PREDICTION OF DRUG RESPONSE IN CANCER USING HYBRID DEEP NEURAL NETWORKS

HİBRİT DERİN SİNİR AĞLARI İLE KANSERDE İLAÇ YANIT TAHMİNİ

BURAKCAN İZMİRLİ

PROF. TUNCA DOĞAN

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment the Requirements

for the Award of the Degree of Master of Science in Computer Engineering

2024

ÖZET

HİBRİT DERİN SİNİR AĞLARI İLE KANSERDE İLAÇ YANIT TAHMİNİ

Burakcan İZMİRLİ

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Tunca Doğan

Ocak 2024, 126 Sayfa

Her hastaya en uygun tedavi seçeneğini belirlemek, tıbbın ana hedefidir. Aynı tanıya sahip hastalar, genetik heterojenlik nedeniyle uygulanan tedaviye özellikle kanserlerde farklı duyarlılık gösterebilir. Bu çalışmada, kanser hücrelerinin ilaç yanıtlarını (duyarlılığını) tahmin eden bir makine öğrenmesi tabanlı sistem olan DeepResponse'u öneriyoruz.

DeepResponse, büyük ölçekli tarama projelerinden elde edilen farklı kanser hücre hatlarının çoklu-omik profillerini ve ilaçların moleküler özelliklerini giriş seviyesinde kullanır ve tümörün çoklu-omik özellikleri ile uygulanan ilaca olan duyarlılığı arasındaki ilişkiyi öğrenmek için hibrit konvolüsyonel ve çizge dönüştürücü derin sinir ağları aracılığıyla işler.

DeepResponse, rastgele bölme, hücre tabakalaştırılmış bölme ve ilaç tabakalaştırılmış bölme test veri setleri için sırasıyla 1.014 ± 0.001, 1.105 ± 0.013 ve 1.142 ± 0.104 kök ortalama kare hatasına ulaşarak ilaç yanıtlarını tahmin etmedeki etkinliğini göstermiştir.

Performans sonuçları, DeepResponse’un kanser hücrelerinin ilaç duyarlılığını başarıyla tahmin ettiğini ve özellikle çoklu-omik yönünün öğrenme sürecinden faydalandığını ve tüm bölümlerde mevcut modellerden daha iyi performansa sahip olduğunu gösterir.

Her bir omik veri türünün DeepResponse’un performansı üzerindeki etkisini değerlendirmek için bir ablasyon çalışması yürütüldü, bu da ilaç yanıtı tahmininde çoklu-omik entegrasyonun önemini kanıtladı. DeepReponse’un kullanımına örnek olarak, Eprinomectin, hepatosellüler karsinoma kanser hücre hatlarına karşı yeniden amaçalanma ilaç adayı olarak önerildi ve ıslak laboratuvar deneylerinde doğrulandı. DeepResponse’un kod tabanı, veri setleri ve sonuçları https://github.com/HUBioDataLab/DeepResponse adresinde paylaşılmıştır. DeepResponse, yeni ilaç adaylarının erken aşama keşfi ve dirençli tümörlere karşı mevcut olanların yeniden amaçlanması için kullanılabilir.

Anahtar Kelimeler: Biyoenformatik, Çoklu-omik analizler, İlaç yanıtı tahmini, Makine öğrenmesi / derin öğrenme, Hibrit mimariler, Çizge dönüştürücüler

ABSTRACT

PREDICTION OF DRUG RESPONSE IN CANCER USING HYBRID DEEP NEURAL NETWORKS

Burakcan İZMİRLİ

Master of Sciences, Department of Computer Engineering

Supervisor: Prof. Tunca Doğan

January 2024, 126 Pages

Assessing the best treatment option for each patient is the main goal of precision medicine. Patients with the same diagnosis may display varying sensitivity to the applied treatment due to genetic heterogeneity, especially in cancers.

Here, we propose DeepResponse, a machine learning-based system that predicts drug responses (sensitivity) of cancer cells. DeepResponse employs multi-omics profiles of different cancer cell-lines obtained from large-scale screening projects, together with drugs' molecular features at the input level, and processing them via hybrid convolutional and graph-transformer deep neural networks to learn the relationship between multi-omics features of the tumor and its sensitivity to the administered drug.

DeepResponse has reached a Root Mean Squared Error (RMSE) of 1.014 ± 0.001 in random split, 1.105 ± 0.013 in cell stratified split, and 1.142 ± 0.104 in drug stratified split test datasets, showcasing its effectiveness in predicting drug responses. Performance results indicated DeepResponse successfully predicts drug sensitivity of cancer cells, and especially the multi-omics aspect benefited the learning process and yielded better performance compared to the state-of-the-art on all the splits.

An ablation study was conducted to assess the impact of each omics data type on the performance of DeepResponse, providing further insights into the importance of multi-omics integration in drug response prediction. As a use case analysis, Eprinomectin was proposed as a drug repurposing candidate against hepatocellular carcinoma cancer cell lines, which was validated in wet lab experiments. The code base, datasets, and results of DeepResponse are openly shared at https://github.com/HUBioDataLab/DeepResponse. DeepResponse can be used for early-stage discovery of new drug candidates and for repurposing the existing ones against resistant tumors.

Keywords: Bioinformatics, Multi-omics analysis, Drug response prediction, Machine / Deep learning, Hybrid architectures, Graph transformers

ACKNOWLEDGEMENTS

I would like to express my gratitude,

To my thesis advisor Prof. Tunca Doğan, who guided me by sharing his valuable knowledge and experiences throughout this challenging process,

To Umut Onur Özcan, with whom we worked together and who made a significant contribution to our project, especially with his work on the data part,

To Dr. Deniz Cansen Kahraman and Etkin Akar, who conducted in vitro experiments to validate the results of our model,

To my family and friends, whose support I have always felt throughout my thesis work.

January 2024, Ankara

Burakcan İzmirli

TABLE OF CONTENTS

FIGURES

TABLES

SYMBOLS AND ABBREVIATIONS

Symbols

$IC50$	Half-maximal inhibitory concentration
Mg+2	Divalent magnesium
$pIC50$	The negative logarithm of the IC50 value

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
aCGH	Array comparative genomic hybridization
ANN	Artificial neural network
AUC	Area under the curve
CCLE	Cancer Cell Line Encyclopedia
CDR	Cancer drug response
CGP	Cancer Genome Project
CNN	Convolutional neural network
CNV	Copy number variation
DepMap	Cancer Dependency Map
DNA	Deoxyribonucleic acid
ENet	Elastic Net
FN	False negatives
FP	False positives
GDSC	Genomics of Drug Sensitivity in Cancer

GTNN	Graph-Transformer neural network
HCC	Hepatocellular carcinoma
KBMF	Kernelized Bayesian matrix factorization
LSTMs	Long short-term memory networks
MAE	Mean absolute error
MCC	Matthew’s Correlation Coefficient
MGH	Massachusetts General Hospital
MLP	Multilayer Perceptron
mRNA	Messenger RNA
MSE	Mean squared error
NCI-60	National Cancer Institute - 60
RMSE	Mean squared error
RNA	Ribonucleic acid
RNN	Recurrent neural networks
PCC	Pearson correlation coefficient
SCC	Spearman correlation coefficient
SMILES	Simplified molecular input line entry system
SRMF	Similarity-Regularized Matrix Factorization
TARGETS	Treatment Response Generalized Elastic-Net Signatures
TN	True negatives
TP	True positives
WTS	Wellcome Trust Sanger Institute

1. INTRODUCTION

The journey to discover effective cancer drugs is a complex task that involves both scientific and economic challenges. Large financial investments are needed to develop new treatments. To address these challenges, researchers are exploring new strategies that are both scientifically effective and cost-efficient. The rapid progress in technology is making biological data more accessible and affordable, which is changing the way we use precision medicine to treat complex diseases.

In cancer research, cell lines are often used because they are efficient and cost-effective for simulating tumor tissues. Cell lines are the most used models for studying cancer biology, validating cancer targets, and for defining drug efficacy. They are collections of cells originating from one cell and are typically kept in a growth medium in tubes, flasks, or dishes, where they can continue to divide indefinitely [1].

However, a deeper understanding is achieved through pharmacogenomic panels. Pharmacogenomics is a type of genetic testing that looks for small variations within genes. These variations may affect whether genes activate or deactivate specific drugs [2]. Pharmacogenomic panels are crucial for studying the complex interactions between drugs and cell lines. They focus on decoding the specific molecular patterns of these cells, which goes beyond what traditional methods can do [2].

This research involves a detailed study of the molecular factors that influence how cancer cells respond to drugs. Using computational methods, researchers can understand the relationships within large datasets of molecular features. This not only allows them to predict how similar cell lines will respond to drugs, but also represents a new approach when empirical data is not available.

Looking forward, this research could have a transformative impact. It could lead to more personalized treatment options for patients and a more efficient process that saves time and money. This research could change the field of cancer therapeutics, moving towards a future where treatments are not only more effective, but also tailored to the unique molecular characteristics of each patient.

The process of drug development is complex and involves significant costs and time as summarized in Figure 1.1. Recent trends indicate a decrease in the approval rates of newly developed drugs. The primary challenge is the limitations of current drug response prediction methods. These methods struggle to incorporate diverse omics data types and to integrate essential drug descriptors effectively. Therefore, it is crucial for researchers to develop innovative solutions that can improve the efficiency and success rates of drug development.

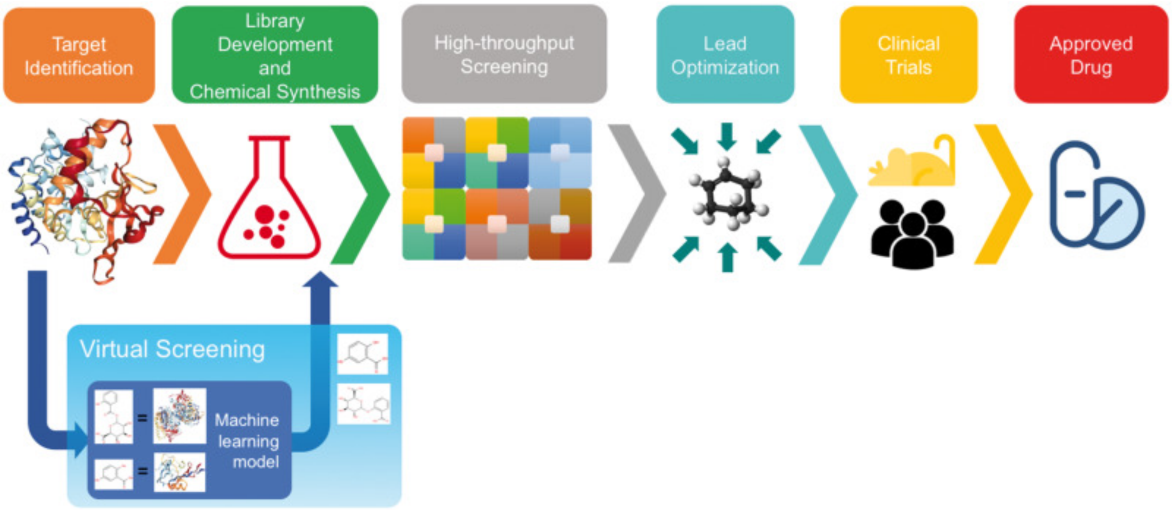

Figure 1.1. Diagram of the drug development process [3]

In vitro experiments conducted on cell lines, although informative about potential effects in patients, are costly and time-consuming. Therefore, it is not feasible to screen all samples taken from all patient and healthy tissues against all drug molecules. For this reason, the approach of expressing and statistically predicting drug effects in a computational environment is adopted. Various statistical tests and analysis methods are used for this purpose. The advantage of such computational analyses is that they can produce desired results quickly and cost-effectively. However, the disadvantage is that if the modeling process is not planned thoroughly, the reliability of the results may be low.

1.1. Unraveling Pharmacogenetics, Pharmacogenomics and Drug Development Challenges

The central objective of pharmacogenetics, which is the study of how an individual's genetic variation affects their response to specific drugs, and its intricate interplay with drug responses, has seen a transformative evolution from its inception to the present day [3]. With the advent of advanced sequencing technologies, the field, focusing particularly on pharmacogenomics in this study, has undergone a significant shift. Pharmacogenomics examines an individual's entire genetic makeup to predict their response to drugs across multiple therapeutic areas. This shift has moved the field from being merely descriptive to embracing a more forward-looking scientific paradigm [4]. Distinguishing between pharmacogenetic and pharmacogenomic fields remains challenging, leading this study to primarily consider pharmacogenomics. The difference between the two lies in their scope. Pharmacogenetics focuses on single gene variations and their impact on the response to a particular drug or group of drugs. Pharmacogenomics, however, takes a broader view by examining the entire genome or multiple genes to understand how they collectively influence drug response.

In the 1990s, groundbreaking studies foresaw a future where patients at the same disease level might exhibit diverse responses to treatments due to unique genetic variations [5]. Despite the proposal of over 150 biomarkers in research articles, the clinical applicability of fewer than a hundred has been reported. Many biomarkers, hindered by the challenge of phenotype variability, are yet to find practical applications in clinical settings [6]. Consequently, efforts have been redirected towards the standardization of consistent phenotype and biomarker definitions. This not only promises a reduction in adverse reactions but also ensures the safe and precise utilization of standardized biomarkers. [7] Moving beyond conventional methods, there is an escalating focus on developing more detailed biomarkers, extracting insights from a variety of molecular data. This strategy seeks to devise treatment alternatives customized to specific patient or tumor phenotypes, promoting more precise drug response predictions [8]. Unraveling the complex association between genetic variations and drug molecules emerges as a primary objective in the pharmacogenomic field.

Investigations in this area not only simplify the identification of drug effectiveness with reduced experimentation and increased precision, but also yield positive impacts on patient welfare, workforce productivity, clinical applications, time, and economic factors. Given the capacity to tackle a wide range of efficacy and safety concerns in drugs, researchers' interest is progressively attracted to these innovative methods [8].

The incorporation of high-throughput sequencing technologies, combining extensive amounts of biological data types, offers potential for accurately characterizing diseases in the field of pharmacogenomics. Despite the optimism for precision medicine strategies for drug treatments, the recent downturn in the approval rates and pace of new drugs has instigated a quest for alternative paths [8]. Translational research, strengthened by pharmacogenetic and pharmacogenomic methods, has infused new energy into the field of drug discovery. As a result, there is an increasing possibility of formulating methods that not only acknowledge patient variations but also design more effective treatments customized to individual profiles [8].

Cancer, historically the leading cause of disease-based deaths globally and a significant obstacle to increased life expectancy, exhibited a notable decrease in mortality between 2011 and 2015 [9] . Researchers attribute this positive trend to early diagnosis and more effectively applied treatment approaches. In this context, despite computational models encountering challenges at the data and algorithmic levels, the significance of predictive models in estimating the disease's response to medication is deemed crucial in determining the best treatment. Recently developed drug response prediction models utilizing advanced algorithm architectures emerge as tools supporting the enhancement of patients' chances of survival, ushering in a new era in the pharmacogenomic landscape [10].

1.2. Drug Response

The cultivation of cells in a controlled environment facilitates detailed examination of the genetic factors influencing drug metabolism and efficacy. These experiments, often conducted using various cell lines that represent different tissues or organs, are crucial

tools for unraveling the intricate molecular mechanisms that underlie individual responses to pharmacological interventions [11].
Furthermore, the incorporation of high-throughput technologies into in vitro studies augments their ability to provide a comprehensive view of cellular responses to drugs. This all-encompassing approach enables the detection of minor variations that may have a significant influence on drug interactions [12]. The collaboration between pharmacogenomics and in vitro experimentation equips researchers with a robust method to decipher the connections between specific variations and cellular responses, thereby aiding in the creation of predictive models for individual drug responses.

To measure drug response, a variety of techniques are employed in these in vitro experiments. Cell viability assays, such as the MTT assay or ATP-based assays, assess the overall health and viability of cells following drug exposure [13]. Transcriptomic analyses, including techniques like microarrays or RNA sequencing, enable the measurement of changes in gene expression patterns in response to drug treatment [14]. Immunoblotting and enzyme activity assays evaluate changes in protein expression levels and activity. Metabolomic profiling studies alterations in the metabolite profile of cells or tissues in response to drug exposure [15]. Advanced microscopy techniques, flow cytometry, and electrophysiological measurements further contribute to a comprehensive understanding of drug-induced effects at the cellular level [16].

Nevertheless, navigating the intricacies of in vitro experiments comes with its challenges. Replicating the complexity of the human body within a cell culture setting is inherently challenging, and researchers must carefully consider factors such as cellular microenvironments, cell types, and experimental conditions to ensure the relevance and reliability of their findings. Overcoming these challenges is crucial for the successful translation of in vitro insights into clinical applications.

The pIC50, a fundamental pharmacological metric, serves as a convenient means to express the potency of a drug in inhibiting a specific biological response [17]. It is derived from the IC50, which signifies the concentration of a drug necessary to achieve a 50% inhibition of a particular biological activity [18].

The transformation to pIC50 is achieved through a simple mathematical formula (1):

$$pIC50 = -log_{10}(IC50) \tag{1}$$

This logarithmic transformation simplifies the representation of data and aligns with the intuitive understanding of drug potency. The IC50 itself is determined through meticulous dose-response experiments, where the concentration of a drug is systematically varied, and the resulting biological response is measured. The dose-response curve is then analyzed to pinpoint the concentration at which the response is reduced by 50%.

Mathematically, the dose-response curve is often fitted using models such as the sigmoidal logistic equation (2):

$$Response = \frac{Max\ Response}{1 + 10^{(log(IC50) - log(Concentration)) \times Hill\ slope}} \tag{2}$$

Here, the sigmoidal logistic equation employs the sigmoid function, which is characterized by an S-shaped curve. The sigmoid function (3) is defined as:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

In the logistic equation, this sigmoid function is modulated by parameters such as the IC50, concentration, and the Hill slope. The IC50 represents the concentration at which the response reaches 50% of the maximum, while the Hill slope determines the steepness of the curve as can be seen in Figure 1.2. The resulting curve provides a comprehensive representation of the dose-response relationship, enabling the precise determination of the IC50 and facilitating a nuanced understanding of a drug's potency [19].

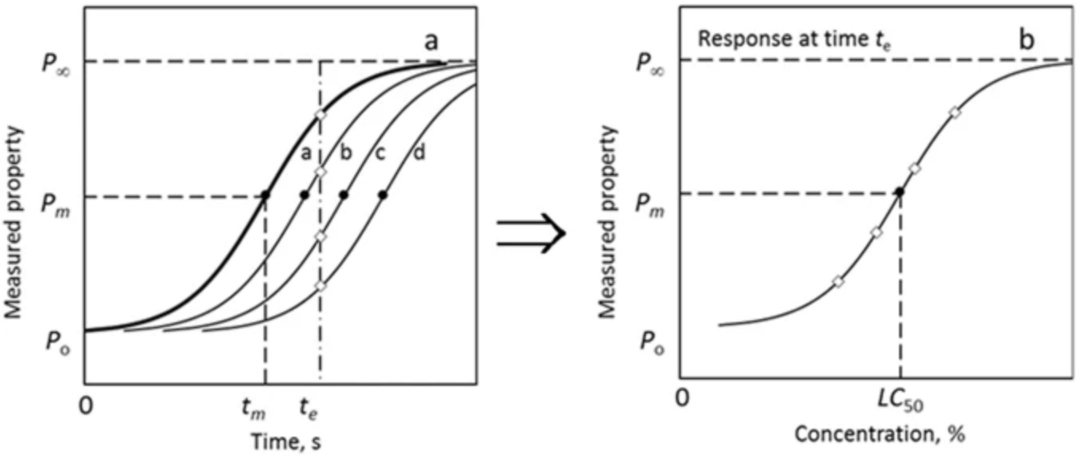

Figure 1.2. Example time-response curves (left), a dose-response curve for a fixed time is derived from these time-response curves (right) [20]

The pIC50 values, derived from this sigmoidal modeling, offer a standardized, dimensionless metric for comparing drug potencies across different compounds and experimental conditions. This logarithmic compression proves particularly valuable in managing the often wide range of IC50 values encountered in pharmacological studies. As a result, pIC50 serves as a powerful tool for researchers and practitioners, aiding in the quantification and communication of a drug's potency with clarity and precision in the context of enzyme inhibition, cellular responses, and other biological activities relevant to drug development.

1.2.1. Drug Response in Cancer Cell Lines

Cancer cell line drug response studies are a cornerstone of cancer research, providing critical insights into the intricate interplay between therapeutic agents and cancer cells. By subjecting diverse cancer cell lines to varying drug concentrations, researchers can conduct an in-depth examination of the responses, illuminating the nuances of treatment outcomes. These studies delve deeper than mere cytotoxicity evaluations [21], investigating the complex dynamics of cell cycle arrest [22], apoptosis [23], and alterations in molecular signaling pathways. Each of these processes plays a significant role in understanding the behavior of cancer cells and the impact of potential treatments. Cytotoxicity refers to a compound’s toxic effects on cells [21] , while cell cycle arrest describes a halt in the cell cycle, often due to DNA damage [22].

Apoptosis, on the other hand shown in Figure 1.3., is a programmed cell death process that is crucial for maintaining cellular balance. Together, these elements contribute to a comprehensive understanding of cancer cell behavior in response to therapeutic agents [23].

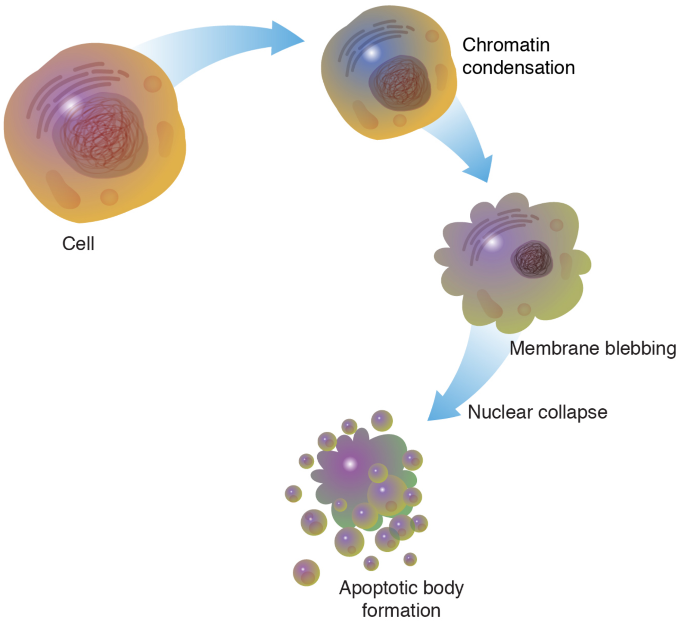

Figure 1.3. The process of apoptosis [24]

The utilization of a wide range of cancer cell lines representing various tissue origins and genetic backgrounds enables a comprehensive exploration of drug sensitivity and resistance patterns across different cancer types. This diversity mirrors the heterogeneity observed in human tumors, allowing for a more realistic emulation of clinical scenarios. Moreover, incorporating advanced technologies, such as high-throughput genomics and multi-omics approaches, facilitates a deeper understanding of the underlying genetic and molecular determinants influencing drug responses [25]. Identification of key biomarkers associated with favorable or adverse outcomes becomes paramount in guiding the development of precision therapies.

These studies also contribute to the ongoing efforts in personalized medicine, aiming to tailor treatments based on individual tumor characteristics [26].

The integration of pharmacogenomic data into cancer cell line studies allows for the identification of genetic variants that influence drug metabolism and response.
This personalized approach holds promise in guiding clinicians toward more effective and less toxic therapeutic regimens [26].

As cancer cell line drug response studies continue to evolve, they not only aid in candidate selection for clinical trials but also contribute valuable data to the broader scientific community. Collaborative initiatives, such as the Cancer Cell Line Encyclopedia (CCLE) [27], the Genomics of Drug Sensitivity in Cancer (GDSC) [28] project, and the NCI-60 [29], aggregate and disseminate large-scale drug response data, fostering a collaborative and open-access environment for researchers worldwide. These resources collectively provide information on a large number of cell lines, representing a wide range of tissues of origin and disease subtypes. The overlap of cell lines across these databases allows for data cross-validation and expansion studies. They provide extensive drug sensitivity data, which is crucial for identifying potential anticancer drugs and understanding their mechanisms of action. The comprehensive genomic data available in these databases can be used to identify cell lines with specific mutations for hypothesis-driven research.

The preparation of these cancer datasets involves several steps: cell line selection, genomic characterization, drug sensitivity testing, and data integration [27,28]. Ultimately, these studies propel the development of novel anti-cancer therapies, ushering in an era where treatment strategies are not only effective but also tailored to the unique characteristics of each patient's cancer. These resources provide a wealth of information for researchers and contribute significantly to the advancement of personalized cancer therapy. They represent a collaborative effort in the scientific community to share data and knowledge, accelerating the pace of cancer research and the development of effective, tailored treatment strategies.

1.2.2. Drug Response Prediction with Omic Data

The integration of omic data, including genomic, transcriptomic, proteomic, and metabolomic analyses, has revolutionized the study of drug responses [30].

These analyses provide a comprehensive view of the genetic basis of drug responses, gene expression patterns, functional molecules involved, and cellular physiology [31]. High-throughput technologies have been pivotal in generating the vast amounts of data needed for these analyses [32].

Genomic data, which explores an individual's DNA sequence, uncovers genetic variations that influence drug metabolism, receptor targeting, and specific response [33]. Transcriptomic analyses provide insights into dynamic gene expression patterns, which are crucial for understanding the body's molecular responses to drug exposure. Metabolomic studies provide a comprehensive view of the end products of cellular processes, elucidating metabolic pathways that are modified by drug interventions [34]. This integration of various omic layers not only enhances our comprehension of molecular complexities but also propels the advancement of personalized therapeutic interventions, which are designed based on each patient's distinct molecular signatures.

The integration of omic data is pivotal in advancing individualized drug treatments in modern medicine. Deepening our understanding of drug responses through omics reveals molecular details that govern efficacy and adverse effects, thereby unlocking potential for innovative strategies in line with precision medicine principles [35]. This shift towards personalized therapeutic interventions, tailored to an individual's unique genetic and molecular profile, signifies a new era in patient-centric healthcare. The advent of cost-effective sequencing technologies has enabled the detection of molecular changes associated with diseases, aiding in predicting drug responses in complex diseases like cancer [36].

The intricate landscape of cancer biology is influenced by a multitude of genetic variations, including mutations in gene expression, changes in methylation, and variations in copy number [37]. These variations, which include genetic deletions, insertions, translocations, and single nucleotide polymorphisms, contribute to the complex architecture of primary tumor structures. Comprehending these variations is essential for progress in areas such as cancer development and drug discovery.

Experiments that screen cancer cell lines and integrate various types of omic data emphasize the critical role these data play in determining the anticancer effects of drugs [38]. The importance of multiple omic data types in characterizing the molecular features of cell lines is further underscored by international collaborations and national research groups.

Omic data types offer invaluable insights into the complexities of cancer biology and guide the selection of therapeutic approaches. The data acquired is not only suitable for use in mechanistic models of cells but also instrumental in generating relational and correlative predictions based on machine learning. This introductory discussion paves the way for a more in-depth exploration of mutations in gene expression, methylation, and copy number variations in the context of cancer research, promising to illuminate the complex interplay of genetic variations in cancer.

1.2.2.1. Gene Expression

Gene expression, a fundamental biological process, is the mechanism by which information from a gene is used to synthesize a functional gene product, typically a protein [39]. This process involves two key steps: transcription, where the genetic information in DNA is copied into RNA, and translation, where that RNA is used to produce proteins as represented in Figure 1. 4. These proteins play diverse roles in cellular functions, influencing everything from signaling pathways to structural components [39].

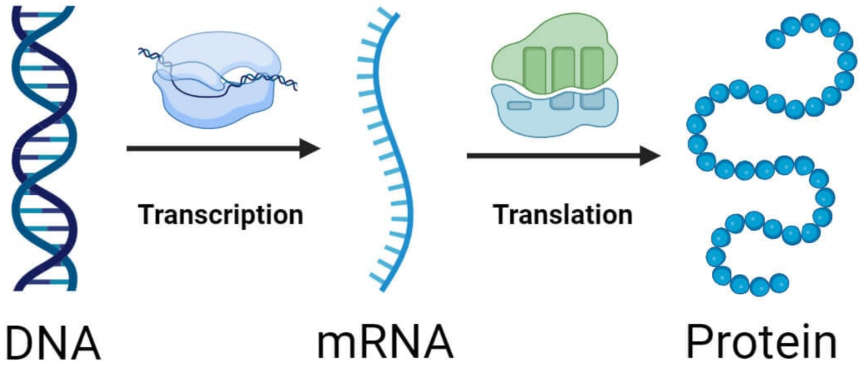

Figure 1.4. Representation of gene expression [40]

The role of gene expression in drug response is underscored by its ability to alter the activity of drug-metabolizing enzymes, thereby affecting the pharmacokinetics of a drug within the body [31]. Moreover, variations in gene expression can influence the abundance and functionality of drug target receptors, thereby affecting the efficacy of pharmacological interventions. Beyond these immediate effects, changes in gene expression can have downstream effects on cellular pathways, influencing cell cycle progression, apoptosis, and DNA repair mechanisms, all of which are crucial in determining the ultimate outcome of drug treatment.

Technologies such as RNA sequencing provide researchers with the means to examine the landscape of gene expression comprehensively [41]. Through large-scale transcriptomic analyses, it is possible to identify specific gene signatures associated with distinct drug responses. This information facilitates the categorization of individuals into responder and non-responder groups, thereby enabling personalized therapeutic strategies.

In the context of precision medicine, where the goal is to tailor treatments to individual patients, understanding gene expression patterns is of paramount importance. This understanding allows clinicians to predict how an individual will metabolize and respond to a particular drug, thereby facilitating the selection of the most effective and least toxic therapeutic regimens. As a result, studies of gene expression not only contribute to our understanding of the molecular basis of drug responses but also have practical applications in the clinical setting, bringing us closer to a future where healthcare interventions are tailored to the unique genetic makeup of each patient healthcare [42].

1.2.2.2. Mutation

Mutations refer to changes in the DNA sequence that can be either inherited or acquired. These alterations in the genetic code are pivotal determinants in shaping an individual's response to drugs represented in Figure 1.5. They can profoundly influence the efficacy, safety, and tolerability of pharmacological interventions [43]. The impact of mutations extends across various facets of drug response, including drug metabolism, target interactions, and cellular signaling pathways.

In the context of drug metabolism, genetic mutations can modulate the activity of enzymes responsible for drug biotransformation, leading to variations in the rates at which drugs are processed and eliminated from the body. Mutations in genes encoding drug target receptors may alter the binding affinity or downstream signaling events, thereby influencing the drug's effectiveness. Furthermore, mutations in key signaling pathways can introduce aberrations in cellular responses to drugs, affecting processes like apoptosis, cell cycle regulation, and DNA repair [44].

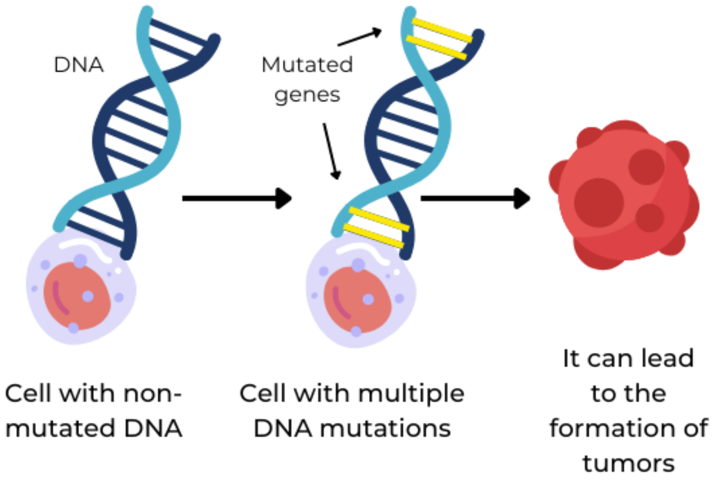
DNA
Mutated genes
Cell with non-mutated DNA
Cell with multiple DNA mutations
It can lead to the formation of tumors

Figure 1.5. Representation of gene mutations [45]

Advanced genomic technologies, such as next-generation sequencing, facilitate the comprehensive exploration of the genetic landscape underlying drug response variability. Through these approaches, researchers can identify specific mutations associated with diverse drug responses [46]. This wealth of genetic information holds promise for the development of targeted therapies tailored to individuals with specific mutation profiles. In the era of precision medicine, understanding the role of mutations in drug response is imperative. This knowledge enables clinicians to predict potential challenges in drug metabolism or target interactions based on an individual's unique mutation profile. Consequently, personalized treatment plans can be devised, optimizing therapeutic outcomes while minimizing adverse effects.

Analyses of mutations not only enhance our comprehension of the complex molecular mechanisms that dictate drug reactions, but also set the stage for a future where medical treatments are precisely adjusted to the genetic alterations unique to each patient. This ushers in a new phase in personalized and efficient healthcare.

1.2.2.3. DNA Methylation

DNA Methylation, an epigenetic modification, influences gene expression without altering the DNA sequence and is a critical factor in determining an individual's response to drugs. This biochemical process involves the addition of a methyl group to DNA as in Figure 1.6., thereby impacting the accessibility of genes for transcription [47]. Methylation patterns are fundamental in regulating gene expression, influencing various cellular functions and pathways involved in drug responses.

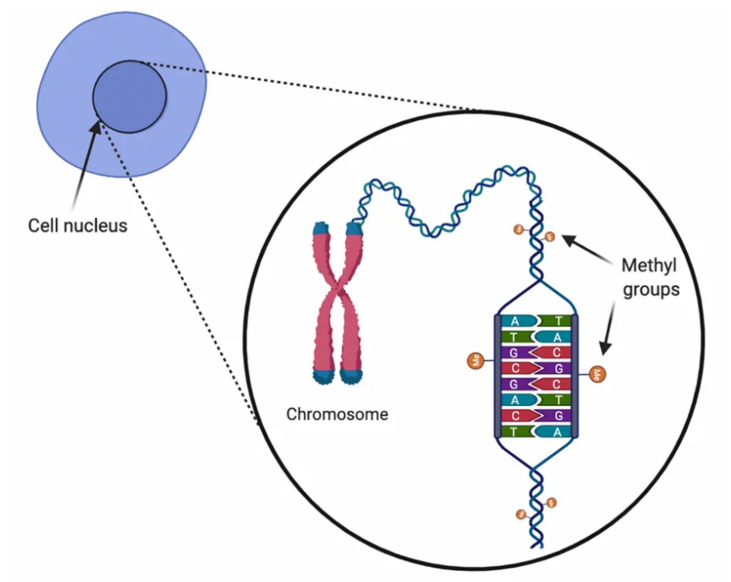

Figure 1.6. Representation of DNA methylation [48]

The significance of methylation in drug response is attributed to its capacity to modulate gene expression profiles. Changes in methylation patterns can silence or activate specific genes associated with drug metabolism, target receptors, and cellular pathways involved in drug action. Alterations in methylation status can affect the expression and functionality of drug-metabolizing enzymes, influencing drug pharmacokinetics.

Moreover, methylation changes in genes encoding drug targets may alter their expression levels or structural conformation, influencing drug efficacy [47]. Advanced epigenomic technologies, such as bisulfite sequencing, allow for comprehensive mapping and analysis of DNA methylation patterns. Deciphering these complex epigenetic signatures associated with different drug responses provides insights into individual variations in drug efficacy and toxicity.

In the context of personalized medicine, understanding methylation dynamics is essential. Methylation patterns can serve as potential biomarkers, providing valuable information for predicting an individual's response to specific drugs. This knowledge enables clinicians to tailor treatments based on an individual's methylation profile, optimizing therapeutic outcomes, and minimizing adverse effects. Therefore, methylation studies contribute to our understanding of the epigenetic mechanisms underlying drug responses and show promise in guiding the development of interventions tailored to each patient's unique epigenetic makeup. This shift towards personalized healthcare strategies represents a significant advancement where medical interventions are finely adjusted to the epigenetic signatures characterizing individual patients.

1.2.2.4. Copy Number Variation

Copy Number Variation (CNV), a type of structural genomic alteration involving changes in the number of copies of specific DNA segments, is a significant factor influencing an individual's response to drugs [49]. These variations, which can include duplications, deletions, or amplifications of genomic regions, contribute to the genetic diversity within populations and can significantly impact drug metabolism, target interactions, and cellular signaling pathways.

The influence of CNV on drug response is complex. In drug metabolism, variations in the copy number of genes encoding drug-metabolizing enzymes can directly affect enzymatic activity, influencing the rate at which drugs are processed and cleared from the body.

Similarly, changes in the copy number of genes encoding drug target receptors may alter the expression levels or structural integrity of these receptors, ultimately affecting the efficacy of pharmacological interventions. Additionally, CNV within key signaling pathways can introduce complexities in cellular responses to drugs, affecting critical processes such as apoptosis, cell cycle regulation, and DNA repair.

The study of CNV, facilitated by advanced genomic techniques like array comparative genomic hybridization (aCGH) and next-generation sequencing, allows for a comprehensive investigation of the genomic landscape underlying drug response variations [49]. By identifying specific CNVs associated with diverse drug responses, researchers can unravel the genetic underpinnings of individual variability in drug efficacy and toxicity.

In the context of precision medicine, understanding the role of CNV in drug response is essential. This knowledge enables clinicians to anticipate potential challenges in drug metabolism or target interactions based on an individual's unique CNV profile. Consequently, personalized treatment strategies can be developed, optimizing therapeutic outcomes while minimizing adverse effects. CNV analyses not only contribute to our understanding of the genetic intricacies governing drug responses but also pave the way for a future where healthcare interventions are finely adjusted to the copy number variations characterizing each patient. This shift towards personalized healthcare strategies represents a significant advancement where medical interventions are finely adjusted to the genetic signatures characterizing individual patients.

1.2.2.5. Simplified Molecular-Input Line-Entry System

In the field of computational chemistry and drug discovery, the representation of molecular structures plays a pivotal role. One common way to represent molecules is through Simplified Molecular-Input Line-Entry System (SMILES) strings [50]. This compact notation provides a string representation of a molecule’s structure. SMILES strings can be imported by most molecular editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

In terms of a graph-based computational procedure, SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. A representation can be seen in Figure 1.7. The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree [50].

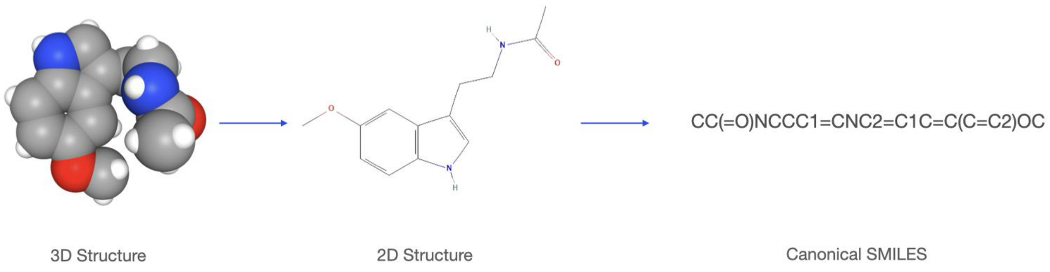

Figure 1.7. SMILES representation of melatonin molecule

SMILES has been useful in modeling quantitative structure–property/activity relationships (QSPRs/QSARs). It has also been used in diverse problems in science, technology, and medicine In conclusion, the SMILES notation system has proven to be a valuable tool in the field of computational chemistry and drug discovery, providing a compact, efficient, and versatile method for representing molecular structures.

1.3. Problem Definition

The process of discovering effective cancer drugs is a complex task that involves both scientific and economic challenges. Large financial investments are required for the development of new treatments. The primary issue lies in the limitations of current drug response prediction methods, which often struggle to incorporate diverse omics data types and effectively integrate essential drug descriptors.

In cancer research, cell lines are frequently used as they provide an efficient and cost-effective means to simulate tumor tissues. These cell lines serve as vital models for studying cancer biology, validating cancer targets, and defining drug efficacy.

However, the use of cell lines alone is not sufficient to capture the complexity of drug-cell interactions. A deeper understanding is achieved through pharmacogenomic panels, which focus on decoding the specific molecular patterns of these cells. These panels look for small variations within genes that may affect whether genes activate or deactivate specific drugs.

Despite the insights gained from these panels, the process of expressing and statistically predicting drug effects in a computational environment presents its own set of challenges. While computational analyses can produce results quickly and cost-effectively, the reliability of these results may be compromised if the modeling process is not meticulously planned. This highlights the need for careful planning and execution in computational drug effect prediction to ensure the reliability and validity of the results. Moreover, the study of the molecular factors that influence how cancer cells respond to drugs involves understanding the relationships within large datasets of molecular features. This not only allows for the prediction of how similar cell lines will respond to drugs, but also represents a new approach when empirical data is not available.

Looking forward, the development of more personalized treatment options for patients and a more efficient process that saves time and money could have a transformative impact on the field of cancer therapeutics. However, the path to this future is filled with challenges, necessitating innovative solutions that can improve the efficiency and success rates of drug development. Therefore, it is crucial to address these challenges to enhance the field of cancer therapeutics.

1.4. Aim and Scope

The central assumption of this thesis is that the response of a cell line to a drug is a complex interaction of numerous variables, with the molecular characteristics of the cell playing a significant role in shaping these responses. This assumption guides the research towards a comprehensive approach that incorporates multiple omics data types, each contributing to the detailed understanding of cellular characteristics.

This thesis proposes an approach that leverages multi-omics cell line data and drug data to enhance drug response predictions. The utilization of multi-omics data provides a more comprehensive representation of the problem, distinguishing this approach from many existing models. The goal extends beyond merely identifying gaps in current knowledge; it is a focused effort to improve drug response prediction models to achieve high levels of accuracy and reliability.

The focus of this work is on cancer data, given the abundance of available information. However, it is recognized that existing models may not be suitable for real-world test cases due to various limitations. Therefore, the objective includes a broader perspective. The aim is to develop not just a technically robust model, but also one that is generalizable, taking into account real-world scenarios and the complexities inherent in pharmaceutical research.

By addressing these challenges, this work could contribute significantly to the field of cancer therapeutics and the development of more personalized treatment options for patients. This signifies a transformative initiative set to advance the field of predictive modeling within the dynamic sphere of pharmaceutical research, reflecting a vision of setting a new standard, a benchmark for accuracy in drug response predictions.

2. RELATED WORK

In the field of drug response prediction, several models have been developed that leverage various types of data and methodologies.

In their research, Park, Lee, and Nam [51] utilized the Elastic Net (ENet) model, a deep learning model that leverages gene expression and mutation profiles of cancer cell lines to predict drug responses. This model is particularly effective when dealing with datasets where predictors are correlated. It combines both L1 and L2 regularization terms to the loss function of linear regression. They introduced an approach utilizing in vitro DNA and RNA sequencing and drug response data to create Treatment Response Generalized Elastic-Net Signatures (TARGETS). They trained TARGETS drug response models using Elastic-Net regression in the publicly available Genomics of Drug Sensitivity in Cancer (GDSC) database. Figure 2.1. shows the architecture of ElasticNet. Their study confirms the applicability of drug response prediction models for individual drugs.

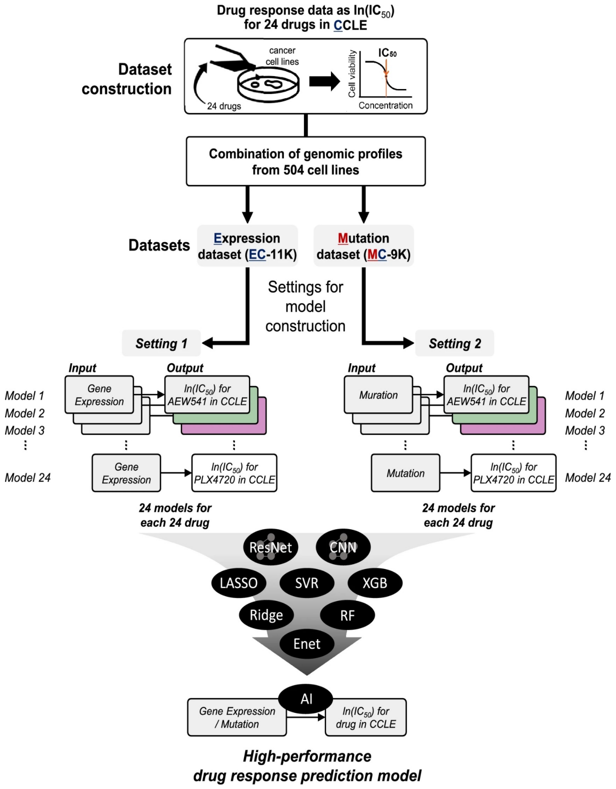
Drug response data as ln(IC50) for 24 drugs in CCLE
Dataset construction
cancer cell lines
24 drugs
Cell viability
IC50
Concentration
Combination of genomic profiles from 504 cell lines
Datasets
Expression dataset (EC-11K)
Mutation dataset (MC-9K)
Settings for model construction
Setting 1
Setting 2
Input
Output
Gene Expression
ln(IC50) for AEW541 in CCLE
Model 1
Model 2
Model 3
Model 24
Gene Expression
ln(IC50) for PLX4720 in CCLE
Input
Output
Muration
ln(IC50) for AEW541 in CCLE
Mutation
ln(IC50) for PLX4720 in CCLE
24 models for each 24 drug
24 models for each 24 drug
ResNet
CNN
LASSO
SVR
XGB
Ridge
RF
Enet
AI
Gene Expression / Mutation
ln(IC50) for drug in CCLE
High-performance drug response prediction model

Figure 2.1. Architecture of ElasticNet [51]

The authors developed a method called Similarity-Regularized Matrix Factorization (SRMF) [52] to predict anticancer drug responses of cell lines using chemical structures of drugs and baseline gene expression levels in cell lines. Structure of the SMRF represented in Figure 2.2. They incorporated chemical structural similarity of drugs and gene expression profile similarity of cell lines as regularization terms into the drug response matrix factorization model. The effectiveness of SRMF was demonstrated using a set of simulation data and compared with two typical similarity-based methods. Furthermore, it was applied to the Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) datasets. The authors also used SRMF to estimate the missing drug response values in the GDSC dataset. The drug response matrix of 23 drugs by 491 cell lines has 11,293 entries, out of which 423 (3.75%) are missing and 10,870 are known. The study concluded that the proposed data integration method improves the accuracy of prediction of anticancer drug responses in cell lines.

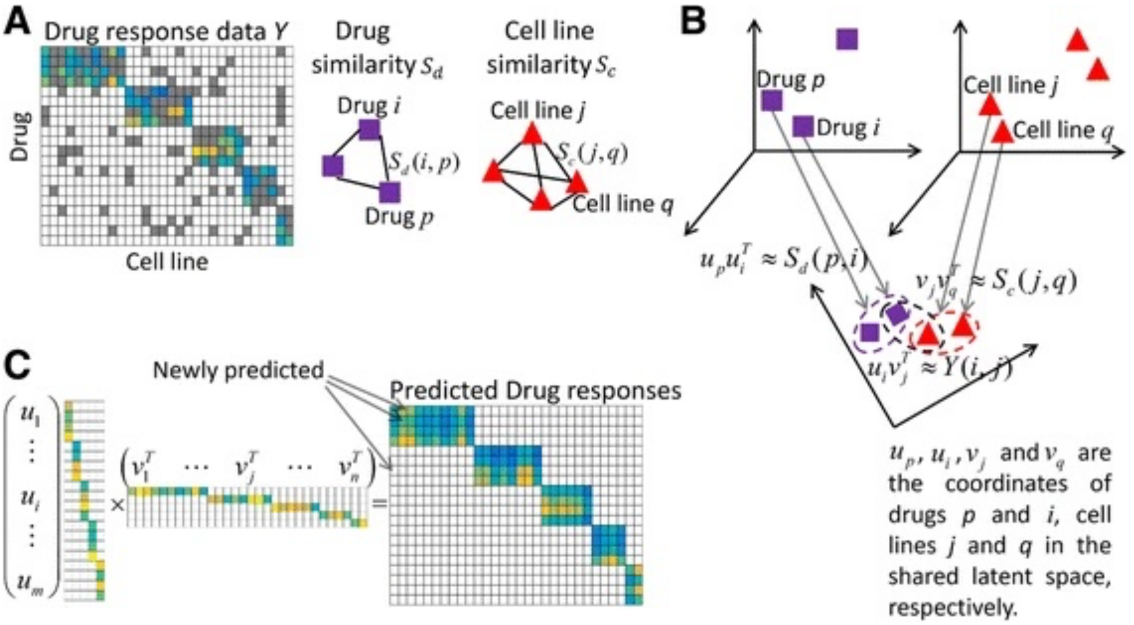
A
Drug response data Y
Drug
Cell line
Drug similarity S_d
Drug i
$S_d(i,p)$
Drug p
Cell line similarity S_c
Cell line j
$S_c(j,q)$
Cell line q
B
Drug p
Drug i
Cell line j
Cell line q
$u_p u_i^T \approx S_d(p,i)$
$v_j v_q^T \approx S_c(j,q)$
$u_i v_j^T \approx Y(i,j)$
u_p, u_i, v_j and v_q are the coordinates of drugs p and i, cell lines j and q in the shared latent space, respectively.
C
Newly predicted
Predicted Drug responses
u_1 ⋮ u_i ⋮ u_m
×
$(v_1^T \cdots v_j^T \cdots v_n^T)$
=

Figure 2.2. Structure of SMRF [52]

DrugCell, a deep learning model developed by Kuenzi BM et al. [53], trained on the responses of 1,235 tumor cell lines to 684 drugs, integrates tumor genotypes and drug structure to predict therapy response and learn the biological mechanisms underlying the drug response. The predictions made by DrugCell were accurate in cell lines and stratified clinical outcomes.

Furthermore, DrugCell was instrumental in designing synergistic drug combinations, which were validated through combinatorial CRISPR, drug-drug screening in vitro, and patient-derived xenografts. This study provides a framework for constructing interpretable models for predictive medicine, demonstrating the potential of deep learning models in predicting drug responses and understanding the underlying biological mechanisms. Architecture of DrugCell represented in Figure 2.3.

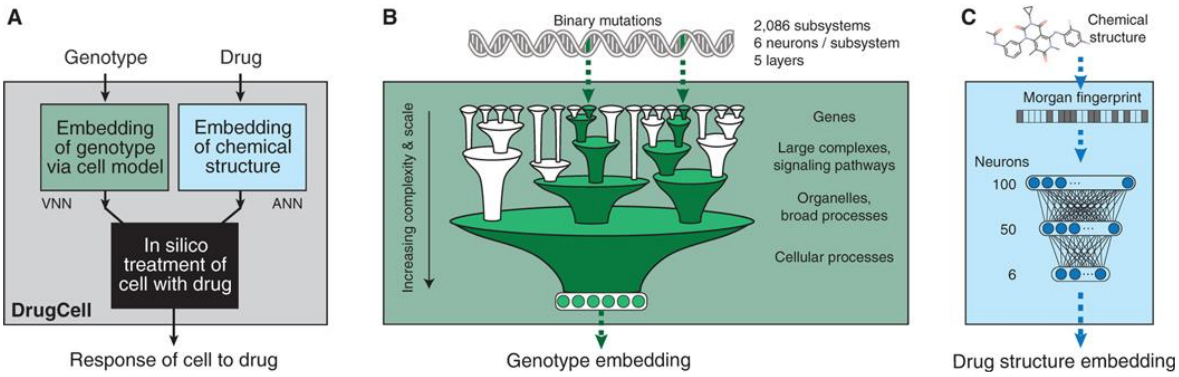

Figure 2.3. Architecture of DrugCell [53]

The paper titled “twin Convolutional Neural Network for drugs in SMILES format (tCNNS) for phenotypic screening” [54] introduces a model for predicting the phenotypic drug response on cancer cell lines. Unlike previous research that relied on molecular fingerprints or physicochemical features of drugs, this study used the simplified molecular input line entry specification (SMILES) format of drugs. The tCNNS model uses one convolutional network to extract features for drugs from their SMILES format and another convolutional network to extract features for cancer cell lines from the genetic feature vectors. A fully connected network is then used to predict the interaction between the drugs and the cancer cell lines as can be seen in Figure 2.4.

However, the performance of tCNNS decreased significantly when the training and testing sets were divided exclusively based on drugs or cell lines. Despite this, the approach was able to predict the drug effects on cancer cell lines with high accuracy, and its performance remained stable with less but high-quality data, and with fewer features for the cancer cell lines. tCNNS also provided insights into phenotypic screening and was able to solve the problem of outliers in other feature spaces. However, the performance of tCNNS dropped in the blind test.

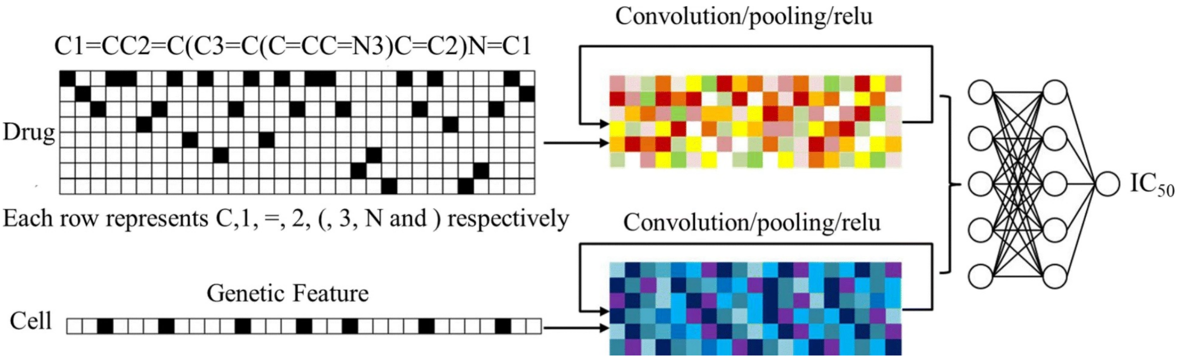
C1=CC2=C(C3=C(C=CC=N3)C=C2)N=C1
Drug
Each row represents C,1, =, 2, (, 3, N and) respectively
Genetic Feature
Cell
Convolution/pooling/relu
Convolution/pooling/relu
IC50

Figure 2.4. Architecture of tCNNs [54]

The study titled "DeepCDR: a hybrid graph convolutional network for predicting cancer drug response" [55] introduces a model called DeepCDR for predicting cancer drug response (CDR). This model integrates multi-omics profiles of cancer cells and explores the intrinsic chemical structures of drugs. Specifically, DeepCDR is a hybrid graph convolutional network consisting of a uniform graph convolutional network and multiple subnetworks. Unlike previous studies that modeled hand-crafted features of drugs, DeepCDR automatically learns the latent representation of topological structures among atoms and bonds of drugs. The authors also evaluated the contribution of different types of omics profiles for assessing drug response. Furthermore, the authors provided an exploratory strategy for identifying potential cancer-associated genes concerning specific cancer types, architecture can be seen in Figure 2.5. The results highlighted the predictive power of DeepCDR and its potential translational value in guiding disease-specific drug design.

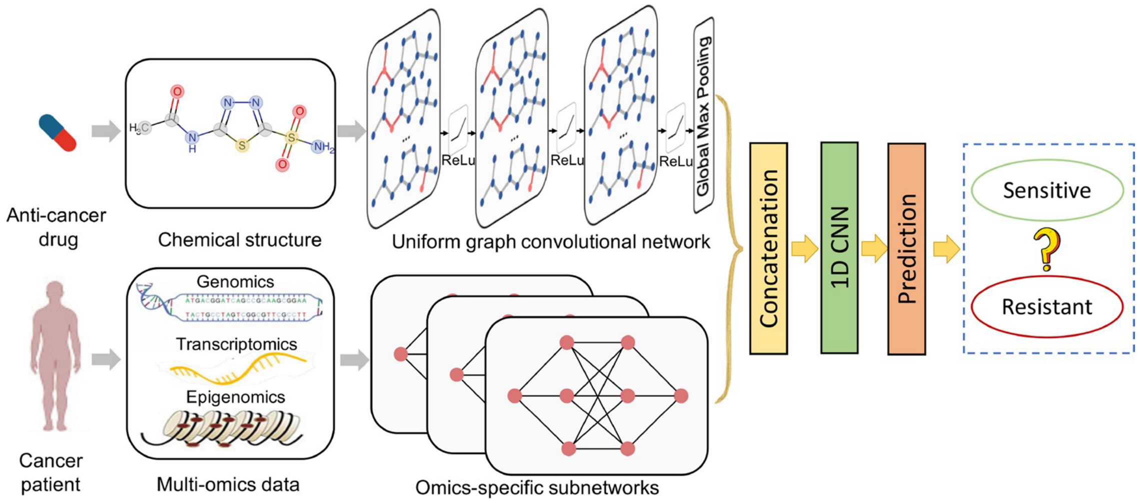

Figure 2.5. Flow of DeepCDR [55]

The paper titled "Graph Convolutional Networks for Drug Response Prediction" [56] presents a novel method, GraphDRP, for predicting drug responses. This method is based on graph convolutional networks and aims to improve upon existing machine-learning-based methods, particularly those that use deep learning. In GraphDRP, drugs are represented as molecular graphs, which directly capture the bonds among atoms. This is a departure from traditional methods that often represent drugs as strings. On the other hand, cell lines are depicted as binary vectors of genomic aberrations.The representative features of drugs and cell lines are learned by convolution layers, and then combined to represent each drug-cell line pair. The response value of each drug-cell line pair is predicted by a fully-connected neural network. The study used four variants of graph convolutional networks for learning the features of drugs. Through saliency maps of the resulting GraphDRP models, they discovered the contribution of the genomic aberrations to the responses. The authors concluded that representing drugs as graphs can improve the performance of drug response prediction.

The study titled “Dr.VAE: Drug Response Variational Autoencoder” [57] introduces a model called Dr.VAE for predicting drug responses. This model integrates binary vectors of genomic aberrations of cell lines and explores the intrinsic chemical structures of drugs. Specifically, Dr.VAE is a Variational Autoencoder that directly captures the underlying gene states before and after drug application.

Unlike previous studies that often represent drugs as strings, Dr.VAE automatically learns the latent representation of topological structures among atoms and bonds of drugs. The authors also evaluated the contribution of different types of genomic aberrations for assessing drug response. Furthermore, the authors provided an exploratory strategy for identifying potential drug-associated genes concerning specific drug-cell line pairs. The results highlighted the predictive power of Dr.VAE and its potential translational value in guiding disease-specific drug design. Architecture of Dr.VAE can be seen in Figure 2.6.

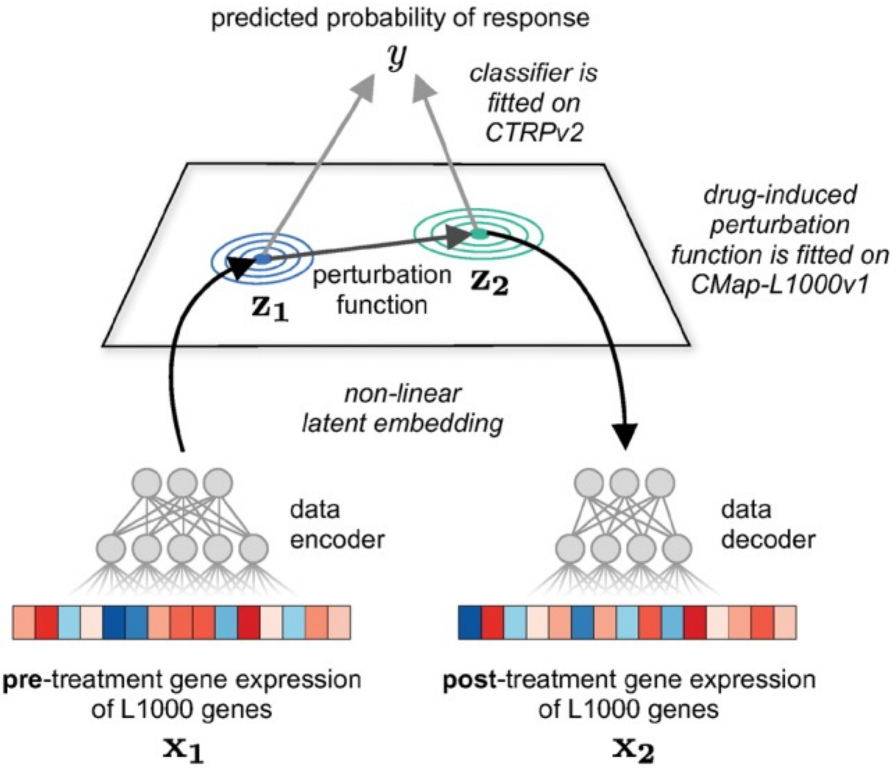
predicted probability of response
y
classifier is fitted on CTRPv2
drug-induced perturbation function is fitted on CMap-L1000v1
perturbation function
$\mathbf{z_1}$
$\mathbf{z_2}$
non-linear latent embedding
data encoder
data decoder
pre-treatment gene expression of L1000 genes
$\mathbf{x_1}$
post-treatment gene expression of L1000 genes
$\mathbf{x_2}$

Figure 2.6. Architecture of Dr.VAE [57]

In another study, the authors developed a deep learning model called CDRscan that predicts the effectiveness of anticancer drugs based on large-scale drug screening assay data. This data encompasses genomic profiles of 787 human cancer cell lines and structural profiles of 244 drugs. The model uses a two-step convolution architecture where the genomic mutational fingerprints of cell lines and the molecular fingerprints of drugs are processed individually.

These are then merged by ‘virtual docking’, an in silico modeling of drug treatment. This research represents the first-time application of a deep learning model in predicting the feasibility of drug repurposing as seen in Figure 2.7. The authors suggest that with further clinical validation, CDRscan could potentially allow for the selection of the most effective anticancer drugs based on the genomic profile of an individual patient [58].

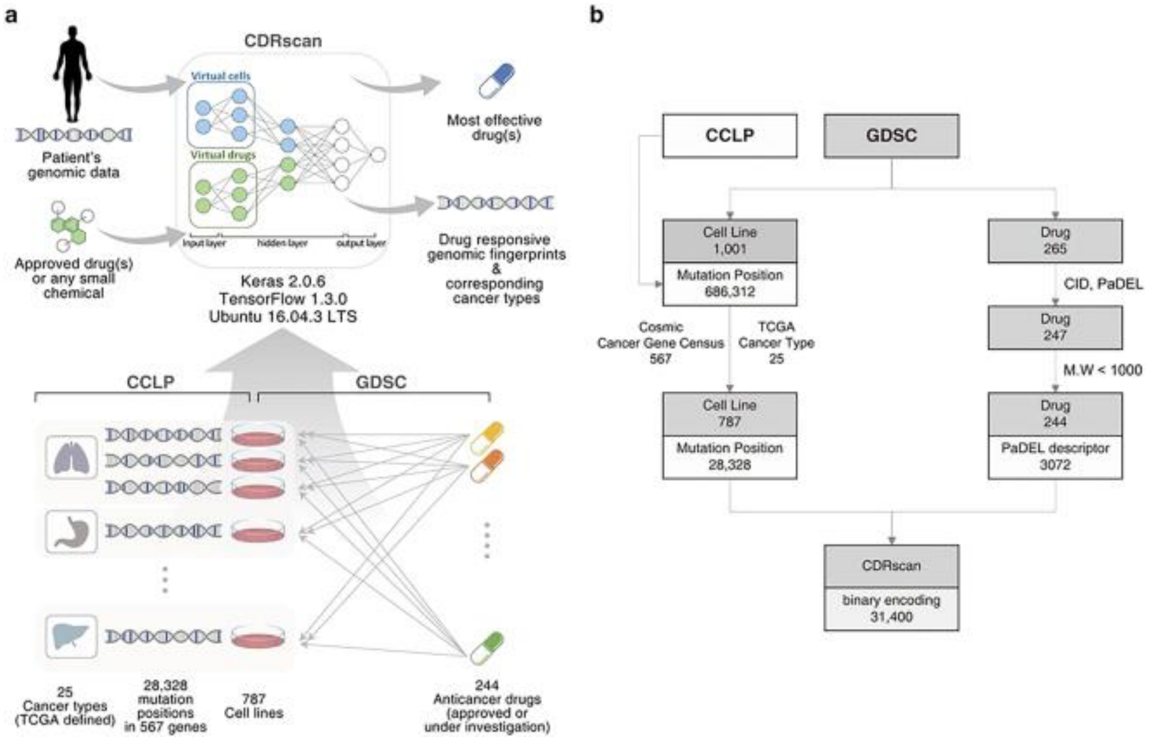

Figure 2.7. Flow of CDRscan [58]

The authors propose MOLI [59], a multi-omics late integration method based on deep neural networks. MOLI takes somatic mutation, copy number aberration, and gene expression data as input, and integrates them for drug response prediction. MOLI uses type-specific encoding sub-networks to learn features for each omics type, concatenates them into one representation, and optimizes this representation via a combined cost function consisting of a triplet loss and a binary cross-entropy loss. The authors validate MOLI on in vitro and in vivo datasets for five chemotherapy agents and two targeted therapeutics. Compared to state-of-the-art single-omics and early integration multi-omics methods, MOLI achieves higher prediction accuracy in external validations.

Moreover, a significant improvement in MOLI's performance is observed for targeted drugs when training on a pan-drug input, i.e., using all the drugs with the same target compared to training only on drug-specific inputs. Detailed architecture can be seen in Figure 2.8. The authors suggest that MOLI's high predictive power may have utility in precision oncology, potentially allowing for the selection of the most effective anticancer drugs based on the genomic profile of an individual patient.

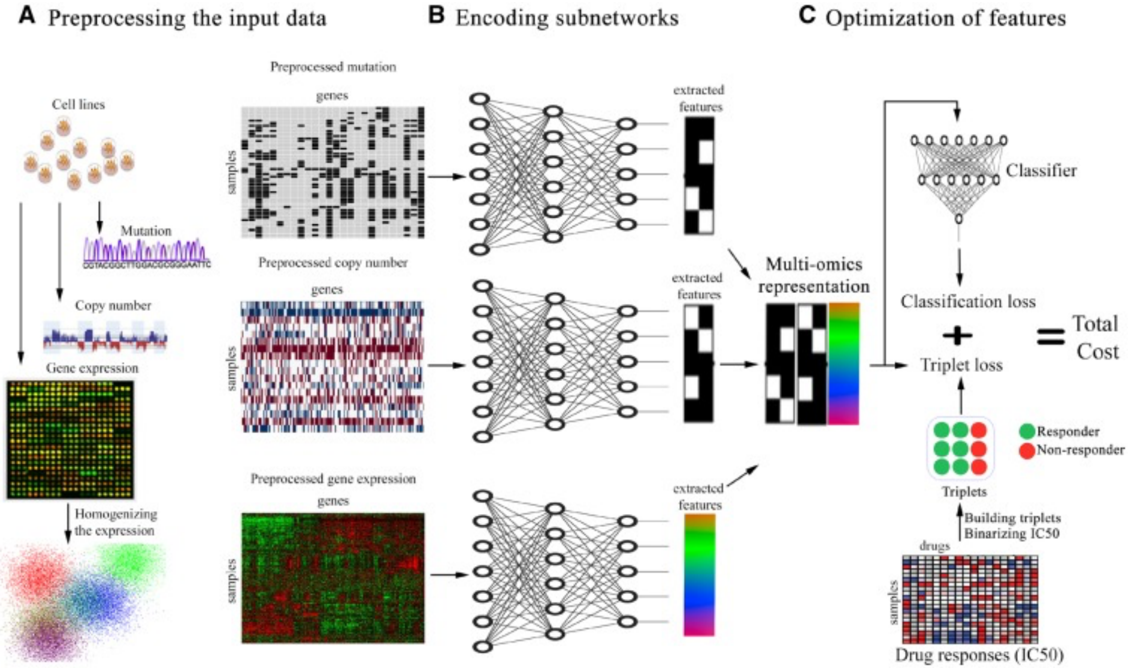

Figure 2.8. Architecture of MOLI [59]

Lastly, the authors propose a novel formulation of multi-task matrix factorization that allows selective data integration for predicting drug responses. The method, called kernelized Bayesian matrix factorization (KBMF) [60], infers pathway-response associations. KBMF uses genomic and other molecular features of samples to predict drug responses for a previously unseen sample. This is particularly valuable in oncology, where the molecular and genetic heterogeneity of the cells has a major impact on the response.

The authors demonstrate that KBMF quantitatively outperforms the state of the art on predicting drug responses in two publicly available cancer datasets as well as on a synthetic dataset. Moreover, the authors introduce a way for incorporating prior biological knowledge, in the form of pathways, for modeling pathway-drug response associations. This opens up the opportunity for elucidating drug action mechanisms and has important implications for the field of computational personalized medicine.

3. METHOD

3.1. Data

In the context of predicting drug responses in cancer cell lines, this study utilizes a diverse range of data sources. These include the GDSC, CCLE, NCI-60, DrugBank, and PubChem. Each of these projects provides a wealth of information on drug responses, genetic variations, molecular profiles, drug characteristics, and chemical compounds. Each project contributes unique and valuable data, enabling a comprehensive analysis of drug responses in cancer cells. In the following sections, the specifics of each data source will be detailed, and their contributions to the DeepResponse, in predicting drug responses in cancer cell lines will be discussed.

3.1.1. GDSC

The GDSC project, a collaborative effort between the Wellcome Trust Sanger Institute (WTS) in the UK and the Massachusetts General Hospital (MGH) in the US, emerged as a result of this partnership. While initially referred to as the Cancer Genome Project (CGP), its name was later changed. The project's data sets are available for download on its internet portal [61], and the database allows for query capabilities. The portal is designed with a user-friendly graphical interface to aid in interpreting query results.

Studies conducted between 2010 and 2015 are referred to as GDSC1, while more recent data sets are termed GDSC2. GDSC1 involved assessing drug response data using Resazurin or Syto60 assays, encompassing 987 cell lines and 367 drugs. In contrast, GDSC2 utilized the CellTiter Glo assay to analyze drug response, including 809 cell lines and 198 drug molecules, resulting in data points [28]. The most recent project version, 8.3, was released in June 2020. For the preparation of GDSC drug response data used in this thesis, the guidelines provided on the project's portal were followed. When both GDSC1 and GDSC2 data were available for the same cell line-drug pair, GDSC2 data points were preferred due to equipment and procedural enhancements.

Regarding the measurement of drug response in the project, GDSC1's procedures for assessing cell viability involved the use of Resazurin, a compound. Enzymatic activity in live cells was determined by the color change resulting from the enzymatic reduction of intracellular Resazurin. GDSC1 also employed the Syto60 assay, based on nucleic acid analysis, which involved a colorimetric approach. Syto60 binds to nucleic acid structures in live cells, causing them to turn red. This method aids in quantifying live cells under a microscope in relation to the intensity of the color emission [28]. Data points for each cell line-drug pair in the project were presented with metrics such as IC50 and the Area Under the Curve (AUC).

The GDSC2 method, known as CellTitreGlo, is an ATP-based analysis technique. This method for assessing cell viability is based on the phenomenon in which cells, following drug treatment, exhibit decreased ATP production as they approach loss of viability. The luciferase enzymes used in the analysis generate luminescence in the presence of Mg+2 (divalent magnesium) and ATP within live cells. The resulting luminescence intensity is used to determine the quantity of live cells [61].

3.1.2. CCLE

The CCLE project is the outcome of a collaborative effort between the Broad Institute and Novartis Institutes of Biomedical Research in the United States, dating back to 2006. The project's initial phases were conducted in three stages from 2008 to 2017. The first data sets were made publicly available on both the project's dedicated internet portal and the Cancer Dependency Map (DepMap) database [62]. In the period from 2018 to the present day, omics data types within the project have been consistently updated and added to the DepMap database, usually on a quarterly basis. The latest version of the project, released in the second quarter of 2022, is identified by the code 22Q2.

Although approximately 1000 cell lines had omics data extracted in CCLE, there are a total of 504 cell lines for which drug response has been measured. In contrast to other panels, CCLE tested a smaller number of drugs and conducted drug response analysis for a total of 24 drugs. Similar to GDSC2, CCLE preferred the use of the CellTiter Glo method for drug response analysis [27].

Using this method, the drug response values for 504 cell lines were measured with 24 drugs. IC50 was employed as the metric for the measured drug response values.

In CCLE, in addition to drug response data, various omics data types such as gene expression, mutation, methylation, and Copy Number Variation (CNV), along with whole-genome sequencing, whole exome sequencing, reverse-phase protein analysis, metabolomics, chromatin profiling, and gene effect analysis (gene knockout through CRISPR), have been integrated into the project. These analysis datasets are publicly available and readily accessible [62].

3.1.3. NCI-60

The NCI-60 project, initiated in the 1980s, was designed to facilitate drug discovery and create a tool that could substitute for animal models. The project's development took place in three stages, focusing on the exploration of in vitro drug response analysis methods, the development of the panel itself, and the establishment of the information technology to be used within the panel [63]. The technologies developed during this project have served as examples for other ongoing drug screening projects. Today, the NCI-60 project has become a valuable resource for researchers studying the mechanisms of growth inhibition in tumor cells.

Within the NCI-60 panel, over 130,000 drug molecules have been tested on 60 human cancer cell lines. Approximately 22,000 drug response data points are publicly available. For drug response analysis within the panel, the Sulforhodamine B (SRB) colorimetric analysis method has been preferred [29]. The amino xanthine dye used in SRB analysis binds to basic amino acids in the protein structures of live cells under low acidic conditions, emitting a pink color. The intensity of this emission is proportional to the quantity of live cells in the analyzed sample. Four different metrics have been used for drug response values within the panel, including IC50, GI50, Total Growth Inhibition (TGI), and Lethal Concentration 50 (LC50) [63].

The NCI-60 panel encompasses various omics data types in addition to drug response data, including gene expression, mutation, methylation, and Copy Number Variation (CNV).

Further, it incorporates data from whole-genome sequencing, whole exome sequencing, reverse-phase protein analysis, metabolomics, chromatin profiling, and gene effect analysis (gene knockout through CRISPR). These analysis datasets are available to the public and readily accessible [63].

3.1.4. DrugBank

The DrugBank project, established in the early 21st century, is a dynamic and comprehensive initiative committed to providing crucial information on drugs, encompassing their pharmacology, pharmacokinetics, and molecular details [64]. Serving as a vital resource for researchers, healthcare professionals, and the pharmaceutical industry, DrugBank has grown into a central repository of drug-related data.

With an impressive repository of drug information, DrugBank boasts a total of approximately 16,600 drugs, categorized as follows: 12,700 small molecule drugs, 3,900 biotech drugs, 4,400 approved drugs, including 2,760 approved small molecule drugs, 140 nutraceutical drugs, 6,720 experimental drugs, 210 illicit drugs, and 320 withdrawn drugs [65]. These structured records provide detailed insights into various aspects, including drug targets, mechanisms of action, pharmacokinetics, adverse effects, drug-drug interactions, and related pathways. The data are meticulously curated from diverse sources and are easily accessible through the DrugBank website [65].

In addition to its wealth of data on approved drugs, DrugBank provides valuable insights into experimental drugs and drug candidates, fostering research into potential new treatments. The platform facilitates exploration into chemical structures, physiological effects, and known drug interactions, broadening its utility across a wide spectrum of applications.

DrugBank's database is regularly updated to incorporate the latest advancements in pharmacology and the development of new drugs. With a substantial number of updates per month, this ensures that the information remains current and reflective of the ever-evolving landscape of drug-related knowledge.

Its user-friendly interface, coupled with this commitment to currency, solidifies DrugBank as an invaluable tool for researchers and healthcare practitioners seeking up-to-date drug-related information for various purposes. The project continues to play a pivotal role in advancing drug discovery and healthcare practices.

3.1.5. PubChem

PubChem is an extensive resource established to facilitate access to comprehensive information about chemical compounds, with a primary focus on their biological activities and applications [66]. PubChem has become a pivotal reference tool for researchers, chemists, and life scientists.

One of the fundamental aspects of PubChem is its vast collection of chemical data. With over 116 million unique chemical structures extracted from contributed PubChem Substance records, it serves as a rich source of information for researchers exploring the intricacies of chemical compounds. Additionally, PubChem provides information about more than 310 million chemical entities, aggregating data from various contributors and ensuring accessibility through the PubChem website [66].

PubChem's database extends beyond chemical structures to encompass a broad spectrum of biological data. With 1.6 million biological experiments in the BioAssays collection and an impressive 293 million data points on biological activities reported in PubChem BioAssays, it proves indispensable forearchers seeking a comprehensive understanding of the biological and pharmacological properties of chemical compounds.

The biological landscape covered by PubChem is vast, including 113,000 gene targets, 186,000 protein targets, and 114,000 organisms tested in PubChem BioAssays and involved in PubChem Pathways. Moreover, PubChem provides insights into the interactions between chemicals, genes, and proteins through its collection of 241,000 pathways [67].

Beyond biological data, PubChem also hosts information about cell lines (2,000 entries) and is a valuable repository of scientific knowledge, with links to 39.6 million scientific publications and 37.9 million patents [66].

The project encompasses 70 data classifications, allowing users to browse the distribution of PubChem data among nodes in the hierarchy of interest, and draws from 940 organizations contributing data to PubChem.

Furthermore, PubChem continually updates its databases, with daily additions and revisions to ensure the data remain current and reflective of the latest advancements in the field of chemistry and biochemistry. The user-friendly interface and extensive data make PubChem an indispensable resource for researchers and professionals in various fields, contributing significantly to the advancement of chemical and life sciences.
The project has established itself as a cornerstone in chemical research and is crucial for the discovery of new compounds, the development of pharmaceuticals, and research into the biological activities of chemicals.

3.1.6. Overview of the Datasets

The raw data that forms the basis of the analyses in this study was initially collected in the context of a separate research project, as referenced in [68]. Additionally, the essential operations and preliminary analysis on the data were performed as part of the same previous study. This study has adopted the same datasets, using them as basis for further exploration and analysis. Table 3.1. provides a comprehensive overview of the datasets used in the research, specifically detailing the number of drugs included in each database, the type of drug response analysis used, the duration of treatment with the drug, and the metrics used to measure drug response. It offers a snapshot of the scope and methodology of each database, namely GDSC, CCLE, and NCI-60.

Metrics / Database	**GDSC**	**CCLE**	**NCI-60**
Number of Drugs	518 (367 in GDSC1, 198 in GDSC2)	24	1054
Type of Drug Response Analysis	Resazurin or Syto 60 (GDSC1), CellTiter Glo (GDSC2)	CellTiter Glo	SRB
Treatment Duration with Drug	72 hours	72 hours	48 hours
Metrics of Drug Response	IC50, AUC	IC50, AUC	IC50, GI50, TGI, LC50

Table 3.1. Data Sources and Features

Table 3.2. presents the types of data and features available in each database, specifically focusing on gene expression, mutation, methylation, and copy number variation. It outlines the number of cell lines and genes included under each data type for each database. This table provides a detailed view of the genetic information available in each database, which is crucial for understanding drug sensitivity in cancer research.

Omic / Database		**GDSC**	**CCLE**	**NCI-60**
Gene Expression	Number of Cell Lines	1018	1088	60
	Number of Genes	17737	19851	23059
Mutation	Number of Cell Lines	1032	1570	60
	Number of Genes	21972	19286	443
Methylation	Number of Cell Lines	1080	1089	60
	Number of Genes	19864	19880	17553
Copy Number Variation	Number of Cell Lines	986	1754	60
	Number of Genes	24502	25368	19951

Table 3.2. Number of cell lines and genes in each database

The tables highlight the distinct characteristics and strengths of the GDSC, CCLE, and NCI-60 databases in cancer research. GDSC, with its extensive collection of cell lines and drugs, appears to be a comprehensive resource for studying a wide range of cellular responses to various drugs. Its two versions, GDSC1 and GDSC2, further add to its versatility. On the other hand, CCLE, despite having fewer drugs, provides a rich source of genetic information across a substantial number of cell lines. This makes it a valuable database for studying the genetic basis of drug responses.NCI-60, while having the least number of cell lines, stands out for its vast array of drugs tested. This, coupled with a unique set of drug response metrics, makes NCI-60 a potent tool for high-throughput screening of drug responses. In terms of genetic data, the tables underscore the depth of information available across different ‘omics’ categories in each database. The variation in the number of genes studied under each category across the databases indicates the diverse focus areas of each resource. Overall, these databases collectively offer a wealth of information, each with its unique strengths, catering to various facets of cancer research. The choice of database would thus depend on the specific requirements of the research question at hand.

3.2. Data Imputation

Data imputation is a crucial step in data preprocessing, especially when the quantity of data is limited. Each piece of data holds potential value for the research, and losing any part of it could lead to a significant loss of information. The imputation process is guided by an examination of the column-wise and row-wise proportions. These proportions represent the ratios of missing values in each column and row, respectively.

Based on these analyses, different imputation approaches are implemented. These approaches are tailored to the specific characteristics of the data and are designed to preserve the underlying data distribution and relationships. The implementation of these approaches is summarized in Figure 3.1, which provides a pseudocode representation of the imputation process.

```
IF column wised proportion $< 0.1$ and row wised proportion $< 0.5$ THEN
  FOR each type of omics DO
    IF type is 'Gene Expression' OR 'Methylation' THEN
      Fill missing values with the mean of the data
    ELSE IF type is 'Mutation' THEN
      Fill missing values with zero
    ELSE IF type is 'Copy Number Variation' THEN
      Fill missing values with the median of the data
    END IF
  END FOR
END IF
```

Figure 3.1. Pseudocode of data imputation

For Gene Expression data, the chosen imputation approach involved filling missing values with the mean of the available data. This method seeks to maintain the statistical properties of the dataset while compensating for incomplete information. Gene expression profiles are highly dynamic, making the mean a suitable proxy for the missing values.

In the context of Mutation data, a different strategy was deployed, wherein missing values were replaced with zero, symbolizing the absence of any mutation. This approach is grounded in the understanding that a lack of mutation data signifies a non-altered state, which is adequately represented by a value of zero. Methylation data, like Gene Expression, underwent imputation through the utilization of mean values. Here, missing values were filled with the mean of the available data points, a method well-suited to preserving the underlying patterns in methylation levels. The Copy Number Variation data, on the other hand, was imputed using the median of the available data. This choice is informed by the nature of copy number variations, which often exhibit skewed distributions. The median, being a robust measure of central tendency, provides a balanced representation of the data in the presence of potential outliers.

3.3. Data Manipulation

In data analysis, especially when working with data from multiple sources, data manipulation is considered a crucial step. This process ensures the reliability, consistency, and suitability of the final dataset for further analysis. In this study, the foundation of the dataset was based on data collected from various sources. This data was organized into three main columns: cell line name, drug name, and pIC50 values.

The initial step in the data manipulation process involved standardizing the data. This was essential to achieve uniformity across the dataset. Variations in data entries, such as the use of lowercase/uppercase or the inclusion of special characters like dashes, underscores between cell line names, were identified. A common format was applied to all data entries to make the data consistent. Following standardization, the data from different sources was merged. This was performed by using common identifiers present in the various data sources. The outcome was a comprehensive dataset that served as the raw version of our dataset. After the raw dataset was created, the next step was to integrate features from both the drugs and cell lines. For the drugs, the Simplified Molecular Input Line Entry System (SMILES) was used as a representation. The SMILES data, which was collected from another source, was then merged with the raw dataset using the drug names as a reference.

Regarding the cell line features, each cell line was associated with multiple genes, which required the application of a nested structure. Each cell line was represented by a dataframe of either (16501,4) or (897,4), depending on the selected gene subset. 16501, 897 is the number of genes in the selected cell line and 4 represents multi omics (Gene Expression, Mutation, Methylation, Copy Number Variation) respectively. This data, also collected externally, was subsequently merged into the raw dataset. The resulting comprehensive dataset, enriched with drug and cell line features can be seen in Table 3.3.

Drug Name	SMILES	Cell Line Name	Cell Line Features	pIC50
Rapamycin	CC1CCC2 .. OC	MHH-ES-1	[[4.188, 0.0, 0.0, 4.0], … [5.45, 0.0, 0.0, 4.0]]	6.966429
FH535	CC1=C(C .. Cl	RERF-LC-Sq1	[[10.05, 0.0, 0.0, 4.0], … [7.89, 0.0, 0.0, 3.0]]	3.806829
Enzastaurin	CN1C=C .. N7	JHU-011	[[10.08, 0.0, 0.0, 4.0], … [8.03, 0.0, 0.0, 3.0]]	5.150244
Refametinib	COC1=CC ..)F	NEC8	[[8.27, 0.0, 0.0, 3.0], … [8.98, 0.0, 0.0, 2.0]]	5.291252
Pazopanib	CC1=C(C ..)N	NCI-H1869	[[8.89, 0.0, 0.0, 3.0], … [9.81, 0.0, 0.0, 2.0]]	3.936815
TWS119	C1=CC(.. 4)O	BPH-1	[[8.625, 0.0, 0.0, 4.0], … [10.29, 0.0, 0.0, 4.0]]	5.625702
Venotoclax	CC1(CCC ..)C	SK-HEP-1	[[9.18, 0.0, 0.0, 3.0], … [9.664, 0.0, 0.0, 2.0]]	5.654595
Trichostatin A	CC(C=C ..)C	SK-PN-DW	[[9.29, 0.0, 0.0, 1.0], … [8.586, 0.0, 0.0, 2.0]]	7.019587
AZD8186	CC(C1 ..)F	KARPAS-1106P	[[8.83, 0.0, 0.0, 2.0], … [10.375, 0.0, 0.0, 2.0]]	5.244440
YK-4-279	COC1=C .. l)O	FADU	[[8.625, 0.0, 0.0, 4.0], … [8.97, 0.0, 0.0, 3.0]]	5.236225

Table 3.3. The first ten rows in the dataset

3.4. Gene Set Organisation

Different data types were required to perform different tests and compare the model to existing ones. Data types can be differentiated by the combinations of data source, gene types, pathway knowledge and tissue type.

The L1000 technology is a cost-effective, high-throughput transcriptomics technology. It has been used to profile a collection of human cell lines for their gene expression response to more than 30,000 chemical and genetic perturbations [69]. In total, there are currently over 3 million available L1000 profiles. The L1000 assay measures the mRNA expression of 978 landmark genes, while 11,350 additional genes are computationally inferred [69]. Therefore, the 'L1000' dataset is created by filtering out only the L1000 gene names from the 'Normal' dataset. By focusing on these key genes, the 'L1000' dataset allows for efficient data handling while still maintaining a high level of precision in model testing and comparison.

Pathway knowledge refers to the understanding of how genes interact in biological pathways. These pathways provide common conceptual models which explain groups of chemical reactions within their biological context [70]. Visual representations of the reactions in biological pathway diagrams provide intuitive ways to study the complex metabolic processes. In order to link (clinical) data to these pathways, they have to be understood by computers. Understanding how to move from a regular pathway drawing to its machine-readable counterpart is pertinent for creating proper models. This is followed by three examples of bioinformatics applications including a pathway enrichment analysis, a biological network extension, and a final example that integrates pathways with clinical biomarker data.

In the pathway-sorted datasets, genes that belong to the same pathway are arranged consecutively. This arrangement is particularly beneficial when using Convolutional Neural Networks (CNNs) [71] for analysis. The CNN can take advantage of this ordering to extract features by looking at a larger window of consecutive genes. This allows the CNN to capture the relationships and patterns within the same pathway, potentially leading to more accurate and insightful model predictions.

The last type of dataset is referred to as the 'Digestive System' dataset. This dataset is derived by filtering the 'Normal' dataset based on tissue types, specifically focusing on tissues from the digestive system. This allows for a more focused study on the genes of digestive system cell lines and their corresponding drugs.

This targeted approach facilitates more precise analysis and model testing within the specific context of the digestive system. All the dataset and corresponding info was summarized in Table 3.4.

Dataset Type	Available Sources	Features
Normal	GDSC, CCLE, NCI-60	All the genes of all the cell lines and corresponding drugs.
L1000	GDSC, CCLE, NCI-60	Only L1000 genes of all the cell lines and corresponding drugs.
Pathway	GDSC	All the genes of all the cell lines ordered by pathway and corresponding drugs.
Pathway Reduced	GDSC	Only L1000 genes of all the cell lines ordered by pathway and corresponding drugs.
Digestive System	GDSC	All the genes of digestive system cell lines and corresponding drugs.

Table 3.4. Data types and features

3.5. Model Construction

3.5.1. Convolutional Neural Network

Convolutional Neural Networks (CNNs) [71] are a type of Artificial Neural Network (ANN) that is particularly suited for spatial data. CNNs are inspired by the biological processes of the visual cortex of living beings. They are composed of multiple layers, including convolutional layers, non-linearity layers, pooling layers, and fully-connected layers as in Figure 3.2. The convolutional and fully-connected layers have parameters, while the pooling and non-linearity layers do not.

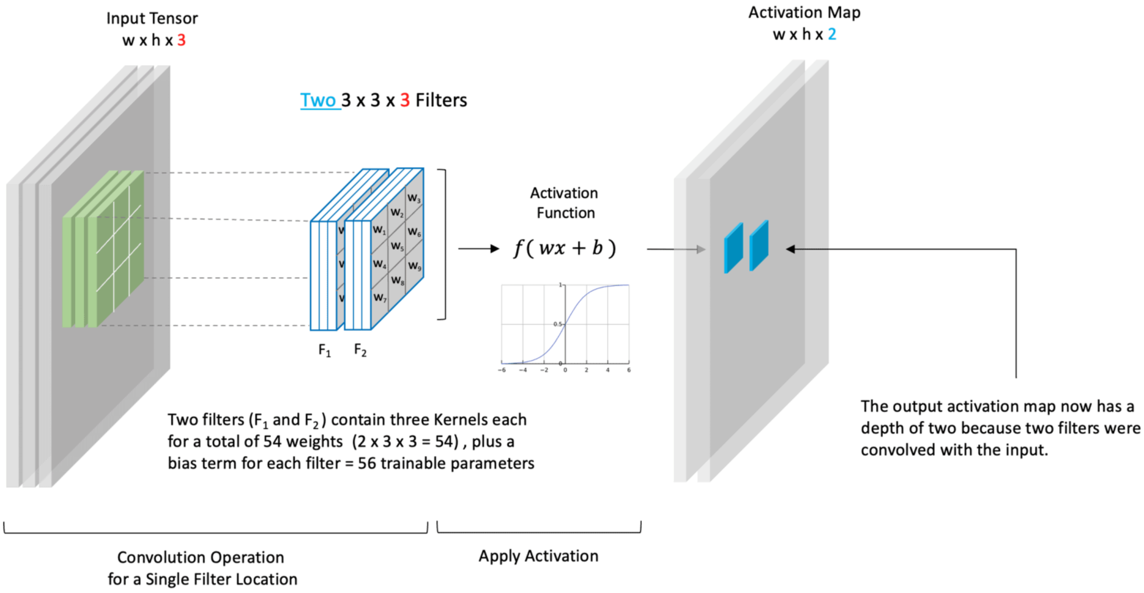
Input Tensor
w x h x 3
Two 3 x 3 x 3 Filters
Activation Map
w x h x 2
Activation
Function
$f(wx + b)$
F_1
F_2
W_1
W_2
W_3
W_4
W_5
W_6
W_7
W_8
W_9
Two filters (F_1 and F_2) contain three Kernels each
for a total of 54 weights (2 x 3 x 3 = 54) , plus a
bias term for each filter = 56 trainable parameters
The output activation map now has a
depth of two because two filters were
convolved with the input.
Convolution Operation
for a Single Filter Location
Apply Activation

Figure 3.2. Representation of a convolutional layer [72]

The Convolutional Neural Network (CNN) is one of the sub deep learning components of DeepResponse, designed to analyze multi-dimensional cell line data effectively. Given the diverse nature of multi-omics and pathway-ordered data for cell lines, the CNN processes gene expression, mutation, methylation, and copy number variation features. Each gene in the cell line is analyzed by the CNN, which uses a series of 2D convolutional layers to extract hierarchical features and patterns within the molecular structures with the help of max pooling layers as can be seen in the (4).

$$Y_{i.j} = max_{m.n}\{X_{m.n}\} \tag{4}$$

$Y_{i.j}$: $Value\ of\ the\ output\ feature\ map\ at\ position(i,j)(i,j)$

$X_{m.n}$: $Value\ of\ the\ input\ feature\ map\ at\ position(m,n)$

$max_{m.n}$: $Maximum\ value\ over\ all\ the\ positions\ m\ and\ n\ within\ the\ pooling\ window$

The strength of CNNs in analyzing cell line data is not only due to their ability to discern complex patterns within the molecular structures, but also their effective use of batch normalization layers. These layers, strategically placed within the network, normalize, and stabilize the activations from the convolutional layers.

This ensures a more reliable and consistent learning process, even when dealing with the diverse nature of multi-omics and pathway-ordered data for cell lines.

Equation (5) calculates the mean value β of the batch, where x_i are the individual elements of the batch and m is the batch size.

$$\mu_\beta = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (5)$$

Then it computes as in equation (6) the variance B^2 of the batch, which measures how far each number in the set is from the mean B.

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \qquad (6)$$

Equation (7) normalizes each element x_i of the batch by subtracting the mean B and dividing by the square root of the variance B^2 plus a small constant ϵ to prevent division by zero.

$$\widehat{x_\iota} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \qquad (7)$$

Last step is scaling and shifting the normalized value ($\widehat{x_\iota}$) using learnable parameters γ (scale) and β (shift) to produce the final output yiof the batch normalization process.

$$y_i = \gamma \widehat{x_\iota} + \beta \qquad (8)$$

In the context of analyzing multi-dimensional cell line data, one distinctive architectural decision in the CNN component of DeepResponse is the utilization of 2D global average pooling in the final layer. This approach goes beyond conventional flattening layers, reflecting a nuanced understanding of the challenges posed by molecular data. Global average pooling not only helps prevent overfitting but also serves as a powerful mechanism for distilling essential information from the diverse nature of multi-omics and pathway-ordered data for cell lines.

By taking the average of each feature map, global average pooling retains critical information while significantly reducing the dimensionality of the data. This contributes to the model's interpretability and generalization capabilities, enhancing its effectiveness in analyzing cell line data.

3.5.2. Graph-Transformer Based Neural Network

Graph-Transformer Based Neural Networks (GTNNs) [73] are a novel approach to representation learning on graphs. GTNNs are capable of generating new graph structures, identifying useful connections between unconnected nodes on the original graph, and learning effective node representation on the new graphs in an end-to-end fashion. This approach allows GTNNs to learn new graph structures based on data and tasks without domain knowledge.

A generalization of transformer neural network architecture for arbitrary graphs has been proposed, which operates on fully connected graphs representing all connections between the words in a sequence [74]. This architecture leverages the graph connectivity inductive bias and can perform well when the graph topology is important and has not been encoded into the node features.

The Graph-Transformer Neural Network (GTNN) within DeepResponse represents an innovative integration of two significant paradigms in deep learning: message passing neural networks and transformer encoders. This integration is not just a combination of methodologies, but a strategic fusion that results in a model capable of understanding the complex structural details inherent in drug molecules. The operation of the GTNN is characterized by a two-step process, each step playing a crucial role in comprehending molecular relationships. The difference between them was represented in Figure 3.3.

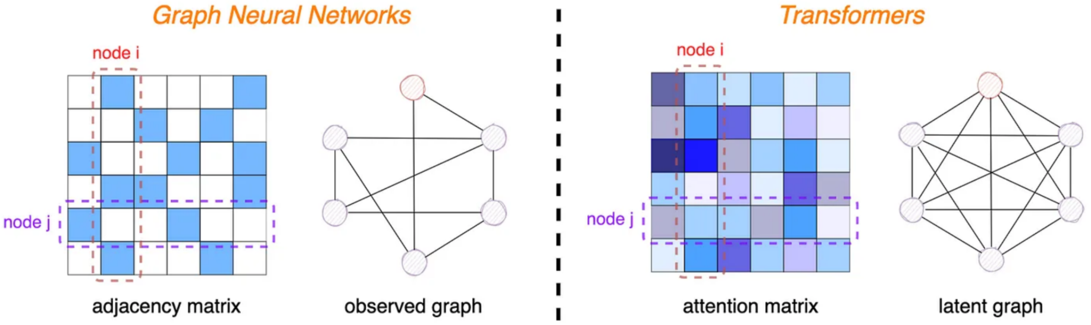

Figure 3.3. Comparison between Graph Neural Networks (GNN) and Graph Transformer Neural Networks (GTNN) [75]

In the first step of the GTNN operation, dynamic message passing layers are employed (9). These layers construct a graph from the drug molecule data, creating sub-edge networks that aggregate information from neighboring atoms. This process allows the GTNN to adapt to the inherent complexities of molecular structures. The subsequent use of gate recurrent unit layers further enhances the model's ability to capture the dynamic nature of molecular interactions. This ensures that the GTNN is not merely a static model, but a responsive and adaptive tool in the field of molecular property prediction, capable of comprehending the intricate structural details inherent in drug molecules.

$$h_i^{(l+1)} = \sigma\left(W^{(l)}h_i^{(l)} + \sum_{j\,\varepsilon\,N_i} M^{(l)}(h_i^{(l)}, h_j^{(l)})\right) \tag{9}$$

Here, $h_i^{(l)}$ represents the feature vector of node i at layer l, N_i is the set of neighboring nodes of node i, $W^{(l)}$ and $M^{(l)}$ are learnable parameters, and σ is a non-linear activation function.

In the second step of the GTNN operation, transformer encoder layers are incorporated, representing a significant advancement in the GTNN's architecture (10). These layers build upon the foundation established by the dynamic message passing layers, enhancing the model's ability to understand intricate dependencies and long-range interactions within drug molecules.

The transformer architecture, well-known for its effectiveness in natural language processing tasks, is applied to the molecular domain, where it serves as a key component in the model's operation. Each transformer encoder layer contributes to a hierarchical abstraction, enabling the GTNN to capture complex relationships and dependencies in a way that surpasses conventional neural network architectures.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

Here, $Q, K,$ and V are the query, key, and value matrices respectively, and d_k is the dimension of the key.

3.5.3. Multi-Layer Perceptron

Multi-Layer Perceptrons (MLPs) are a class of artificial neural networks that have gained significant attention in the field of machine learning [75]. MLPs consist of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The nodes, or "neurons", in these layers apply a set of weights to the inputs and pass them through a non-linear activation function (11,12).

$$z_k = \sum_{i=1}^{n} w_{ik} \times x_i + b_j \tag{11}$$

$$y_k = f(z_k) \tag{12}$$

z_k: $Sum\ of\ weighted\ inputs\ and\ bias\ to\ the\ neuron\ k$

n: $Neuron\ count\ of\ the\ previous\ layer$

w_{ik}: $Weight\ of\ the\ connection\ between\ neuron\ i\ and\ k$

x_i: $Output\ of\ neuron\ i\ that\ will\ be\ input\ to\ neuron\ k$

y_k: $Output\ of\ neuron\ k$

f: $Activation\ function$

MLPs are known for their ability to solve problems that are not linearly separable, and they are widely used in applications such as image recognition, speech recognition, and machine translation.

The backpropagation algorithm is commonly used for training MLPs, adjusting the weights of the neurons by minimizing the error between the predicted and actual outputs, and MLP was illustrated in Figure 3.4.

The Multi-Layer Perceptron (MLP) is the last but not least sub deep learning algorithm in DeepResponse. Following the Convolutional Neural Network (CNN) and the Graph-Transformer Neural Network (GTNN), the MLP is responsible for synthesizing and consolidating the information processed by the preceding stages. The Multi-Layer Perceptron (MLP) in DeepResponse utilizes a combination of hidden layers, dropout mechanisms, and selected activation functions. This allows the MLP to process the information from the molecular landscape.

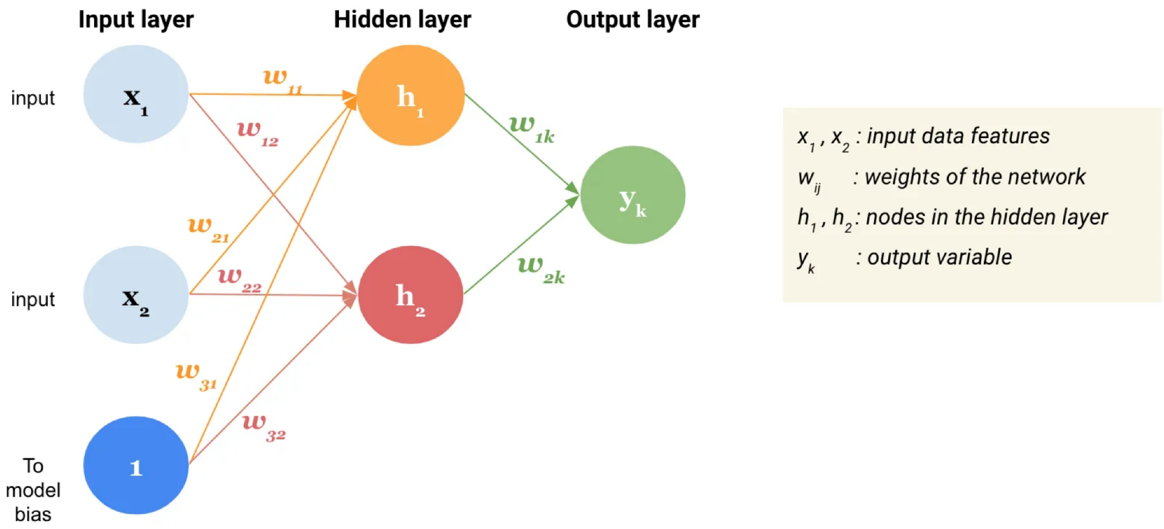

Figure 3.4. An illustration of Multi-Layer Perceptron [76]

The Multi-Layer Perceptron (MLP) in DeepResponse employs hidden layers to process the complex patterns within the molecular data. Each hidden layer uses Rectified Linear Unit (ReLU) activation functions (13), introducing non-linearity into the model. This allows the MLP to capture intricate relationships and nonlinear dependencies, enabling it to effectively process the information from the molecular landscape.

$$f(x) = max(0, x) \tag{13}$$

In addition to hidden layers, the MLP incorporates dropout mechanisms to prevent overfitting. These mechanisms randomly deactivate a fraction of neurons during training, introducing an element of robustness into the model. This prevents the MLP from relying too heavily on specific neurons, enhancing its ability to generalize.

3.5.4. Architecture of the DeepResponse

The development of a generic and well-performing algorithm for predicting drug responses presents a significant challenge. The ideal model should possess a high degree of complexity while also maintaining the ability to generalize across real-life scenarios. Shallow machine learning models fall short in meeting these requirements due to their limited capacity for complexity. Hence, a complex hybrid deep learning architecture has been proposed. This architecture leverages the capabilities of deep learning to provide the necessary complexity and generalization for effective drug response prediction. The reason behind designing a complex architecture is to effectively utilize unique strengths of the different deep learning models. Therefore, these strengths have been combined to construct a comprehensive model, aiming to provide an optimal solution for the problem at hand.

In the proposed approach, Convolutional Neural Networks (CNNs) [71] are employed to process cell line data. This is achieved by convolving a filter over the cell line matrix, a process that allows for the extraction of relevant features. This method takes advantage of the spatial relationships in the data, capturing local patterns that can be crucial for understanding the cell lines.

On the drug side, Graph Transformer Neural Networks (GTNNs) are utilized. These networks are particularly suited for representing drug molecules accurately using a graph representation. The GTNNs are capable of capturing the complex, non-Euclidean structure of the molecules, which allows for a more detailed comprehension of their properties.

The outputs from both the CNN and GTNN models are then concatenated. This combined output represents a fusion of information from both the cell line and the drug molecule. It encapsulates the intricate relationships and patterns that the individual models have captured.

This combined output is subsequently fed into a Multi-Layer Perceptron (MLP) for the prediction of pIC50 values. The MLP acts as the concluding stage, processing the combined information from the CNN and GTNN to generate the final pIC50 prediction as can be seen in Figure 3.5.

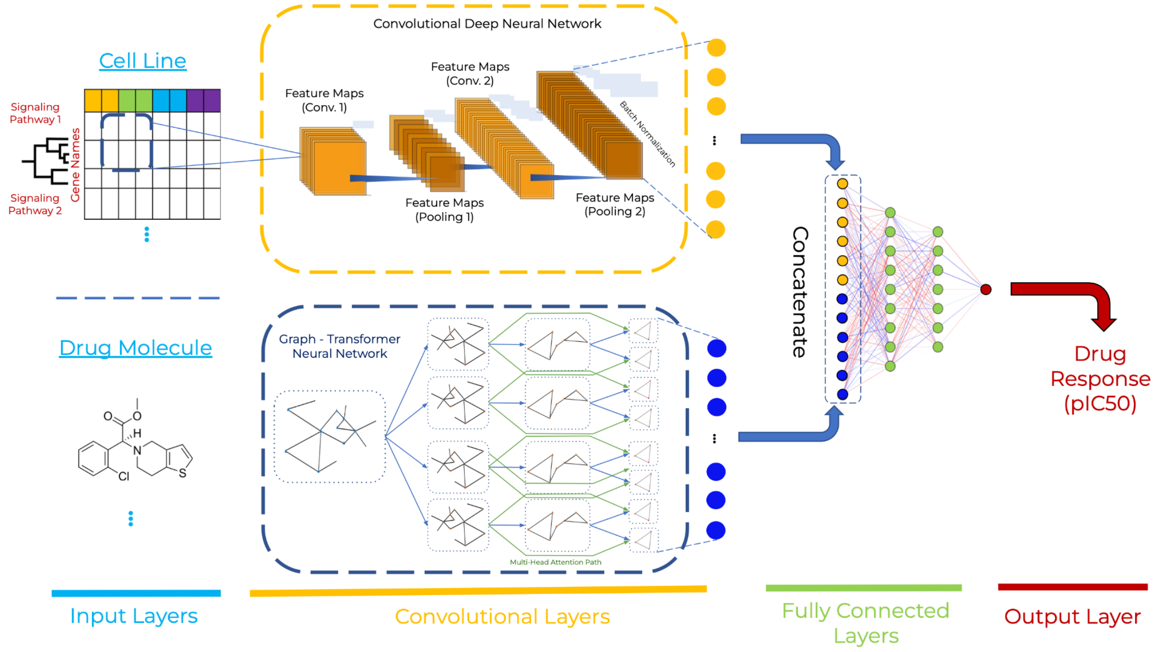

Figure 3.5. Architecture of DeepReponse

3.6. Training the Model

3.6.1. Creating Tensorflow Dataset

In the field of machine learning, particularly when dealing with complex data such as molecular structures and cell line features, the preparation and organization of data is a crucial step. This process begins with the initialization of our data, which includes both independent and dependent variables.

The batch size is also specified at this stage, which determines the number of samples that will be propagated through the network at once. This is a key parameter that can significantly impact the learning process and the resulting model performance.

Two specific datasets are also introduced, one related to molecular properties and another related to convolutional features. These datasets are constructed to capture the essential characteristics of the molecules and cell lines we are studying. The independent variables, which could be a wide range of molecular or cellular properties, are converted into a structured format. This structured data is then combined with the two specific datasets, effectively integrating the diverse types of information into a unified data structure.

One of the unique aspects of this process is the conversion of molecular properties into a graph representation. This transformation allows us to capture the structural information of the molecules, which is often critical for understanding their behaviors and interactions. This graph-based dataset provides a rich source of information for the machine learning model to learn from.

In parallel, the features related to cell lines are processed into a dataset suitable for convolutional operations. Convolutional neural networks have proven to be highly effective for tasks involving spatially structured data, and by preparing our cell line data in this way, we are able to leverage these powerful models for our task. Once these datasets are prepared, they are combined into a TensorFlow dataset. This dataset is then organized into batches, a process that involves grouping the data into subsets of the specified batch size. Batching the data is a common practice in machine learning that can help to make the training process more efficient and stable as illustrated in Figure 3.6.

To further enhance the efficiency of our model training, the dataset is also prefetched. Prefetching is a technique where the data needed for future steps is prepared while the current step is still being processed. This can significantly reduce the idle time of the computational resources, leading to faster training times.

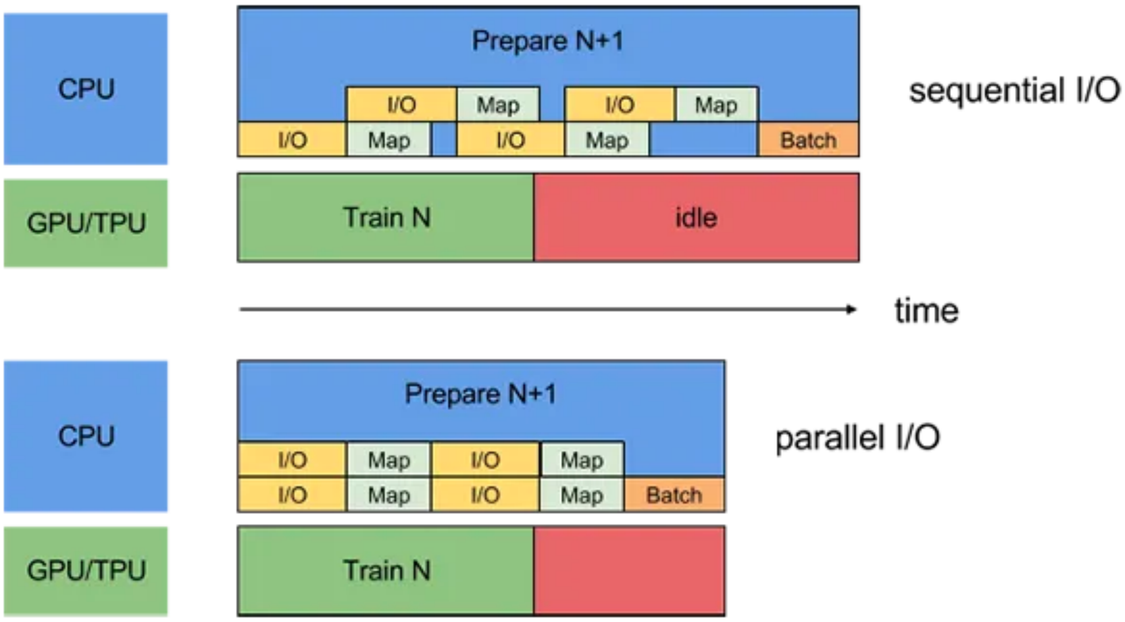

Figure 3.6. Utilization of computational resources by time with TensorFlow operations [77]

The function returns several pieces of information, including the dimensions of the molecular features, the shape of the convolutional dataset, and the batched and prefetched TensorFlow dataset. These outputs not only provide useful information about the prepared data, but also serve as a confirmation that the data has been correctly processed and is ready for model training.

In conclusion, this method of data preparation ensures that our data is appropriately preprocessed, integrated, batched, and prefetched, setting the stage for effective and efficient training of our machine learning model. This meticulous preparation of data is a testament to the importance of data management in machine learning and serves as a solid foundation for the subsequent steps of model training and evaluation.

3.6.1.1. Preparing Cell Line Data

In the process of preparing data for machine learning models, a crucial step involves optimizing the format of a specific dataset. This dataset could be in any format and may contain a variety of features related to cell lines.

The process begins by initializing an empty list, which will be used to store the processed data. It then iterates over each row in the input dataset. For each row, the function simply appends it to the list. This step might seem trivial, but it is actually a key part of the conversion process. By iterating over the data in this way, the process ensures that the data is in a consistent and ordered format, which is important for the subsequent steps of the machine learning pipeline.

Once all the data has been processed and appended to the list, the function converts the list into a NumPy array. NumPy arrays are a popular data structure in the field of data science and machine learning, known for their efficiency and versatility. By converting the data into this format, the process ensures that the dataset is optimized for the computational operations that will be performed on it during the model training process.

In conclusion, this process plays a vital role in the data preprocessing pipeline, transforming the raw dataset into an optimized format that is ready for model training. This exemplifies the importance of proper data management in machine learning, ensuring that the data is in the right format and structure for the subsequent steps of the pipeline.

3.6.1.2. Preparing Drug Data

In the field of computational chemistry and drug discovery, the representation of molecular structures plays a pivotal role. One common way to represent molecules is through Simplified Molecular-Input Line-Entry System (SMILES) strings, a compact notation that provides a string representation of a molecule's structure.

To address this, a critical step in the data preprocessing pipeline involves converting these SMILES strings into graph representations. This conversion allows the machine learning model to better understand and learn from the structural information of the molecules. The graph representation captures the atoms as nodes and the bonds between them as edges, effectively transforming the complex 3D structure of a molecule into a simplified 2D graph.

The process begins by iterating over a list of SMILES strings. Each string is processed individually to generate the components of a molecule graph. This processing involves several steps, starting with the conversion of the SMILES string into a molecule object. This conversion is performed using a specialized function that interprets the SMILES notation and constructs a corresponding molecule object.

Once the molecule object is created, the next step is to extract features from the atoms and bonds in the molecule. These features could include various properties such as the atom type, the bond type, the number of connected atoms, and so on. The extraction of these features is performed using specific encoding functions, which transform the properties of the atoms and bonds into numerical representations. These numerical features capture the essential characteristics of the atoms and bonds, providing valuable information for the machine learning model.

In addition to the atom and bond features, the process also constructs pair indices that represent the connections between atoms. These pair indices are crucial for understanding the structure of the molecule, as they indicate which atoms are bonded together. To ensure that every atom is considered during the learning process, self-loops are also added to the graph, which involve connections of an atom to itself.

Once all the components of the molecule graph are generated, they are organized into arrays. These arrays are then combined into a tuple, which includes ragged tensors for atom features, bond features, and pair indices. The use of ragged tensors is particularly suitable for this task, as they can efficiently handle data with varying lengths, which is often the case with molecular structures.

In conclusion, this process transforms the raw SMILES strings into a structured graph representation that captures the structural information of the molecules. This graph-based representation is highly suitable for machine learning models, allowing them to effectively learn from the structural information of the molecules. This meticulous preparation of data is a testament to the importance of data management in machine learning and serves as a solid foundation for the subsequent steps of model training and evaluation.

By ensuring that the data is in the right format and structure, we can facilitate the learning process and potentially improve the performance of our machine learning models as represented in Figure 3.7.

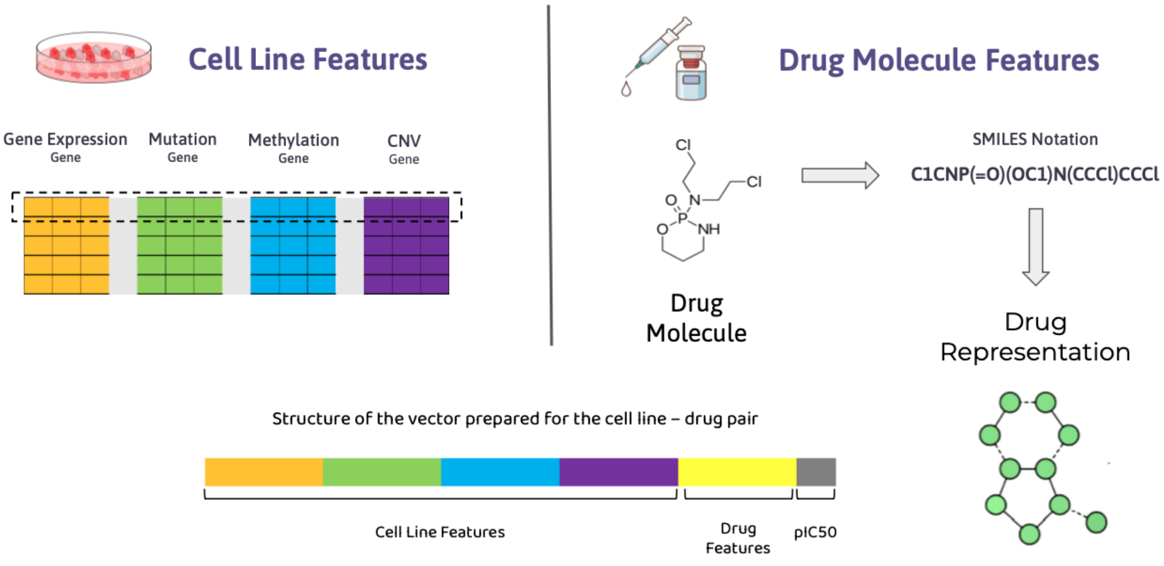

Figure 3.7. Representation of data processing

3.6.2. Data Split Strategies

3.6.2.1. Random Split Strategy

The Random Split Strategy is the most basic and straightforward approach to dividing the dataset. It randomly assigns each data point to the training, validation, or testing set. This randomness ensures a diverse range of data in each set, which can help the model learn a broad set of features during training and then test those features on a wide variety of data. However, this strategy does not take into account the specific characteristics of the data, such as the type of cell line or drug. Therefore, while it provides a good baseline for model performance, it may not fully reflect the model's ability to generalize to unseen cell lines or drugs. Despite its simplicity, the Random Split Strategy is a powerful tool in the machine learning toolbox, providing a straightforward and effective means of evaluating a model's performance. It is often the first strategy used when training a new model, providing a baseline against which more complex strategies can be compared.

3.6.2.2. Cell Stratified Split

The Cell Stratified Strategy takes the biological aspect of the data into account by grouping the data based on cell line. Each unique cell line is used as a test set exactly once, while the remaining data is used for training. This strategy is particularly useful for assessing how well the model can generalize to new cell lines. By ensuring that the model is tested on every unique cell line, we can gain a better understanding of its performance across different cell types. This strategy is especially relevant in the field of precision medicine, where treatments are often tailored to the specific characteristics of a patient's cells. By stratifying the data by cell line, we can ensure that our model is capable of handling the wide variety of cell types that it may encounter in a real-world setting. This strategy also allows us to identify any cell lines that the model struggles with, which can provide valuable insights for further model development and improvement.

3.6.2.3. Drug Stratified Split

The Drug Stratified Strategy is similar to the Cell Stratified Strategy, but it groups the data based on drug instead of cell line. Each unique drug is used as a test set exactly once. This strategy provides a rigorous test of the model's ability to generalize to new drugs. It is particularly useful in the field of precision medicine, where the ability to predict the response of a specific drug is crucial. By testing the model on every unique drug, we can gain a better understanding of its ability to predict drug responses accurately and reliably. This strategy is especially important in the context of drug discovery and development, where the ability to predict a drug's efficacy and safety profile is of paramount importance. It allows us to identify any drugs that the model struggles with, providing valuable insights for further model development and improvement.

3.6.2.4. Drug-Cell Stratified Split

The Drug-Cell Stratified Strategy is the most challenging and realistic test scenario. In this strategy, the model is tasked with predicting drug-cell pairs that it has not seen before. This is akin to the real-world situation where we often need to predict the response of a specific cell line to a specific drug, both of which the model may not have encountered during training.

This strategy provides the most accurate measure of how well the model will perform in practice. It is a stringent test of the model's ability to generalize and is particularly relevant in the context of personalized medicine, where treatments are often tailored to the specific characteristics of a patient's cells and the specific drugs being used. This strategy also allows us to identify any drug-cell pairs that the model struggles with, providing valuable insights for further model development and improvement. Figure 3.8. shows the different split strategies.

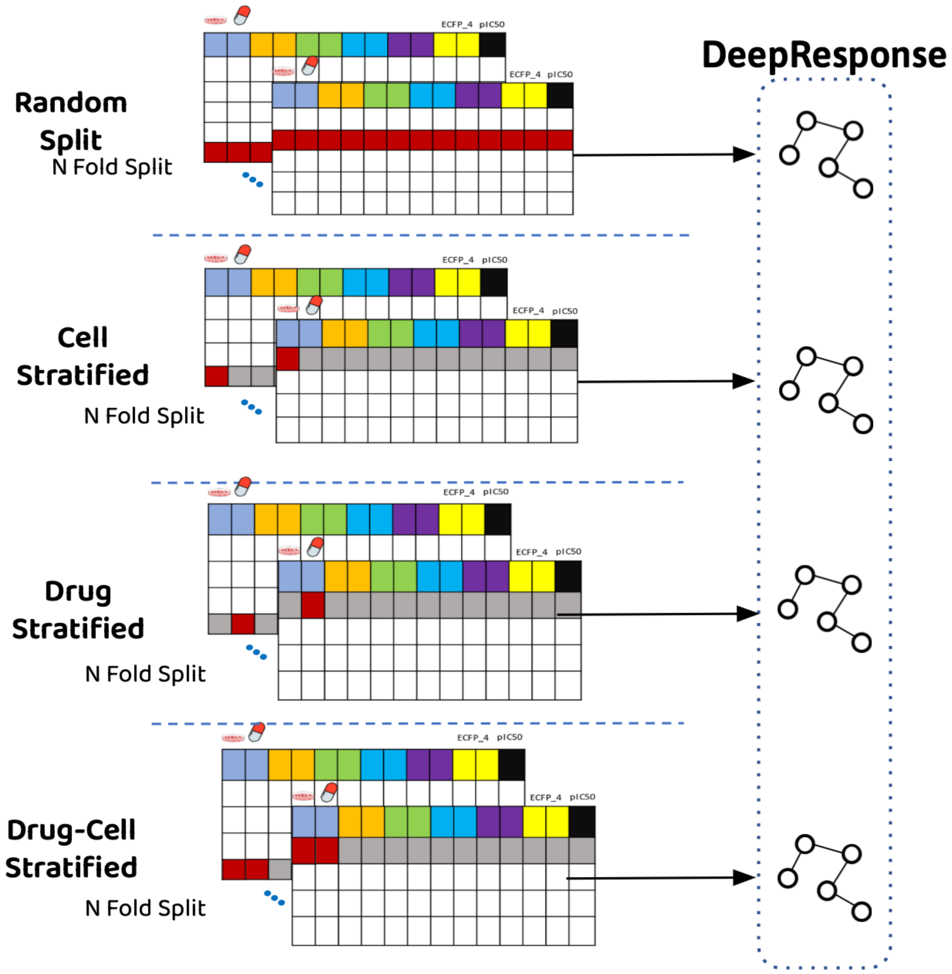

Figure 3.8. Comparison of different split strategies

3.6.2.5. Cross Domain Strategy

The Cross-Domain Strategy involves training the model on one database and evaluating it on another. For example, the model could be trained on the GDSC database and tested on the CCLE database. This strategy tests the model's ability to generalize across different domains, which is crucial in a field like precision medicine where new data is constantly being generated. By evaluating the model on a separate database, we can gain a better understanding of its performance on data that is fundamentally different from the data it was trained on. This strategy is particularly relevant in the context of large-scale biomedical research, where data is often collected from multiple sources, and it is important to ensure that the model can handle the variability and complexity of these different data sources. This strategy also allows us to identify any domains that the model struggles with, providing valuable insights for further model development and improvement.

3.6.3. Loss Function and Optimizer

In this research, the Huber loss is employed as the loss function (14). Huber loss is often used in robust regression scenarios, presenting a combination of the mean squared error and mean absolute error loss functions. It is particularly effective in mitigating the impact of outliers on the model's performance. For errors smaller than a certain threshold, it behaves quadratically like the mean squared error, but for larger errors, it behaves linearly like the mean absolute error. This dual behavior makes Huber loss a more robust choice for many machine learning tasks, as it can handle outliers without being overly sensitive to them. The threshold at which the loss function changes from quadratic to linear is a tunable parameter, offering flexibility in controlling the robustness of the model.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & for\ |y - f(x)| \le \delta, \\ \delta\ |y - f(x)| - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{14}$$

y: $Actual\ value$

$f(x)$: $Predicted\ value$

δ: $Threshold\ to\ decide\ whether\ to\ use\ squared\ loss\ or\ absolute\ loss\ functions$

The optimizer used is Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions. The Adam optimizer uses an algorithm for first-order gradient-based optimization of stochastic objective functions. It maintains an exponential moving average of the gradient and the squared gradient, and the parameters β_1 and β_2 control the decay rates of these moving averages.

The first moment estimate is updated as in (15):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{15}$$

And the second moment estimate is updated as in (16):

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) {g_t}^2 \tag{16}$$

These estimates are biased towards zero, especially during the initial time steps, and this bias is corrected as follows (17,18):

$$\widehat{m_t} = \frac{m_t}{1-\beta_1^t} \tag{17}$$

$$\widehat{v_t} = \frac{v_t}{1-\beta_2^t} \tag{18}$$

The learning rate, a critical hyperparameter of the Adam optimizer, determines the step size at each iteration while moving toward a minimum of the loss function. This learning rate is carefully chosen to ensure that the model converges to a solution efficiently without overshooting the minimum. The updated parameters are computed as in (19):

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\widehat{v_t}}+\epsilon} \widehat{m_t} \tag{19}$$

m_t, v_t: *First and second moment estimates*

g_t: *Gradient of loss function*

$\widehat{m_t}, \widehat{v_t}$: *Bias corrected moment estimates*

β_1, β_2: *Exponential decay rates*

ϵ: *Small value to prevent division by zero*

η: *Learning rate*

θ_t: *Optimized parameter at time t*

The model is then trained on the training dataset for a specified number of epochs. An epoch is one complete pass through the entire training dataset. During each epoch, the model learns to adjust its weights and biases to minimize the loss function. The validation dataset is used to evaluate the model's performance at the end of each epoch, providing a check on overfitting.

3.7. Evaluating the Model

The model's performance is evaluated on the test dataset. This evaluation is crucial as it provides an unbiased estimate of the model's performance on new, unseen data. The evaluation metrics used include MSE, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). These metrics provide different perspectives on the model's performance. For instance, RMSE can be interpreted in the same units as the response variable, making it more interpretable. In the context of drug response prediction, these metrics have specific interpretations.

A lower MSE, MAE and RMSE indicates that the model's predicted pIC50 values are closer to the actual values, suggesting better model performance. However, it's important to remember that these metrics should not be viewed in isolation. They should be considered in conjunction with the biological and clinical significance of the predictions. A small error in the predicted pIC50 value might lead to a significant difference in the interpretation of a compound's potency. Therefore, the choice and interpretation of these metrics should also consider the practical implications in the field of drug discovery. This rigorous training and evaluation process ensures the robustness and reliability of the machine learning model, making it a valuable tool in the drug discovery process.

3.7.1. Performance Metrics

In the context of predicting drug response as pIC50 values, the performance of the machine learning model is evaluated using a variety of metrics, each providing a unique perspective on the model's performance.

- **Mean Squared Error (MSE):** This metric quantifies the average squared difference between the predicted and actual pIC50 values. It is particularly useful as it penalizes larger errors more due to the squaring operation. In the context of drug response prediction, a lower MSE indicates (20) that the model's predicted pIC50 values are closer to the actual values, suggesting better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{20}$$

n: *Total number of data points*
Y_i: *Actual value*
$\hat{Y}_i$: *Predicted value*

- **Root Mean Squared Error (RMSE):** This is the square root of the MSE (21) and can be interpreted in the same units as the response variable, making it more interpretable. A lower RMSE indicates a better fit of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{21}$$

n: *Total number of data points*
Y_i: *Actual value*
$\hat{Y}_i$: *Predicted value*

- **Mean Absolute Error (MAE):** This measures the average magnitude of the errors in a set of predictions (22), without considering their direction. It is less sensitive to outliers compared to MSE and RMSE.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{22}$$

n: *Total number of data points*
y_i: *Predicted value*
x_i: *Actual value*

In addition to these regression metrics, several classification metrics are calculated after binarizing the data based on a specified threshold. These include:

- **Accuracy:** This measures the proportion of true results (both true positives and true negatives) among the total number of cases examined (23).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

- **Precision:** This quantifies the number of true positive predictions divided by the total number of positive predictions. It is a measure of a classifier's exactness (24).

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

- **Recall:** Also known as sensitivity, this measures the proportion of actual positives that are correctly identified. It is a measure of a classifier's completeness (25).

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

- **F1 Score:** This provides a balance between precision and recall. It is the harmonic mean of precision and recall and gives equal weight to both metrics (26).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{26}$$

- **Matthew's Correlation Coefficient (MCC):** This is a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes (27).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN+FN)}} \tag{27}$$

These metrics collectively provide a comprehensive evaluation of the model’s performance, allowing for the identification of areas of strength and potential improvement. It’s important to note that these metrics should be interpreted in the context of the specific task of predicting pIC50 values, and in conjunction with the biological and clinical significance of the predictions. This rigorous evaluation process ensures the robustness and reliability of the machine learning model, making it a valuable tool in the field of drug discovery.

4. RESULTS

4.1. Prediction Performance of the Model

The model’s predictive performance, particularly in predicting pIC50 values for drug response, is a crucial aspect of its evaluation. Metrics such as Huber Loss, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are employed to assess the closeness between the predicted drug response and the true drug response.

In addition to these, classification metrics like accuracy, precision, recall, and the F1 score are used to determine the effectiveness of the drugs on the cancer cell line. These metrics provide the binary perspective, categorizing the drugs as effective or not based on a certain threshold.

The model’s performance was evaluated under a variety of conditions, each representing a unique method of data splitting. These data splits are not arbitrary; they are designed to mimic real-world scenarios and challenges in predicting drug responses. For instance, they may represent different patient groups, various types of cancer, or a range of drug compounds. This rigorous evaluation process, therefore, not only tests the model’s robustness and reliability across different data splits but also its potential effectiveness in real-world applications. By succeeding under these diverse conditions, the model demonstrates its readiness to handle the complexity and variability inherent in cancer treatment.

4.1.1. Random Split

In the random split scenario, the dataset was divided into training, validation, and test sets without considering any structure in the data. This provides a baseline performance metric for the model as summarized in Table 4.1.

	Data Source		
Performance Metrics	**GDSC**	**CCLE**	**NCI-60**
Mean Squared Error	**1.028**	1.182	1.213
Root Mean Squared Error	**1.014**	1.166	1.186
Mean Absolute Error	**0.812**	0.933	0.966
Accuracy	**0.852**	0.721	0.682
Precision	**0.838**	0.711	0.712
Recall	**0.821**	0.697	0.673
F1 Score	**0.825**	0.701	0.684

Table 4.1. Performance metrics on random split on test dataset. Best scores are shown in bold font.

The performance metrics indicate that the model's predictions are generally close to the actual values, suggesting its capability to accurately predict drug responses when the data is split randomly. The model also demonstrates high accuracy, precision, recall, and F1 score, particularly for the GDSC data source. This implies that the model is correctly classifying a high percentage of the drugs while maintaining a balanced ratio of precision and recall, which is crucial in real-world settings.

However, the performance metrics for the CCLE data source are slightly lower than those for the GDSC. This could be attributed to the fact that the CCLE dataset is smaller than the GDSC dataset, which might affect the model's learning and prediction capabilities. It underscores the importance of using diverse and sufficiently large datasets for training and evaluating the model to ensure its robustness and generalizability.

In conclusion, the random split evaluation provides a solid baseline for the model's performance. It demonstrates the model's capability to handle randomness in data splits, a common scenario in real-world applications.

However, it also highlights the need for further evaluations under different data split scenarios to fully assess the model's robustness and reliability.

4.1.2. Cell Stratified Split

In this scenario, the dataset was divided in such a way that the model is tested on unseen cell lines. This means that all instances of a particular cell line are either in the training set or in the validation/test sets, but not both. This approach is designed to ensure that the model is evaluated on a diverse set of cell lines, testing its ability to generalize across different cell types as summarized in Table 4.2.

	Data Source		
Performance Metrics	**GDSC**	**CCLE**	**NCI-60**
Mean Squared Error	**1.300**	1.447	1.534
Root Mean Squared Error	**1.105**	1.257	1.326
Mean Absolute Error	**1.026**	1.166	1.200
Accuracy	**0.648**	0.432	0.518
Precision	**0.639**	0.568	0.543
Recall	**0.627**	0.557	0.514
F1 Score	**0.631**	0.562	0.523

Table 4.2. Performance metrics on cell stratified split on test dataset. Best scores are shown in bold font.

This approach is designed to test the model's ability to generalize across different cell types. The performance metrics in this scenario were lower than those in the random split scenario. This indicates that maintaining the same distribution of cell lines in the splits is a more challenging task for the model. It suggests that the model may be overfitting to specific cell lines in the training data and struggling to generalize to new cell lines in the validation and test sets.

This scenario highlights the importance of considering the structure of the data when splitting it for model training and evaluation. It shows that while a model may perform well under random splits, it may struggle when the data is split in a way that reflects the real-world complexity and variability of the data.

In conclusion, the cell stratified split scenario provides valuable insights into the model's ability to generalize across different cell types. It highlights the need for models that can handle the inherent variability in the data and underscores the importance of rigorous model evaluation under different data split scenarios.

4.1.3. Drug Stratified Split

This method tests the model's adaptability to new drugs. The dataset is arranged so that the model does not see certain drugs during training, and then it is evaluated on how well it predicts the responses to these unseen drugs in the validation/test sets as summarized in Table 4.3.

	Data Source		
Performance Metrics	**GDSC**	**CCLE**	**NCI-60**
Mean Squared Error	**1.241**	1.359	1.526
Root Mean Squared Error	**1.142**	1.221	1.370
Mean Absolute Error	**0.979**	1.072	1.185
Accuracy	**0.684**	0.612	0.527
Precision	**0.675**	0.603	0.526
Recall	**0.662**	0.595	0.530
F1 Score	**0.665**	0.595	0.545

Table 4.3. Performance metrics on drug stratified split on test dataset. Best scores are shown in bold font.

The performance metrics for this scenario were a bit better than the cell stratified split but not as high as the random split, indicating that predicting responses to new drugs is a complex task for the model. This suggests that the model might be learning too much from the specific drugs in the training data, which hampers its ability to predict responses to new drugs. This scenario emphasizes the need to consider the real-world complexity and variability of the data when preparing it for model training and evaluation. It shows that a model's performance can vary significantly depending on how the data is split.

4.1.4. Drug-Cell Stratified Split

This scenario is arguably the most rigorous test of the model's predictive capabilities. It involves splitting the dataset in a way that the model is evaluated on unseen drug-cell combinations, which is a significant challenge. This approach mirrors real-world situations where a model has to predict responses for new drug-cell pairs that it has not encountered during training. These results in Table 4.4. underscore the complexity of predicting drug responses for new drug-cell combinations. It highlights the need for models that can effectively handle the variability and complexity inherent in the data. Despite the lower performance in this scenario, it provides valuable insights into the model's robustness and its ability to generalize to new, unseen data. This scenario serves as a reminder of the challenges involved in drug response prediction and underscores the importance of using rigorous evaluation methods to fully assess a model's performance.

	Data Source		
Performance Metrics	**GDSC**	**CCLE**	**NCI-60**
Mean Squared Error	**2.968**	3.412	3.769
Root Mean Squared Error	**1.723**	1.981	2.154
Mean Absolute Error	**1.378**	1.585	1.696
Accuracy	**0.759**	0.645	0.554
Precision	**0.731**	0.621	0.548
Recall	**0.715**	0.607	0.558
F1 Score	**0.726**	0.617	0.559

Table 4.4. Performance metrics on drug-cell stratified split on test dataset. Best scores are shown in bold font.

While the model shows promising results in less challenging scenarios, its performance in the drug-cell stratified split scenario indicates that there is room for improvement. Future work could focus on improving the model's ability to generalize to new drug-cell combinations, which is crucial for its applicability in real-world settings. This could involve exploring different model architectures, incorporating more diverse data, or using advanced training techniques to better capture the complex relationships in the data.

4.1.5. Cross Domain Split

This scenario pushes the boundaries of the model’s adaptability. It involves training the model on one domain (e.g., GDSC) and then assessing its performance on a different domain (e.g., CCLE). This is a tough test as it requires the model to apply its learning from one domain to another, which can be quite challenging due to potential differences in the underlying data distributions and characteristics between domains.

Results in Table 4.5. suggest that the model's performance varies significantly depending on the training and testing domains. The model performs reasonably well when trained on GDSC and tested on CCLE, but its performance decreases when the training and testing domains are reversed. This could be due to the model being too closely fitted to the specific characteristics of the training domain, thereby limiting its ability to generalize to the test domain. This scenario emphasizes the need for models that can effectively transfer their learning from one domain to another, a critical requirement for real-world applications.

	Data Source	
Performance Metrics	**GDSC/CCLE**	**CCLE/GDSC**
Mean Squared Error	**1.212**	1.551
Root Mean Squared Error	**1.101**	1.398
Mean Absolute Error	**0.865**	1.089
Accuracy	**0.859**	0.618
Precision	**0.833**	0.641
Recall	**0.769**	0.607
F1 Score	**0.800**	0.592

Table 4.5. Performance results of the cross-domain analysis. Best scores are shown in bold font.

Despite the challenges and lower performance in this scenario, it provides valuable insights into the model's robustness and adaptability, underscoring the importance of rigorous evaluation methods to fully assess a model's performance. Future work could focus on improving the model's cross-domain generalization capabilities, potentially through techniques such as domain adaptation or transfer learning. All the results are summarized in Figure 4.1.

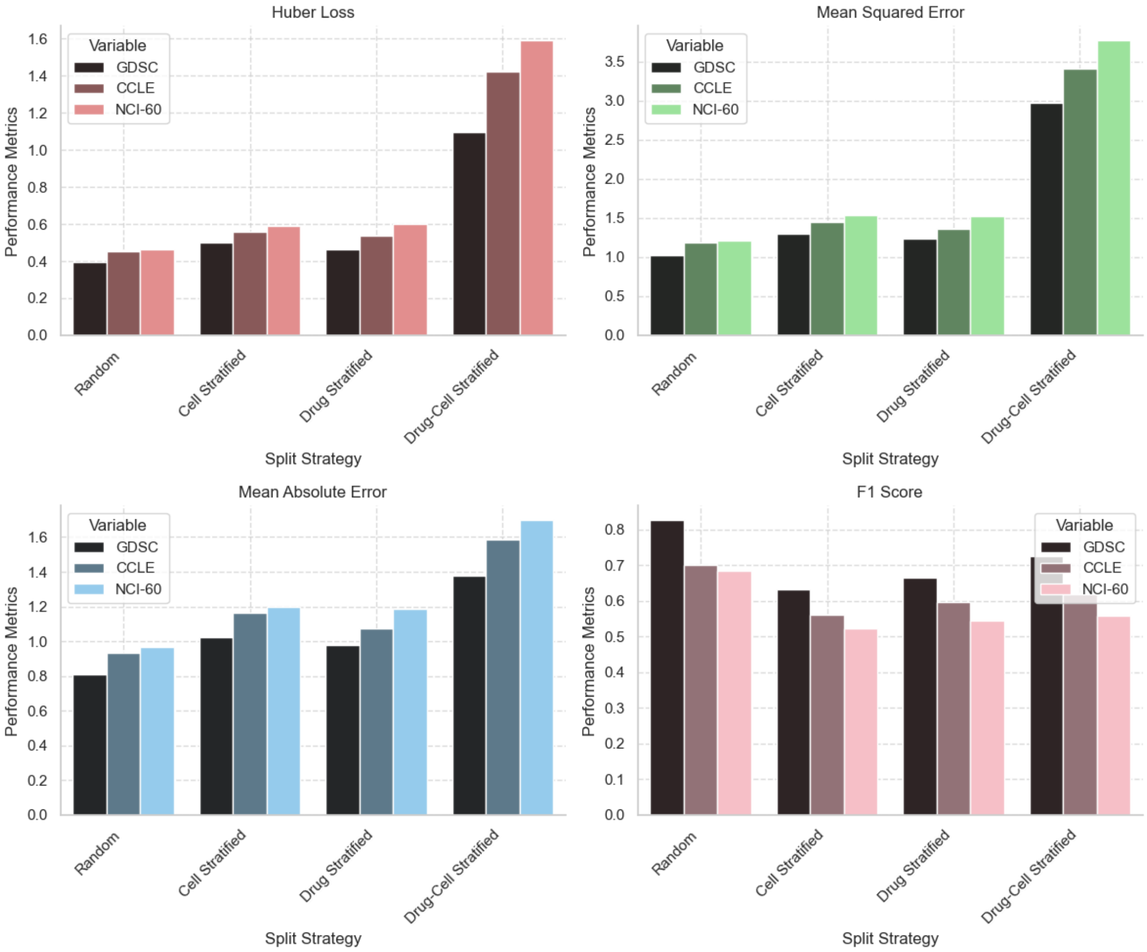

Figure 4.1. Comparison of performance metrics by split strategy for each dataset on test dataset

4.2. Comparison of Prediction Performance with State-of-the-Art

In the field of drug response prediction, a variety of models have been developed, each with its unique approach and different data sources and types as discussed in detail in the related work section. Despite their differences, all models share a common goal: to predict drug responses that are as close as possible to the actual values. A universally accepted performance metric for these models is the Root Mean Square Error (RMSE). RMSE provides a quantifiable measure of how much the predictions deviate from the actual values. The lower the RMSE, the closer the predicted responses are to the actual values, indicating a better performing model.

In the following section, an in-depth comparison of these models can be found, specifically focusing on their performance on the GDSC dataset. The performance is measured using RMSE values, and to ensure the robustness of these measures, a 10-fold cross-validation method is employed. This method enhances the reliability of the performance estimates by averaging the results over multiple testing rounds.

Among the models evaluated, DeepResponse stands out for its superior performance. It consistently achieves lower RMSE values across all splits, indicating its predictions are closer to the actual values compared to other models. This superior performance of DeepResponse is not a one-off occurrence but is consistent across all test splits. This consistent performance of DeepResponse highlights its potential as a powerful tool in the field of drug response prediction. By providing more precise predictions, it can contribute to the development of more effective therapeutic strategies. This could potentially lead to better patient outcomes, making DeepResponse a valuable addition to the toolkit of researchers and clinicians alike.

4.2.1. Comparison of Model Performances on Random Split

The Table 4.6. presents the RMSE values of various models when applied to the GDSC dataset using a random split. DeepResponse outperforms all other models, achieving the lowest RMSE value of 1.014 ± 0.001. This indicates that DeepResponse’s predictions are closer to the actual values compared to other models. The low standard deviation of DeepResponse further highlights its strong generalization capacity, regardless of the randomness of the dataset.

Model	RMSE
ENET	2.368
MOLI	2.282
DrugCell	1.998
CDRScan	1.982
tCNNs	1.782
SRMF	1.731
KBMF	1.590
GraphDRP	1.111
DualGCN	1.079
DeepCDR	1.058
DeepResponse	**1.014 ± 0.001**

Table 4.6. Model performance comparison in terms of RMSE on random split. Best score is shown in bold font.

4.2.2. Comparison of Model Performances on Cell Stratified Split

The performance of various models on cell stratified splits of the dataset was examined as in Table 4.7. Again, DeepResponse stands out with the lowest RMSE value of 1.105 ± 0.013, indicating its superior precision in predicting drug responses. The relatively low standard deviation of DeepResponse underscores its strong generalization capacity across different cell stratified splits. This consistent performance of DeepResponse, regardless of how the data is split, validates its effectiveness in drug response prediction tasks.

Model	**RMSE**
DrugCell	2.392
ENET	2.216
SRMF	1.865
GraphDRP	1.561
tCNNs	1.519
VAE+MLP	1.406
DeepCDR	1.127
DeepResponse	**1.105 ± 0.013**

Table 4.7. Model performance comparison matrix on cell stratified split. Best score is shown in bold font.

4.2.3. Comparison of Model Performances on Drug Stratified Split

This section presents the performance of various models on drug stratified splits of the dataset. As with the previous splits, DeepResponse outperforms all other models, achieving the lowest RMSE value of 1.142 ± 0.104. This demonstrates that DeepResponse's superior precision in predicting drug responses is maintained even when the data is split based on drug stratification.

The standard deviation of DeepResponse, although slightly higher in this case, still indicates a strong generalization capacity across different drug stratified splits. This consistent performance of DeepResponse across all types of data splits underscores its effectiveness and reliability in drug response prediction tasks as summarized in Table 4.8

Model	**RMSE**
GraphDRP	2.894
tCNNs	2.393
DrugCell	2.388
VAE+MLP	2.369
DeepCDR	1.999
SRMF	1.828
DeepResponse	**1.142 ± 0.104**

Table 4.8. Model performance comparison matrix on drug stratified split. Best score is shown in bold font.

The results demonstrate that the complexity of test scenarios has a negligible impact on the performance of DeepResponse. This is noteworthy as it suggests that DeepResponse's performance remains consistent regardless of the test scenario's difficulty. Conversely, other models exhibit a decline in performance as the complexity of the test scenarios increases. All the comparison results were illustrated in Figure 4.2.

DeepResponse's adaptability to real-world test cases is a key attribute. It is engineered to handle a diverse range of situations, which enhances its utility in testing. While certain models exhibit satisfactory performance in less complex test cases, their performance deteriorates in more intricate scenarios.

DeepResponse, on the other hand, maintains a consistent performance level even under challenging conditions. This underscores the model's robustness and versatility, and highlights its superior performance across a broad spectrum of test conditions. The ability of DeepResponse to deliver reliable results in both controlled and complex real-world scenarios distinguishes it from other models.

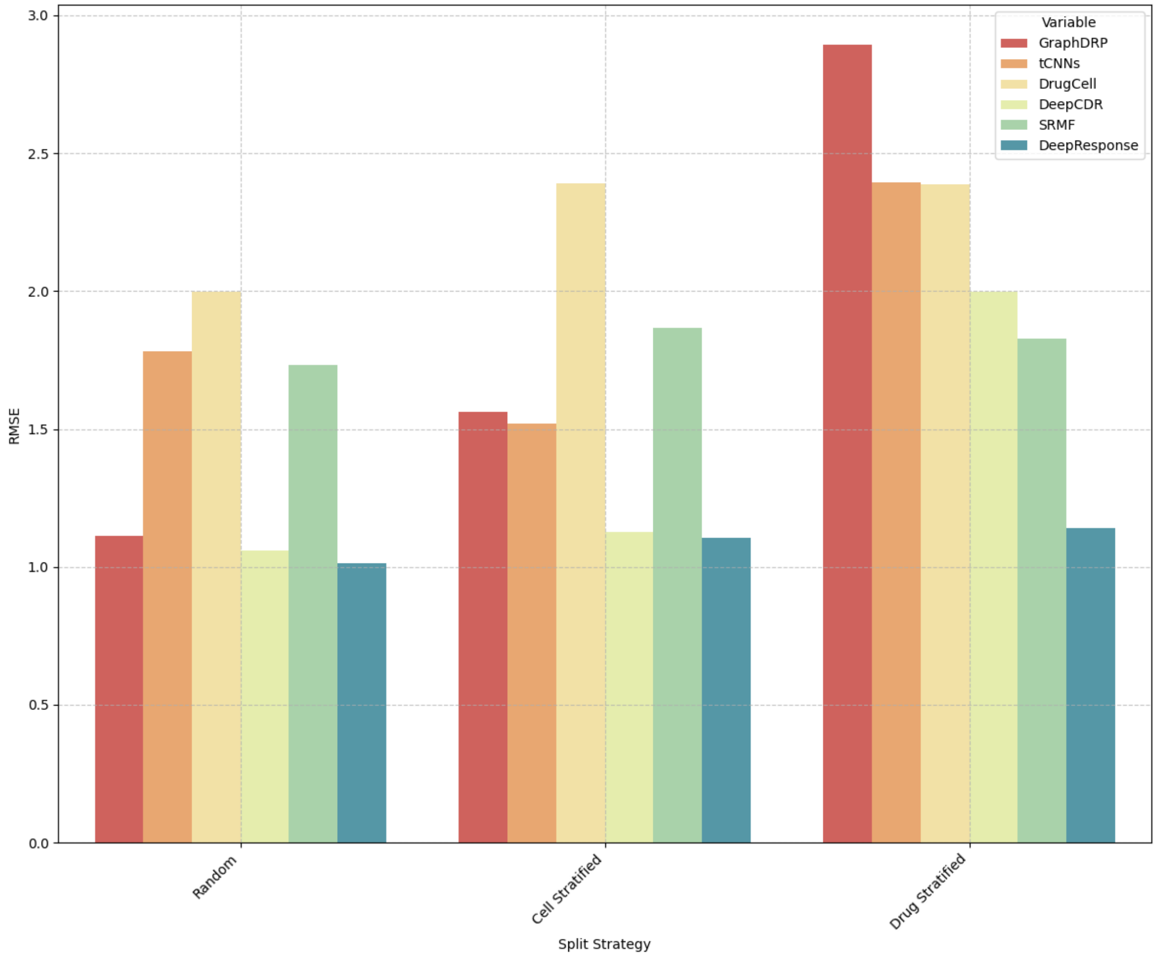

Figure 4.2. Comparison of prediction performance with state-of-the-art

4.3. Ablation Study

An ablation study is a systematic approach used in machine learning research to understand the contribution of different components of a model towards its overall performance. This method involves selectively removing or "ablating" individual components, and observing the effect on the model's performance. The aim is to identify which components are crucial for the model's performance and which ones have minimal or no impact.

An ablation study was conducted using cell line omic features to understand the individual contributions of each feature to the overall performance of the model. In the initial trial, the gene expression feature was removed, and the impact on the model's performance was observed. This approach allowed for an assessment of the importance of gene expression in the model. In subsequent trials, other features such as mutation, methylation, and copy number variation were individually removed. By systematically removing these features, the individual significance of each feature in the model's predictive capabilities was discerned. This comprehensive ablation study yielded valuable insights into the role of each omic feature in the performance of the model. All the ablation study results were shown in Table 4.9.

	Omitted Omic				
Performance Metrics	**None**	**Gene Expression**	**Mutation**	**Methylation**	**Copy Number Variation**
MSE	**1.028**	1.069	1.040	1.049	1.059
RMSE	**1.014**	1.055	1.025	1.034	1.045
MAE	**0.812**	0.844	0.820	0.829	0.835
Accuracy	**0.852**	0.818	0.840	0.826	0.820
Precision	**0.838**	0.804	0.825	0.812	0.805
Recall	**0.821**	0.788	0.810	0.796	0.790
F1 Score	**0.825**	0.792	0.815	0.800	0.795

Table 4.9. Performance metrics for ablation study by omitting single omic. Best scores are shown in bold font.

The ablation study provides valuable insights into the role of each omic feature in the performance of the model. When the gene expression feature was removed, there was a noticeable decrease in all performance metrics, indicating that gene expression plays a significant role in the model's predictive capabilities.

Similarly, the removal of mutation, methylation, and copy number variation also resulted in a decrease in performance metrics, albeit to a lesser extent than gene expression. This suggests that while these features contribute to the model's performance, their individual impact is less than that of gene expression. All the performance metrics were represented in Figure 4.3.

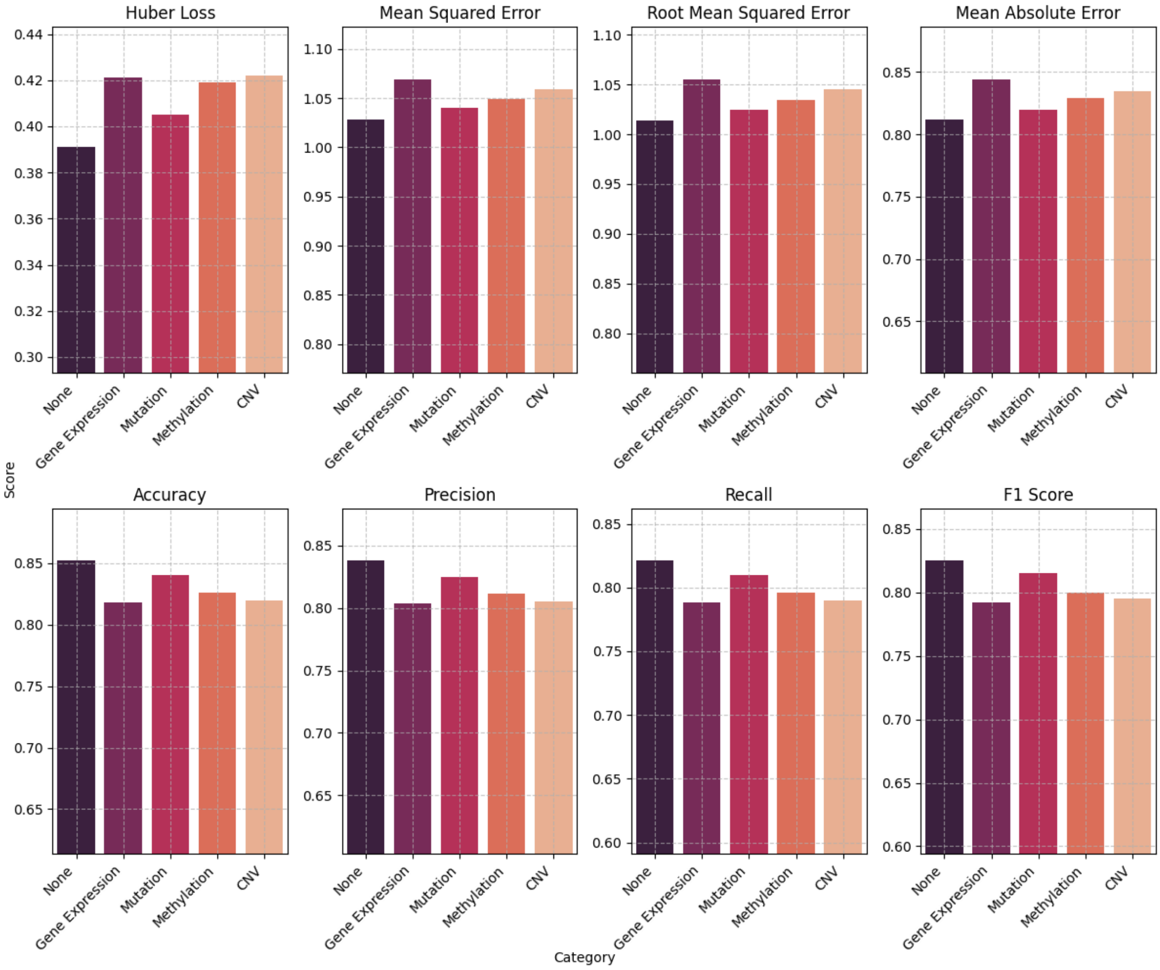

Figure 4.3. Comparison of the performance by single omitted omics

Another iteration of ablation study, as detailed in Tables 4.10 and 4.11, was conducted by removing combinations of two omic features at a time to understand their collective contributions to the overall performance of the model.

The removal of both gene expression and mutation features resulted in a more pronounced decrease in all performance metrics than when these features were removed individually, indicating that the combination of gene expression and mutation plays a significant role in the model's predictive capabilities.

	Omitted Omic			
Performance Metrics	**None**	**Gene Expression - Mutation**	**Methylation - Mutation**	**Copy Number Variation - Mutation**
MSE	**1.028**	1.143	1.101	1.185
RMSE	**1.014**	1.149	1.065	1.129
MAE	**0.812**	0.895	0.878	0.919
Accuracy	**0.852**	0.753	0.793	0.721
Precision	**0.838**	0.748	0.772	0.732
Recall	**0.821**	0.717	0.748	0.703
F1 Score	**0.825**	0.744	0.776	0.729

Table 4.10. Performance metrics for ablation study by omitting double omic in combination with mutation. Best scores are shown in bold font.

Similarly, the removal of methylation and mutation, as well as copy number variation and mutation, also resulted in a decrease in performance metrics. This suggests that while these features contribute to the model's performance, their combined impact is less than that of gene expression and mutation.

When both gene expression and methylation were removed, there was a substantial decrease in all performance metrics, suggesting that these two features collectively play a crucial role in the model's predictive capabilities. The removal of copy number variation and methylation, as well as gene expression and copy number variation, resulted in even more significant decreases in performance metrics.

This indicates that these combinations of features are vital for the model's performance, with the combination of gene expression and copy number variation having the most significant impact.

	Omitted Omic			
Performance Metrics	**None**	**Gene Expression - Methylation**	**Copy Number Variation - Methylation**	**Gene Expression - Copy Number Variation**
MSE	**1.028**	1.283	1.536	1.853
RMSE	**1.014**	1.213	1.409	1.723
MAE	**0.812**	0.996	1.253	1.503
Accuracy	**0.852**	0.678	0.574	0.246
Precision	**0.838**	0.651	0.523	0.201
Recall	**0.821**	0.662	0.435	0.277
F1 Score	**0.825**	0.650	0.477	0.318

Table 4.11. Performance metrics for ablation study by omitting double omic in combination without mutation. Best scores are shown in bold font.

These comprehensive ablation studies provide valuable insights into the role of each omic feature and their combinations in the performance of the model. They highlight the importance of considering the collective impact of multiple features in the model's predictive capabilities.

The results as summarized in Figure 4.4. suggest that the model is not overly reliant on any single feature or pair of features, but rather, it benefits from the synergistic effect of multiple features. This underscores the complexity of biological systems and the need for multi-omic approaches in predictive modeling. It also points to the potential for further optimization, perhaps by identifying and incorporating additional relevant features or by refining the model architecture to better capture the interactions between features.

Ultimately, these findings contribute to our understanding of the model's workings and guide future efforts to improve its performance.

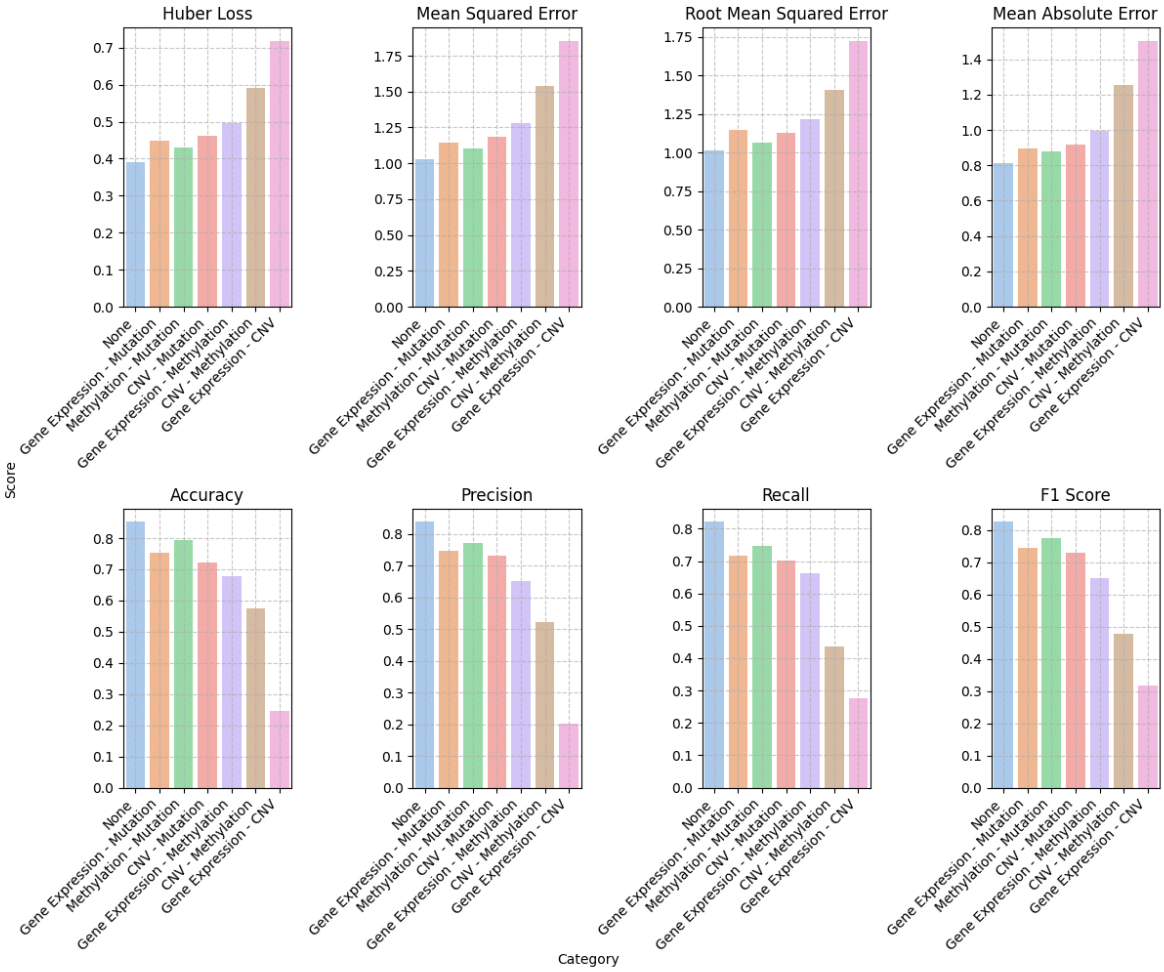

Figure 4.4. Comparison of the performance by multi omitted omics

4.4. Use Case Analysis

In the process of conducting the use case analysis, the deep response model was initially applied to various tissue data. Among all the tissues examined, the model exhibited the most optimal performance on the digestive system. This can be attributed to the unique patterns inherent in each type of tissue data. By training and testing on these specific tissues, the model was able to adapt and thus, perform more effectively.

The following are some of the predictions and actual values of drug response, which further illustrate the model's performance in Table 4.12. It is important to note that these results underscore the potential of using tissue-specific models in predicting drug responses, thereby paving the way for more personalized and effective therapeutic strategies.

Cell Line	Drug Name	True pIC50 Value	Predicted pIC50 Value
CAMA-1	Camptothecin	6.274	6.027
CAMA-1	Cisplatin	4.385	4.924
HCT-116	Cisplatin	5.222	5.483
HCT-116	Dactolisib	6.813	6.512
HCT-116	Fludarabine	3.952	3.541

Table 4.12. Comparison of true value and predicted value in selected cell lines

In the quest for effective anti-cancer drugs, the DeepResponse model was employed to evaluate a multitude of drug candidates for repurposing against hepatocellular carcinoma (HCC), the second deadliest cancer globally. The model was able to predict the activity of several inhibitors across various HCC cell lines, including Huh7, Hep3B, SNU 387/423/475. A diverse array of drug candidates was considered in this process, each with its unique properties and mechanisms of action. However, amidst this vast pool of potential therapeutics, Eprinomectin emerged as a particularly promising candidate. Eprinomectin, an approved avermectin currently used as a veterinary topical endectocide, demonstrated high predicted activity across all tested HCC cell lines.

This high activity can be seen in Table 4.13, coupled with the fact that absence of previous studies investigating its repurposing against HCC, made Eprinomectin an intriguing choice for further experimental analysis.

The cytotoxicity of Eprinomectin was evaluated using the SRB assay and real-time monitoring of HCC cells. These wet lab experiments were conducted by the Cancer Systems Laboratory at Middle East Technical University.

Additional analyses were conducted to understand the mechanisms involved in its cytotoxicity against HCC cells, including cell cycle, apoptosis, and western blot analyses. The results indicated that Eprinomectin has a comparable, if not superior, inhibitory potential to the approved HCC drug Sorafenib. Eprinomectin was found to induce G1 arrest and apoptosis in HCC cell lines. At the protein level, the apoptotic marker cleaved-PARP increased upon treatment with Eprinomectin in Huh7 and Mahlavu cells. Furthermore, cell cycle proteins such as CDK2 and CDK4 decreased, further supporting Eprinomectin's effect on cell cycle progression. The experimental validation results (pIC50) of Eprinomectin on HCC cells is shown in Table 4.13.

These findings highlight the potential of Eprinomectin as a treatment for HCC. However, further analysis is required to better assess the effects of this drug on both cancerous and healthy human cells. This study demonstrates the utility of the DeepResponse model in identifying promising candidates for drug repurposing in the treatment of cancer.

Cell Line	**Drug Name**	**Predicted pIC50 Value**	**Experimental Validation Results (pIC50)**
Hep3B2-1-7	Eprinomectin	6.425	5.377 ± 0.40
HuH-7	Eprinomectin	6.757	5.443 ± 0.23
SNU-387	Eprinomectin	6.924	4.936 ± 0.16
SNU-423	Eprinomectin	6.329	4.978 ± 0.16
SNU-475	Eprinomectin	6.112	5.136 ± 0.52

Table 4.13. Predictions and experimental validation results on HCC cell lines

4.5. Model Implementation

The code base, datasets, and results of DeepResponse are openly shared at https://github.com/HUBioDataLab/DeepResponse, reflecting a commitment to open science and collaboration in the pursuit of advancing personalized medicine.
The implementation of DeepResponse adheres to several best practices in computer science, ensuring the development of a reliable and error-free project.

- **Modularity:** The code is organized into distinct modules or classes, each with a specific role. This separation of concerns makes the code easier to understand, test, and maintain.

- **Use of Abstract Classes and Strategy Pattern:** Abstract classes define a common interface for various strategies, including strategies for using Comet, handling datasets, and training the model. The strategy pattern allows the algorithm to select the appropriate strategy at runtime, providing flexibility and making it easy to introduce new strategies in the future.

- **Parameterization:** The algorithm can take various parameters while running on the terminal, enhancing the flexibility of the code. These parameters include whether to use Comet, the data source, the evaluation source, the data type, the split type, the random state, the batch size, the number of epochs, and the learning rate.

- **Error Handling and Logging:** Proper error handling mechanisms are in place to ensure the robustness of the code. Logging is used extensively to track the flow of execution, making it easier to diagnose and fix issues.

- **Reproducibility:** The use of a random seed ensures that the results of the model are reproducible. This is particularly important in machine learning projects, where the randomness in splitting the dataset and initializing the model can lead to variations in the results.

- **Code Readability and Documentation:** The code is written in a clear and concise manner, making it easy to read and understand. Each method and class is documented with comments, providing valuable context and explanation for the code.

In summary, the implementation of DeepResponse demonstrates a commitment to reliability and adherence to best practices in computer science. By focusing on these principles, the project ensures the development of a robust, reliable, and efficient machine learning model for drug response prediction.

The DeepResponse project is designed for ease of use. It utilizes a Conda environment, which can be effortlessly set up with the provided environment files. All necessary codes and datasets are openly shared, enabling replication or extension of the work. The model can be run directly from the terminal with various customizable parameters, allowing control over aspects like data source, evaluation source, and more. Additionally, the project includes support for Comet, a platform for tracking machine learning experiments, further enhancing its usability and functionality. The code can be runned via terminal with the following statement:

```
python3 -m deep_response [--use_comet --data_source --evaluation_source --data_type --
split_type --random_state --batch_size --epoch --learning_rate]
```

5. DISCUSSION

It has been observed that the Genomics of Drug Sensitivity in Cancer (GDSC) dataset outperformed the Cancer Cell Line Encyclopedia (CCLE) and National Cancer Institute (NCI) datasets across all splits. This superior performance could be primarily attributed to the larger size of the GDSC dataset, which encompasses approximately 339,000 rows, compared to the CCLE and NCI datasets, which contain around 13,000 and 11,000 rows respectively. However, it’s important to note that the size of the dataset may not be the sole determinant of performance. It is plausible that the GDSC dataset provides a more comprehensive and diverse representation of cell lines and pIC50 values, which could contribute to its enhanced performance. This hypothesis is further supported by the observation that despite having similar row sizes, the CCLE and NCI datasets exhibit different performances. This suggests that factors other than size, such as the diversity and representativeness of the data, could play a significant role in determining the performance of these datasets.

Given the nature of biological data, it is also crucial to consider the impact of the source of the data on the experimental outcomes. Biological experiments are inherently complex and can yield varying outcomes based on a multitude of factors, including the experimental conditions, the techniques used, and the source of the biological samples. Therefore, it is possible that the disparate results observed across the GDSC, CCLE, and NCI datasets could be due to differences in the experimental data from different sources. Therefore, while the larger size of the GDSC dataset likely contributes to its superior performance, other factors such as the quality and diversity of its experimental data could also play a significant role. This underscores the importance of considering multiple factors when evaluating the performance of datasets in the context of biological research. It also highlights the need for careful and thorough experimental design and data collection to ensure the reliability and validity of the results.

Shallow machine learning models, despite their simplicity, have shown comparable results in less complex test scenarios, such as random splits. This is likely because in a random split, the data is divided arbitrarily, allowing the model to converge more easily due to the lack of inherent structure or stratification in the split.

This randomness in data division reduces the complexity of the learning task, making it more manageable for shallow models. On the other hand, deep learning architectures, like the DeepResponse, are typically more advantageous in complex test scenarios, such as stratified splits. These architectures have the ability to learn more abstract representations and generalize the problem better, making them more suited for handling the complexity introduced by stratification.

However, the performance of the proposed deep learning models in stratified splits and cross-domain analysis was not as robust as random split. While these models are inherently capable of handling complex scenarios, the increased complexity introduced by stratified splits and cross-domain analyses posed significant challenges. The performance in these scenarios was not solely a reflection of the model's capabilities but also indicative of other factors such as data quality and the representations used. Stratified splits, by their nature, are more challenging than random splits due to the structured division of data. This structure introduces an additional layer of complexity that the model needs to navigate. Similarly, cross-domain analyses involve dealing with data from different domains, each with its unique characteristics and complexities. These scenarios demand not just a robust model, but also high-quality data and effective representations.
In the context of DeepResponse, a hybrid model that requires multi-input drug and cell line data, it was observed that learning primarily stems from the drug side. This could be attributed to the fact that Graph Transformer Neural Networks (GTNNs), used for the drug data, might be a better choice for representing the data and learning capacity compared to Convolutional Neural Networks (CNNs) used for cell line data. GTNNs are particularly adept at handling graph-structured data, which is often the case with drug data. Drugs can be represented as molecular graphs, where atoms are nodes and bonds are edges. This representation allows GTNNs to capture the intricate relationships and properties of drugs, such as their 3D conformation, chemical properties, and potential interactions with other molecules. On the other hand, CNNs, traditionally used for images, might not be as effective in capturing and representing cell line data, which may not have some of these properties.

Furthermore, this discussion underscores the importance of not only focusing on prediction performance but also understanding the underlying learning process. It highlights the need for continuous investigation into model interpretability, which can lead to more robust and trustworthy models. As machine learning models become more complex and are used in more critical applications, the need for interpretability becomes even more important. By continuously investigating model interpretability, it can be ensured that the models are not only high-performing but also transparent, and reliable.
In the context of the ablation study conducted, the systematic removal of individual omic features from the model provided valuable insights into their individual contributions to the model's overall performance. The gene expression feature emerged as a significant contributor, as evidenced by the noticeable decrease in all performance metrics when it was removed. This suggests that gene expression data, likely due to its higher variance, provides a more informative dataset for the model to learn from. In contrast, mutation data, which is predominantly zero, offers limited information gain due to its low variance. However, when mutation data was combined with other omic features such as gene expression, methylation, and copy number variation, and then removed, the performance metrics decreased even further. This indicates that while mutation data may have a lesser impact on its own, its combination with other omic features can significantly influence the model's predictive capabilities.

Similarly, the removal of methylation and copy number variation also resulted in a decrease in performance metrics, albeit to a lesser extent than gene expression. This underscores the significance of gene expression data in the model and highlights the potential benefits of incorporating diverse data types with higher variance. The results from the ablation study underscore the importance of understanding the role of each omic feature in the performance of the model. It also highlights the need for continuous investigation into model interpretability, which can lead to the development of more robust models capable of capturing the complex interplay of factors that influence drug response, ultimately enhancing the accuracy and utility of our predictions. In conclusion, the ablation study provides a deeper understanding of the individual and combined effects of different omic features on the model's performance.

It emphasizes the importance of gene expression data and the potential benefits of incorporating diverse data types with higher variance. This expanded approach could lead to the development of more robust models, capable of capturing the complex interplay of factors that influence drug response, ultimately enhancing the accuracy and utility of our predictions.

6. CONCLUSION

This research introduces DeepResponse, a deep learning-based system designed to predict drug responses in cancer cell lines, paving the way for personalized treatment strategies in oncology.

This research utilized three databases, GDSC, CCLE, and NCI-60, each with unique strengths in drug response analysis, genetic information, and drug variety. The data was meticulously processed, with missing values imputed based on the type of data and the proportions of missing values. The data manipulation process involved standardizing the data, merging data from different sources, and integrating features from both drugs and cell lines. The final dataset, organized into cell line name, drug name, and pIC50 values, was enriched with drug and cell line features, including gene expression, mutation, methylation, and copy number variation data.

The DeepResponse model, a complex hybrid deep learning architecture, was developed to predict drug responses in cancer cell lines. It leverages the strengths of Convolutional Neural Networks (CNNs) and Graph Transformer Neural Networks (GTNNs) to process cell line data and drug molecule data respectively. The CNNs capture local patterns in the cell line data, while the GTNNs comprehend the complex structures of drug molecules. The outputs from both models are then fused and fed into a Multi-Layer Perceptron (MLP) which generates the final pIC50 prediction. This architecture effectively combines the unique strengths of different deep learning models, providing an optimal solution for drug response prediction. The model training process involved meticulous data preparation and organization. The cell line data was processed into a format suitable for Convolutional Neural Networks (CNNs), while the drug data, represented as SMILES strings, was converted into graph representations for the Graph Transformer Neural Networks (GTNNs). This data was then combined into a TensorFlow dataset, batched, and prefetched to enhance the efficiency of model training. The Multi-Layer Perceptron (MLP) synthesized and consolidated the information processed by the CNNs and GTNNs to generate the final pIC50 prediction. This comprehensive approach to data management and model training sets the stage for effective and efficient prediction of drug responses in cancer cell lines.

The model's performance was evaluated using a variety of metrics. The Huber loss function was used for its robustness in handling outliers. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were used to quantify the difference between the predicted and actual pIC50 values. In addition to these regression metrics, several classification metrics were calculated after binarizing the data. These included Accuracy, Precision, Recall, F1 Score, and Matthew's Correlation Coefficient (MCC). Each of these metrics provided a unique perspective on the model's performance, contributing to a comprehensive evaluation of its ability to predict drug responses.

The comprehensive evaluation of the predictive model across various data split scenarios provided significant insights into its performance. The model demonstrated robust performance in the Random Split scenario, where the data was divided without considering any structure. This suggests that the model can accurately predict drug responses when the data is split randomly, providing a strong baseline for its performance. However, when the model was evaluated under more challenging scenarios, such as the Cell Stratified and Drug Stratified splits, its performance metrics were lower. These scenarios tested the model's ability to generalize across different cell types and to new drugs, respectively. The lower performance in these scenarios indicates that predicting responses to new cell types or drugs is a complex task for the model, suggesting that the model might be learning too much from the specific characteristics in the training data, thereby limiting its ability to generalize to new instances. The Drug-Cell Stratified Split scenario, which involved evaluating the model on unseen drug-cell combinations, presented an even more rigorous test of the model's predictive capabilities. The lower performance in this scenario underscores the complexity of predicting drug responses for new drug-cell combinations. It suggests that the model may be overfitting to specific drug-cell combinations in the training data and struggling to generalize to new combinations in the validation and test sets. Finally, the Cross Domain Split scenario tested the model's adaptability to new domains. The model's performance varied significantly depending on the training and testing domains, emphasizing the need for models that can effectively transfer their learning from one domain to another.

This scenario underscores the importance of considering the real-world complexity and variability of the data when preparing it for model training and evaluation. These findings highlight the importance of rigorous evaluation methods to fully assess a model's performance, ensuring its robustness and reliability. They also underscore the complexity and variability inherent in predicting drug responses, emphasizing the need for models that can effectively handle these challenges. Despite the lower performance in more complex scenarios, the model's promising results in less challenging scenarios demonstrate its potential for accurately predicting drug responses.

DeepResponse has demonstrated superior performance compared to existing models in the field of drug response prediction. Across all data split scenarios, including Random Split, Cell Stratified Split, Drug Stratified Split, and Cross Domain Split, DeepResponse consistently achieved the lowest RMSE values. This indicates that DeepResponse's predictions are consistently closer to the actual drug responses compared to other models. Furthermore, the relatively low standard deviation of DeepResponse in all scenarios underscores its robustness and strong generalization capacity, regardless of the randomness or complexity of the data splits. This consistent performance of DeepResponse, regardless of how the data is split, validates its effectiveness in drug response prediction tasks. Therefore, DeepResponse not only outperforms other models across a broad spectrum of test conditions but also delivers reliable results in both controlled and complex real-world scenarios. This distinguishes DeepResponse from other models and highlights its potential as a powerful tool in the field of drug response prediction.

In the use case analysis, the DeepResponse model was applied to various tissue data and exhibited optimal performance on the digestive system. This suggests the potential of using tissue-specific models in predicting drug responses. The model was then used to evaluate drug candidates for repurposing against hepatocellular carcinoma (HCC). Among the evaluated drugs, Eprinomectin, an approved avermectin, demonstrated high predicted activity across all tested HCC cell lines. Subsequent wet lab experiments confirmed the cytotoxicity of Eprinomectin against HCC cells, indicating its potential as a treatment for HCC. The results showed that Eprinomectin induced G1 arrest and apoptosis in HCC cell lines and affected the levels of certain cell cycle proteins.

These findings highlight the potential of Eprinomectin as a treatment for HCC and demonstrate the utility of the DeepResponse model in identifying promising candidates for drug repurposing in cancer treatment. However, further analysis is required to better assess the effects of Eprinomectin on both cancerous and healthy human cells.

In conclusion, the DeepResponse model offers a promising approach for predicting drug sensitivity of cancer cells, with potential applications in the early-stage discovery of new drug candidates and the repurposing of existing ones against resistant tumors. The project demonstrates the power of combining artificial learning techniques with multi-omics data in the field of drug response prediction.

However, the project also highlighted several areas for future research. One such area is the exploration of other machine learning models. The current project utilizes a Graph Transformer Neural Network for learning from drug data and a Convolutional Neural Network for cell line data. However, the potential for improved performance could be offered by other machine learning models and architectures. For instance, recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) could be explored for their ability to capture sequential information in the data.

Another area of interest is the incorporation of additional data types. The project currently uses drug descriptors and cell line data, but the model's predictive power could be enhanced, and a more comprehensive understanding of drug response could be provided by including additional types of data, such as genetic data, clinical data, or real-world patient data. The development of an interactive tool or platform is also a promising direction. Such a tool could allow researchers to utilize the model for predicting drug response and could include features for uploading custom data, adjusting model parameters, and visualizing the results, thereby making the model more accessible for practical use.

While the focus of the current project is on prediction performance, the interpretability of machine learning models is another important aspect. Future work could investigate why the model makes certain predictions, which can provide valuable insights and lead to more trust in the model's predictions.

The current project focuses on hepatocellular carcinoma, but the approach could be extended to other types of cancer or even other diseases. This would involve adjusting the model to handle the specific characteristics of these other conditions, potentially broadening the impact of the work.

Finally, a longitudinal study could be conducted to validate the model's predictions over time. This would involve using the model to make predictions, conducting experiments to test these predictions, and then refining the model based on the results. This could provide a robust validation of the model's predictive power. These areas of future research highlight the potential for further development and refinement of the current model.

REFERENCES

[1] J.-P. Gillet, S. Varma, M.M. Gottesman, The clinical relevance of cancer cell lines, J. Natl. Cancer Inst. 105 **(2013)** 452–458.

[2] K.K. Filipski, L.E. Mechanic, R. Long, A.N. Freedman, Pharmacogenomics in oncology care, Front. Genet. 5 **(2014)** 73.

[3] A.D. Roses, Pharmacogenetics and the practice of medicine, Nature 405 **(2000)** 857–865.

[4] D.M. Roden, R.B. Altman, N.L. Benowitz, D.A. Flockhart, K.M. Giacomini, J.A. Johnson, R.M. Krauss, H.L. McLeod, M.J. Ratain, M.V. Relling, H.Z. Ring, A.R. Shuldiner, R.M. Weinshilboum, S.T. Weiss, Pharmacogenetics Research Network, Pharmacogenomics: challenges and opportunities, Ann. Intern. Med. 145 **(2006)** 749–757.

[5] W.W. Weber, Pharmacogenetics: from description to prediction, Clin. Lab. Med. 28 **(2008)** 499–511.

[6] J. Cook, G. Hunter, J.A. Vernon, The future costs, risks and rewards of drug development: the economics of pharmacogenomics, PharmacoEconomics 27 **(2009)** 355–363.

[7] D. Wang, J. Hensman, G. Kutkaite, T.S. Toh, A. Galhoz, GDSC Screening Team, J.R. Dry, J. Saez-Rodriguez, M.J. Garnett, M.P. Menden, F. Dondelinger, A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates, eLife 9 **(2020)** e60352.

[8] W.H. Dere, T.S. Suto, The role of pharmacogenetics and pharmacogenomics in improving translational medicine, Clin. Cases Miner. Bone Metab. Off. J. Ital. Soc. Osteoporos. Miner. Metab. Skelet. Dis. 6 **(2009)** 13–16.

[9] D.R. Withrow, A. Berrington de González, S. Spillane, N.D. Freedman, A.F. Best, Y. Chen, M.S. Shiels, Trends in Mortality Due to Cancer in the United States by Age and County-Level Income, 1999-2015, J. Natl. Cancer Inst. 111 **(2019)** 863–866.

[10] R. Rafique, S.M.R. Islam, J.U. Kazi, Machine learning in the prediction of cancer therapy, Comput. Struct. Biotechnol. J. 19 **(2021)** 4003–4017.

[11] A. Carrel, M.T. Burrows, CULTIVATION OF TISSUES IN VITRO AND ITS TECHNIQUE, J. Exp. Med. 13 **(1911)** 387–396.

[12] X. Wu, J. Su, J. Wei, N. Jiang, X. Ge, Recent Advances in Three-Dimensional Stem Cell Culture Systems and Applications, Stem Cells Int. 2021 **(2021)** 1–13.

[13] S. Kamiloglu, G. Sari, T. Ozdal, E. Capanoglu, Guidelines for cell viability assays, Food Front. 1 **(2020)** 332–349.

[14] E. Healing, C.F. Charlier, L.B. Meira, R.M. Elliott, A panel of colorimetric assays to measure enzymatic activity in the base excision DNA repair pathway, Nucleic Acids Res. 47 **(2019)** e61.

[15] H. Kim, C.M. Rebholz, Metabolomic Biomarkers of Healthy Dietary Patterns and Cardiovascular Outcomes, Curr. Atheroscler. Rep. 23 **(2021)** 26.

[16] Y.-R.A. Yu, E.G. O'Koren, D.F. Hotten, M.J. Kan, D. Kopin, E.R. Nelson, L. Que, M.D. Gunn, A Protocol for the Comprehensive Flow Cytometric Analysis of Immune Cells in Normal and Inflamed Murine Non-Lymphoid Tissues, PloS One 11 **(2016)** e0150606.

[17] T. Kalliokoski, C. Kramer, A. Vulpetti, P. Gedeck, Comparability of mixed IC_{50} data - a statistical analysis, PloS One 8 **(2013)** e61007.

[18] A. Thakur, A. Kumar, V. Sharma, V. Mehta, PIC50: An open source tool for interconversion of PIC_{50} values and IC_{50} for efficient data representation and analysis, Bioinformatics, **(2022)**.

[19] D.A. Volpe, S.S. Hamed, L.K. Zhang, Use of Different Parameters and Equations for Calculation of IC50 Values in Efflux Assays: Potential Sources of Variability in IC50 Determination, AAPS J. 16 **(2014)** 172–180.

[20] W.W. Focke, I. Van Der Westhuizen, N. Musee, M.T. Loots, Kinetic interpretation of log-logistic dose-time response curves, Sci. Rep. 7 **(2017)** 2234.

[21] I. Bácskay, D. Nemes, F. Fenyvesi, J. Váradi, G. Vasvári, P. Fehér, M. Vecsernyés, Z. Ujhelyi, Role of Cytotoxicity Experiments in Pharmaceutical Development, in: T.A. Çelik (Ed.), Cytotoxicity, InTech, **(2018)**.

[22] H. Zhu, N.J. Gooderham, Mechanisms of Induction of Cell Cycle Arrest and Cell Death by Cryptolepine in Human Lung Adenocarcinoma A549 Cells, Toxicol. Sci. 91 **(2006)** 132–139.

[23] S. Elmore, Apoptosis: A Review of Programmed Cell Death, Toxicol. Pathol. 35 **(2007)** 495–516.

[24] Apoptosis, (n.d.). https://www.genome.gov/genetics-glossary/apoptosis (accessed **January 15, 2024**).

[25] J.M. Zielinski, J.J. Luke, S. Guglietta, C. Krieg, High Throughput Multi-Omics Approaches for Clinical Trial Evaluation and Drug Discovery, Front. Immunol. 12 **(2021)** 590742.

[26] V. Gambardella, N. Tarazona, J.M. Cejalvo, P. Lombardi, M. Huerta, S. Roselló, T. Fleitas, D. Roda, A. Cervantes, Personalized Medicine: Recent Progress in Cancer Therapy, Cancers 12 **(2020)** 1009.

[27] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, C.J. Wilson, J. Lehár, G.V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M.F. Berger, J.E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F.A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I.H. Engels, J. Cheng, G.K. Yu, J. Yu, P. Aspesi, M. De Silva, K. Jagtap, M.D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R.C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J.P. Mesirov, S.B. Gabriel, G. Getz, K. Ardlie, V. Chan, V.E. Myer, B.L. Weber, J. Porter, M. Warmuth, P. Finan, J.L. Harris, M. Meyerson, T.R. Golub, M.P. Morrissey, W.R. Sellers, R. Schlegel, L.A. Garraway, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature 483 **(2012)** 603–607.

[28] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, S. Ramaswamy, P.A. Futreal, D.A. Haber, M.R. Stratton, C. Benes, U. McDermott, M.J. Garnett, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 **(2012)** D955–D961.

[29] R.H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen, Nat. Rev. Cancer 6 **(2006)** 813–823.

[30] M. Iwata, L. Yuan, Q. Zhao, Y. Tabei, F. Berenger, R. Sawada, S. Akiyoshi, M. Hamano, Y. Yamanishi, Predicting drug-induced transcriptome responses of a wide range of human cell lines by a novel tensor-train decomposition algorithm, Bioinformatics 35 **(2019)** i191–i199.

[31] G. Emilien, Impact of genomics on drug discovery and clinical medicine, QJM 93 **(2000)** 391–423.

[32] W. Tansey, K. Li, H. Zhang, S.W. Linderman, R. Rabadan, D.M. Blei, C.H. Wiggins, Dose–response modeling in high-throughput cancer drug screenings: an end-to-end approach, Biostatistics 23 **(2022)** 643–665.

[33] CNCB-NGDC Members and Partners, X. Bai, Y. Bao, S. Bei, C. Bu, R. Cao, Y. Cao, H. Cen, J. Chao, F. Chen, H. Chen, K. Chen, M. Chen, M. Chen, M. Chen, Q. Chen, R. Chen, S. Chen, T. Chen, X. Chen, X. Chen, Y. Cheng, Y. Chu, Q. Cui, L. Dong, Z. Du, G. Duan, S. Fan, Z. Fan, X. Fang, Z. Fang, Z. Feng, S. Fu, F. Gao, G. Gao, H. Gao, W. Gao, X. Gao, X. Gao, X. Gao, J. Gong, J. Gong, Y. Gou, S. Gu, A.-Y. Guo, G. Guo, X. Guo, C. Han, D. Hao, L. Hao, Q. He, S. He, S. He, W. Hu, K. Huang, T. Huang, X. Huang, Y. Huang, P. Jia, Y. Jia, C. Jiang, M. Jiang, S.

Jiang, T. Jiang, X. Jiang, E. Jin, W. Jin, H. Kang, H. Kang, D. Kong, L. Lan, W. Lei, C.-Y. Li, C. Li, C. Li, H. Li, J. Li, J. Li, L. Li, P. Li, R. Li, X. Li, Y. Li, Y. Li, Z. Li, X. Liao, S. Lin, Y. Lin, Y. Ling, B. Liu, C.-J. Liu, D. Liu, G.-H. Liu, L. Liu, S. Liu, W. Liu, X. Liu, X. Liu, Y. Liu, Y. Liu, M. Lu, T. Lu, H. Luo, H. Luo, M. Luo, S. Luo, X. Luo, L. Ma, Y. Ma, J. Mai, J. Meng, X. Meng, Y. Meng, Y. Meng, W. Miao, Y.-R. Miao, L. Ni, Z. Nie, G. Niu, X. Niu, Y. Niu, R. Pan, S. Pan, D. Peng, J. Peng, J. Qi, Y. Qi, Q. Qian, Y. Qin, H. Qu, J. Ren, J. Ren, Z. Sang, K. Shang, W.-K. Shen, Y. Shen, Y. Shi, S. Song, T. Song, T. Su, J. Sun, Y. Sun, Y. Sun, Y. Sun, B. Tang, D. Tang, Q. Tang, Z. Tang, D. Tian, F. Tian, W. Tian, Z. Tian, A. Wang, G. Wang, G. Wang, J. Wang, J. Wang, P. Wang, P. Wang, W. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Z. Wang, H. Wei, Y. Wei, Z. Wei, D. Wu, G. Wu, S. Wu, S. Wu, W. Wu, W. Wu, Z. Wu, Z. Xia, J. Xiao, L. Xiao, Y. Xiao, G. Xie, G.-Y. Xie, J. Xie, Y. Xie, J. Xiong, Z. Xiong, D. Xu, S. Xu, T. Xu, T. Xu, Y. Xue, Y. Xue, C. Yan, D. Yang, F. Yang, F. Yang, H. Yang, J. Yang, K. Yang, N. Yang, Q.-Y. Yang, S. Yang, X. Yang, X. Yang, X. Yang, Y.-G. Yang, W. Ye, C. Yu, F. Yu, S. Yu, C. Yuan, H. Yuan, J. Zeng, S. Zhai, C. Zhang, F. Zhang, G. Zhang, M. Zhang, P. Zhang, Q. Zhang, R. Zhang, S. Zhang, W. Zhang, W. Zhang, W. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y.E. Zhang, Y. Zhang, Z. Zhang, Z. Zhang, D. Zhao, F. Zhao, G. Zhao, M. Zhao, W. Zhao, W. Zhao, X. Zhao, Y. Zhao, Y. Zhao, Z. Zhao, X. Zheng, Y. Zheng, C. Zhou, H. Zhou, X. Zhou, X. Zhou, Y. Zhou, Y. Zhou, J. Zhu, L. Zhu, R. Zhu, T. Zhu, W. Zong, D. Zou, Z. Zuo, Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024, Nucleic Acids Res. 52 **(2024)** D18–D32.

[34] I. Tzoulaki, T.M.D. Ebbels, A. Valdes, P. Elliott, J.P.A. Ioannidis, Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies, Am. J. Epidemiol. 180 **(2014)** 129–139.

[35] L. Wen, F. Tang, Recent advances in single-cell sequencing technologies, Precis. Clin. Med. 5 **(2022)** pbac002.

[36] Y. Li, D.M. Umbach, J.M. Krahn, I. Shats, X. Li, L. Li, Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines, BMC Genomics 22 **(2021)** 272.

[37] S. Savas, G. Liu, Studying Genetic Variations in Cancer Prognosis (and Risk): A Primer for Clinicians, The Oncologist 14 **(2009)** 657–666.

[38] B.N. Chao, D.M. Carrick, K.K. Filipski, S.A. Nelson, Overview of Research on Germline Genetic Variation in Immune Genes and Cancer Outcomes, Cancer Epidemiol. Biomarkers Prev. 31 **(2022)** 495–506.

[39] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van De Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, Nat. Genet. 24 **(2000)** 227–235.

[40] S. Tamang, Gene Expression: Stages, Regulations, Methods, **(2023)**. https://microbenotes.com/gene-expression/ (accessed **January 15, 2024**).

[41] Y.-Y. Wang, H. Kang, T. Xu, L. Hao, Y. Bao, P. Jia, CeDR Atlas: a knowledgebase of cellular drug response, Nucleic Acids Res. 50 **(2022)** D1164–D1171.

[42] S. Quazi, Artificial intelligence and machine learning in precision and genomic medicine, Med. Oncol. 39 **(2022)** 120.

[43] R. Nussinov, H. Jang, C.-J. Tsai, F. Cheng, Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers, PLOS Comput. Biol. 15 **(2019)** e1006658.

[44] R. Hu, H. Xu, P. Jia, Z. Zhao, KinaseMD: kinase mutations and drug response database, Nucleic Acids Res. 49 **(2021)** D552–D561.

[45] Treating Mutations in Cancer Research | LIDE Biotech, (n.d.). https://www.lidebiotech.com/blog/cancer-mutations (accessed **January 15, 2024**).

[46] A. Tafazoli, H.-J. Guchelaar, W. Miltyk, A.J. Kretowski, J.J. Swen, Applying Next-Generation Sequencing Platforms for Pharmacogenomic Testing in Clinical Practice, Front. Pharmacol. 12 **(2021)** 693453.

[47] R. Sun, C. Du, J. Li, Y. Zhou, W. Xiong, J. Xiang, J. Liu, Z. Xiao, L. Fang, Z. Li, Systematic Investigation of DNA Methylation Associated With Platinum Chemotherapy Resistance Across 13 Cancer Types, Front. Pharmacol. 12 **(2021)** 616529.

[48] E. Shantsila, Predicting Age with DNA methylation data, Medium **(2021)**. https://towardsdatascience.com/predicting-age-with-dna-methylation-data-99043406084 (accessed **January 15, 2024**).

[49] A. Valsesia, A. Macé, S. Jacquemont, J.S. Beckmann, Z. Kutalik, The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation, Front. Genet. 4 **(2013)**.

[50] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 **(1988)** 31–36.

[51] A. Park, Y. Lee, S. Nam, A performance evaluation of drug response prediction models for individual drugs, Sci. Rep. 13 **(2023)** 11911.

[52] L. Wang, X. Li, L. Zhang, Q. Gao, Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization, BMC Cancer 17 **(2017)** 513.

[53] B.M. Kuenzi, J. Park, S.H. Fong, K.S. Sanchez, J. Lee, J.F. Kreisberg, J. Ma, T. Ideker, Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells, Cancer Cell 38 **(2020)** 672-684.e6.

[54] P. Liu, H. Li, S. Li, K.-S. Leung, Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network, BMC Bioinformatics 20 **(2019)** 408.

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

[56] T. Nguyen, G.T.T. Nguyen, T. Nguyen, D.-H. Le, Graph Convolutional Networks for Drug Response Prediction, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 **(2022)** 146–154.

[57] L. Rampasek, D. Hidru, P. Smirnov, B. Haibe-Kains, A. Goldenberg, Dr.VAE: Drug Response Variational Autoencoder, **(2017)**.

[58] Y. Chang, H. Park, H.-J. Yang, S. Lee, K.-Y. Lee, T.S. Kim, J. Jung, J.-M. Shin, Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature, Sci. Rep. 8 **(2018)** 8857.

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

[60] M. Ammad-ud-din, S.A. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio, S. Kaski, Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization, Bioinformatics 32 **(2016)** i455–i463.

[61] Home page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer, (n.d.). https://www.cancerrxgene.org/ (accessed **January 14, 2024**).

[62] Cancer Cell Line Encyclopedia (CCLE), (n.d.). https://sites.broadinstitute.org/ccle/ (accessed **January 14, 2024**).

[63] NCI-60 Human Tumor Cell Lines Screen | Discovery & Development Services | Developmental Therapeutics Program (DTP), (n.d.). https://dtp.cancer.gov/discovery_development/nci-60/ (accessed **January 15, 2024**).

[64] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 **(2018)** D1074–D1082.

[65] DrugBank Online, (n.d.). https://go.drugbank.com (accessed **January 15, 2024**).

[66] PubChem, PubChem, (n.d.). https://pubchem.ncbi.nlm.nih.gov/ (accessed **January 15, 2024**).

[67] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2023 update, Nucleic Acids Res. 51 **(2023)** D1373–D1380.

[68] U.O. Ozcan, N. Mohammadvand, B. Izmirli, E. Akar, D.C. Kahraman, T. Doğan, A Multi-Omics and Machine Learning-Based Predictor of Drug Sensitivity in Cancer, in: Int. Symp. Health Inform. Bioinforma., Orta Doğu Teknik Üniversitesi Enformatik Enstitüsü, 2022. https://open.metu.edu.tr/handle/11511/101343 (accessed **February 20, 2024**).

[69] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, J. Gould, J.F. Davis, A.A. Tubelli, J.K. Asiedu, D.L. Lahr, J.E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I.C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O.M. Enache, F. Piccioni, S.A. Johnson, N.J. Lyons, A.H. Berger, A.F. Shamji, A.N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D.Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N.S. Gray, P.A. Clemons, S. Silver, X. Wu, W.-N. Zhao, W. Read-Button, X. Wu, S.J. Haggarty, L.V. Ronco, J.S. Boehm, S.L. Schreiber, J.G. Doench, J.A. Bittker, D.E. Root, B. Wong, T.R. Golub, A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles, Cell 171 **(2017)** 1437-1452.e17.

[70] D.N. Slenter, M. Kutmon, E.L. Willighagen, WikiPathways: Integrating Pathway Knowledge with Clinical Data, in: N. Blau, C. Dionisi Vici, C.R. Ferreira, C. Vianey-Saban, C.D.M. Van Karnebeek (Eds.), Physicians Guide Diagn. Treat. Follow- Inherit. Metab. Dis., Springer International Publishing, Cham, **(2022)**: pp. 1457–1466.

[71] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 Int. Conf. Eng. Technol. ICET, IEEE, Antalya, **(2017)**: pp. 1–6.

[72] Understanding Convolutional Neural Network: A Complete Guide, **(2023)**. https://learnopencv.com/understanding-convolutional-neural-networks-cnn/ (accessed **February 13, 2024**).

[73] S. Yun, M. Jeong, R. Kim, J. Kang, H.J. Kim, Graph Transformer Networks, **(2019)**.

[74] V.P. Dwivedi, X. Bresson, A Generalization of Transformer Networks to Graphs, **(2020)**.

[75] Q. Wu, How to Build Graph Transformers with O(N) Complexity, Medium **(2023)**. https://towardsdatascience.com/how-to-build-graph-transformers-with-o-n-complexity-d507e103d30a (accessed **February 13, 2024**).

[76] AIML.com, What is a Multilayer Perceptron (MLP) or a Feedforward Neural Network (FNN)?, AIML.Com **(2022)**. https://aiml.com/what-is-a-multilayer-perceptron-mlp/ (accessed **February 13, 2024**).

[77] A. Agarwal, Building efficient data pipelines using TensorFlow, Medium **(2019)**. https://towardsdatascience.com/building-efficient-data-pipelines-using-tensorflow-8f647f03b4ce (accessed **February 14, 2024**).