Performance evaluation of different IC50 Prediction Models

Dissertation submitted in partial fulfilment of the

Requirements for the degree of

Master of Technology

in

Computer Science

by

Abhishek Bale
CS2202

Under the guidance of

Professor Utpal Garain

Computer Vision and Pattern Recognition unit

Indian Statistical Institute

June 2024

Computer Vision and Pattern Recognition
Unit,
Indian Statistical Institute
Kolkata – 700108, India

CERTIFICATE

This is to certify that we have examined the thesis entitled **“Performance evaluation of different IC50 Prediction Models”** submitted by **Abhishek Bale** (Roll Number: *CS2202*), a postgraduate student of the Indian Statistical Institute in partial fulfilment for the award of degree of Master of Technology (Computer Science). We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment for the Post Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

Professor Utpal Garain
Computer Vision and Pattern Recognition
Unit
Indian Statistical Institute
Kolkata – 700108, India

DECLARATION

I, Abhishek Bale, a registered M. Tech student in Computer Science at the Indian Statistical Institute, hereby declare that I have fulfilled all the requirements set forth by the institute for the submission of my dissertation.

I further declare that the project presented is original and represents the outcome of my independent investigations and research. It is free from any form of plagiarism and has contributed to the development of new techniques. This work has not been submitted to any other university or institution in pursuit of a degree, diploma, or any other academic recognition.

I also declare that any text, diagrams, or other materials obtained from external sources, including but not limited to books, journals, and the internet, have been properly acknowledged, referenced, and cited to the best of my knowledge and understanding.

Date: 21/06/24

Abhi

Abhishek Bale
M.Tech (CS), CS2202
Indian Statistical institute

ACKNOWLEDGEMENTS

I owe my deep sense of gratitude to my research supervisor **Prof. Utpal Garain**, ISI, Kolkata for giving me an opportunity to work on this project and for providing me with invaluable guidance, feedback and suggestions. I enjoyed the discussions I had with him during the course of this project. I am indeed inspired by his attitude and academic discipline which made this project possible.

I would also like to express my sincere thanks to **Prof. Swagatam Das, Prof. Mandar Mitra, Prof. Shyamal K. De** and **Prof. Debapriyo Majumdar** for teaching me the important concepts necessary for the implementation of this project.

I am hugely thankful to **Mr. Udbhas Garai**, IISER, Mohali for helping me during the project. With his depth in understanding the task from a chemistry perspective, he provided me with constant help whenever required.

I am equally grateful to the Indian Statistical Institute, Kolkata, for allowing me to undertake this project and for providing me with all the required facilities. Lastly, I would like to thank my friends and family for their help and emotional support.

ABSTRACT

In recent years, there has been a surge of interest in using machine learning algorithms (MLAs) in oncology, particularly for biomedical applications such as drug discovery, drug repurposing, diagnostics, clinical trial design, and pharmaceutical production. The accurate prediction of the half-maximal inhibitory concentration (IC50) of chemical compounds is pivotal for advancing personalized medicine and accelerating drug discovery. This dissertation presents a comprehensive performance evaluation of several state-of-the-art IC50 prediction models, including PaccMann, Precily, tCNN, AGMI, and DeepCDR. Each model employs distinct methodologies ranging from graph neural networks and convolutional neural networks to deep learning architectures tailored for multi-omics data integration. The primary objective of this study is to compare these models' predictive capabilities, robustness, and applicability across diverse datasets of cancer cell lines and chemical compounds. We employ rigorous cross-validation techniques and various performance metrics such as mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient to assess each model's effectiveness. Additionally, we analyze the models' performance in terms of computational efficiency and scalability, as these factors are crucial for practical implementation in large-scale drug screening processes. Our findings highlight the strengths and limitations of each approach, providing critical insights into their potential clinical and pharmacological applications. This evaluation aims to guide future research in selecting and optimizing predictive models for IC50, ultimately contributing to more effective and personalized cancer treatments.

CONTENTS

List of Figures

Chapter 1

INTRODUCTION

The prediction of the half-maximal inhibitory concentration (IC50) of chemical compounds is a fundamental task in the fields of pharmacology and oncology. IC50 values are a crucial measure of a drug's potency, representing the concentration required to inhibit a biological process by 50%. Accurate IC50 predictions can significantly enhance the drug discovery process, enabling the identification of promising therapeutic candidates and the optimization of personalized treatment strategies. This dissertation focuses on evaluating the performance of several state-of-the-art IC50 prediction models, namely PaccMann, Precily, tCNN, AGMI and DeepCDR.

1.1 Cell Line and its representations:

A cell line is a population of cells derived from a single cell and grown in a controlled laboratory environment. These cells are often used in biomedical research to study biological processes, drug responses, and disease mechanisms. Cancer cell lines are critical for screening potential anticancer compounds and understanding their effects on different types of cancer cells.

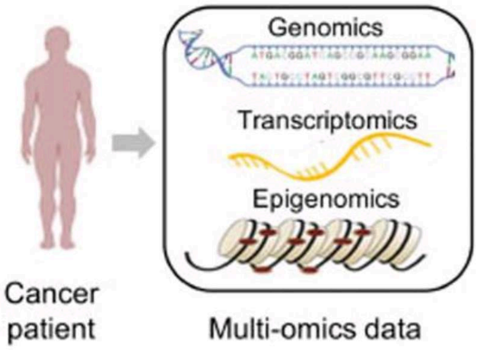
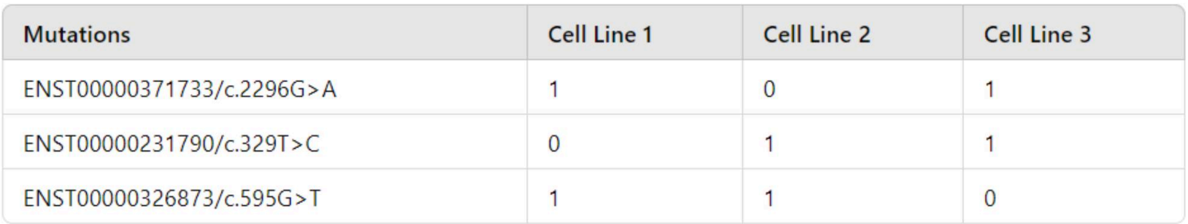


Figure 1: Cancer cell line representations

Cell lines are extensively characterized using various "-omic" profiles, providing comprehensive data crucial for IC50 prediction models. These profiles include genomic, epigenetic, transcriptomic, and proteomic information, each contributing unique insights into the biological state of the cells.

Genomics:

The genomic profile of a disease state is identified through genetic sequencing, revealing key mutations that influence disease onset and outcomes. These mutations can include single-nucleotide variants, insertions, deletions, copy number variations, and translocations. The genomic mutational profile serves as a critical feature for machine learning models, with mutational status and copy number variations frequently used to predict the efficacy of new therapeutics. Example Mutational Profile of three cell lines described below:

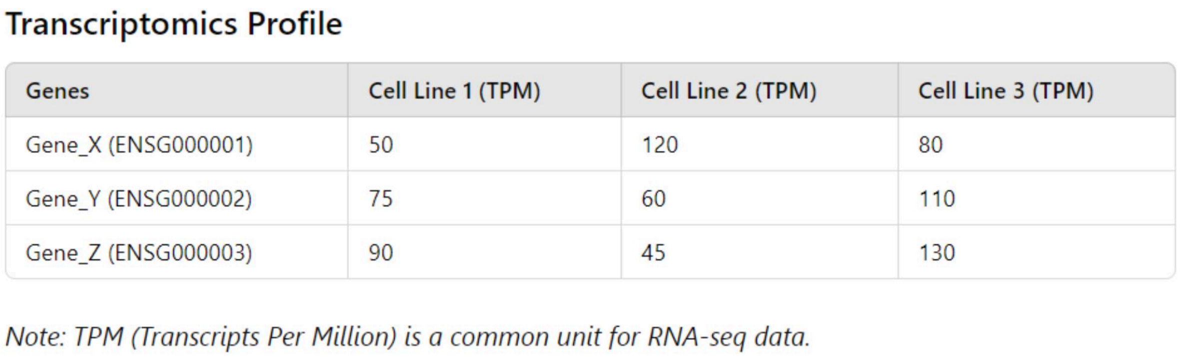


Mutations	Cell Line 1	Cell Line 2	Cell Line 3
ENST00000371733/c.2296G>A	1	0	1
ENST00000231790/c.329T>C	0	1	1
ENST00000326873/c.595G>T	1	1	0

Figure 2: Genomics Profile

Transcriptomics:

Transcriptomic profiles, captured through RNA sequencing (RNA-seq), are widely used in computational bioinformatics. These profiles reveal the degree of mRNA expression, indicating which genes are active or inhibited in each cell. RNA-seq can be conducted on bulk populations or single cells, with high-throughput methods providing spatiotemporal data on mRNA expression changes over time and across different cell regions.



Transcriptomics Profile

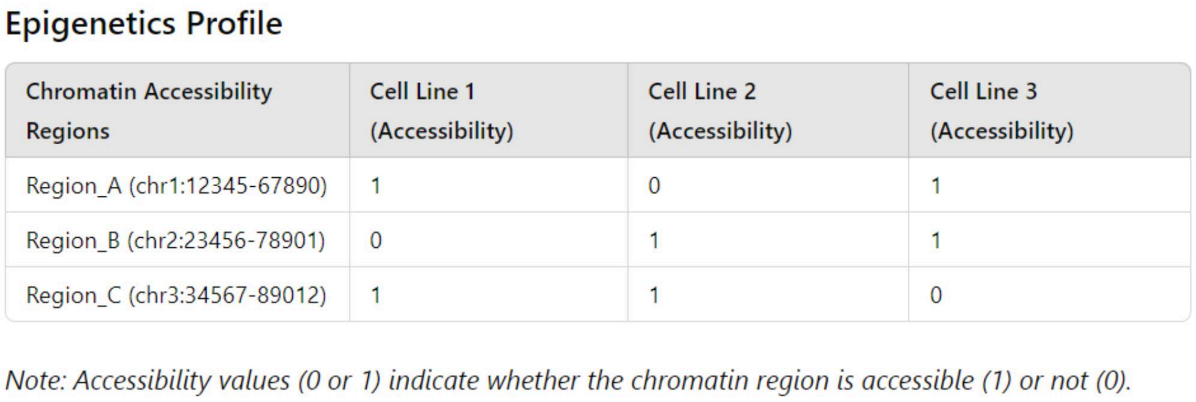
Genes	Cell Line 1 (TPM)	Cell Line 2 (TPM)	Cell Line 3 (TPM)
Gene_X (ENSG000001)	50	120	80
Gene_Y (ENSG000002)	75	60	110
Gene_Z (ENSG000003)	90	45	130

Note: TPM (Transcripts Per Million) is a common unit for RNA-seq data.

Figure 3: Transcriptomics Profile

Epigenetics:

Epigenetic modifications provide deeper insights into the biological processes underlying a disease state. One significant epigenetic feature is chromatin accessibility, assessed using methods like human assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq). Databases of epigenetic information are rapidly growing, offering valuable data for bioinformatics analyses.



Epigenetics Profile

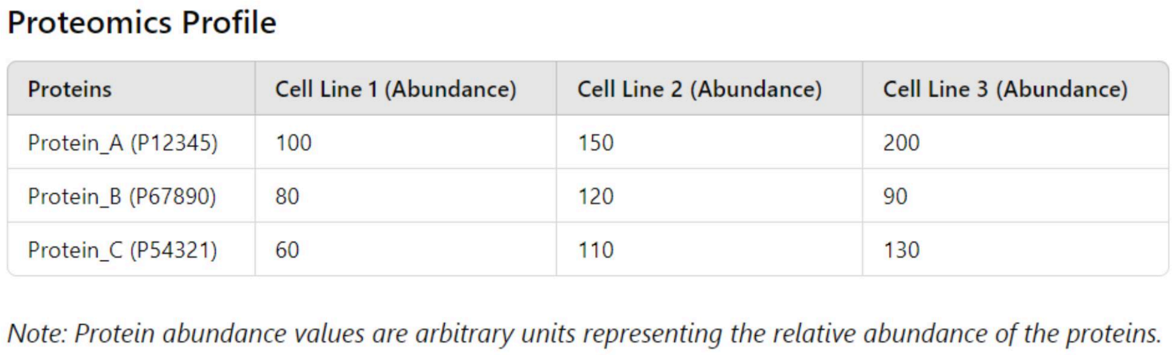
Chromatin Accessibility Regions	Cell Line 1 (Accessibility)	Cell Line 2 (Accessibility)	Cell Line 3 (Accessibility)
Region_A (chr1:12345-67890)	1	0	1
Region_B (chr2:23456-78901)	0	1	1
Region_C (chr3:34567-89012)	1	1	0

Note: Accessibility values (0 or 1) indicate whether the chromatin region is accessible (1) or not (0).

Figure 4: Epigenetics Profile

Proteomics:

Proteomic profiles encompass data on protein structure, properties, interactions, and abundances. Resources like UNIPROT provide structural properties and amino acid sequences of proteins, aiding in the creation of protein embeddings and therapeutic target features. Databases such as CHEMBL and ProteomicsDB offer key features and mass spectrometry data, respectively, which inform drug prediction models and provide a proteomic overview of the disease state.



Proteomics Profile

Proteins	Cell Line 1 (Abundance)	Cell Line 2 (Abundance)	Cell Line 3 (Abundance)
Protein_A (P12345)	100	150	200
Protein_B (P67890)	80	120	90
Protein_C (P54321)	60	110	130

Note: Protein abundance values are arbitrary units representing the relative abundance of the proteins.

Figure 5: Proteomics Profile

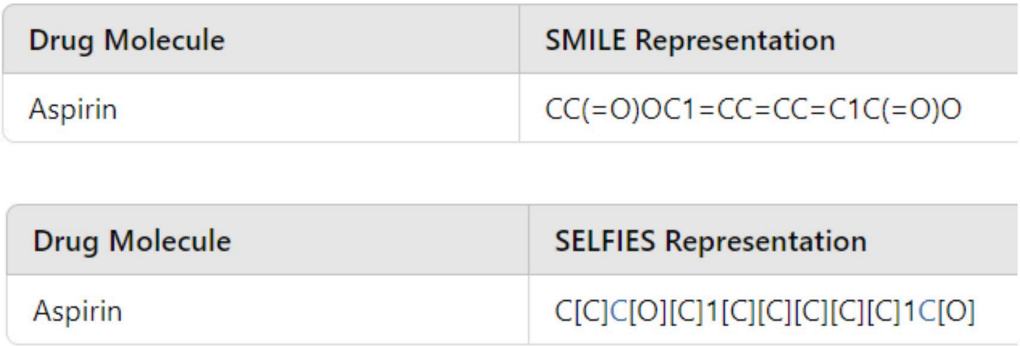
1.2 Drug and its representations:

A drug refers to any chemical compound used for therapeutic purposes, specifically those intended to treat cancer. These compounds can vary widely in their chemical structures and mechanisms of action. Understanding how different drugs interact with various cancer cell lines is essential for developing effective treatments.

Constructing a machine learning (ML) algorithm to connect a molecular state reflecting a disease with the response to a therapeutic intervention requires selecting the best computer-readable form to represent the therapeutic agent. Various methods are available for representing small molecules in a computer-readable manner, each with its own advantages and challenges.

SMILE (Simplified Molecular Input Line Entry System):

SMILE is a chemical annotation system that represents molecular structures using characters for atoms and special symbols for bonds and higher-order structural properties. However, SMILE strings might not always correspond to valid molecules, leading to the development of SELFIES (Self-referencing Embedded Strings) to ensure all generated strings refer to valid molecules. These character representations often need to be converted into numerical forms for ML models.



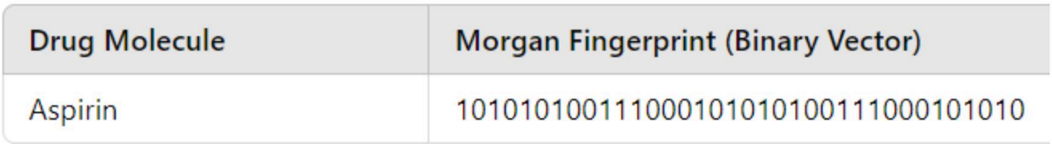
Drug Molecule	SMILE Representation
Aspirin	CC(=O)OC1=CC=CC=C1C(=O)O

Drug Molecule	SELFIES Representation
Aspirin	C[C]C[O][C]1[C][C][C][C][C]1C[O]

Figure 6: SMILES and STRINGS representation

Fingerprinting:

Fingerprinting converts a chemical structure into a binary vector of predetermined size, capturing the structural information of the compound. Morgan fingerprinting is a widely used technique that allows ML models expecting binary vector input to process chemical structures effectively.

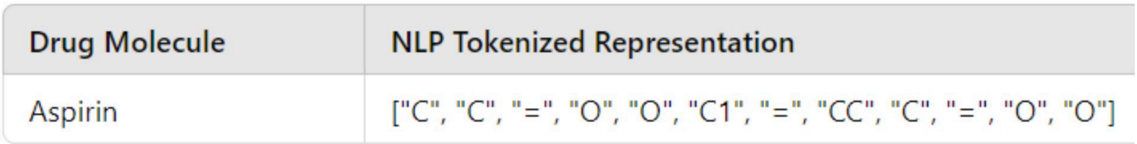


Drug Molecule	Morgan Fingerprint (Binary Vector)
Aspirin	101010100111000101010100111000101010

Figure 7: Fingerprint Representation

Natural Language Processing (NLP):

NLP approaches to chemical structure embedding involve tokenizing SMILE/SELFIE strings and training a specific language model to embed the chemical structure. This method captures higher-order relationships within the molecule and can outperform traditional fingerprinting in various classification tasks.

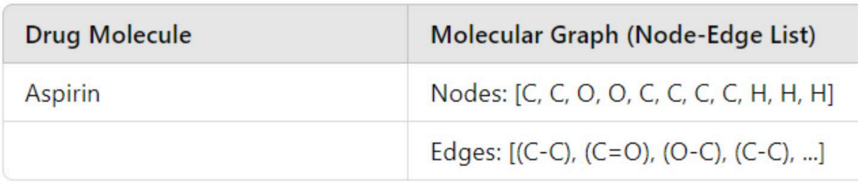


Drug Molecule	NLP Tokenized Representation
Aspirin	["C", "C", "=", "O", "O", "C1", "=", "CC", "C", "=", "O", "O"]

Figure 8: Cell line Representation using NLP Techniques

Molecular Graphs:

This representation is particularly useful for larger therapeutics such as proteins and peptides.



Drug Molecule	Molecular Graph (Node-Edge List)
Aspirin	Nodes: [C, C, O, O, C, C, C, C, H, H, H]
	Edges: [(C-C), (C=O), (O-C), (C-C), ...]

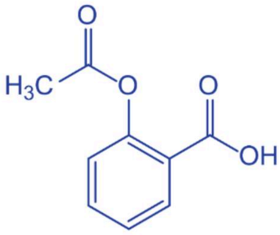
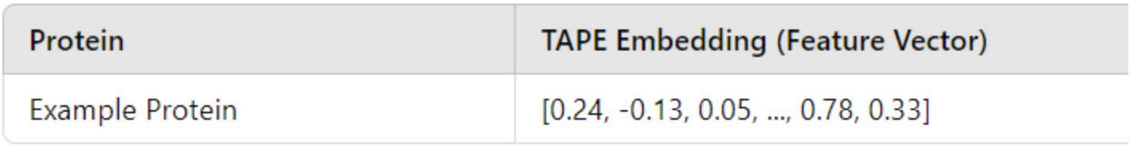


Figure 9: Molecular Graph Representation of a drug

Task-Assisted Protein Embeddings (TAPE):

TAPE builds on NLP and semi-supervised ML paradigms, creating protein embeddings from amino acid sequences through biologically relevant tasks such as structure prediction, homology detection, and protein engineering. These embeddings have been widely adopted in higher-order models like IBM's PaccMannRL.



Protein	TAPE Embedding (Feature Vector)
Example Protein	[0.24, -0.13, 0.05, ..., 0.78, 0.33]

Figure 10: TAPE Representation sample

1.3 Introducing IC50:

Half-maximal inhibitory concentration (IC50) is the most widely used and informative measure of a drug's efficacy. It indicates how much concentration of a drug is needed to inhibit a biological process by half, thus providing a measure of potency.

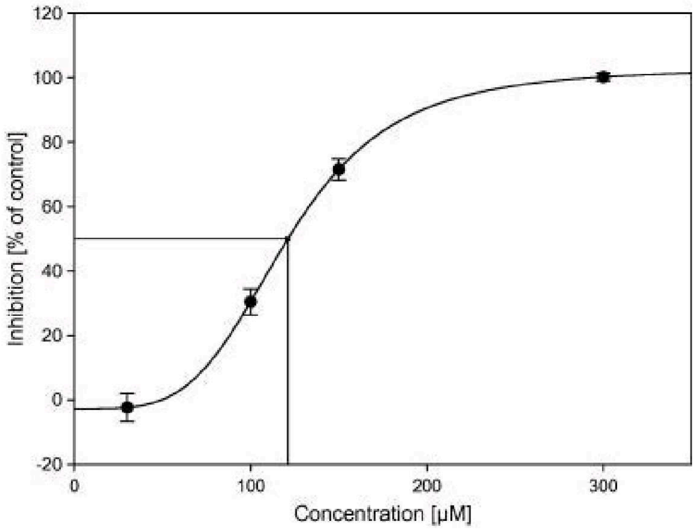


Figure 11: IC50 vs Concentration graph

LN_IC50 stands for the natural logarithm of the half-maximal inhibitory concentration (IC50). The IC50 value represents the concentration of a drug required to inhibit a biological process or the growth of cells by 50%. Taking the natural logarithm (LN) of the IC50 value helps in stabilizing the variance and normalizing the distribution of these values, which is particularly useful for statistical analyses and modelling.

Why Use LN_IC50?

Normalization:

Raw IC50 values can span several orders of magnitude, and taking the natural logarithm helps in compressing this range, making the data more manageable and suitable for various machine learning algorithms.

Stabilizing Variance:

Biological data often exhibit heteroscedasticity, where the variance changes across different levels of an independent variable. The logarithmic transformation helps in stabilizing the variance.

Improving Model Performance:

Many predictive models perform better when the input data follow a normal distribution or have less skewness, which is often achieved through logarithmic transformation.

Chapter 2

PROBLEM DEFINITION

In the rapidly advancing field of drug discovery, predicting the half-maximal inhibitory concentration (IC50) of drug molecules is crucial for understanding their efficacy against specific cell lines. Accurate prediction of IC50 values can significantly streamline the drug development process, reducing time and costs associated with experimental testing. Each model leverages different methodologies and representations of molecular and cellular data to predict drug efficacy. By systematically comparing these models, we aim to identify their strengths and weaknesses, ultimately contributing to the improvement of computational approaches in drug discovery.

2.1 What is the need?

Traditional experimental approaches to determine IC50 values are time-consuming, labour-intensive, and costly.

1. Personalized Medicine:

   Accurate IC50 predictions tailored to specific cell lines can aid in developing personalized treatment plans, improving patient outcomes by identifying the most effective drugs for individual genetic profiles.

2. Ethical Considerations:

   Improving IC50 prediction models reduces the need for animal testing, addressing ethical concerns and minimizing the use of animals in the drug development process.

3. Resource Optimization:

   By improving prediction models, we can optimize the use of resources in pharmaceutical research, focusing efforts on the most promising compounds and reducing the attrition rate of drug development projects.

Chapter 3

DATASET DETAILS

In this study, we utilized publicly available gene expression and drug IC50 data from well-known datasets in cancer research to train and evaluate our models. These datasets provide comprehensive information on the genetic profiles of various cancer cell lines and their responses to a wide range of anticancer compounds, facilitating the development of predictive models for drug sensitivity.

3.1 Genomics of Drug Sensitivity in Cancer (GDSC) Dataset

The GDSC database is a rich resource that includes screening results for over a thousand genetically profiled human pan-cancer cell lines treated with a diverse array of anticancer compounds. These compounds include both traditional chemotherapeutic agents and modern targeted therapeutics from various sources

- It includes 969 Cell Lines data
- Over 290 wide range of anticancer compounds
- GDSC2 dataset includes 2,43,466 Drug-Cell Line pairs with its IC50 values, gene expression, mutations, and copy number variations.

The GDSC dataset has been instrumental in understanding drug response mechanisms and identifying potential therapeutic targets.

3.2 Cancer Cell Line Encyclopedia (CCLE) Dataset

The CCLE dataset provides detailed genomic profiles of a large collection of cancer cell lines. It includes data on gene expression, gene mutations, and copy number variations (CNV), which are crucial for understanding the molecular basis of cancer and predicting drug responses. The CCLE dataset complements the GDSC dataset by providing additional genomic information that enhances the accuracy of predictive models.

- It includes comprehensive genomic profiles of cancer cell lines
- Data includes gene expression, mutations, and CNV
- Widely used for cancer research and drug discovery

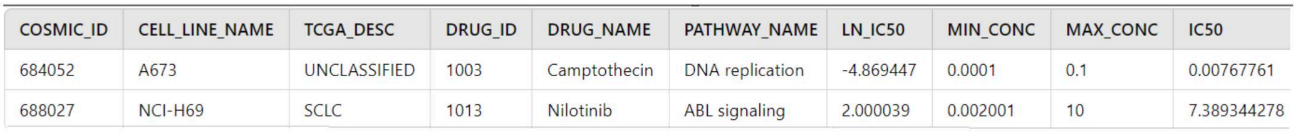
3.3 Drug Datasets

In addition to the GDSC and CCLE datasets, we leveraged various drug datasets that include detailed information on the chemical properties and biological activities of anticancer compounds. These datasets are essential for constructing drug representations that can be used in predictive models.

- PubChem: A large-scale bioactivity database providing information on drug-like compounds and their biological activities.
- DrugBank: A comprehensive resource that combines detailed drug data with drug-target interaction information.

3.4 Data Preparation

GDSC2 includes 243,466 drug-cell line pairs below is the sample dataset, which consist of 969 cancer cell lines and 297 drugs.



COSMIC_ID	CELL_LINE_NAME	TCGA_DESC	DRUG_ID	DRUG_NAME	PATHWAY_NAME	LN_IC50	MIN_CONC	MAX_CONC	IC50
684052	A673	UNCLASSIFIED	1003	Camptothecin	DNA replication	-4.869447	0.0001	0.1	0.00767761
688027	NCI-H69	SCLC	1013	Nilotinib	ABL signaling	2.000039	0.002001	10	7.389344278

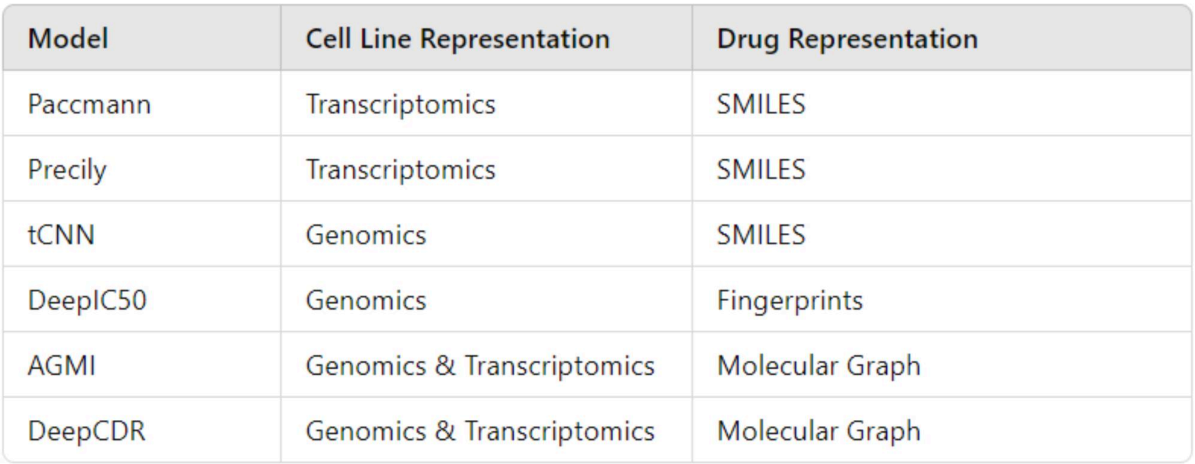
Table 1: GDSC2 Sample dataset

From the initial dataset containing 243,466 drug-cell line pairs, we applied a filtering criterion based on the IC50 values. Specifically, we considered only those drug-cell line pairs where the IC50 value lies between the minimum concentration (MIN_CONC) and the maximum concentration (MAX_CONC) recorded in the dataset. This filtering step is crucial to ensure the reliability of the IC50 values used in our models.

After applying this filter, we retained approximately 51,652 drug-cell line pairs with 969 cell lines and 202 drugs, which provided a robust dataset for training and evaluating our IC50 prediction models.

We will utilize the filtered dataset of 51,652 drug-cell line pairs, where IC50 values lie within the specified concentration ranges, to evaluate the performance of Paccmann, Precily, tCNN, AGMI, and DeepCDR, ensuring reliable and meaningful comparisons across these different approaches to IC50 prediction.

The table below provides a summary of the various predictive models evaluated in this study, highlighting their respective representations for both cell lines and drug molecules. Each model employs distinct methods for encoding biological and chemical information, which is crucial for their performance in predicting drug sensitivity (LN_IC50 values). The cell line representations range from transcriptomics to genomics, while drug representations include SMILES strings, molecular fingerprints, and molecular graphs.



Model	Cell Line Representation	Drug Representation
Paccmann	Transcriptomics	SMILES
Precily	Transcriptomics	SMILES
tCNN	Genomics	SMILES
DeepIC50	Genomics	Fingerprints
AGMI	Genomics & Transcriptomics	Molecular Graph
DeepCDR	Genomics & Transcriptomics	Molecular Graph

Table 2: Cell and Drug Representations of the models

Each of this model predicts LN_IC50 values i.e the natural logarithm of the half-maximal inhibitory concentration (IC50). Taking the natural logarithm (LN) of the IC50 value helps in stabilizing the variance and normalizing the distribution of these values, which is particularly useful for statistical analyses and modelling.

Chapter 4

METHODOLOGY

In this section, we outline the methodology employed to evaluate the performance of various predictive models using a common dataset of 51,652 drug-cell line pairs, filtered to ensure IC50 values lie within specified concentration ranges. This dataset includes gene expression, mutations, and copy number variations for cell lines, along with drug descriptors such as SMILES strings and molecular fingerprints. The use of this common dataset ensures that all models are evaluated on the same basis, allowing for direct comparisons of their performance.

Each predictive model will be evaluated individually, documenting key aspects such as drug embedding, cell line embedding, model structure, and performance of each model. The model structure section will detail the architecture and design of each model, highlighting notable features or techniques such as convolutional neural networks or attention mechanisms. Performance metrics will be assessed using Root Mean Square Error (RMSE), correlation, and R-squared (R^2), providing a comprehensive view of each model's predictive accuracy and reliability.

4.1 PaccMann:

PaccMann is a novel approach for predicting the sensitivity of anticancer compounds using multi-modal attention-based neural networks. This method integrates three critical aspects of drug sensitivity: the molecular structure of compounds, transcriptomic profiles of cancer cells, and prior knowledge about protein interactions within cells. PaccMann processes a drug-cell pair, which consists of the SMILES encoding of a compound and the gene expression profile of a cancer cell, to predict an IC50 sensitivity value.

The PaccMann framework includes three different encoders for SMILES strings: bidirectional recurrent, convolutional, and attention-based encoders. These diverse encoders are designed to capture various structural features of the compounds.

4.1.2 Drug Embedding:

PaccMann employs SMILES drug embedding, utilizing the text encodings to represent the structural information of the compounds. To encode these SMILES strings they have used the attention-based encoders which not only captures the intricate details of the molecular structures effectively but also enhances the overall predictive accuracy of the model.

4.1.3 Cell Line Embedding:

In PaccMann, cell line embedding relies on gene expression data, The STRING protein-protein interaction (PPI) network is employed to incorporate intracellular interactions, Through a weighting and propagation scheme, relevant genes are identified, and the top 20 genes across all compounds are pooled to create a subset of 2,128 informative genes. Now, Cell Lines are represented by the gene expression values of a subset of these 2,128 genes.

4.1.4 Model Architecture:

Cells are represented by the gene expression values of a subset of 2,128 genes, selected for having the highest weights following the network propagation. Compound structures are represented in the SMILES formats. The gene-vector is fed into an attention-based encoder that assigns higher weights to the most informative genes. SMILES encoding of compounds is employed by an array of encoders that are combined with a representation of gene expression to obtain a drug sensitivity prediction.

PaccMann predicts normalised IC50 values which lies between 0 & 1. So it uses the below formula to convert these normalised IC50 value(ỹ) to LN_IC50 value(y).

$$y = \tilde{y} + (\text{ic50\_max} - \text{ic50\_min}) + \text{ic50\_min}$$

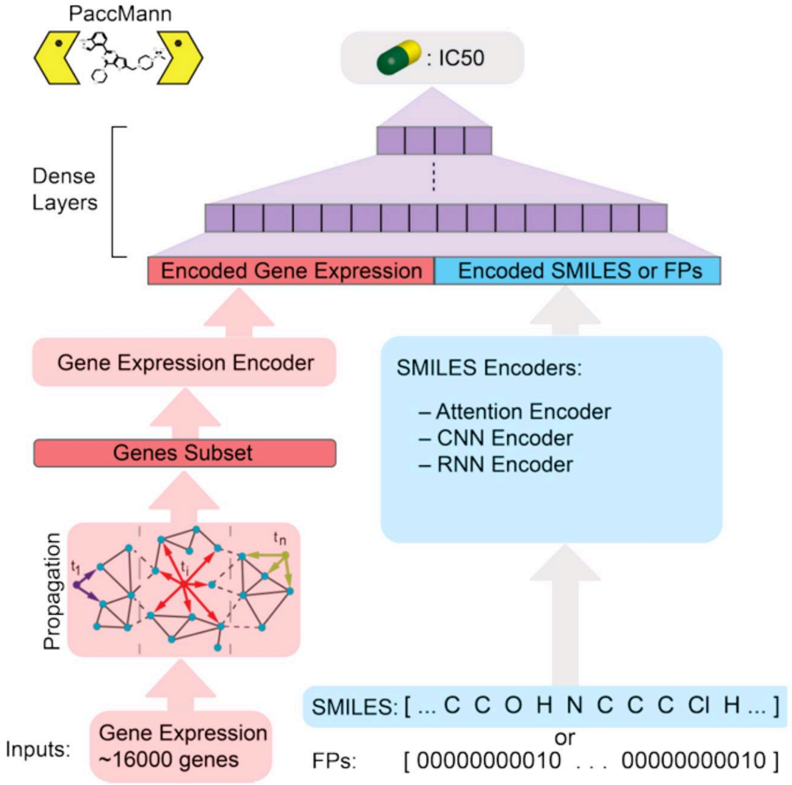


Figure 12: PaccMann Model Architecture

4.2 Precily:

Precily introduces a novel predictive modelling approach for inferring treatment response in cancers using gene expression data. This framework emphasizes the incorporation of pathway activity estimates alongside drug descriptors as features. Utilizing a deep neural network (DNN)-based framework, Precily predicts the response to cancer therapy based on gene expression profiles and drug descriptors, providing insights into the biological mechanisms underlying drug resistance through the explicit use of pathway enrichment scores.

4.2.1 Drug Embedding:

In Precily, Drug Embeddings are numeric molecular descriptors for anti-cancer compounds are obtained using SMILESVec, by supplying simplified molecular-input line-entry system (SMILES) notation.

4.2.2 Cell Embedding:

In Precily, they have considered the 500 top highly variable gene expressions for each cell line.

4.2.3 Model Architecture:

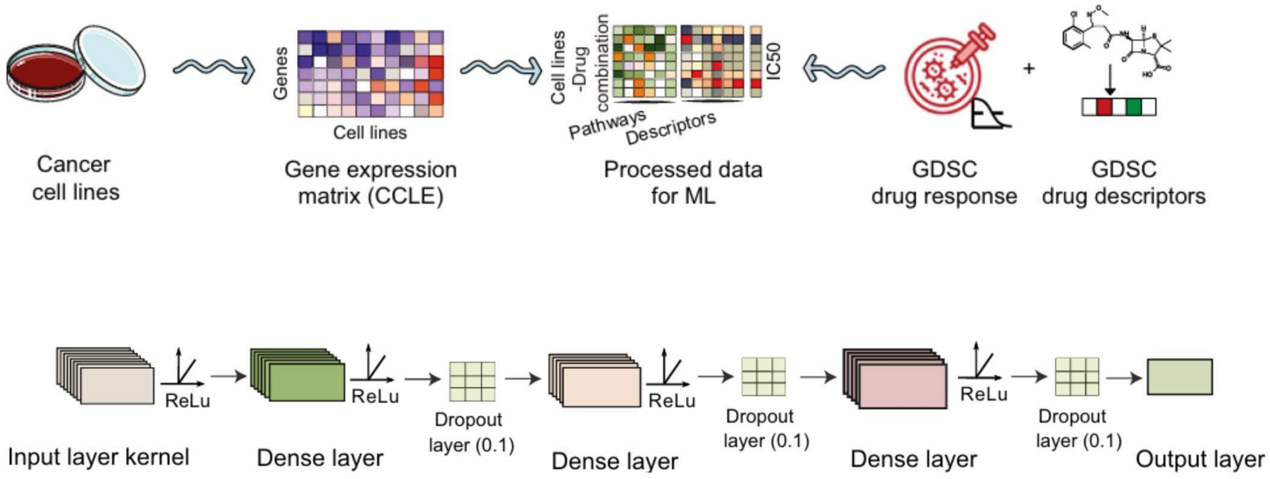


Figure 13: Precilt model architecture

A deep neural network (DNN) was trained using the Keras framework. The DNN architecture comprised one input layer entailing all the 600 features (500 gene features and drug descriptors of size 100), followed by one hidden layer of size 512, with Rectified Linear Unit(RELU) as an activation function was used to predict the LN_IC50 values.

4.3 tCNN:

tCNN employs a convolutional network to extract features from the simplified molecular input line entry specification (SMILES) format for drugs. Simultaneously, another convolutional network is utilized to extract features from genetic feature vectors of cancer cell lines. These extracted features are then combined in a fully connected network to predict interactions between drugs and cancer cell lines. However, the model's performance decreases significantly when the training and testing sets are divided exclusively based on either drugs or cell lines, resulting in R^2 values that are barely above zero.

4.3.1 Drug Embedding:

In tCNN, the SMILES format for drugs, containing 72 different symbols, is converted into a one-hot matrix, where each drug is represented as a 72 x 188 matrix with binary values (0 or 1) where 188 is the size of longest SMILES. In the one-hot matrix for a drug, a value 1 at row i and column j means that the i^{th} symbol appears at j^{th} position in the SMILES format for the drug. The 1D convolutional operation is applied along each row, confining the operation within the same chemical element, thus enabling the model to extract relevant features from the drug's SMILES representation.

4.3.2 Cell Line Embedding:

In tCNN, The cell line features are acquired from GDSC which represents each cell line by a 735 feature vector where each feature either belongs to mutation state or copy number alteration. A 1D CNN is applied along the 1D feature vectors for cell lines.

4.3.3 Model Architecture:

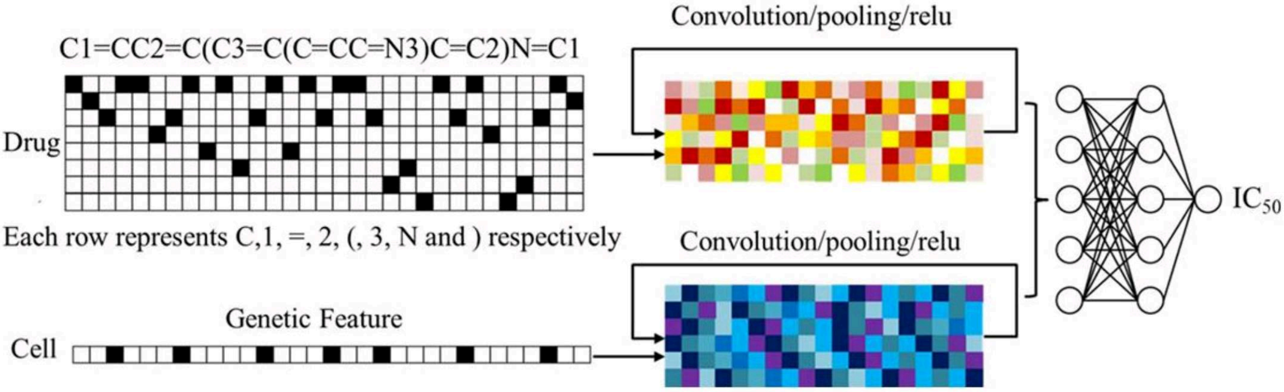

C1=CC2=C(C3=C(C=CC=N3)C=C2)N=C1
Drug
Each row represents C,1, =, 2, (, 3, N and) respectively
Genetic Feature
Cell
Convolution/pooling/relu
Convolution/pooling/relu
IC50

Figure 14: tCNN model architecture

The left-hand side is the input data of one-hot representations for drugs and the feature vectors for cell lines. The black square stands for ‘1’ and empty square stands for ‘0’. In the middle, there are a CNN branch to process the drug inputs and a CNN branch to process cell lines inputs, respectively. They take the one-hot representations and feature vectors as input data respectively, and their outputs can be interpreted as the abstract features for drugs and cell lines. The structures of the two convolution neural networks are similar. The two CNN’s are then connected to a fully-connected neural network to predict IC50 values.

For training, tCNN normalizes the logarithmic IC50 values into the (0,1) interval. Given a logarithmic IC50 value *x,* first it takes the exponential of it to get the real IC50 value $y=e^{x}$ and then use the following function to normalize:

$$y \longrightarrow \frac{1}{1+y^{-0.1}}$$

Usually y is very small ($<10^{-3}$), and the parameter −0.1 is chosen to distribute the result more uniformly on the interval (0,1).

4.4 AGMI:

The Attention-Guided Multi-omics Integration (AGMI) approach introduces a novel method for IC50 prediction. AGMI constructs a Multi-edge Graph (MeG) for each cell line and integrates multi-omics features using a unique structure called the Graph edge-aware Network (GeNet). This approach is groundbreaking as it explores gene constraint-based multi-omics integration for prediction of IC50 across genome the entire using Graph Neural Networks (GNNs), offering a new dimension in predictive modelling for drug response.

4.4.1 Drug Embedding:

AGMI uses a *Graph Isomorphism Network* (GIN) to generate drug embedding, It collect SMILES from PubChem and uses RDKit package to construct molecular graphs where atoms are described as nodes and bonds between any two atoms are described as edges. The GIN network generates a 128 size vector embedding for each drug.

4.4.2 Cell Line Embedding:

The AGMI approach constructs a Multi-edge Graph (MeG) for each cell line, where each node represents a gene with its expression level, mutation state, and copy number variation (CNV) as features. Edges represent different types of relations between genes, such as protein-protein interactions (proteomics), gene pathway relations (metabolomics), and PCC of gene expression. Now, they introduced node-level GRU (nGRU) after this they introduce a graph-level GRU (gGRU) this will map the node features and edge features of the whole graph to a cell line feature vector of size 128.

4.4.3 Model Architecture:

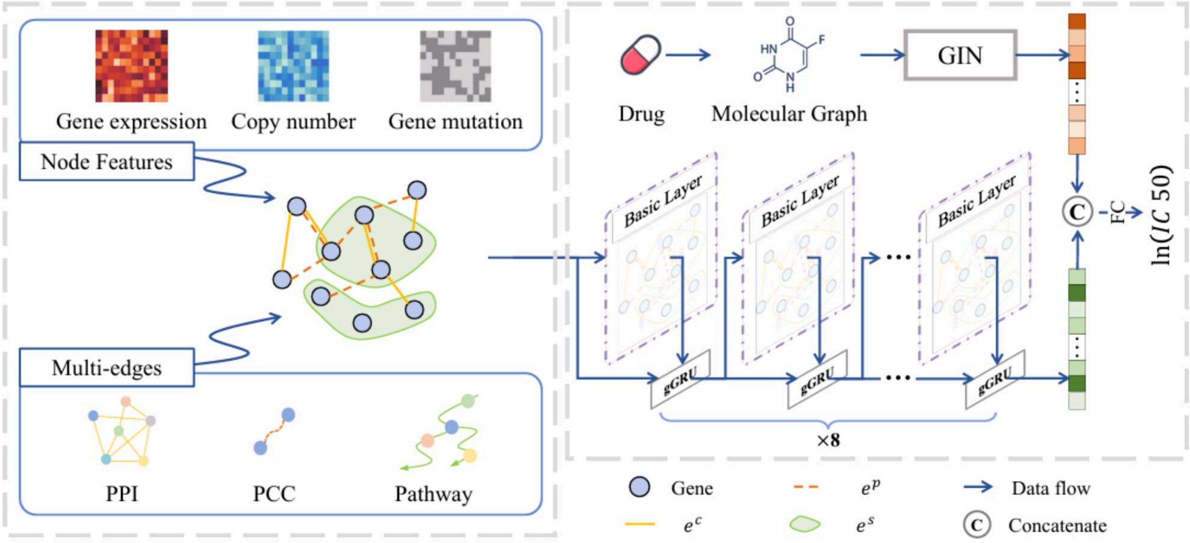


Figure 15: AGMI model architecture

AGMI integrates multi-omics data by modelling a cell line as a graph with multiple types of edges (MeG) by introducing a node-level GRU and a graph-level GRU to capture the complex features of MeG which gives a feature vector of size 128. It uses Graph Isomorphism Network (GIN) to generate drug embedding which takes he molecular graph of the drug and converts it to a feature vector of size 128.

Finally, the cell line and drug features are concatenated for the final prediction, which is made using three fully connected (FC) layers and predicts LN_IC50 value.

4.5 DeepCDR:

DeepCDR is a novel approach for predicting Cancer Drug Response (CDR) by integrating multi-omics profiles of cancer cells and intrinsic chemical structures of drugs. It employs a hybrid graph convolutional network architecture, including a uniform graph convolutional network and multiple subnetworks. It features several subnetworks for multi-omics profile extraction from cancer cells. Notably, DeepCDR introduces a Uniform Graph Convolutional Network (GCN) for innovative drug feature extraction, automatically capturing drug structures by considering atom interactions within compounds.

4.5.1 Drug Embedding:

Each drug has its unique chemical structure which can be naturally represented as a graph where the vertices and edges denote chemical atoms and bonds, respectively. Each drug is represented by a graph which contains the feature matrix and adjacent matrix of the respective drug. The Uniform Graph Convolutional Network (GCN) ensures that drugs will be embedded into a fixed dimensional vector of size 100.

4.5.2 Cell Line Embedding:

In DeepCDR, they designed omics-specific subnetworks to integrate the information of multi-omics profiles. They used a 1D convolutional network for processing genomic data as the mutation positions are distributed linearly along the chromosome. For transcriptomic and epigenomic data, we directly used fully connected networks for feature representation. The dimension of latent space *d* is set to 100 of each cell line.

4.5.3 Model Architecture:

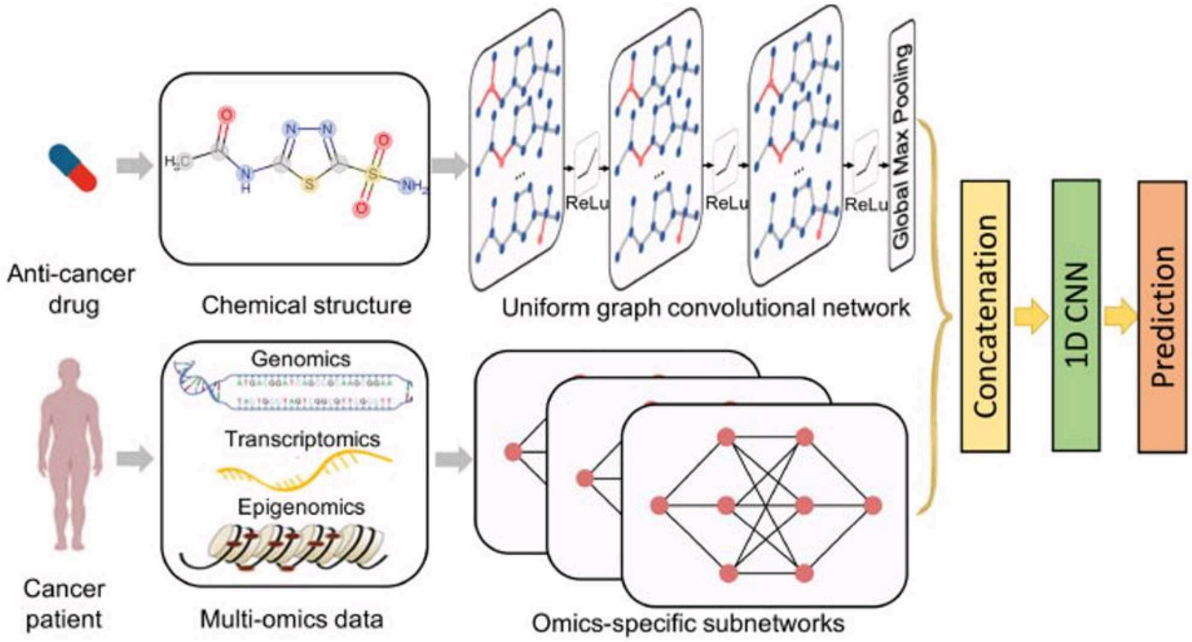


Figure 16: DeepCDR model architecture

DeepCDR contains a UGCN and three subnetworks for processing drug structures and cancer cell profiles (genomic mutation, gene expression and DNA methylation data) respectively. The drug will be represented as a graph based on the chemical structure before transformed into a high-level latent representation by the UGCN. Omics featured learned by subnetworks will be concatenated to the drug feature. DeepCDR concatenates the high-level features of drug and multiple omics data and were fed to a 1D CNN and predicts the drug sensitivity (LN_IC50).

Chapter 5

RESULTS

. In this section, we present the evaluation results of the models using a dataset consisting of 51,652 drug-cell line pairs, encompassing 969 cell lines and 202 drugs. To ensure the robustness and reliability of our findings, we trained the models five times. We use the following three different splitting strategies to comprehensively assess the models:

5.1 Random Split

In random split the entire dataset is divided randomly into training, validation and test sets. We divided the dataset into using 70-15-15 split i.e. training set (70%) will consist of 36,156 drug-cell line pairs, validation set (15%) and testing (15%) will consist of 7,748 drug-cell line pairs each. Below table compares the test performance of the models.

Models	**Average**	
	Pearson Corelation	R^2
PaccMann	0.803 ± 0.021	0.650 ± 0.005
Precily	0.791 ± 0.013	0.642 ± 0.012
tCNN	**0.816 ± 0.024**	**0.681 ± 0.018**
AGMI	0.771 ± 0.035	0.598 ± 0.025
DeepCDR	0.801 ± 0.025	0.652 ± 0.008

Table 3: Results for Random Split

tCNN model has significantly outperformed all other models in random split.

5.2 Cell Line Split

A cell line split refers to the division of cancer cell lines into separate groups for training and testing purposes of predictive models. This process is crucial to ensure that the models are robust, generalize well to unseen data, and accurately predict drug responses. The results for the cell line split can be seen below where PaccMann outperforms the all other models:

Models	**Average**	
	Pearson Corelation	R^2
PaccMann	**0.792 ± 0.021**	**0.641 ± 0.011**
Precily	0.733 ± 0.041	0.614 ± 0.025
tCNN	0.771 ± 0.011	0.639 ± 0.014
AGMI	0.721 ± 0.041	0.521 ± 0.039
DeepCDR	0.762 ± 0.018	0.638 ± 0.001

Table 4: Results for Cell line split

5.3 Drug Split

Models	**Average**	
	Pearson Corelation	R^2
PaccMann	0.197 ± 0.087	-1.151 ± 0.425
Precily	0.291 ± 0.091	-0.452 ± 0.382
tCNN	**0.301 ± 0.084**	**-0.378 ± 0.294**
AGMI	0.174 ± 0.032	-1.473 ± 0.521
DeepCDR	0.281 ± 0.062	-0.563 ± 0.241

Table 5: Results for Drug Split

A drug split refers to the division of drugs into separate groups for training and testing purposes of predictive models. Drug split allows to rigorously test and validate the models. It ensures that the predictive model isn't just tailored to a specific subset of drugs but can generalize to a broader range of drugs. This approach is crucial for developing reliable and robust models that can accurately predict drug responses in real-world scenarios, where new and diverse drugs are constantly being developed and tested. The Table shows tCNN outperforming all other models when tested on unseen drugs data.

Chapter 7

CONCLUSION & FUTURE SCOPE

7.1 Conclusion

In this work, we presented evaluation of different state-of-the-art models which follows the same core idea of having drug and cell line representations with the goal of predicting the IC50 value of a drug and cancer cell combination. These models use different representations of drug compounds as well as the cell lines and the underlying architecture of the neural networks.

We found that representing the cell line with the genomics profiles or using the multi-omic data has proved to be effective in representation of a cell line which ultimately improves prediction accuracy. Representing the drug with SMILES is found effective as it captures sequential as well as the structural data of a drug. Representing the drug using graphs is more meaningful as it will captures the complex structures of molecules, which has provided significant benefits.

While the model performed exceptionally well in random and cell-line splits, it showed a relatively lower performance in drug-wise splits. We found that tCNN has significantly performed well by achieving a corelation of around 0.81 and 0.30 in random and drug split. We have also observed that PaccMann performed comparatively well for unseen cell lines by achieving a pearson corelation 0.79. This indicates a potential area for further improvement, particularly in enhancing the model's ability to generalize to entirely new drugs.

7.2 Future Work

The following are the factors which may further enhance the performance of the prediction models.

1. **Improved Drug Representation**:

   Explore alternative drug representations that capture additional information beyond SMILES strings. This could include incorporating 3D structure, physicochemical properties, or target protein interactions.

2. **Analysis of drug properties**:

   Since, the performances of the predictive models decreased significantly for unseen drugs, we hypothesize that intrinsic drug properties like molecular weight, number of atoms, solubility and tissue type will have any significant impact on the accuracy of IC50 predictions.

3. **Expanding the dataset:**

   We should also focus on expanding the dataset, particularly by incorporating more drug data which may further improve drug embedding, which ultimately improves prediction accuracy of the models.

BIBLIOGRAPHY

[1] “Liu, Q., Hu, Z., Jiang, R., and Zhou, M. (2020). DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 36, i911–i918. https://doi.org/10.1093/bioinformatics/btaa822.”.

[2] “Feng, R., Xie, Y., Lai, M., Chen, D.Z., Cao, J., and Wu, J. (2021). AGMI: attention-guided multi-omics integration for drug response prediction with graph neural networks, pp. 1295–1298.”.

[3] “Ammad-Ud-Din, M. et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics 32, i455–i463 (2016).”.

[4] “Liu, P., Li, H., Li, S., and Leung, K.-S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinf. 20, 408. https://doi.org/10.1186/s12859-019-2910-6.”.

[5] “Kim, H., Lee, J., Ahn, S. et al. A merged molecular representation learning for molecular properties prediction with a web-based service. Sci Rep 11, 11028 (2021). https://doi.org/10.1038/s41598-021-90259-7”.

[6] “Garrett B Goh, Nathan O Hodas, Charles Siegel, and Abhinav Vishnu. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. arXiv preprint arXiv:1712.02034, 2017.”.

[7] “Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.”.

[8] “Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the”.

[9] “Chawla, S., Rockstroh, A., Lehman, M. et al. Gene expression based inference of cancer drug sensitivity. Nat Commun 13, 5680 (2022). https://doi.org/10.1038/s41467-022-33291-z.”.

[10] “Born, J., Manica, M., Oskooei, A., Cadow, J., Markert, G., and Rodriguez Martinez, M. (2021). PaccMannRL: de novo generation of hit-like anti-cancer molecules from transcriptomic data via reinforcement learning. iScience 24, 102269. https://doi.org/10.10”.

[11] “Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607 (2012).”.

[12] “Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.”.

[13] “Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol. 4, 19 (2020).”.