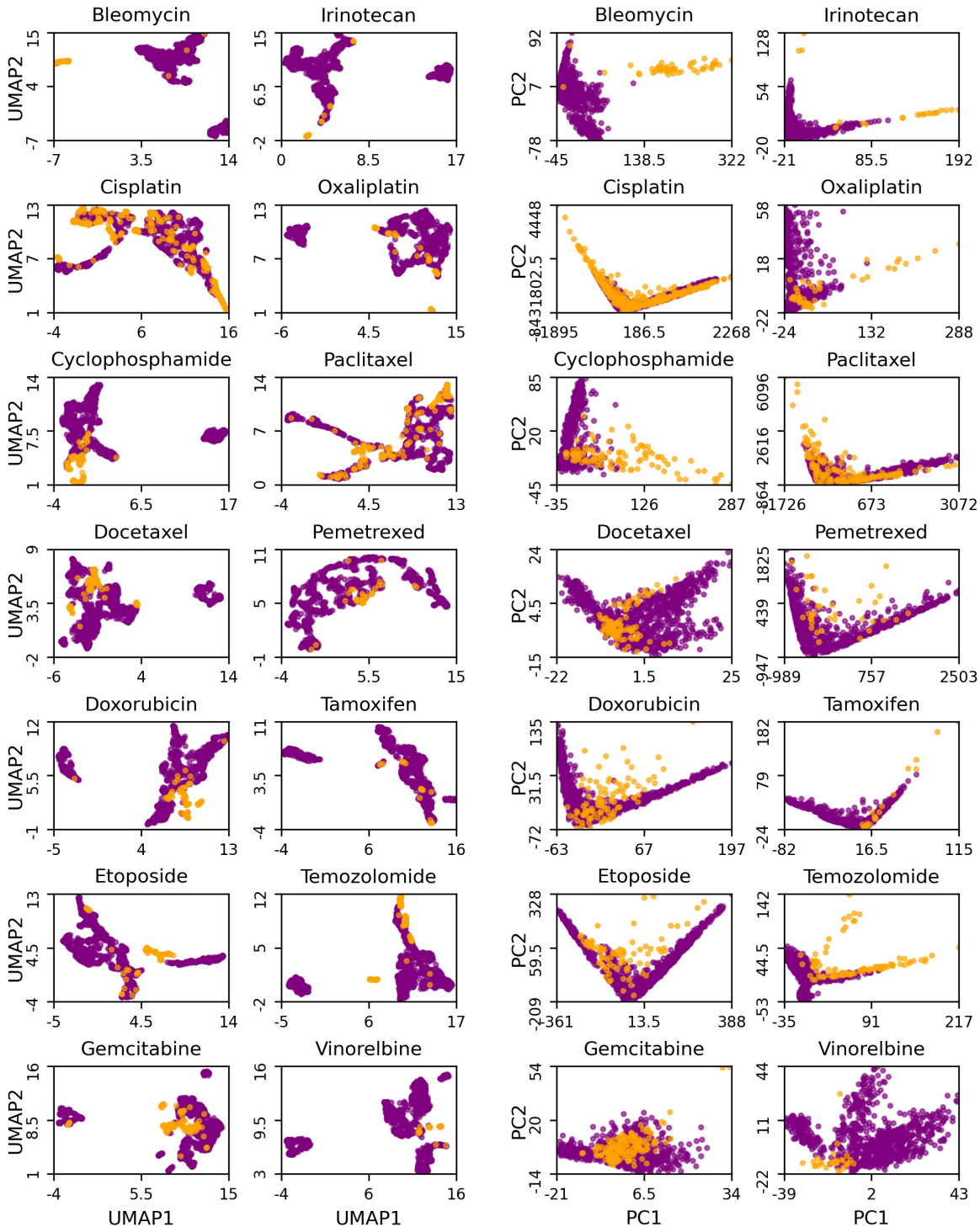

Figure A.4: PCA and UMAP plots of the CCLs (purple) and tumors (orange) for different drugs learned by DANN-DL.

A.4 Supplementary Files for Chapter 3

The following tables correspond to the supplementary files of our publication [12] that were referenced in this thesis. These tables are attached as spreadsheets, instead of printed, due to their extensive length. Alternatively, you may access them online through the links below.

Publication link: `https://doi.org/10.1016/j.gpb.2023.01.006`

File A.1 List of top genes identified by TINDL per drug (Table S4)
`https://ars.els-cdn.com/content/image/1-s2.0-S1672022923000323-mmc18.xlsx`

File A.2 Pathways associated to the top-identified genes of each drug (Table S6)
`https://ars.els-cdn.com/content/image/1-s2.0-S1672022923000323-mmc20.xlsx`

Appendix B

B.1 Supplementary Tables for Chapter 4

Table B.1: P values of the one-sided Wilcoxon signed-rank test comparing the drug-wise SCC values of BiG-DRP+ and the baseline methods. Here, the alternative is that the median of the population differences (SCCs of BiG-DRP+ *minus* SCCs of baseline) > 0.

Baseline	LCO-CV	LPO-CV
SVR-Linear	1.79E-39	6.22E-41
SVR-Linear (w/ RFE)	1.36E-39	6.22E-41
SVR-RBF	7.42E-36	6.22E-41
SVR-RBF (w/ RFE)	2.07E-33	6.22E-41
tCNN [6]	6.22E-41	6.22E-41
NRL2DRP [2]	NA	6.22E-41
PathDNN [5]	1.93E-40	1.93E-40
MLP	8.52E-22	6.30E-41
BiG-DRP (inverted)	NA	1.87E-40
BiG-DRP	3.00E-23	2.26E-36

Table B.2: Treatment responses of tumor samples (including multi-drug and sequential treatments) in TCGA and statistical test results when compared to BiG-DRP+ predictions.

Drug	Sensitive (S)		Resistant (R)		One-sided MWU P value *(alternative: greater)*			SC† P value	Normality P value
	CR	PR	SD	PD	R vs S	Others vs CR	PD vs Others		
Cisplatin	270	36	29	63	2.19E-07	2.48E-04	2.15E-03	2.06E-05	0.004
Doxorubicin	140	15	18	35	8.80E-03	6.75E-02	2.03E-02	2.25E-02	0.828
Gemcitabine	86	19	20	101	3.40E-02	1.02E-02	6.99E-02	3.47E-02	1.000
Paclitaxel	148	17	19	49	4.35E-01	7.90E-01	9.11E-01	4.33E-01	0.199

† SC of log IC50 (continuous) and CDR (ordinal with 4 categories)

Legend: MWU: Mann-Whitney U test, SC: Spearman correlation, Normality: D'Agostino and Pearson's test of normality, CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease

Table B.3: Single-drug treatment responses of tumor samples in TCGA and statistical test results when compared to BiG-DRP+ predictions.

Drug	Sensitive (S)		Resistant (R)		One-sided T test P value *(alternative: greater)*			SC† P value	Normality P value
	CR	PR	SD	PD	R vs S	Others vs CR	PD vs Others		
Cisplatin	107	8	6	28	1.82E-02	4.93E-02	3.17E-02	0.08613	0.303
Doxorubicin	16	7	3	22	3.02E-02	4.29E-02	1.10E-01	0.11698	0.255
Gemcitabine	33	10	12	58	1.95E-01	2.13E-01	1.61E-01	0.30132	0.254
Paclitaxel	64	7	8	23	6.15E-01	9.25E-01	8.99E-01	0.16866	0.082

† SC of log IC50 (continuous) and CDR (ordinal with 4 categories)

Legend: SC: Spearman correlation, Normality: D'Agostino and Pearson's test of normality, CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease

Table B.4: List of CCL clusters significantly enriched for some characteristics.

Cluster ID†	Property type	Property	P value	FDR
1	Origin	Haematopoietic and Lymphoid	6.78E-08	3.80E-06
4	Origin	Haematopoietic and Lymphoid	1.49E-05	5.57E-04
5	Origin	Haematopoietic and Lymphoid	2.64E-04	7.39E-03
6	Origin	Haematopoietic and Lymphoid	5.28E-09	5.92E-07
1	Cancer type	B-Lymphoblastic Leukemia	1.78E-04	1.54E-02
4	Cancer type	Chronic Myelogenous Leukemia	1.04E-06	1.81E-04
1	Driver mutation	RBM38	4.75E-07	3.95E-04
1	Driver mutation	GNA13	8.50E-08	1.41E-04
2	Driver mutation	POLQ	3.11E-04	3.97E-02
2	Driver mutation	BRCA1	2.89E-04	3.97E-02
3	Driver mutation	RASA2	2.41E-04	3.65E-02
4	Driver mutation	CBL	3.77E-04	4.17E-02
4	Driver mutation	ASXL1	5.29E-04	4.89E-02
5	Driver mutation	KMT2D	3.16E-05	1.61E-02
5	Driver mutation	SGK1	1.03E-04	2.18E-02
6	Driver mutation	KMT2D	9.88E-05	2.18E-02
6	Driver mutation	CREBBP	2.42E-04	3.65E-02
7	Driver mutation	FBN2	4.30E-04	4.21E-02
7	Driver mutation	NF1	1.05E-04	2.18E-02
7	Driver mutation	FAT4	4.30E-04	4.21E-02
7	Driver mutation	MARK2	3.41E-04	4.05E-02
10	Driver mutation	TET2	4.69E-05	1.61E-02
11	Driver mutation	FLT3	4.83E-05	1.61E-02
11	Driver mutation	LEF1	2.12E-04	3.65E-02

† Cluster membership of CCLs can be accessed in the Supplementary File B.4

B.2 Supplementary Files for Chapter 4

The following tables correspond to the supplementary files of our publication [13] that were referenced in this thesis. These tables are attached as spreadsheets, instead of printed, due to their extensive length. Alternatively, you may access them online through the link below.

Publication link: `https://doi.org/10.1093/bioinformatics/btac383`

File B.1 List of top genes identified to be relevant for BiG-DRP(+) in top-performing drugs (Table S7)

File B.2 Pathways associated to the top-identified genes of each top-performing drug (Table S8)

File B.3 Results of statistical tests regarding predicted drug response and mutations in TCGA (Table S9)

File B.4 Cluster assignments of CCLs and drugs (Table S5)

Appendix C

C.1 Sink Node Trick

When quantifying the graph-based importance scores, I remove nodes and edges from the knowledge graph. Depending on the normalization factor $c_{u,v}$ (see Equation 4.1), the output of the H-GCN might depend on the in/out/total node degree. As such, prediction changes due to node degree reductions, may not be as meaningful because they do not provide insight on the information propagated by/to the node. To address this, I created a sink node trick to preserve the out-degree of the nodes prior to node/edge removal.

When removing a node, say u, I add a new node $\hat{u}$ with the same node type as u. All edges that were pointing to u will be redirected to $\hat{u}$. By doing this, all parents of u will maintain their out-degree.

When removing an edge, say (t, u, v), where t is the edge type, u is the source, and v is the destination, I add a new node $\hat{v}$ with the same node type as v. I then add a new edge $(t, u, \hat{v})$ so that the out-degree of u is maintained.

Since the new node is a sink node, it does not pass information to other nodes. The attributes can just be set to a zero-vector. I do note that it may not be appropriate to preserve the in-degree of the nodes because preserving the in-degree while actually receiving information from one less neighbor would alter the magnitude of the output. For example, if the inner summation of Equation 4.1 is a sum of only two elements (post node/edge removal), then using the original in-degree of three would induce a change in magnitude, which may not be handled well by the subsequent layers of the model.

Bibliography

[1] J. C. Costello, L. M. Heiser, E. Georgii, M. Gonen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud din, P. Hintsanen, S. A. Khan, *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, vol. 32, no. 12, pp. 1202–12, 2014.

[2] J. Yang, A. Li, Y. Li, X. Guo, and M. Wang, "A novel approach for drug response prediction in cancer cell lines via network representation learning," *Bioinformatics*, vol. 35, no. 9, p. 1527–1535, 2019.

[3] C. Liu, D. Wei, J. Xiang, F. Ren, L. Huang, J. Lang, G. Tian, Y. Li, and J. Yang, "An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression," *Molecular Therapy-Nucleic Acids*, vol. 21, pp. 676–686, 2020.

[4] M. Li, Y. Wang, R. Zheng, X. Shi, Y. Li, F.-X. Wu, and J. Wang, "DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 2, pp. 575–582, 2019.

[5] L. Deng, Y. Cai, W. Zhang, W. Yang, B. Gao, and H. Liu, "Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity," *Journal of Chemical Information and Modeling*, vol. 60, no. 10, pp. 4497–4505, 2020.

[6] P. Liu, H. Li, S. Li, and K. S. Leung, "Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network," *BMC Bioinformatics*, vol. 20, no. 1, p. 408, 2019.

[7] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS Computational Biology*, vol. 11, no. 9, p. e1004498, 2015.

[8] J. Hao, Y. Kim, T.-K. Kim, and M. Kang, "PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data," *BMC Bioinformatics*, vol. 19, pp. 1–13, 2018.

[9] B. M. Kuenzi, J. Park, S. H. Fong, K. S. Sanchez, J. Lee, J. F. Kreisberg, J. Ma, and T. Ideker, "Predicting drug response and synergy using a deep learning model of human cancer cells," *Cancer Cell*, vol. 38, no. 5, pp. 672–684, 2020.

[10] K. Preuer, R. P. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer, "DeepSynergy: predicting anti-cancer drug synergy with deep learning," *Bioinformatics*, vol. 34, no. 9, pp. 1538–1546, 2018.

[11] E. W. Huang, A. Bhope, J. Lim, S. Sinha, and A. Emad, "Tissue-guided LASSO for prediction of clinical drug response using preclinical samples," *PLoS Computational Biology*, vol. 16, no. 1, p. e1007607, 2020.

[12] D. E. Hostallero, L. Wei, L. Wang, J. Cairns, and A. Emad, "Preclinical-to-clinical anti-cancer drug response prediction and biomarker identification using TINDL," *Genomics, Proteomics & Bioinformatics*, vol. 21, no. 3, pp. 535–550, 2023.

[13] D. E. Hostallero, Y. Li, and A. Emad, "Looking at the BiG picture: incorporating bipartite graphs in drug response prediction," *Bioinformatics*, vol. 38, no. 14, pp. 3609–3620, 2022.

[14] A. L. Roy and R. S. Conroy, "Toward mapping the human body at a cellular resolution," *Molecular Biology of the Cell*, vol. 29, no. 15, pp. 1779–1785, 2018.

[15] J. S. Brown, S. R. Amend, R. H. Austin, R. A. Gatenby, E. U. Hammarlund, and K. J. Pienta, "Updating the definition of cancer," *Molecular Cancer Research*, vol. 21, no. 11, pp. 1142–1147, 2023.

[16] National Cancer Institute, "What is cancer?." https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[17] World Health Organization, "Cancer." https://www.who.int/en/news-room/fact-sheets/detail/cancer.

[18] National Cancer Institute, "Types of cancer treatment." https://www.cancer.gov/about-cancer/treatment/types.

[19] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *European Journal of Cancer*, vol. 45, no. 2, pp. 228–247, 2009. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

[20] L. H. Schwartz, S. Litière, E. De Vries, R. Ford, S. Gwyther, S. Mandrekar, L. Shankar, J. Bogaerts, A. Chen, J. Dancey, *et al.*, "RECIST 1.1—Update and clarification: From the RECIST committee," *European Journal of Cancer*, vol. 62, pp. 132–137, 2016.

[21] P. Larsson, H. Engqvist, J. Biermann, E. Werner Rönnerman, E. Forssell-Aronsson, A. Kovács, P. Karlsson, K. Helou, and T. Z. Parris, "Optimization of cell viability assays to improve replicability and reproducibility of cancer drug sensitivity screens," *Scientific Reports*, vol. 10, no. 1, p. 5798, 2020.

[22] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, *et al.*, "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules," *Cell*, vol. 154, no. 5, p. 1151–1161, 2013.

[23] Z. Ding, S. Zu, and J. Gu, "Evaluating the molecule-based prediction of clinical drug responses in cancer," *Bioinformatics*, vol. 32, no. 19, p. 2891–5, 2016.

[24] 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 Biology*, vol. 15, no. 3, p. R47, 2014.

[25] N. G. Walter, "Are non-protein coding RNAs junk or treasure? An attempt to explain and reconcile opposing viewpoints of whether the human genome is mostly transcribed into non-functional or functional RNAs," *BioEssays*, p. 2300201, 2024.

[26] C. M. Rands, S. Meader, C. P. Ponting, and G. Lunter, "8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage," *PLoS Genetics*, vol. 10, no. 7, p. e1004525, 2014.

[27] H. Lodish, A. Berk, P. Matsudaira, K. C. A., M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Cell Biology*. Macmillan, 2008.

[28] I. Miko and L. LeJeune, *Essentials of Genetics*. Cambridge, MA: NPG Education, 2009.

[29] D. P. Nusinow, J. Szpyt, M. Ghandi, C. M. Rose, E. R. McDonald III, M. Kalocsay, J. Jané-Valbuena, E. Gelfand, D. K. Schweppe, M. Jedrychowski, *et al.*, "Quantitative proteomics of the cancer cell line encyclopedia," *Cell*, vol. 180, no. 2, pp. 387–402, 2020.

[30] Y. Liu, A. Beyer, and R. Aebersold, "On the dependency of cellular protein levels on mRNA abundance," *Cell*, vol. 165, no. 3, pp. 535–550, 2016.

[31] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 227–232, 2012.

[32] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, *et al.*, "Genome-scale CRISPR-Cas9 knockout screening in human cells," *Science*, vol. 343, no. 6166, pp. 84–87, 2014.

[33] C. Bock, P. Datlinger, F. Chardon, M. A. Coelho, M. B. Dong, K. A. Lawson, T. Lu, L. Maroc, T. M. Norman, B. Song, *et al.*, "High-content CRISPR screening," *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–23, 2022.

[34] J. M. Dempster, I. Boyle, F. Vazquez, D. E. Root, J. S. Boehm, W. C. Hahn, A. Tsherniak, and J. M. McFarland, "Chronos: a cell population dynamics model of CRISPR

experiments that improves inference of gene fitness effects," *Genome Biology*, vol. 22, pp. 1–23, 2021.

[35] Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–20, 2013.

[36] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, *et al.*, "Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research*, vol. 41, no. D1, pp. D955–D961, 2012.

[37] D. van der Meer, S. Barthorpe, W. Yang, H. Lightfoot, C. Hall, J. Gilbert, H. E. Francies, and M. J. Garnett, "Cell model passports-a hub for clinical, genetic and functional datasets of preclinical cancer models," *Nucleic Acids Research*, vol. 47, no. D1, pp. D923–D929, 2019.

[38] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, p. 603–7, 2012.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[40] W. L. Hamilton, *Graph representation learning*. Morgan & Claypool Publishers, 2020.

[41] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[42] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *International Conference on Learning Representations*, 2019.

[43] Q. Wan and R. Pal, "An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge," *PloS One*, vol. 9, no. 6, p. e101183, 2014.

[44] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," in *Biocomputing 2014*, pp. 63–74, World Scientific, 2014.

[45] A. Emad, J. Cairns, K. R. Kalari, L. Wang, and S. Sinha, "Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance," *Genome Biology*, vol. 18, no. 1, pp. 1–21, 2017.

[46] Q. T. Trac, Y. Pawitan, T. Mou, T. Erkers, P. Östling, A. Bohlin, A. Österroos, M. Vesterlund, R. Jafari, I. Siavelis, *et al.*, "Prediction model for drug response of acute myeloid leukemia patients," *NPJ Precision Oncology*, vol. 7, no. 1, p. 32, 2023.

[47] G. T. Nguyen and D.-H. Le, "A matrix completion method for drug response prediction in personalized medicine," in *Proceedings of the 9th International Symposium on Information and Communication Technology*, pp. 410–415, 2018.

[48] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[49] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077, 2015.

[50] 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, no. 1, pp. 1–12, 2017.

[51] N. N. Guan, Y. Zhao, C. C. Wang, J. Q. Li, X. Chen, and X. Piao, "Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization," *Molecular Therapy-Nucleic Acids*, vol. 17, pp. 164–174, 2019.

[52] C. Suphavilai, D. Bertrand, and N. Nagarajan, "Predicting cancer drug response using a recommender system," *Bioinformatics*, vol. 34, no. 22, pp. 3907–3914, 2018.

[53] D. Baptista, P. G. Ferreira, and M. Rocha, "Deep learning for drug response prediction in cancer," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 360–379, 2021.

[54] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[55] M. Joo, A. Park, K. Kim, W.-J. Son, H. S. Lee, G. Lim, J. Lee, D. H. Lee, J. An, J. H. Kim, *et al.*, "A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients," *International Journal of Molecular Sciences*, vol. 20, no. 24, p. 6276, 2019.

[56] F. Xia, J. Allen, P. Balaprakash, T. Brettin, C. Garcia-Cardona, A. Clyde, J. Cohn, J. Doroshow, X. Duan, V. Dubinkina, *et al.*, "A cross-study analysis of drug response prediction in cancer cell lines," *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab356, 2022.

[57] I. Jin and H. Nam, "HiDRA: hierarchical network for drug response prediction with attention," *Journal of Chemical Information and Modeling*, vol. 61, no. 8, pp. 3858–3867, 2021.

[58] H. Zhang, Y. Chen, and F. Li, "Predicting anticancer drug response with deep learning constrained by signaling pathways," *Frontiers in Bioinformatics*, vol. 1, p. 639349, 2021.

[59] K. Koras, E. Kizling, D. Juraeva, E. Staub, and E. Szczurek, "Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines," *Scientific Reports*, vol. 11, no. 1, p. 15993, 2021.

[60] M. R. El Khili, S. A. Memon, and A. Emad, "MARSY: a multitask deep-learning framework for prediction of drug combination synergy scores," *Bioinformatics*, vol. 39, no. 4, p. btad177, 2023.

[61] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T. S. Kim, J. Jung, and J.-M. Shin, "Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature," *Scientific Reports*, vol. 8, no. 1, p. 8857, 2018.

[62] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.

[63] 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.

[64] T. Nguyen, G. T. T. Nguyen, T. Nguyen, and D.-H. Le, "Graph convolutional networks for drug response prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, p. 146–154, 2022.

[65] X. Liu, C. Song, F. Huang, H. Fu, W. Xiao, and W. Zhang, "GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction," *Briefings in Bioinformatics*, vol. 23, p. bbab457, 11 2021.

[66] J. Shin, Y. Piao, D. Bang, S. Kim, and K. Jo, "DRPreter: interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer," *International Journal of Molecular Sciences*, vol. 23, no. 22, p. 13919, 2022.

[67] X. Yan, Y. Liu, and W. Zhang, "Deep graph and sequence representation learning for drug response prediction," in *International Conference on Artificial Neural Networks*, pp. 97–108, Springer, 2022.

[68] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[69] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, "Graph convo-

lutional networks for computational drug development and discovery," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 919–935, 2020.

[70] B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen, and J. Tang, "Comparative analysis of molecular fingerprints in prediction of drug combination effects," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab291, 2021.

[71] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *International Conference on Learning Representations*, 2019.

[72] P. Jia, R. Hu, G. Pei, Y. Dai, Y.-Y. Wang, and Z. Zhao, "Deep generative neural network for accurate drug response imputation," *Nature Communications*, vol. 12, no. 1, p. 1740, 2021.

[73] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young, and X. Lu, "Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Molecular Cancer Research*, vol. 16, no. 2, pp. 269–278, 2018.

[74] H. Liu, Y. Zhao, L. Zhang, and X. Chen, "Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal," *Molecular Therapy-Nucleic Acids*, vol. 13, pp. 303–311, 2018.

[75] L. Zhang, X. Chen, N. N. Guan, H. Liu, and J. Q. Li, "A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction," *Frontiers in Pharmacology*, vol. 9, p. 1017, 2018.

[76] Y.-C. Tang and A. Gottlieb, "Explainable drug sensitivity prediction through cancer pathway enrichment," *Scientific Reports*, vol. 11, no. 1, p. 3128, 2021.

[77] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, p. D607–D613, 2019.

[78] Y. Li, D. E. Hostallero, and A. Emad, "Interpretable deep learning architectures for improving drug response prediction performance: myth or reality?," *Bioinformatics*, vol. 39, p. btad390, 06 2023.

[79] P. Bertin, M. Hashir, M. Weiss, V. Frappier, T. J. Perkins, G. Boucher, and J. P. Cohen, "Analysis of gene interaction graphs as prior knowledge for machine learning models," *arXiv preprint arXiv:1905.02295*, 2019.

[80] P. Geeleher, Z. Zhang, F. Wang, R. F. Gruener, A. Nath, G. Morrison, S. Bhutra, R. L. Grossman, and R. S. Huang, "Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies," *Genome Research*, vol. 27, no. 10, pp. 1743–1751, 2017.

[81] S. Mourragui, M. Loog, M. A. Van De Wiel, M. J. Reinders, and L. F. Wessels, "PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors," *Bioinformatics*, vol. 35, no. 14, pp. i510–i519, 2019.

[82] S. M. Mourragui, M. Loog, D. J. Vis, K. Moore, A. G. Manjon, M. A. van de Wiel, M. J. Reinders, and L. F. Wessels, "Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 49, p. e2106682118, 2021.

[83] T. Sakellaropoulos, K. Vougas, S. Narang, F. Koinis, A. Kotsinas, A. Polyzos, T. J. Moss, S. Piha-Paul, H. Zhou, E. Kardala, *et al.*, "A deep learning framework for predicting response to therapy in cancer," *Cell Reports*, vol. 29, no. 11, pp. 3367–3373, 2019.

[84] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, "MOLI: multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, 2019.

[85] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, p. 118–127, 2007.

[86] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2960–2967, 2013.

[87] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[88] I. Anastopoulos, L. Seninge, H. Ding, and J. Stuart, "Patient informed domain adaptation improves clinical drug response prediction," *BioRxiv*, pp. 2021–08, 2021.

[89] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, pp. 97–105, PMLR, 2015.

[90] R. Peres da Silva, C. Suphavilai, and N. Nagarajan, "TUGDA: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from in vitro to in vivo settings," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i76–i83, 2021.

[91] H. Sharifi-Noghabi, S. Peng, O. Zolotareva, C. C. Collins, and M. Ester, "AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics," *Bioinformatics*, vol. 36, no. Supplement_1, p. i380–i388, 2020.

[92] Y.-C. Chiu, H.-I. H. Chen, T. Zhang, S. Zhang, A. Gorthi, L.-J. Wang, Y. Huang, and Y. Chen, "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," *BMC Medical Genomics*, vol. 12, pp. 143–155, 2019.

[93] Z. Wang, R. Li, M. Wang, and A. Li, "GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction," *Bioinformatics*, p. 2963–70, 2021.

[94] Z. Lv, Y. Lin, R. Yan, Y. Wang, and F. Zhang, "TransSurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorec-

tal cancer," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 6, pp. 3411–3420, 2023.

[95] P. Jiang, W. R. Sellers, and X. S. Liu, "Big data approaches for modeling response and resistance to cancer drugs," *Annual Review of Biomedical Data Science*, vol. 1, p. 1–27, 2018.

[96] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomics*, vol. 22, no. 1, p. 214, 2021.

[97] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, pp. 1180–1189, PMLR, 2015.

[98] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

[99] P. Schwab and W. Karlen, "Cxplain: Causal explanations for model interpretation under uncertainty," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[100] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, pp. 424–438, 1969.

[101] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *International Conference on Distributed Computing Systems Workshops*, p. 166–171, 2011.

[102] C. Blatti III, A. Emad, M. J. Berry, L. Gatzke, M. Epstein, D. Lanier, P. Rizal, J. Ge, X. Liao, O. Sobh, *et al.*, "Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform," *PLoS Biology*, vol. 18, no. 1, p. e3000583, 2020.

[103] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, pp. 1–14, 2013.

[104] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[105] A. Behdenna, J. Haziza, C.-A. Azencott, and A. Nordor, "pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods," *BioRxiv*, 2021.

[106] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[107] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[108] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[109] M. L. Liu, F. Zang, and S. J. Zhang, "RBCK1 contributes to chemoresistance and stemness in colorectal cancer (CRC)," *Biomedicine & Pharmacotherapy*, vol. 118, p. 109250, 2019.

[110] T. J. Chen, C. L. Chou, Y. F. Tian, C. F. Yeh, T. C. Chan, H. L. He, W. S. Li, H. H. Tsai, C. F. Li, and H. Y. Lai, "High FRMD3 expression is prognostic for worse survival in rectal cancer patients treated with CCRT," *International Journal of Clinical Oncology*, vol. 26, no. 9, p. 1689–1697, 2021.

[111] E. J. Kim, S. H. Kim, X. Jin, X. Jin, and H. Kim, "KCTD2, an adaptor of Cullin3 E3 ubiquitin ligase, suppresses gliomagenesis by destabilizing c-Myc," *Cell Death Differ*, vol. 24, no. 4, p. 649–659, 2017.

[112] A. Longatto-Filho, J. H. Fregnani, A. Mafra da Costa, P. S. de Araujo-Souza, C. Scapulatempo-Neto, S. Herbster, E. Boccardo, and L. Termini, "Evaluation of elafin immunohistochemical expression as marker of cervical cancer severity," *Acta Cytologica*, vol. 65, no. 2, p. 165–174, 2021.

[113] L. Y. Li, Q. Yang, Y. Y. Jiang, W. Yang, Y. Jiang, X. Li, M. Hazawa, B. Zhou, G. W. Huang, X. E. Xu, *et al.*, "Interplay and cooperation between SREBF1 and master transcription factors regulate lipid metabolism and tumor-promoting pathways in squamous cancer," *Nature Communications*, vol. 12, no. 1, p. 4362, 2021.

[114] J. Deng, X. Chen, T. Zhan, M. Chen, X. Yan, and X. Huang, "CRYAB predicts clinical prognosis and is associated with immunocyte infiltration in colorectal cancer," *PeerJ*, vol. 9, p. e12578, 2021.

[115] R. Fredriksson, S. Sreedharan, K. Nordenankar, J. Alsio, F. A. Lindberg, A. Hutchinson, A. Eriksson, S. Roshanbin, D. M. Ciuculete, A. Klockars, *et al.*, "The polyamine transporter Slc18b1(VPAT) is important for both short and long time memory and for regulation of polyamine content in the brain," *PLoS Genetics*, vol. 15, no. 12, p. e1008455, 2019.

[116] Y. Liu, Q. Li, L. Zhou, N. Xie, E. C. Nice, H. Zhang, C. Huang, and Y. Lei, "Cancer drug resistance: redox resetting renders a way," *Oncotarget*, vol. 7, no. 27, p. 42740–42761, 2016.

[117] H. M. Abdallah, A. M. Al-Abd, R. S. El-Dine, and A. M. El-Halawany, "P-glycoprotein inhibitors of natural origin as potential tumor chemo-sensitizers: A review," *Journal of Advanced Research*, vol. 6, no. 1, p. 45–62, 2015.

[118] K. G. Chen, J. C. Valencia, J. P. Gillet, V. J. Hearing, and M. M. Gottesman, "Involvement of ABC transporters in melanogenesis and the development of multidrug resistance of melanoma," *Pigment Cell & Melanoma Research*, vol. 22, no. 6, p. 740–9, 2009.

[119] F. M. Barzak, S. Harjes, M. V. Kvach, H. M. Kurup, G. B. Jameson, V. V. Filichev, and E. Harjes, "Selective inhibition of APOBEC3 enzymes by single-stranded dnas containing 2'-deoxyzebularine," *Organic & Biomolecular Chemistry*, vol. 17, no. 43, p. 9435–9441, 2019.

[120] F. Liu, J. Wei, Y. Hao, J. Lan, W. Li, J. Weng, M. Li, C. Su, B. Li, M. Mo, *et al.*, "Long intergenic non-protein coding RNA 02570 promotes nasopharyngeal carcinoma

progression by adsorbing microrna mir-4649-3p thereby upregulating both sterol regulatory element binding protein 1, and fatty acid synthase," *Bioengineered*, vol. 12, no. 1, p. 7119–7130, 2021.

[121] G. Hu, J. Zhang, F. Xu, H. Deng, W. Zhang, S. Kang, and W. Liang, "Stomatin-like protein 2 inhibits cisplatin-induced apoptosis through MEK/ERK signaling and the mitochondrial apoptosis pathway in cervical cancer cells," *Cancer Science*, vol. 109, no. 5, p. 1357–1368, 2018.

[122] A. M. Green, K. Budagyan, K. E. Hayer, M. A. Reed, M. R. Savani, G. B. Wertheim, and M. D. Weitzman, "Cytosine deaminase APOBEC3A sensitizes leukemia cells to inhibition of the DNA replication checkpoint," *Cancer Research*, vol. 77, no. 17, pp. 4579–4588, 2017.

[123] M. Petljak, A. Dananberg, K. Chu, E. N. Bergstrom, J. Striepen, P. von Morgen, Y. Chen, H. Shah, J. E. Sale, L. B. Alexandrov, M. R. Stratton, and J. Maciejowski, "Mechanisms of APOBEC3 mutagenesis in human cancer cells," *Nature*, vol. 607, no. 7920, pp. 799–807, 2022.

[124] R. A. Burrell, S. E. McClelland, D. Endesfelder, P. Groth, M. C. Weller, N. Shaikh, E. Domingo, N. Kanu, S. M. Dewhurst, E. Gronroos, *et al.*, "Replication stress links structural and numerical cancer chromosomal instability," *Nature*, vol. 494, no. 7438, p. 492–496, 2013.

[125] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, p. 646–74, 2011.

[126] J. Murai, A. Thomas, M. Miettinen, and Y. Pommier, "Schlafen 11 (SLFN11), a restriction factor for replicative stress induced by DNA-targeting anti-cancer therapies," *Pharmacology & Therapeutics*, vol. 201, p. 94–102, 2019.

[127] Y. Deng, Y. Cai, Y. Huang, Z. Yang, Y. Bai, Y. Liu, X. Deng, and J. Wang, "High SLFN11 expression predicts better survival for patients with kras exon 2 wild type colorectal cancer after treated with adjuvant oxaliplatin-based treatment," *BMC Cancer*, vol. 15, p. 833, 2015.

[128] C. Winkler, J. Armenia, G. N. Jones, L. Tobalina, M. J. Sale, T. Petreus, T. Baird, V. Serra, A. T. Wang, A. Lau, *et al.*, "SLFN11 informs on standard of care and novel treatments in a wide range of cancer models," *British Journal of Cancer*, vol. 124, no. 5, p. 951–962, 2021.

[129] E. E. Gardner, B. H. Lok, V. E. Schneeberger, P. Desmeules, L. A. Miles, P. K. Arnold, A. Ni, I. Khodos, E. de Stanchina, T. Nguyen, *et al.*, "Chemosensitive relapse in small cell lung cancer proceeds through an EZH2-SLFN11 axis," *Cancer Cell*, vol. 31, no. 2, p. 286–299, 2017.

[130] C. M. Fillmore, C. Xu, P. T. Desai, J. M. Berry, S. P. Rowbotham, Y. J. Lin, H. Zhang, V. E. Marquez, P. S. Hammerman, K. K. Wong, and C. F. Kim, "EZH2 inhibition sensitizes BRG1 and EGFR mutant lung tumours to TopoII inhibitors," *Nature*, vol. 520, no. 7546, p. 239–42, 2015.

[131] A. E. Gelman, J. Zhang, Y. Choi, and L. A. Turka, "Toll-like receptor ligands directly promote activated CD4+ T cell survival," *The Journal of Immunology*, vol. 172, no. 10, p. 6065–73, 2004.

[132] L. Alexopoulou, A. C. Holt, R. Medzhitov, and R. A. Flavell, "Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3," *Nature*, vol. 413, no. 6857, p. 732–8, 2001.

[133] J.-K. Li, J. J. Balic, L. Yu, and B. Jenkins, *TLR Agonists as Adjuvants for Cancer Vaccines*, pp. 195–212. Singapore: Springer Singapore, 2017.

[134] A. K. Nowak, B. W. Robinson, and R. A. Lake, "Synergy between chemotherapy and immunotherapy in the treatment of established murine solid tumors," *Cancer Research*, vol. 63, no. 15, p. 4490–6, 2003.

[135] S. Rakoff-Nahoum and R. Medzhitov, "Toll-like receptors and cancer," *Nature Reviews Cancer*, vol. 9, no. 1, p. 57–63, 2009.

[136] M. Kawashima, S. O. Suzuki, K. Doh-ura, and T. Iwaki, "Alpha-synuclein is expressed

in a variety of brain tumors showing neuronal differentiation," *Acta Neuropathologica*, vol. 99, no. 2, p. 154–60, 2000.

[137] Y. Ge and K. Xu, "Alpha-synuclein contributes to malignant progression of human meningioma via the Akt/mTOR pathway," *Cancer Cell International*, vol. 16, p. 86, 2016.

[138] S. Shekoohi, S. Rajasekaran, D. Patel, S. Yang, W. Liu, S. Huang, X. Yu, and S. N. Witt, "Knocking out alpha-synuclein in melanoma cells dysregulates cellular iron metabolism and suppresses tumor growth," *Scientific Reports*, vol. 11, no. 1, p. 5267, 2021.

[139] G. Tzivion, Z. Luo, and J. Avruch, "A dimeric 14-3-3 protein is an essential cofactor for Raf kinase activity," *Nature*, vol. 394, no. 6688, p. 88–92, 1998.

[140] J. W. Clancy, Y. Zhang, C. Sheehan, and C. D'Souza-Schorey, "An ARF6-Exportin-5 axis delivers pre-miRNA cargo to tumour microvesicles," *Nature Cell Biology*, vol. 21, no. 7, p. 856–866, 2019.

[141] J. W. Clancy, C. J. Tricarico, D. R. Marous, and C. D'Souza-Schorey, "Coordinated regulation of intracellular fascin distribution governs tumor microvesicle release and invasive cell capacity," *Molecular and Cellular Biology*, vol. 39, no. 3, p. e00264–18, 2019.

[142] R. Li, C. Peng, X. Zhang, Y. Wu, S. Pan, and Y. Xiao, "Roles of Arf6 in cancer cell invasion, metastasis and proliferation," *Life Sciences*, vol. 182, p. 80–84, 2017.

[143] Z. Hu, R. Xu, J. Liu, Y. Zhang, J. Du, W. Li, W. Zhang, Y. Li, Y. Zhu, and L. Gu, "GEP100 regulates epidermal growth factor-induced MDA-MB-231 breast cancer cell invasion through the activation of Arf6/ERK/uPAR signaling pathway," *Experimental Cell Research*, vol. 319, no. 13, p. 1932–1941, 2013.

[144] B. D. Hopkins, C. Pauli, X. Du, D. G. Wang, X. Li, D. Wu, S. C. Amadiume, M. D. Goncalves, C. Hodakoski, M. R. Lundquist, *et al.*, "Suppression of insulin feedback enhances the efficacy of PI3K inhibitors," *Nature*, vol. 560, no. 7719, p. 499–503, 2018.

[145] H. Hua, Q. Kong, J. Yin, J. Zhang, and Y. Jiang, "Insulin-like growth factor receptor signaling in tumorigenesis and drug resistance: a challenge for cancer therapy," *Journal of Hematology & Oncology*, vol. 13, no. 1, p. 64, 2020.

[146] S. Agrawal, M. Wozniak, M. Luc, S. Makuch, E. Pielka, A. K. Agrawal, J. Wietrzyk, J. Banach, A. Gamian, M. Pizon, and P. Ziolkowski, "Insulin enhancement of the antitumor activity of chemotherapeutic agents in colorectal cancer is linked with downregulating PIK3CA and GRB2," *Scientific Reports*, vol. 9, no. 1, p. 16647, 2019.

[147] B. O. Bodemann and M. A. White, "Ral GTPases and cancer: linchpin support of the tumorigenic platform," *Nature Reviews Cancer*, vol. 8, no. 2, p. 133–40, 2008.

[148] N. F. Neel, T. D. Martin, J. K. Stratford, T. P. Zand, D. J. Reiner, and C. J. Der, "The RalGEF-Ral effector signaling network: the road less traveled for anti-ras drug discovery," *Genes Cancer*, vol. 2, no. 3, p. 275–87, 2011.

[149] A. M. Gocher, C. J. Workman, and D. A. A. Vignali, "Interferon-gamma: teammate or opponent in the tumour microenvironment?," *Nature Reviews Immunology*, 2021.

[150] N. Coleman, B. Zhang, L. A. Byers, and T. A. Yap, "The role of Schlafen 11 (SLFN11) as a predictive biomarker for targeting the DNA damage response," *British Journal of Cancer*, vol. 124, no. 5, p. 857–859, 2021.

[151] J. Luan, X. Gao, F. Hu, Y. Zhang, and X. Gou, "SLFN11 is a general target for enhancing the sensitivity of cancer to chemotherapy (DNA-damaging agents)," *Journal of Drug Targeting*, vol. 28, no. 1, p. 33–40, 2020.

[152] Y. Li, M. Wang, M. Yang, Y. Xiao, Y. Jian, D. Shi, X. Chen, Y. Ouyang, L. Kong, X. Huang, *et al.*, "Nicotine-induced ILF2 facilitates nuclear mRNA export of pluripotency factors to promote stemness and chemoresistance in human esophageal cancer," *Cancer Research*, p. 3525–3538, 2021.

[153] H. Kikuchi, N. Maishi, D. A. Annan, M. T. Alam, R. I. H. Dawood, M. Sato, M. Morimoto, R. Takeda, K. Ishizuka, R. Matsumoto, *et al.*, "Chemotherapy-induced IL8

upregulates MDR1/ABCB1 in tumor blood vessels and results in unfavorable outcome," *Cancer Research*, vol. 80, no. 14, p. 2996–3008, 2020.

[154] R. Kubiliute, I. Januskeviciene, R. Urbanaviciute, K. Daniunaite, M. Drobniene, V. Ostapenko, R. Daugelavicius, and S. Jarmalaite, "Nongenotoxic ABCB1 activator tetraphenylphosphonium can contribute to doxorubicin resistance in MX-1 breast cancer cell line," *Scientific Reports*, vol. 11, no. 1, p. 6556, 2021.

[155] L. Li, B. L. Fridley, K. Kalari, N. Niu, G. Jenkins, A. Batzler, R. P. Abo, D. Schaid, and L. Wang, "Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines," *BMC Genomics*, vol. 15, p. 93, 2014.

[156] A. U. Lindner, C. G. Concannon, G. J. Boukes, M. D. Cannon, F. Llambi, D. Ryan, K. Boland, J. Kehoe, D. A. McNamara, F. Murray, *et al.*, "Systems analysis of BCL2 protein family interactions establishes a model to predict responses to chemotherapy," *Cancer Research*, vol. 73, no. 2, p. 519–28, 2013.

[157] B. Zagidullin, J. Aldahdooh, S. Zheng, W. Wang, Y. Wang, J. Saad, A. Malyutina, M. Jafari, Z. Tanoli, A. Pessia, and J. Tang, "DrugComb: an integrative cancer drug combination data portal," *Nucleic Acids Research*, vol. 47, no. W1, p. W43–W51, 2019.

[158] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng, "DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy," *Nucleic Acids Research*, vol. 48, no. D1, p. D871–D881, 2020.

[159] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1388–D1395, 2021.

[160] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, 2017.

[161] "RDKit: Open-source cheminformatics." `http://www.rdkit.org`.

[162] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[163] D. Weininger, "SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Modeling*, vol. 28, no. 1, pp. 31–36, 1988.

[164] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[165] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[166] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[167] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data," *Nucleic Acids Research*, vol. 44, no. D1, pp. D380–4, 2016.

[168] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, no. D1, pp. D498–D503, 2020.

[169] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, p. 011047, 2014.

[170] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, 1971.

[171] G. M. Sizemore, J. R. Pitarresi, S. Balakrishnan, and M. C. Ostrowski, "The ets family of oncogenic transcription factors in solid tumours," *Nature Reviews Cancer*, vol. 17, no. 6, pp. 337–351, 2017.

[172] L. Lopez-Delisle, C. Pierre-Eugene, C. Louis-Brennetot, D. Surdez, V. Raynal, S. Baulande, V. Boeva, S. Grossetete-Lalami, V. Combaret, M. Peuchmaur, O. Delattre, and I. Janoueix-Lerosey, "Activated ALK signals through the ERK-ETV5-RET pathway to drive neuroblastoma oncogenesis," *Oncogene*, vol. 37, no. 11, pp. 1417–1429, 2018.

[173] M. Ranzani, C. Alifrangis, D. Perna, K. Dutton-Regester, A. Pritchard, K. Wong, M. Rashid, C. D. Robles-Espinoza, N. K. Hayward, U. McDermott, M. Garnett, and D. J. Adams, "BRAF/NRAS wild-type melanoma, NF1 status and sensitivity to trametinib," *Pigment Cell Melanoma Res*, vol. 28, no. 1, pp. 117–9, 2015.

[174] B. Wang, E. B. Krall, A. J. Aguirre, M. Kim, H. R. Widlund, M. B. Doshi, E. Sicinska, R. Sulahian, A. Goodale, G. S. Cowley, *et al.*, "ATXN1L, CIC, and ETS transcription factors modulate sensitivity to MAPK pathway inhibition," *Cell Reports*, vol. 18, no. 6, pp. 1543–1557, 2017.

[175] J. P. Gustin, D. P. Cosgrove, and B. H. Park, "The PIK3CA gene as a mutated target for cancer therapy," *Current Cancer Drug Targets*, vol. 8, no. 8, pp. 733–40, 2008.

[176] A. S. Alzahrani, "PI3K/Akt/mTOR inhibitors in cancer: At the bench and bedside," *Seminars in Cancer Biology*, vol. 59, pp. 125–132, 2019.

[177] K. A. West, S. S. Castillo, and P. A. Dennis, "Activation of the PI3K/Akt pathway and chemotherapeutic resistance," *Drug Resistance Updates*, vol. 5, no. 6, pp. 234–48, 2002.

[178] R. Liu, Y. Chen, G. Liu, C. Li, Y. Song, Z. Cao, W. Li, J. Hu, C. Lu, and Y. Liu, "PI3K/AKT pathway as a key link modulates the multidrug resistance of cancers," *Cell Death & Disease*, vol. 11, no. 9, p. 797, 2020.

[179] C. Dong, J. Wu, Y. Chen, J. Nie, and C. Chen, "Activation of PI3K/AKT/mTOR pathway causes drug resistance in breast cancer," *Frontiers in Pharmacology*, vol. 12, p. 628690, 2021.

[180] J. Yang, J. Nie, X. Ma, Y. Wei, Y. Peng, and X. Wei, "Targeting PI3K in cancer: mechanisms and advances in clinical trials," *Molecular Cancer*, vol. 18, no. 1, p. 26, 2019.

[181] M. Wang, J. Li, J. Huang, and M. Luo, "The predictive role of PIK3CA mutation status on PI3K inhibitors in HR+ breast cancer therapy: a systematic review and meta-analysis," *BioMed Research International*, vol. 2020, p. 1598037, 2020.

[182] E. Jokinen and J. P. Koivunen, "MEK and PI3K inhibition in solid tumors: rationale and evidence to date," *Therapeutic Advances in Medical Oncology*, vol. 7, no. 3, pp. 170–80, 2015.

[183] E. Halilovic, Q. B. She, Q. Ye, R. Pagliarini, W. R. Sellers, D. B. Solit, and N. Rosen, "PIK3CA mutation uncouples tumor growth and cyclin D1 regulation from MEK/ERK and mutant KRAS signaling," *Cancer Research*, vol. 70, no. 17, pp. 6804–14, 2010.

[184] J. Zorea, M. Prasad, L. Cohen, N. Li, R. Schefzik, S. Ghosh, B. Rotblat, B. Brors, and M. Elkabets, "IGF1R upregulation confers resistance to isoform-specific inhibitors of PI3K in PIK3CA-driven ovarian cancer," *Cell Death Discovery*, vol. 9, no. 10, p. 944, 2018.

[185] C. Leroy, P. Ramos, K. Cornille, D. Bonenfant, C. Fritsch, H. Voshol, and M. Bentires-Alj, "Activation of IGF1R/p110β/AKT/mTOR confers resistance to α-specific PI3K inhibition," *Breast Cancer Research*, vol. 18, no. 1, p. 41, 2016.

[186] T. C. Beadnell, K. W. Nassar, M. M. Rose, E. G. Clark, B. P. Danysh, M. C. Hofmann, N. Pozdeyev, and R. E. Schweppe, "Src-mediated regulation of the PI3K pathway in advanced papillary and anaplastic thyroid cancer," *Oncogenesis*, vol. 7, no. 2, p. 23, 2018.

[187] F. Crispo, T. Notarangelo, M. Pietrafesa, G. Lettini, G. Storto, A. Sgambato, F. Maddalena, and M. Landriscina, "BRAF inhibitors in thyroid cancer: Clinical impact, mechanisms of resistance and future perspectives," *Cancers (Basel)*, vol. 11, no. 9, 2019.

[188] U. McDermott, S. V. Sharma, L. Dowell, P. Greninger, C. Montagut, J. Lamb, H. Archibald, R. Raudales, A. Tam, D. Lee, *et al.*, "Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19936–19941, 2007.

[189] D. B. Solit, L. A. Garraway, C. A. Pratilas, A. Sawai, G. Getz, A. Basso, Q. Ye, J. M. Lobo, Y. She, I. Osman, *et al.*, "BRAF mutation predicts sensitivity to mek inhibition," *Nature*, vol. 439, no. 7074, pp. 358–62, 2006.

[190] A. Koleti, R. Terryn, V. Stathias, C. Chung, D. J. Cooper, J. P. Turner, D. Vidovic, M. Forlin, T. T. Kelley, A. D'Urso, *et al.*, "Data portal for the library of integrated network-based cellular signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data," *Nucleic Acids Research*, vol. 46, no. D1, pp. D558–D566, 2018.

[191] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[192] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan, "Selformer: Molecular representation learning via selfies language models," *Machine Learning: Science and Technology*, 2023.

[193] M. Swain, "PubChemPy: Python wrapper for the PubChem PUG REST API." `https://github.com/mcs07/PubChemPy`.

[194] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[195] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath, "Fastshap: Real-time shapley value estimation," in *International Conference on Learning Representations*, 2021.

[196] H. Sharifi-Noghabi, S. Jahangiri-Tazehkand, P. Smirnov, C. Hon, A. Mammoliti, S. K. Nair, A. S. Mer, M. Ester, and B. Haibe-Kains, "Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab294, 2021.

[197] G. Zoppoli, M. Regairaz, E. Leo, W. C. Reinhold, S. Varma, A. Ballestrero, J. H. Doroshow, and Y. Pommier, "Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents," *Proceedings of the National Academy of Sciences*, vol. 109, no. 37, pp. 15030–15035, 2012.

[198] Y. Mu, J. Lou, M. Srivastava, B. Zhao, X.-h. Feng, T. Liu, J. Chen, and J. Huang, "SLFN 11 inhibits checkpoint maintenance and homologous recombination repair," *EMBO reports*, vol. 17, no. 1, pp. 94–109, 2016.

[199] J. Murai, S.-W. Tang, E. Leo, S. A. Baechler, C. E. Redon, H. Zhang, M. Al Abo, V. N. Rajapakse, E. Nakamura, L. M. M. Jenkins, *et al.*, "SLFN11 blocks stressed replication forks independently of ATR," *Molecular Cell*, vol. 69, no. 3, pp. 371–384, 2018.

[200] Y.-p. Yin, L.-y. Ma, G.-z. Cao, J.-h. Hua, X.-t. Lv, and W.-c. Lin, "FK228 potentiates topotecan activity against small cell lung cancer cells via induction of SLFN11," *Acta Pharmacologica Sinica*, vol. 43, no. 8, pp. 2119–2127, 2022.

[201] J.-Y. Yang, X.-Y. Deng, Y.-S. Li, X.-C. Ma, J.-X. Feng, B. Yu, Y. Chen, Y.-L. Luo, X. Wang, M.-L. Chen, *et al.*, "Structure of Schlafen13 reveals a new class of tRNA/rRNA-targeting RNase engaged in translational control," *Nature Communications*, vol. 9, no. 1, p. 1165, 2018.

[202] X. Liu and R. L. Erikson, "Polo-like kinase (Plk) 1 depletion induces apoptosis in cancer cells," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5789–5794, 2003.

[203] R. J. Youle and A. Strasser, "The BCL-2 protein family: opposing activities that mediate cell death," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 1, pp. 47–59, 2008.

[204] D. R. Premkumar, E. P. Jane, S. Thambireddy, P. A. Sutera, J. M. Cavaleri, and I. F. Pollack, "Mitochondrial dysfunction RAD51, and Ku80 proteolysis promote apoptotic effects of Dinaciclib in Bcl-xL silenced cells," *Molecular Carcinogenesis*, vol. 57, no. 4, pp. 469–482, 2018.

[205] M. M. Szwarc, A. L. Guarnieri, M. Joshi, H. N. Duc, M. C. Laird, A. Pandey, S. Khanal, E. Dohm, A. K. Bui, K. D. Sullivan, *et al.*, "FAM193A is a positive regulator of p53 activity," *Cell Reports*, vol. 42, no. 3, 2023.

[206] C. Bornstein, R. Brosh, A. Molchadsky, S. Madar, I. Kogan-Sakin, I. Goldstein, D. Chakravarti, E. R. Flores, N. Goldfinger, R. Sarig, *et al.*, "SPATA18, a spermatogenesis-associated gene, is a novel transcriptional target of p53 and p63," *Molecular and Cellular Biology*, vol. 31, no. 8, pp. 1679–1689, 2011.

[207] E. A. Harrington, D. Bebbington, J. Moore, R. K. Rasmussen, A. O. Ajose-Adeogun, T. Nakayama, J. A. Graham, C. Demur, T. Hercend, A. Diu-Hercend, *et al.*, "VX-680, a potent and selective small-molecule inhibitor of the aurora kinases, suppresses tumor growth in vivo," *Nature Medicine*, vol. 10, no. 3, pp. 262–267, 2004.

[208] A. Sharma, S. V. Madhunapantula, R. Gowda, A. Berg, R. I. Neves, and G. P. Robertson, "Identification of aurora kinase B and Wee1-like protein kinase as downstream targets of V600EB-RAF in melanoma," *The American Journal of Pathology*, vol. 182, no. 4, pp. 1151–1162, 2013.

[209] F. J. Giles, J. Cortes, D. Jones, D. Bergstrom, H. Kantarjian, and S. J. Freedman, "MK-0457, a novel kinase inhibitor, is active in patients with chronic myeloid leukemia or acute lymphocytic leukemia with the T315I BCR-ABL mutation," *Blood*, vol. 109, no. 2, pp. 500–502, 2007.

[210] Z.-J. Kang, Y.-F. Liu, L.-Z. Xu, Z.-J. Long, D. Huang, Y. Yang, B. Liu, J.-X. Feng, Y.-J. Pan, J.-S. Yan, *et al.*, "The Philadelphia chromosome in leukemogenesis," *Chinese Journal of Cancer*, vol. 35, pp. 1–15, 2016.

[211] S. Adnan-Awad, D. Kim, H. Hohtari, K. K. Javarappa, T. Brandstoetter, I. Mayer, S. Potdar, C. A. Heckman, S. Kytölä, K. Porkka, *et al.*, "Characterization of p190-

Bcr-Abl chronic myeloid leukemia reveals specific signaling pathways and therapeutic targets," *Leukemia*, vol. 35, no. 7, pp. 1964–1975, 2021.

[212] N. Parry, C. Busch, V. Aßmann, J. Cassels, A. Hair, G. V. Helgason, H. Wheadon, and M. Copland, "BH3 mimetics in combination with nilotinib or ponatinib represent a promising therapeutic strategy in blast phase chronic myeloid leukemia," *Cell Death Discovery*, vol. 8, no. 1, p. 457, 2022.

[213] K. Alhazzani, A. Almangour, A. Alsalem, M. Alqinyah, A. S. Alhamed, H. N. Alhamami, and A. Z. Alanazi, "Examining the effects of dasatinib, sorafenib, and nilotinib on vascular smooth muscle cells: Insights into proliferation, migration, and gene expression dynamics," *Diseases*, vol. 11, no. 4, 2023.

[214] H. Wang, B. Hong, X. Li, K. Deng, H. Li, V. W. Y. Lui, and W. Lin, "JQ1 synergizes with the Bcl-2 inhibitor ABT-263 against MYCN-amplified small cell lung cancer," *Oncotarget*, vol. 8, no. 49, p. 86312, 2017.

[215] C. Zhang, Z.-Y. Su, L. Wang, L. Shu, Y. Yang, Y. Guo, D. Pung, C. Bountra, and A.-N. Kong, "Epigenetic blockade of neoplastic transformation by bromodomain and extra-terminal (BET) domain protein inhibitor JQ-1," *Biochemical Pharmacology*, vol. 117, pp. 35–45, 2016.

[216] I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *Journal of Machine Learning Research*, vol. 22, no. 209, pp. 1–90, 2021.

[217] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, pp. i457–i466, 06 2018.

[218] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, vol. 70, pp. 3145–3153, PMLR, 06–11 Aug 2017.

[219] C. Suphavilai, S. Chia, A. Sharma, L. Tu, R. P. Da Silva, A. Mongia, R. DasGupta, and

N. Nagarajan, "Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures," *Genome Medicine*, vol. 13, pp. 1–14, 2021.

[220] J. Chen, X. Wang, A. Ma, Q.-E. Wang, B. Liu, L. Li, D. Xu, and Q. Ma, "Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data," *Nature Communications*, vol. 13, no. 1, p. 6494, 2022.

[221] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1997.

[222] S. Fatima, S. Ali, and H.-C. Kim, "A comprehensive review on multiple instance learning," *Electronics*, vol. 12, no. 20, p. 4323, 2023.

[223] A. Malyutina, M. M. Majumder, W. Wang, A. Pessia, C. A. Heckman, and J. Tang, "Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer," *PLoS Computational Biology*, vol. 15, no. 5, p. e1006752, 2019.

[224] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*, pp. 2376–2384, PMLR, 2019.

[225] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.