Prediction of IC50 value using Drug & Cell-line pair

Dissertation submitted in partial fulfilment of the

Requirements for the degree of

Master of Technology

in

Computer Science

by

Sneha Tiwari

CS2229

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 **“Prediction of IC50 values using Drug & Cell line pairs”** submitted by **Sneha Tiwari** (Roll Number: *CS2229*), 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 fulfilment 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, Sneha Tiwari, 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

Sneha Tiwari

Sneha Tiwari
M.Tech (CS), CS2229
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. 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. Furthermore, I am grateful to **Neeraj Kumar Singh** and **Siva K. Lakkoju** for his support and guidance 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 the pursuit of personalized cancer treatment, predicting drug response in specific cell lines is a critical challenge. This study presents a novel approach for predicting the half-maximal inhibitory concentration (IC50) of drugs on various cell lines using a Convolutional Neural Network (CNN) architecture. The proposed model, IC50Predictor, integrates drug and cell line features through separate CNN pipelines, combining the extracted features to make robust predictions. We evaluated the model using three data splitting strategies: random split, cell-line split, and drug split, to assess its generalization capabilities. Additionally, an analysis of drug properties was conducted to determine their impact on prediction accuracy and error. The performance of the model was measured using metrics such as R^2, RMSE, and Pearson correlation coefficient. Our results demonstrate the efficacy of the CNN-based approach in accurately predicting drug responses and highlight the significance of drug properties in refining these predictions, showcasing its potential in the advancement of personalized medicine. The model achieved an Pearson correlation of 0.941 in the random split, 0.895 in cell line split and 0.421 in drug split, indicating its robustness and effectiveness.

CONTENTS

List of Figures

Chapter 1

INTRODUCTION

Cancer, a multifaceted and heterogeneous disease, continues to pose a significant challenge to healthcare systems worldwide. Traditional cancer treatments often employ a generalized approach, leading to varied outcomes due to the unique genetic and molecular profiles of individual tumors. Personalized medicine seeks to address this variability by tailoring treatments based on the specific characteristics of each patient's cancer, promising more effective and less toxic therapies.

A critical aspect of personalized cancer therapy is the ability to predict how different cancer cell lines will respond to various drugs. The half-maximal inhibitory concentration (IC50) is a key measure in this context, indicating the concentration of a drug required to inhibit cell growth by 50%. Accurate prediction of IC50 values can streamline drug development, optimize treatment regimens, and improve patient outcomes by ensuring the most effective drugs are used at appropriate dosages.

However, predicting IC50 values is a complex task due to the intricate relationships between drug properties and cell line responses. Traditional machine learning methods often struggle to capture these complexities, necessitating the development of more advanced predictive models. This research aims to introduce a novel approach to address these challenges and improve the accuracy of IC50 predictions.

1.1 What is IC50?

IC50, or the half maximal inhibitory concentration, measures the concentration of an inhibitor needed to inhibit a specific biological process by 50%, and is expressed in molarity (M), typically in nM or μM. In our context, it essentially gauges the potency of a drug for a given cell line. IC50 is determined by plotting a dose-response curve of the drug for the cell line, with drug concentrations usually plotted on a log scale.

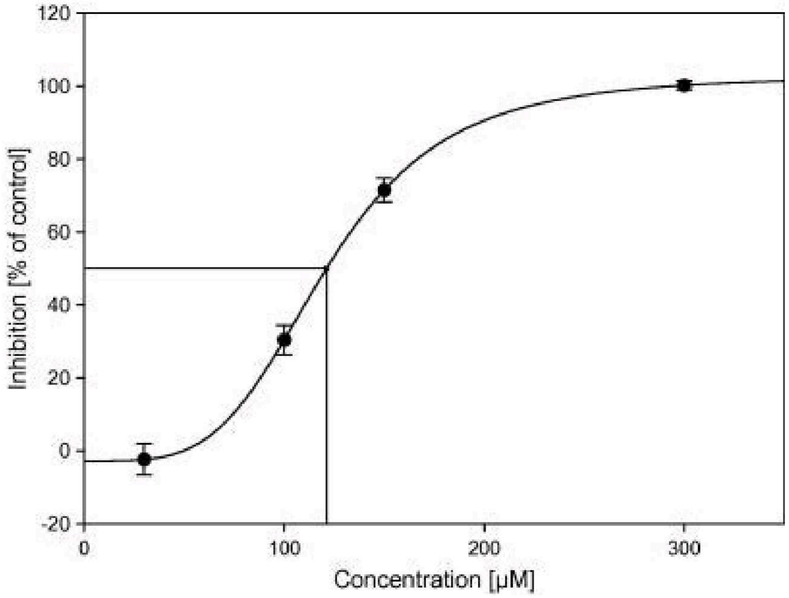


Figure 1: IC50 vs Concentration graph

LN_IC50 represents the natural logarithm of the half-maximal inhibitory concentration (IC50). The IC50 value indicates the concentration of a drug needed to inhibit a biological process or cell growth by 50%. Taking the natural logarithm (LN) of the IC50 value helps stabilize variance and normalize the distribution, making it particularly useful for statistical analyses and modelling.

1.2 Drug Representation

A drug refers to any chemical compound used for therapeutic purposes, particularly for treating cancer. These compounds can vary widely in their chemical structures and mechanisms of action. Understanding the interactions between different drugs and various cancer cell lines is crucial for developing effective treatments.

- SMILE (Simplified Molecular Input Line Entry System):

  It is a chemical annotation system that represents molecular structures using characters for atoms and special symbols for bonds and higher-order structural properties.

- Fingerprinting:

  Fingerprinting transforms a chemical structure into a binary vector of a fixed size, encapsulating the compound's structural information.

- Molecular Graphs:

  Graphical representations depict each atom as a node and each bond as an edge within a graph. Software tools like RDKit in Python enable the creation of these molecular graphs.

1.3 Cell Representation

Cell lines undergo extensive characterization using various "-omic" profiles, which provide comprehensive data crucial for IC50 prediction models. These profiles encompass genomic, epigenetic, transcriptomic, and proteomic information, each offering distinct insights into the biological state of the cells.

- Mutation Profile:

  The mutation profile of a cancer cell line refers to the specific genetic alterations present within its genome. This includes identifying mutations such as single-nucleotide variants (SNVs), insertions, deletions, copy number variations (CNVs), and structural variations like translocations. The mutation profile helps researchers determine the genetic drivers of cancer in specific cell lines and may guide personalized treatment strategies.

- Gene Expressions:

  Gene expression profiling involves measuring the activity (expression) levels of thousands of genes within a cancer cell line. Gene expression profiles can provide insights into the biological processes that are dysregulated in cancer cells compared to normal cells. It helps in identifying biomarkers, predicting drug responses, and understanding the molecular mechanisms underlying cancer progression.

- Protein Embeddings:

  Protein embeddings represent a computational method to encode and analyze the characteristics of proteins based on their sequences or structures. This approach helps in understanding how proteins interact within cancer cells, identifying potential drug targets, and predicting protein functions related to cancer biology.

Chapter 2

PROBLEM DEFINITION

Personalized cancer treatment depends heavily on accurately predicting how different cancer cell lines will respond to various drugs. The IC50 value is a crucial metric for assessing drug efficacy, but its prediction is complicated by the diverse and complex nature of biological systems and chemical structures. Accurate IC50 predictions can significantly enhance drug development and treatment personalization, ultimately improving patient outcomes.

Problem Statement

The primary objective of this research is to develop a reliable and accurate model to predict IC50 values for drug-cell line pairs. The specific challenges and goals of this research include:

1. **Understanding Complex Data**: Integrating and analysing the diverse and complex features of both drugs and cell lines to make accurate predictions.
2. **Ensuring Generalization**: Evaluating the model's ability to generalize to new, unseen data by testing it with different data splitting strategies, such as random splits, cell-line splits, and drug splits.
3. **Incorporating Drug Properties**: Investigating the impact of specific chemical and physical properties of drugs on prediction accuracy and incorporating these insights to improve the model’s performance.

Research Questions

To achieve these objectives, the research focuses on answering the following key questions:

1. **Prediction Accuracy**: Can the proposed model accurately predict IC50 values, outperforming traditional machine learning approaches?
2. **Generalization**: How well does the model generalize across different data splitting strategies, ensuring robustness and reliability?
3. **Drug Properties**: What is the influence of specific drug properties on prediction accuracy, and how can these properties be effectively integrated into the model to enhance its performance?

By addressing these questions, this research aims to develop a model that not only improves the accuracy of IC50 predictions but also enhances our understanding of the factors influencing drug efficacy. This will contribute significantly to the advancement of personalized cancer therapy, optimizing treatment regimens and improving patient outcomes.

Chapter 3

LITERATURE REVIEW

To establish a benchmark for evaluating the performance of the IC50Predictor model, three established baseline models were implemented: the Paccmann model, the Precily model, and the tCNN model. These models are widely recognized in the field for their effectiveness in predicting drug response, providing a solid foundation for comparison. To Evaluate the baseline models we have used the data from GDSC2 database. After applying the filter, we have approximately 51,652 drug-cell line pairs which consist of 969 cancer cell lines and 202 drugs, which provided a robust dataset for training and evaluating our IC50 prediction models.

1. **PaccMann Model**

   The PaccMann is a noval approach which uses multi-modal attention-based neural networks. PaccMann takes SMILES encoding of the drug compound and gene expression of a cancer cell to predict the IC50 value. Below are the results after 5-Fold cross validation for random, cell line and drug split:

   - **Random Split**:
     - RMSE = [0.0922 ± 0.0035],
     - Pearson correlation = [0.7714 ± 0.021]
   - **Cell-line Split**:
     - RMSE = [0.0784 ± 0.004],
     - Pearson correlation = [0.792 ± 0.02]

- **Drug Split**:
    - RMSE = [0.169 ± 0.013],
    - Pearson correlation = [0.197 ± 0.087]

2. **Precily Model**

    Precily utilizes a deep neural network (DNN)-based framework which predicts the IC50 based on gene expression profiles for each cell line and drug Embeddings are numeric molecular descriptors for anti-cancer compounds are obtained from SMILESVec by taking SMILES as input.

    - **Random Split**:
        - RMSE = [1.132 ± 0.153],
        - Pearson correlation = [0.7912 ± 0.013]
        - R^2 = [0.642 ± 0.012]
    - **Cell-line Split**:
        - RMSE = [1.315 ± 0.214],
        - Pearson correlation = [0.733 ± 0.041]
        - R^2 = [0.614 ± 0.025]
    - **Drug Split**:
        - RMSE = [2.232 ± 0.652],
        - Pearson correlation = [0.291 ± 0.091]
        - R^2 = [-0.451 ± 0.38]

3. tCNN Model

tCNN uses a convolutional network to extract features from the simplified molecular input line entry specification (SMILES) format for drugs. At the same time, another convolutional network extracts features from genetic feature vectors of cancer cell lines. These features are then combined in a fully connected network to predict interactions between the drugs and the cancer cell lines.

- **Random Split**:
  - RMSE = [0.259 ± 0.003],
  - Pearson correlation = [0.826 ± 0.024]
  - R^2 = [0.681 ± 0.018]
- **Cell-line Split**:
  - RMSE = [0.0297 ± 0.002],
  - Pearson correlation = [0.814 ± 0.011]
  - R^2 = [0.654 ± 0.014]
- **Drug Split**:
  - RMSE = [0.061 ± 0.026],
  - Pearson correlation = [0.280 ± 0.084]
  - R^2 = [-0.378 ± 0.281]

These baseline models provide a crucial reference point for assessing the performance of the IC50Predictor model. By comparing the results across different data splitting strategies, the advantages and improvements of the CNN-based approach can be clearly demonstrated. The detailed results of these baseline models will help in highlighting the relative strengths and areas of enhancement of the IC50Predictor.

Chapter 4

DATASET DETAILS

Our main dataset was obtained from Genomics of Drug Sensitivity in Cancer (GDSC2), which has around 243,466 drug-cell line IC50 pairs, with 297 drugs and 969 cell lines.

COSMIC ID	CELL_LINE NAME	DRUG ID	DRUG NAME	LN_IC50	MIN_CONC	MAX_CONC	IC50
684052	A673	1003	Camptothecin	-4.8694	0.0001	0.1	0.0076
688027	NCI-H69	1013	Nilotinib	2.0001	0.0020	10	7.3893
924239	L-363	2145	PBD-288	-1.0939	0.0050	5	0.3348

Table 1: GDSC2 Dataset Sample

In the above example:

- **COSMIC ID**: Unique identifier given to each cell line.
- **CELL_LINE NAME**: Name of the cancer cell line
- **DRUG ID**: Unique identifier for the drug being tested.
- **DRUG NAME**: Name of the drug being tested
- **LN_IC50**: Natural logarithm of the half-maximal inhibitory concentration (IC50), which stabilizes variance and normalizes the distribution of IC50 values for statistical analysis.
- **MIN_CONC**: Minimum concentration of the drug tested.
- **MAX_CONC**: Maximum concentration of the drug tested.
- **IC50**: Half-maximal inhibitory concentration, representing the concentration of the drug required to inhibit a biological process or cell growth by 50%

We applied a filtering method based on the IC50 values. We have considered only those drug-cell line pairs where its IC50 values lies between MIN_CONC (minimum concentration) and MAX_CONC (maximum concentration) in the dataset. This is a very crucial step which ensures the practical relevance and reliability of the IC50 values used in our model.

After this filtering step, our dataset has approximately 51,652 drug-cell line pairs out of which 202 are drugs and contains 969 cell lines. Now we utilize this dataset to train, validate and test our model.

4.1 GENE EMBEDDING:

We based our models on gene expression data, as it has been shown to be more predictive of drug sensitivity than genomics (i.e., copy number variation and mutations) or epigenomics (i.e., methylation) data. STRING protein-protein interaction (PPI) network was used to incorporate intracellular interactions in our model which was used by PaccMann. STRING PPI network reduces the dimensionality of the cell line gene expression profile, consisting of 16,000 genes to smaller subset containing the most informative 2128 genes.

4.2 DRUG EMBEDDING:

The three models employed used SMILES embeddings, trained on approximately 200 drugs. To improve prediction accuracy, we opted to utilize ChemBERT embeddings, which are trained on a significantly larger dataset of drug SMILES representations.

ChemBERT, a transformer-based model, is specifically designed for molecular property prediction. It leverages large-scale pretraining on extensive datasets like PubChemPy, which includes millions of compounds. This extensive pretraining enables ChemBERT to capture more nuanced chemical features and relationships, thereby improving its ability to generalize across diverse molecular structures.

This choice was driven by the hypothesis that embeddings trained on a vast number of drug molecules would better capture the underlying chemical properties, leading to more accurate IC50 value predictions even for previously unseen drugs. ChemBERT's robust performance on various molecular property prediction tasks supports this approach.

Chapter 5

Analysis of Drug Properties and IC50 Prediction Error

The baseline models performed well on splits with drug-cell line pairs seen during training but showed significant performance decline on splits with unseen drugs. This indicated a limitation in the models' generalizability to new drugs.

This motivated us to investigate whether drug properties impact prediction accuracy, hypothesizing that intrinsic drug properties might crucially affect the models' ability to predict IC50 values accurately.

In this section, we investigate the relationship between various drug properties and the error in predicting IC50 values using machine learning models. The aim is to understand whether properties such as molecular weight, number of atoms, solubility and tissue type will have any significant impact on the accuracy of IC50 predictions. Additionally, we analyse the effects of these properties on prediction errors for specific cell lines interacting with multiple drugs.

5.1 Methods

This section outlines the methodologies employed in developing the IC50Predictor model, including property extraction, error calculation, and correlation analysis. These methods are integral to the process of understanding and improving the model's performance in predicting IC50 values for drug-cell line pairs.

5.1.1 Property Extraction

Using the SMILES representation, we calculated the following properties for each drug:

- *Molecular Weight*
- *Number of Atoms*
- *Solubility*

5.1.2 Error Calculation

The percentage error for each drug-cell line pair was computed. The use of percentage error allows for a standardized measure of prediction accuracy, enabling us to compare errors across drugs with varying IC50 value ranges. This metric provides a clear understanding of the model's performance relative to the true values, making it easier to identify significant discrepancies.

The percentage error for drug-each cell line pair was computed as:

$$\text{Error} = \left| \frac{\text{IC50}_{\text{actual}} - \text{IC50}_{\text{predicted}}}{\text{IC50}_{\text{actual}}} \right| \times 100$$

5.1.3 Correlation Analysis

We performed correlation analysis to investigate the relationship between the percentage error and each of the drug properties. This analysis was carried out both on the overall dataset and on a cell-line-specific basis.

5.2. Results

5.2.1 Overall Correlation Analysis

The results of the correlation analysis between drug properties and prediction error are summarized in Table 1. Figures 1 and 2 present the graphical representation and heat map of these correlations, respectively.

Property	Correlation Coefficient
Molecular Weight	-0.0043
Number of Atoms	-0.0098
Solubility	-0.093

Table 2: Correlation between Drug Properties and Prediction Error

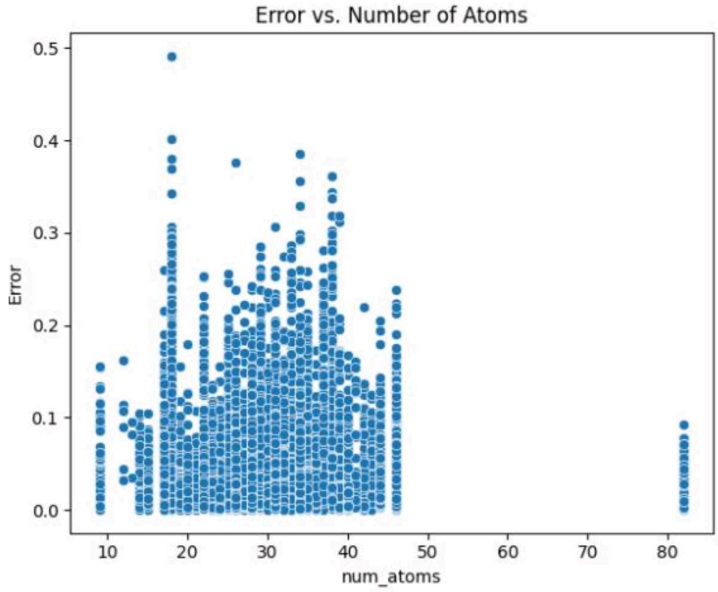


Figure 2: Scatter Plot of No. of atoms in the drug vs. Prediction Error

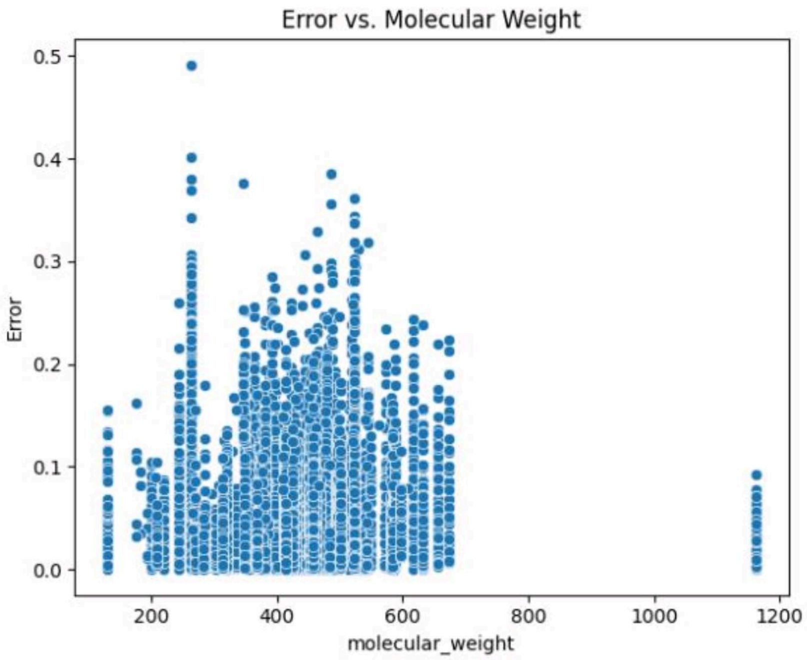


Figure 3: Scatter Plot of Molecular weight in the drug vs. Prediction Error

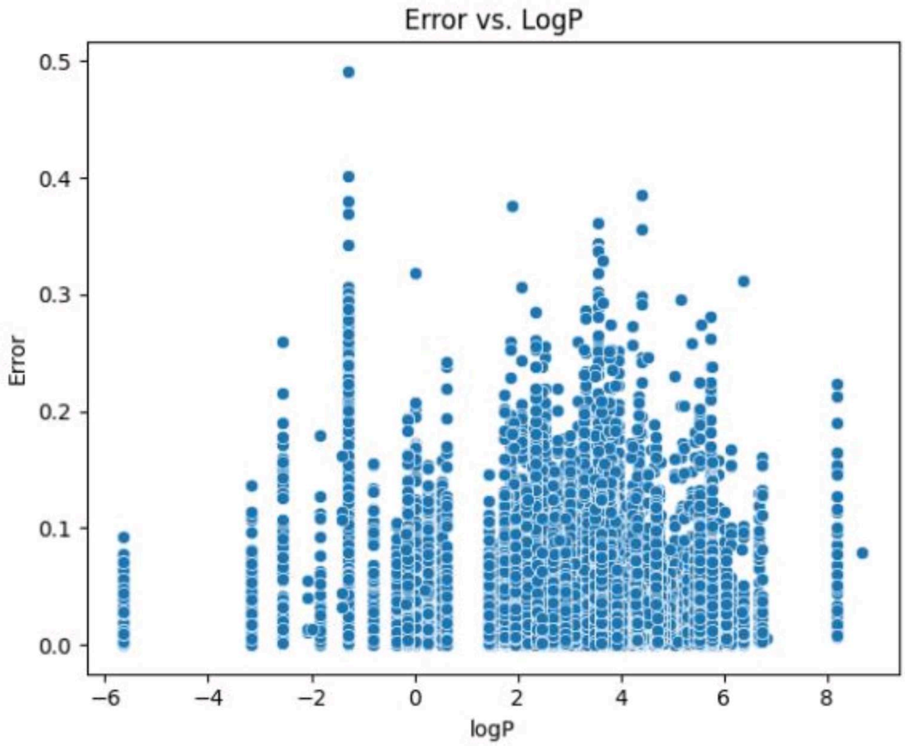


Figure 4: Scatter Plot of Molecular weight in the drug vs. Prediction Error

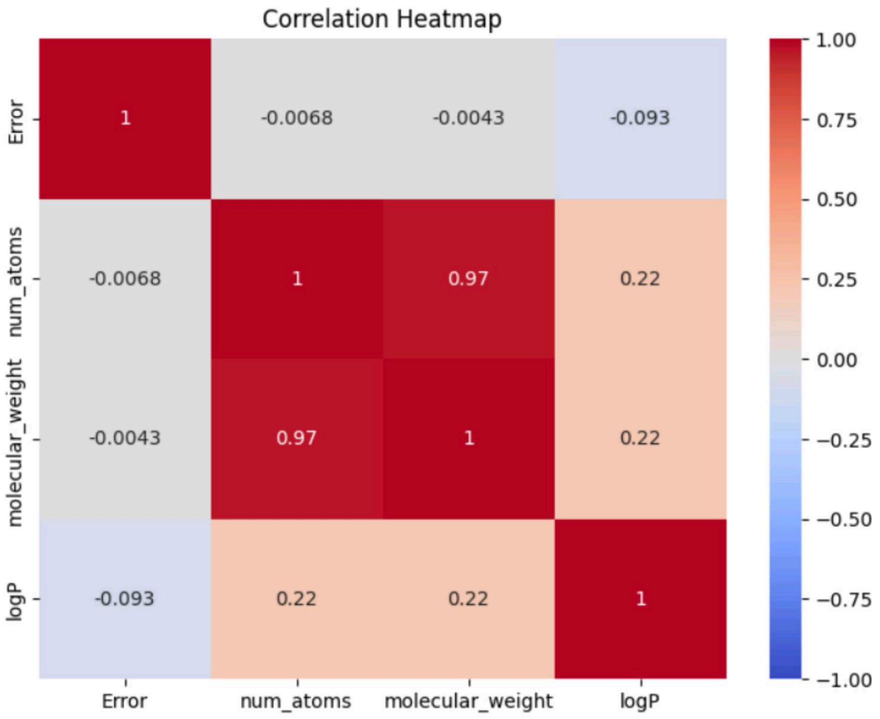


Figure 5: Heat Map of Correlation between Drug Properties and Prediction Error

As shown in Table 2, the correlation coefficients for all properties are close to zero, indicating no significant correlation between these drug properties and the prediction error.

5.2.2 Cell-Line Specific Analysis

We also examined whether the properties of drugs have a significant impact on prediction errors for specific cell lines. The analysis was carried out for a couple of cell lines interacting with multiple drugs. Figures 3 and 4 show the results for one representative cell line.

The cell lines for this experiment were selected based on the following criteria:

1. The first cell line interacts with the maximum number of varied-length drugs("IM-9").
2. The second cell line interacts with the maximum number of drugs (Here "A3-KAW").

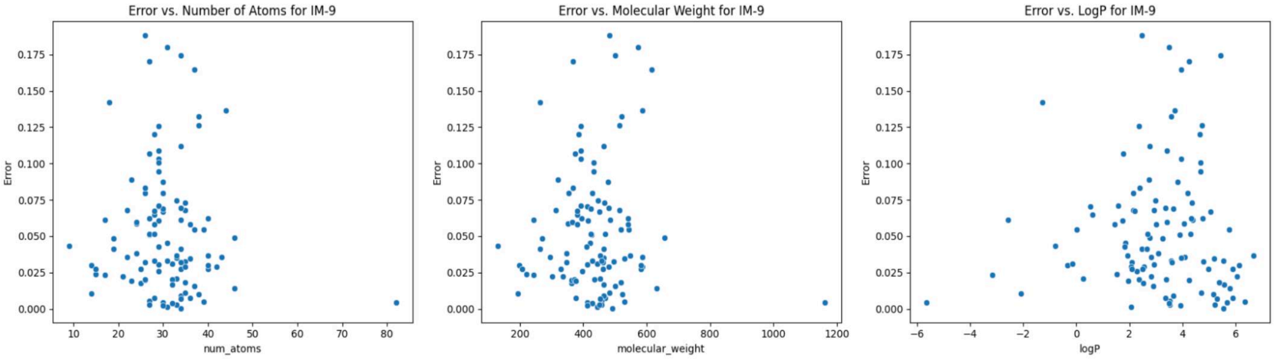


Figure 6: Scatter Plot of Drug Properties vs. Prediction Error for Cell Line "IM-9"

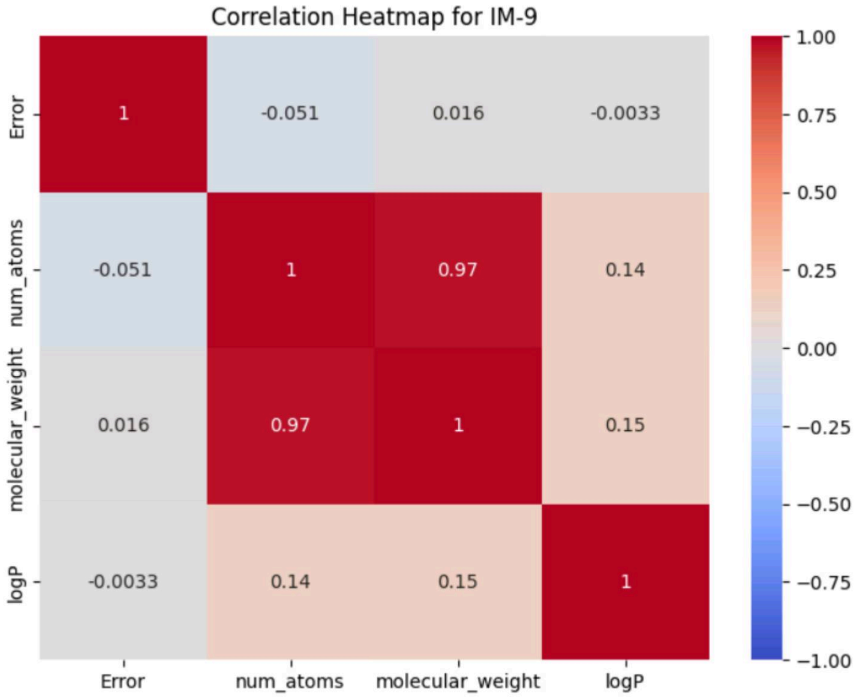


Figure 7: Heat Map of Correlation between Drug Properties and Prediction Error for Cell Line "IM-9"

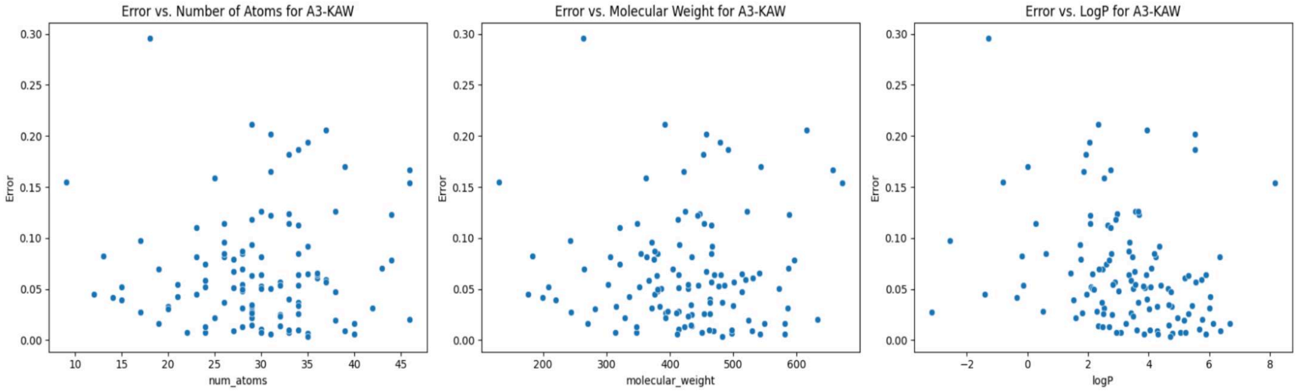


Figure 8: Scatter Plot of Drug Properties vs. Prediction Error for Cell Line "A3-KAW"

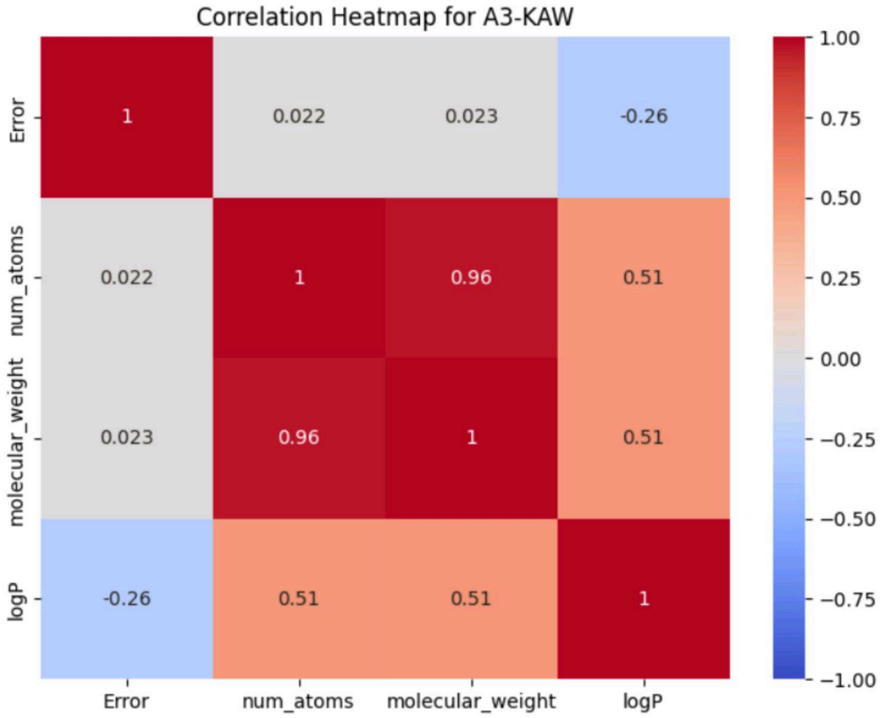


Figure 9: Heat Map of Correlation between Drug Properties and Prediction Error for Cell Line "A3-KAW"

Similar to the overall analysis, the cell-line-specific analysis revealed no significant correlation between drug properties and prediction errors for any of the cell lines studied.

5.2.3 Impact of Tissue Type on Prediction Accuracy

Given that the response of cancer cells to drugs can be influenced by the tissue of origin, it was essential to investigate whether tissue type affects the model's prediction accuracy. Violin plots were used to visualize the distribution of prediction errors across different tissue types. Violin plots are particularly useful for this task as they combine aspects of box plots and density plots, providing a detailed view of the error distribution.

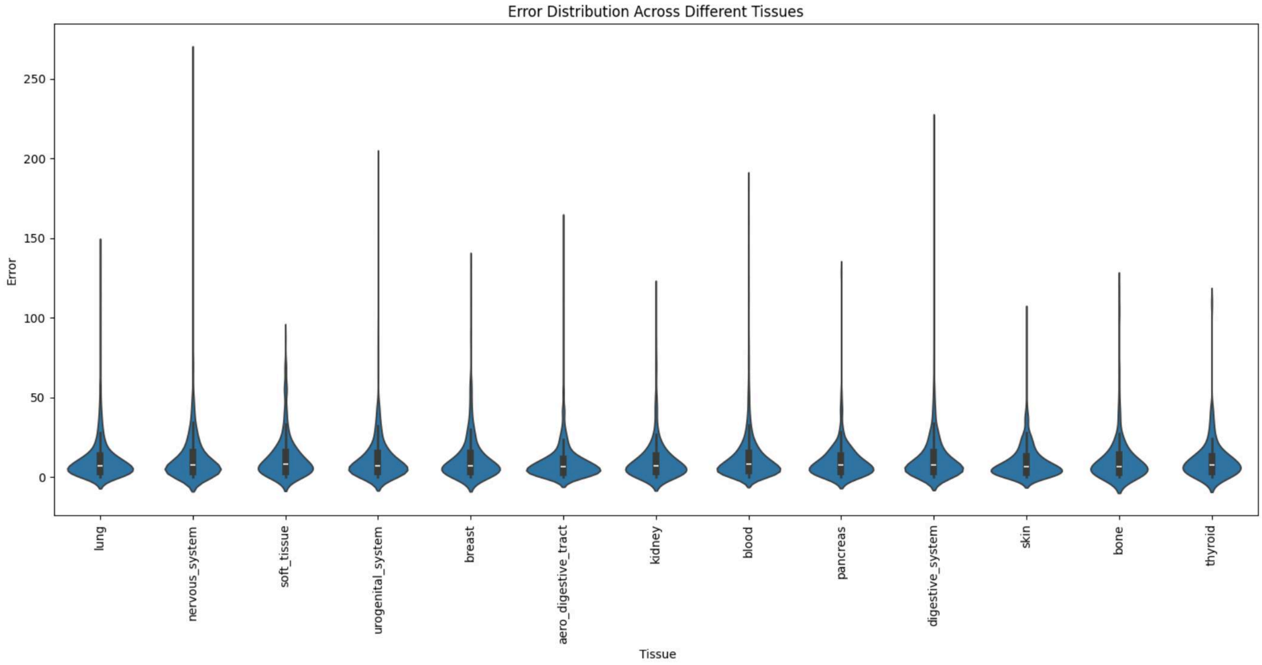


Figure 10: Violin Plots Showing Prediction Error Distribution Across Different Tissue Types

The violin plots showed that the prediction errors were consistent and uniformly distributed across different tissue types. This suggests that tissue type does not significantly impact the model's prediction accuracy, indicating minimal bias and robust performance across various cancer types.

Chapter 6:

Model Development & Implementation

After analysing the impact of tissue type and other properties on prediction accuracy, we concluded that improving the representation of drug and cell-line features could enhance the model's performance. Therefore, we decided to update the drug embedding technique to better capture the intricate information inherent in drug properties. Specifically, we adopted ChemBERTa for drug embedding, which leverages advanced natural language processing techniques tailored for chemical data. For cell-line representation, we utilized gene embeddings comprising 2,128 features, ensuring a comprehensive and detailed capture of the biological characteristics of the cell lines. These changes aim to provide the model with richer and more informative input features, ultimately improving its predictive accuracy and robustness.

To predict the IC50 values, we developed a sophisticated deep learning model that leverages both drug and cell line representations. These embeddings were processed through separate Convolutional Neural Networks (CNNs) before being combined to predict the IC50 value.

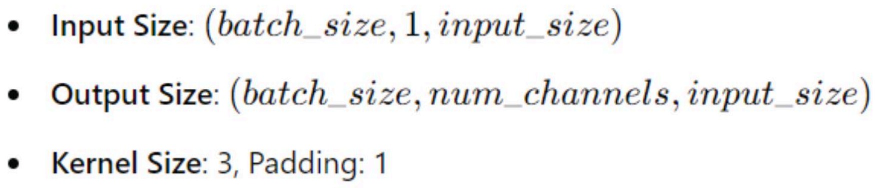
6.1 Model Architecture

The IC50Predictor model is designed to predict the half-maximal inhibitory concentration (IC50) values for drug-cell line pairs using a convolutional neural network (CNN) based approach. The model consists of two main components: separate CNNs for drug and cell-line feature extraction, and a fully connected network for combining these features to predict the IC50 value. Here, we provide a detailed explanation of the model architecture and the functionality of each layer.

6.1.1 CNN Module

The core component of the model is a custom CNN module, which processes both drug and cell-line features. This module is designed to capture intricate patterns within the input data through several key layers:

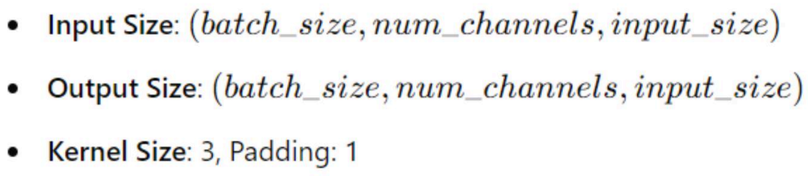
1. **Convolutional Layers**:
   - **First Convolutional Layer**:



- **Input Size**: $(batch\_size, 1, input\_size)$
- **Output Size**: $(batch\_size, num\_channels, input\_size)$
- **Kernel Size**: 3, Padding: 1

The input data is passed through the first convolutional layer, which applies a set of filters to detect local patterns. This layer has a kernel size of 3 and maintains the spatial dimensions of the input through padding.

   - **Second Convolutional Layer**:



- **Input Size**: $(batch\_size, num\_channels, input\_size)$
- **Output Size**: $(batch\_size, num\_channels, input\_size)$
- **Kernel Size**: 3, Padding: 1

The output from the first layer is passed through a second convolutional layer with the same kernel size and padding. This further refines the feature extraction by capturing more complex patterns.

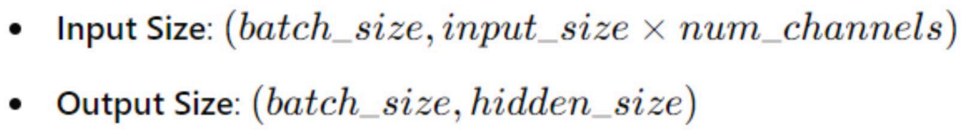
- **Activation Function & Dropout Layer**:

  Input and Output Sizes: These layers do not alter the dimensions of the data.

  - **ReLU Activation**: After each convolutional layer, a ReLU (Rectified Linear Unit) activation function is applied to introduce non-linearity, enabling the model to learn and represent complex relationships in the data.

  - **Dropout**: To prevent overfitting, a dropout layer is included after the convolutional layers with dropout rate 0.5. This randomly sets a fraction of input units to zero during training, promoting robust feature learning by preventing reliance on specific neurons.

2. **Fully Connected Layer**:



- Input Size: $(batch\_size, input\_size \times num\_channels)$
- Output Size: $(batch\_size, hidden\_size)$

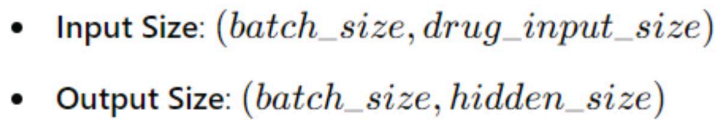
  - **Flattening**: The output from the convolutional layers is flattened into a single vector, transforming the multi-dimensional tensor into a 1D vector suitable for a fully connected layer.

  - **First Fully Connected Layer**: This layer transforms the flattened vector into a hidden representation of the data, enabling further abstraction and feature learning.

6.1.2 IC50 Predictor

The IC50Predictor integrates the outputs from the drug and cell-line CNN modules and further processes them to predict the IC50 value. The architecture involves the following layers:

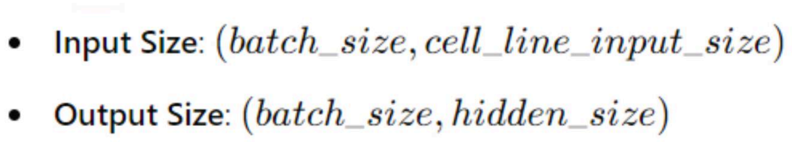
1. **Feature Extraction**:
   - **Drug CNN**:



- **Input Size**: $(batch\_size, drug\_input\_size)$
- **Output Size**: $(batch\_size, hidden\_size)$

Extracts features from drug data using the previously described CNN module.

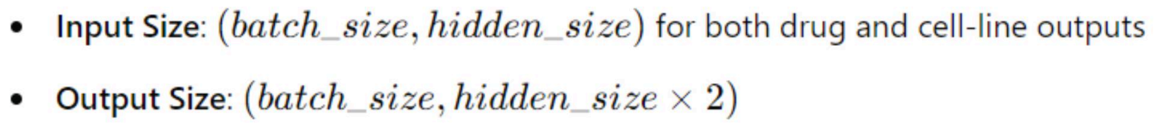
   - **Cell-Line CNN**:



- **Input Size**: $(batch\_size, cell\_line\_input\_size)$
- **Output Size**: $(batch\_size, hidden\_size)$

Extracts features from cell-line data using an identical CNN module.

2. **Combination of Features**:

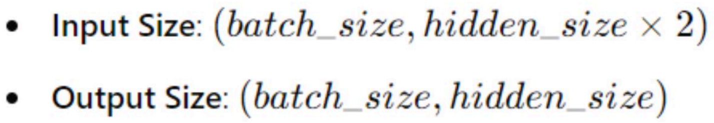


- **Input Size**: $(batch\_size, hidden\_size)$ for both drug and cell-line outputs
- **Output Size**: $(batch\_size, hidden\_size \times 2)$

   - **Concatenation**: The features extracted from the drug and cell-line CNNs are concatenated to form a single combined feature vector, capturing information from both sources.

3. **Fully Connected Layers**:

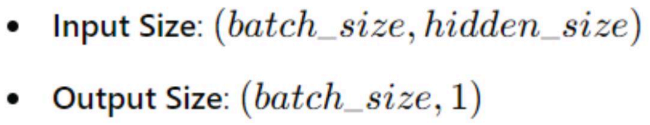
   - **First Combined Fully Connected Layer**:



- Input Size: $(batch\_size, hidden\_size \times 2)$
- Output Size: $(batch\_size, hidden\_size)$

The combined feature vector is passed through a fully connected layer, which reduces the dimensionality and learns a high-level representation of the combined features.

   - **Dropout**: A dropout layer is applied to this representation to prevent overfitting.
   - **Second Combined Fully Connected Layer**:



- Input Size: $(batch\_size, hidden\_size)$
- Output Size: $(batch\_size, 1)$

   - This final layer further refines the high-level representation and outputs the predicted IC50 value as a single continuous value.

The careful design of each layer in the IC50Predictor ensures that the model effectively captures and integrates complex patterns in drug and cell-line data, leading to accurate predictions of IC50 values. This architecture is robust and capable of generalizing across different datasets, making it a valuable tool for advancing personalized cancer therapy.

6.1.3 Training and Optimization Framework

- The IC50Predictor model is trained using three distinct data splits to ensure robustness and generalizability: random split, cell-line split, and drug-wise split.

  Loss Function:

  - **Mean Squared Error (MSE)**: Measures the squared difference between predicted and actual IC50 values during training, optimizing model parameters for accurate predictions.

  Optimizer:

  - **Adam Optimizer**: Utilized for optimizing model parameters.
    - **Learning Rate**: 0.00001
    - Adjusts learning rate dynamically based on validation performance.

  Learning Rate Scheduler:

  - **ReduceLROnPlateau**: Scheduler that adjusts the learning rate when a metric has stopped improving.
    - **Mode**: Minimizing validation loss.
    - **Factor**: Reduces learning rate by a factor of 0.1.
    - **Patience**: Waits for 5 epochs before reducing learning rate.
    - **Verbose**: Prints updates on learning rate adjustments.

Training Across Different Splits:

1. **Random Split:**
   - Randomly divides drug-cell line pairs into training and validation sets.
   - Evaluates model performance across diverse data distributions.
2. **Cell-line Split:**
   - Separates data based on specific cell lines to assess model adaptability.
   - Validates model performance across varying biological contexts.
3. **Drug-wise Split:**
   - Segregates data by drugs, ensuring evaluation on entirely new drug compounds.
   - Tests model generalization capabilities to novel therapeutic agents.

Training Progress Visualization:

- **Plot:** Displays epoch-wise training and validation losses for each split, showcasing model convergence and performance improvements.

6.1.4 Evaluation metrics

After training, the IC50Predictor model was evaluated on test data using the following performance metrics:

RMSE	**R²**	**Pearson Correlation Coefficient**
Indicates the proportion of the variance in the dependent variable (IC50) that is predictable from the independent variables (predictions).	It also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variable	Measures the linear correlation between predicted and actual IC50 values.
$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$\text{Correlation} = \frac{\text{Cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}}$

Where $\text{Cov}(y, \hat{y})$ is the covariance between y and $\hat{y}$, and σ_y and $\sigma_{\hat{y}}$ are the standard deviations of y and $\hat{y}$, respectively.

Table 3: Evaluation Metrics

Chapter 7

RESULTS

In this study, we present IC50Predictor, a model that integrates outputs from the drug and cell-line CNN modules. These modules utilize gene expression features for cell lines and drug features generated from ChemBERTa using SMILES representations. The combined features are then passed through a fully connected layer to predict the LN_IC50 value for drug-cell line pairs.

We trained and evaluated our model on the GDSC 2 dataset, which consists of 51,652 drug-cell line pairs, encompassing 969 cell lines and 202 drugs. To ensure the robustness of our findings, we employed 5-fold cross-validation. This technique involves splitting the data into five folds and using each fold as a validation set once, ensuring that every data point is tested and the results are less likely to be skewed by a particular train-validation split. This approach improves the model’s ability to generalize to unseen drug and cell line data. Additionally, the model was evaluated across three different data splits to assess its generalizability.

7.1 Random Split:

In the random split, the entire dataset is divided randomly into training, validation, and test sets for 5-fold cross-validation. We used a 70-15-15 split: the training set (70%) comprised 36,156 drug-cell line pairs, while the validation and test sets (15% each) comprised 7,748 drug-cell line pairs each. The table below compares the test performance of our model with and without dropout.

Model	Pearson Corelation	
	Best	Average
IC50Predictor	**0.972**	**0.941 ± 0.012**
IC50Predictor with dropout	0.948	0.912 ± 0.026

Table 4: Result of Random Split

The model without dropout performed better than the one with dropout, indicating that dropout might not be necessary for preventing overfitting in this scenario.

7.2 Cell-line Split:

A cell line split involves dividing cancer cell lines into separate groups for training and testing purposes. This process is crucial to ensure that the models are robust, generalize well to unseen data, and accurately predict drug responses. The results for this split are as follows:

Model	Pearson Corelation	
	Best	Average
IC50Predictor		**0.895 ± 0.031**
IC50Predictor with dropout (0.5)		0.862 ± 0.019

Table 5: Result of Cell Line Split

The results indicate that the IC50Predictor model performs well in predicting drug responses for cell lines that were not part of the training set.

7.3 Drug-wise 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.

Model	**Pearson Corelation**	
	Best	Average
IC50Predictor		0.407 ± 0.024
IC50Predictor (lr = 0.00001)		**0.421 ± 0.015**
IC50Predictor with dropout (0.5)		0.368 ± 0.028

Table 6: Result of Drug Split

While the model's performance in the drug-wise split is lower compared to the random and cell-line splits, it highlights areas for potential improvement, particularly in enhancing the model's ability to generalize to new drugs.

Chapter 8

CONCLUSION & FUTURE WORK

8.1 Conclusion:

In this study, we successfully developed the IC50Predictor model, a robust and accurate tool for predicting IC50 values for drug-cell line pairs. The model leverages advanced drug embedding techniques, specifically utilizing ChemBERT embeddings trained on an extensive dataset of drug SMILES representations. This approach significantly improved the model's performance, achieving an average Pearson correlation of 0.941 in the random split, 0.895 in cell line split and 0.421 in drug split after a 5-fold cross validation, which is notably higher than existing baseline models.

The superior results can be attributed to the comprehensive nature of the ChemBERT embeddings, which capture intricate chemical properties more effectively than traditional methods. A key feature of our model is its convolutional neural network (CNN) which includes separate CNNs for drug and cell-line feature extraction. This design enables the model to efficiently process and integrate complex features from both drugs and cell lines, leading to enhanced prediction accuracy.

While the model performed exceptionally well in random and cell-line splits, it showed a relatively lower performance in drug-wise splits. This indicates a potential area for further improvement, particularly in enhancing the model's ability to generalize to entirely new drugs. Despite this, the overall performance of the IC50Predictor model highlights its significant advancements over baseline models and its potential as a valuable tool in drug discovery and cancer research.

8.2 Future Work:

Looking ahead, several avenues can be explored to further enhance the performance and applicability the of IC50Predictor model.

1. Exploration of more advanced drug embedding techniques:

   ChemBERT embeddings have shown significant improvements, there is potential to achieve even better results by integrating additional drug properties i.e. incorporating structural information, and other relevant chemical and physical properties can provide more comprehensive embeddings. The limitations of SMILES representations, such as their inability to fully capture certain structural nuances, underscore the need for richer and more detailed drug descriptors.

2. Expanding the dataset:

   One primary focus should be on expanding the dataset, particularly by incorporating more drug data. Increasing the size and diversity of the dataset will not only improve the model's accuracy but also its ability to generalize across a broader spectrum of drug-cell line interactions

3. Sophisticated network architectures:

   Exploring ensemble approaches that combine multiple machine learning techniques or incorporating attention mechanisms to better capture the interactions between drug and cell line features. Additionally, investigating the use of graph neural networks, which can more naturally represent the complex structures of molecules, could provide significant benefits.

BIBLIOGRAPHY

1. *Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol. 4, 19 (2020). (n.d.).*

2. *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. (n.d.).*

3. *Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607 (2012). (n.d.).*

4. *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. (n.d.).*

5. *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. (n.d.).*

6. *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 . (n.d.).*

7. *Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. (n.d.).*

8. *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. (n.d.).*

9. *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. (n.d.).*

10. *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. (n.d.).*

11. *Swain, M. PubChemPy: A way to interact with PubChem in Python. (2014). (n.d.).*

12. *Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucl. Acids Res. 41, D955–D961 (2013). (n.d.).*