Machine Learning aided cancer drug response prediction using chemo-genomic data

A Project Report

submitted by

PRASHANT SHARMA

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

Department of Computational Biology
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI
NEW DELHI- 110020

July 6th, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Machine Learning aided cancer drug response prediction using chemo-genomic data**, submitted by **Prashant Sharma**, to the Indraprastha Institute of Information Technology, New Delhi, for the award of the degree of **M.Tech. Computational Biology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Debarka Sengupta
Thesis Supervisor
Associate Professor
Dept . of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 6th July 2023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to those who have played an instrumental role in the successful completion of my thesis. First and foremost, I am deeply grateful to my advisor, Dr. Debarka Sengupta, for his exceptional guidance, unwavering support, and belief in me throughout the entire project. His invaluable insights, expert advice, and constructive criticism have been pivotal in shaping the direction of my research and enhancing the quality of my work. I am truly fortunate to have had the opportunity to work under his mentorship. I would also like to extend my heartfelt appreciation to Dr. Prashant Gupta, my senior colleague and mentor. His extensive knowledge, technical expertise, and willingness to share his insights have been instrumental in helping me navigate through critical technicalities and overcome various challenges. I would like to express my gratitude to my friend and project partner, Maheswari, for her constant support, collaboration, and motivation throughout the duration of this project. Her dedication, enthusiasm, and willingness to assist me during all the highs and lows have been truly remarkable. Working alongside her has not only made the project more enjoyable but also immensely rewarding. Furthermore, I would like to extend my thanks to everyone else who has contributed in any way during this project. Whether it was providing assistance, offering valuable suggestions, or simply being there as a source of motivation and encouragement, your support has been greatly appreciated. I am indebted to all those who have generously shared their knowledge, provided resources, and offered their assistance whenever needed.

To all those mentioned above and many others who have been a part of my academic and personal growth, I offer my heartfelt thanks. Your contributions have been invaluable, and I am truly fortunate to have had such a supportive network of individuals around me.

ABSTRACT

The intricate nature of tumor heterogeneity and the personalized responses of cancer patients to drug treatments present significant challenges in the field of oncology. In this thesis, I aim to address such challenges by developing a predictive modeling approach to forecast drug responses in cancer cell lines using a combination of drug and somatic mutations-based features. A comprehensive dataset for every drug-cell-line pair was constructed by concatenating drug descriptors derived from SMILES representations with mutation features obtained through the application of the Personalized PageRank (PPR) algorithm on a Protein-Protein Interaction Network (PPIN). This approach provides a deeper insight into the propagation of gene deleteriousness within the PPIN. The objective of this study is to provide a modeling approach to predict the drug responses for these drug-cell-line pairs. Through rigorous model selection and hyperparameter tuning, the most effective model for the prediction task was selected. The final model - Precily 2.0 exhibited promising performance during evaluation and generated reliable results. Precily 2.0 achieved a Pearson Correlation Coefficient of 0.83, which is better than the other proposed models with a similar methodology. The study highlights the potential of incorporating drug descriptors and network-based propagation of mutation as predictive features for determining drug responses in cancer cell lines. This model would empower oncologists to make informed treatment decisions for individual patients and thereby contributing to the advancement of precision oncology.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

IIITD	Indraprastha Institute of Information Technology, New Delhi
DRP	Drug Response Prediction
PPIN	Protein-Protein Interaction Network
CCLE	Cancer Cell Line Encyclopedia
GDSC	Genomics of Drug Sensitivity in Cancer
CTRP	Cancer Therapeutics Response Portal
NCI	National Cancer Institute
SIFT	Sorting Intolerant From Tolerant
SMILES	Simplified Molecular Input Line Entry System
PPR	Personalized Page Rank
FDA	Food and Drug Administration
VNN	Virtual Neural Network
FFPE	Formalin-Fixed Paraffin-Embedded
qPCR	Quantitative Polymerase Chain Reaction

CHAPTER 1

Introduction

1.1 Precision Medicine

Precision medicine is a part of a novel approach to treat diseases that differs from the traditional medicine by considering the factors such as genes, environment, and lifestyle of a patient rather than considering just the superficial physical symptoms. Precision medicine aims to provide tailored treatments based to each patient based on their specific needs. Precision medicine moves away from a generalized approach and defies the concept of one-size-fits-all. It can be attributed to the development of newer and more effective therapies and interventions. It also helps to understand the depths of origins of diseases and the assocaited mechanisms of action. Next-generation sequencing (NGS) technology is a key player in enabling the pursuit of personalized medicine. NGS technology brings about the possibility to look into an individual's genome, identifying genetic variations that may impact their healthy and indicate how susceptible would they be to diseases. NGS generates high-throughput data, which is analysed to understand genetic profiles in tumours, infections, and other biological samples. In the current scenario NGS tests have become more and more accessible as well as affordable, they provide valuable information of patients for the physicians and researchers. The FDA is responsible in making sure that the quality and reliability of NGS tests, supporting innovation and access are up to the mark. This enables the patients and physicians to confidently rely on the NGS based data, which is a necessity for successful personalized treatments. The prospective of precision medicine revolutionizing healthcare and improve patient outcomes across various diseases and conditions is phenomenally exciting.

1.2 Precision Oncology

Precision oncology is a type of personalized medicine that deals with cancer. Benefits of precision oncology include enhanced accuracy and efficiency in cancer diagnosis and prognosis through the use of molecular markers and advanced imaging techniques. It also upsurges the probability of finding effective treatments by matching patients with therapies that target their explicit genetic or molecular alterations. Furthermore, precision oncology reduces toxicity and side effects by circumventing redundant or ineffective treatments that may damage healthy cells. It plays a vital role in the discovery and development of novel drugs and biomarkers through the exploitation of advanced technologies like NGS and CRISPR/Cas-9. In the long run, precision oncology improves the quality of life and survival rates for cancer patients by providing personalized and precise solutions. However, precision oncology faces some challenges. Performing wide-ranging genomic testing and analysis for each patient can be pricy and time-consuming. Moreover, the understanding and application of genomic testing results are complicated by the varied and multifarious nature of cancer biology. Approachability, accessibility and delivery of precision oncology services can be difficult due to the scarcity of specialized personnel, infrastructure, and equipment. Furthermore, ethical, legal, and social concerns arise, including issues related to informed consent and data privacy. Likewise, clinical benefits may vary or be limited for some patients due to response variability and development of resistance. Precision oncology holds great potential for advancing cancer care. However, addressing challenges related to cost, interpretation, accessibility, and ethical considerations will be crucial for its successful incorporation into routine clinical practice.

1.3 Drug Response Prediction (DRP) models

Drug response prediction models are computational algorithms designed to anticipate the reaction of a patient or cell line to a specific drug or combination of drugs. These models, commonly referred to as drug response prediction (DRP) models, serve various purposes, including treatment prioritization, exploration of drug repurposing, and validation of existing biomarkers. Additionally, DRP models aid in the discovery and development of new drugs and biomarkers. The drug response prediction problem could

be understood as follows. DRP model work by integrating drug molecular and cell line specific omics features to predict the drug response. The quantification of drug response entails measuring the impact of a drug or drug combination on cell growth or viability in patient samples or cell lines.

Example of a DRP model

Drug response prediction models aim to anticipate the reaction of a patient or cell line to a specific drug or combination of drugs. These models can be formulated using various mathematical equations. One example is the equation for a basic feedforward neural network:

$$y = f(W \cdot x + b)$$

where x represents the input features, W represents the weight matrix, b represents the bias vector, and f represents the activation function.

Approach towards modeling for a DRP problem

What follows is a generalized approach toward modeling for a DRP problem.

Step 1: Gathering and Preprocessing Input Data To begin, we need to collect and pre-process the input data required for training and testing the model. This data can be sourced from various outlets, such as public databases, high-throughput screenings, or clinical trials. It encompasses different types of information, including molecular profiles (e.g., gene expression, mutation, copy number, methylation), drug structures (e.g., SMILES strings, molecular fingerprints), pathway activities (e.g., gene set enrichment scores, pathway scores), as well as other pertinent features like cancer type, cell line name, or drug name. It is crucial to pre-process the input data to ensure its quality, consistency, and compatibility. For instance, molecular profiles need to be normalized, filtered, imputed, or transformed to minimize noise and variations. Drug structures should be standardized, encoded, or transformed to capture chemical properties and similarities. Pathway activities ought to be computed, aggregated, or scaled to reflect biological functions and interactions. Additionally, other features should be encoded, categorized, or mapped to facilitate the input and output of the model.

Step 2: Defining and Measuring the Output Variable The second step involves defining and measuring the output variable, which is used to assess the performance of the model. This variable can be a continuous or discrete measure of drug response or synergy. For instance, it could be drug sensitivity indicators like IC50, AUC, or GI50; synergy scores such as Bliss excess, Loewe excess, ZIP excess, HSA excess, or Chou-Talalay combination index; or survival outcomes like overall survival, progression-free survival, or disease-free survival. To measure the output variable accurately, appropriate methods need to be employed that account for dose-response relationships, drug interactions, and statistical significance. For instance, drug sensitivity can be measured using dose-response curves that fit sigmoidal models to observed cell viability or growth inhibition data. Synergy scores can be determined using isobolograms that compare the observed combination effect with a reference model of additive or independent effect. Survival outcomes can be measured using Kaplan-Meier curves that estimate the survival probability over time.

Step 3: Selecting and Training an Appropriate Machine Learning Model In the third step, it is necessary to choose and train a suitable machine learning model using the through process of model selection of trying various different models. The model that we want could be a neural network like a multilayer perceptron, convolutional neural network, or any other advanced neural network based models. Additionally, It could also include various tree-based model like a random forest, gradient boosting machine, or advanced implementations of such models like XGBoost, CatBoost, LightGBM. The machine learning model should be trained using appropriate optimization methods that minimize a loss function and maximize an accuracy metric. Optionally, the model architecture or regularization can be enhanced with prior biological knowledge to improve performance and interpretability. This could involve incorporating pathway structure or target function into the model parameters or guiding feature selection.

In machine learning models, training involves optimizing the model parameters using an optimization method. One common method is gradient descent, which can be represented by the following update equation:

$$\theta = \theta - \alpha \cdot \nabla J(\theta)$$

where θ represents the model parameters, α represents the learning rate, $J(\theta)$ represents

the loss function, and ∇ represents the gradient operator.

Step 4: Evaluating and Validating Model Performance For evaluation, if the problem is posed as a regression problem metrics can include correlation coefficient or coefficient of determination (R2). For a classification problem the metrics could be F1-score (F1), area under the curve (AUC), or Matthews correlation coefficient (MCC). These would be the most preferred metrics in both the cases. Model performance can be validated using datasets independent of the training data. This can involve using cross-validation sets, external validation sets, or clinical validation sets. Cross-validation sets are subsets of the training data reserved for testing the model's generalization. External validation sets consist of datasets from different sources or platforms to assess the model's transferability. Clinical validation sets comprise datasets from clinical trials or patient cohorts to determine the model's applicability. The model's performance can be compared to existing models or null hypotheses to evaluate improvements or significance. Existing models can include baseline models like linear regression, logistic regression, or k-nearest neighbors, state-of-the-art models like DrugCell, Precily, or DrugSynergy, or alternative models employing different machine learning algorithms or architectures. Null hypotheses can assume pharmacological independence, pharmacological additivity, or statistical independence, which imply no interaction or additive effect between drugs. Step 5: Interpreting and Analyzing the Model Results

Ste 5: In the fifth and final step, it is crucial to interpret and analyze the model's results using various methods that provide insights into the model's predictions and underlying mechanisms. The model's results may include predicted output such as drug sensitivity, synergy score, or survival outcome, as well as model parameters like weights, biases, or coefficients. Feature importance measures like variable importance, permutation importance, or SHAP values, activation patterns like hidden layer outputs, attention weights, or pathway scores, pathway enrichment analyses like gene set enrichment analysis (GSEA), hypergeometric test, or Fisher's exact test, and network visualizations like graphs, heatmaps, or isobolograms. These results can help identify molecular mechanisms, pathways, or targets that explain drug response or synergy. For example, they can highlight gene expression changes, pathway activity alterations, or drug structure modifications that correlate with drug response or synergy. Additionally, they can shed light on biological processes, molecular functions, or cellular components that mediate drug response or synergy. Furthermore, they can identify proteins, genes, or metabolites

that modulate drug response or synergy. The model's results can also generate testable hypotheses or actionable recommendations based on the insights gained. Testable hypotheses might involve causal relationships, functional interactions, or mechanistic explanations that can be verified through experiments such as CRISPR-Cas9 knockout, RNA interference knockdown, in vitro drug-drug screening, in vivo patient-derived xenografts, or human clinical trials. Actionable recommendations could include potential drug candidates, drug combinations, or patient stratification approaches that can be applied for therapy selection, optimization, or personalization.

Model performance can be evaluated using various metrics. For example, the Pearson correlation coefficient (r) can be calculated to measure the linear relationship between two variables. The formula for Pearson correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i represent the individual data points, $\bar{x}$ and $\bar{y}$ represent the means of the x and y values, and n represents the number of data points.

Another commonly used metric is the R-squared (R^2) score, also known as the coefficient of determination. It measures the proportion of the variance in the dependent variable that can be explained by the independent variable(s). The formula for R-squared is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i represents the observed values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ represents the mean of the observed values.

1.4 Importance of cell line data in DRP models

A major hurdle in creating drug response prediction models is the limited availability of human omics data. This type of data encompasses high-throughput biochemical assays that comprehensively and simultaneously measure molecules of the same kind of omics. Human omics data provides a holistic understanding of the biological system and uncovers the molecular mechanisms underlying drug response. While there are public databases like the The Cancer Genome Atlas, the International Cancer Genome Con-

sortium, the ProteomeXchange Consortium, the Proteomics Identifications Database, and the Human Metabolome Database that store human omics data from various studies, these databases have certain limitations when it comes to drug response prediction models. For instance: The human omics data often lack completeness, consistency, or compatibility across different sources, platforms, or assays, which makes it difficult to integrate and analyze the data in a systematic and standardized manner. The human omics data frequently exhibit sparsity, noise, or heterogeneity across different samples, tissues, or individuals, posing challenges in capturing and modeling the variability and complexity of drug response. Obtaining human omics data from clinical trials or patient cohorts is often expensive, time-consuming, or ethically complex, hindering large-scale and timely access and sharing of the data. As an alternative data source, cell line data plays a crucial role in the development of drug response prediction models. Cell line data refers to omics data acquired from cultured cells derived from human tumors or tissues. Leveraging cell line data offers several advantages for drug response prediction models, including: Cell line data is more readily available, consistent, and compatible across different sources, platforms, or assays, facilitating systematic integration and analysis of the data. Cell line data is more abundant, robust, and homogeneous across different samples, tissues, or individuals, making it easier to capture and model the variability and complexity of drug response. Obtaining cell line data from high-throughput screenings or experimental manipulations is more affordable, efficient, and ethical, enhancing access and sharing of the data on a large scale and in a timely manner. The Cancer Cell Line Encyclopedia (CCLE) stands out as one of the most comprehensive sources of cell line data. However, it is important to acknowledge that cell line data also has limitations when it comes to drug response prediction models. Some of these limitations include: Cell line data may not adequately represent the diversity and heterogeneity of human cancers or patients, posing difficulties in generalizing and personalizing the model predictions across different subtypes or populations.To overcome these limitations, a potential solution is to validate results obtained from these models in a wet lab setting.+

1.5 Precily

Precily: A Framework for Drug Response Prediction Precily is a state-of-the-art and fairly advanced model that utilizes deep neural networks to forecast how cancers will respond to various drugs. It leverages gene expression data from CCLE, which encompasses transcriptomic data of cancer cell lines, to train the network. Additionally, it incorporates pathway activity scores and drug features as inputs to capture the molecular mechanisms behind drug response. This versatile framework can be applied to different scenarios, including in vitro and in vivo experiments, as well as patient data from TCGA and separate clinical studies. The primary stages of the Precily framework include: Preprocessing: The gene expression data is normalized and imputed using R packages EDASeq and impute. Pathway activity scores are calculated using the R package GSVA. Drug features are obtained from PubChem via the Python package PubChemPy. Training: A deep neural network with three hidden layers is trained on the CCLE data using Python packages Keras and Keras-tuner. This network takes pathway activity scores and drug features as inputs and generates drug response values (represented by the area under the dose-response curve). The network is optimized using Bayesian optimization and cross-validation techniques. Prediction: The trained network is employed to predict drug response for new samples based on their pathway activity scores and drug features. These predictions can be compared with experimental or clinical data to evaluate the framework's performance. Interpretation: Analysis of the network weights helps identify the pathways and drugs that play a crucial role in drug response prediction. The network architecture is visualized using a workflow diagram.

1.6 DrugCell

DrugCell: A Model for Predicting Drug Response in Tumors DrugCell is an advanced virtual neural network (VNN) that forecasts the impact of anti-cancer drugs on tumors by modeling the hierarchical organization of human cancer cells. It takes genotypes and drug structures as inputs and generates drug response values (represented by the half-maximal inhibitory concentration). The model also captures the patterns of activity on cellular subsystems that explain drug response and synergy. The key steps

involved in the DrugCell model are as follows: Preprocessing: The genotype data is encoded as binary vectors that indicate the presence or absence of mutations or copy number alterations in 15,000 genes. The drug structure data is encoded as binary vectors that indicate the presence or absence of 881 molecular fingerprints. Drug response data is obtained from the GDSC database. Training: A VNN with three hidden layers is trained on GDSC data using the Python package PyTorch. The network takes genotype and drug structure vectors as inputs and generates drug response values. The architecture of the network reflects the hierarchical organization of human cancer cells, with each layer representing a different level of cellular subsystems (e.g., pathways, complexes, proteins). The network is optimized using stochastic gradient descent and cross-validation. Prediction: The trained network is utilized to predict drug response for new samples based on their genotype and drug structure vectors. These predictions can be compared with experimental or clinical data to evaluate the model's performance. Interpretation: Analysis of the network weights helps identify the cellular subsystems that are most activated or inhibited by different drugs or genotypes. The network architecture is visualized using a dedicated VNN browser, allowing interactive exploration of drug response pathways and synergistic targets.

1.7 Need of a new model and problem with using transcriptomics data

The objective of genomic-based drug response prediction models is to determine the most effective treatment for each patient by considering their unique genomic profile. However, this task is challenging due to the complexity and heterogeneity of cancer, limited data availability, and imperfect methodologies employed. Formalin-fixed paraffin-embedded (FFPE) tissue samples are commonly used as a data source for drug response prediction. These samples are easily accessible, abundant, and provide insights into the tumour microenvironment. Nevertheless, FFPE samples have certain drawbacks that impact the quantity and quality of extracted nucleic acids. One notable issue with FFPE samples is RNA degradation, which occurs during the fixation process and long-term storage. This degradation leads to shorter and less intact RNA molecules, affecting their suitability for downstream applications like reverse transcription, PCR,

and sequencing. Moreover, RNA degradation introduces biases and artefacts, altering gene expression levels and diminishing the accuracy and reproducibility of drug response prediction models. Several factors influence the extent of RNA degradation in FFPE samples, including fixation time, storage temperature, humidity, specimen size, thickness, and tissue type. Therefore, optimizing pre-analytical procedures and employing standardized protocols for FFPE sample processing and storage are crucial. Additionally, quantifying the amount and quality of extracted RNA through techniques like electrophoresis, spectrophotometry, or qPCR is essential before conducting any molecular analysis. RNA degradation in FFPE samples poses a significant hurdle for drug response prediction based on genomic data. This indicated the need of a model that can work with a genomics based model, and that is why in Precily - 2.0 we work on building a model based on the genomics data of a cell line instead of transcriptomic or any other omics data. One such state-of-the-art model that already exists is DrugCell. We need a new genomics based model because we believe a model can perform better than how DrugCell is performing.

CHAPTER 2

Data

2.1 Drug Data

2.1.1 SMILES

SMILES, an abbreviation for Simplified Molecular Input Line Entry System, is a system that utilizes concise strings of ASCII characters to represent the structure of chemical substances. It serves as a universal and compact method for describing drugs and other compounds. As an illustration, the SMILES representation for diphenhydramine is: CN(C)CCOC(C1=CC=CC=C1)C1=CC=CC=C1. The application of SMILES offers valuable advantages in drug design, mining, and repurposing through the utilization of machine learning and natural language processing techniques. For instance, SMILES can aid in the creation of novel drug candidates, estimation of drug-drug interactions, and assessment of similarity between different drugs. Various tools and libraries are available to convert SMILES into alternative formats, such as molecular graphs or 3D structures. While SMILES provides a simple and efficient approach to encoding drug information as text, it does possess certain limitations and challenges. One drawback is its lack of uniqueness, as the same molecule can have multiple valid SMILES representations. Additionally, SMILES does not encompass all aspects of molecular structure, such as stereochemistry or conformation. Hence, when working with drug data, caution and verification are necessary when employing SMILES.

2.1.2 Need of standardizing the smiles using open babel

The necessity for standardizing SMILES notation using Open Babel arises due to the inherent limitations and variations in SMILES representation. The problem is that multiple SMILES strings can represent the same molecule, leading to potential inconsistencies and ambiguities. To address this issue, Open Babel, a freely available open-

source software, offers a solution. It enables the conversion of chemical formats, including SMILES, and provides the capability to generate canonical SMILES. Canonical SMILES ensures that each molecule has a unique and reproducible representation, promoting consistency and facilitating accurate comparisons and searches. Canonical SMILES are particularly valuable for indexing, hashing, and promoting interoperability across different software and databases. In addition to standardization, Open Babel offers various operations on SMILES strings. It allows the addition or removal of explicit hydrogens, isotopic or chiral markings, atom classes, or coordinates. Open Babel can also rearrange SMILES strings, enabling the specification of the first or last atom and supporting the concatenation or branching of molecules. Open Babel serves as a versatile and powerful tool to enhance the usability and reliability of SMILES notation. It can be utilized through a command-line interface, a user-friendly graphical interface, or as a library for different programming languages. By leveraging Open Babel's capabilities, researchers and practitioners can ensure standardized SMILES representation, thereby promoting accurate and efficient analysis and manipulation of chemical structures. It is needed that SMILES collected be standardized by OpenBabel.

2.2 CCLE Mutations data

2.2.1 Data in CCLE

The Cancer Cell Line Encyclopedia (CCLE) project, a collaborative effort between the Broad Institute and the Novartis Institutes for Biomedical Research, aims to characterize the genetic and molecular attributes of approximately 1000 cancer cell lines, along with their responses to diverse drugs and perturbations. One of the datasets generated by CCLE is the information on mutations, which details the somatic mutations found in the coding regions of the cancer cell lines.

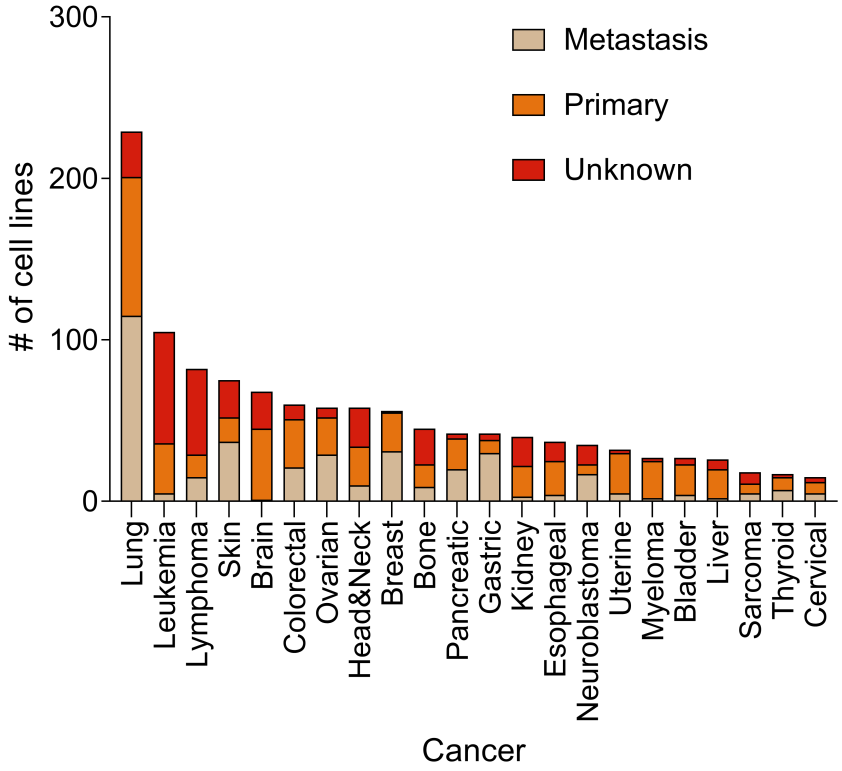

Figure 2.1: Number of cell lines emerging from different cancer types and information on the primary and metastatic state of cancer

The mutation data is acquired through hybrid capture sequencing, which focuses on a set of genes known or suspected to be involved in cancer. It is provided in MAF (Mutation Annotation Format) files, offering specifics like gene name, chromosome, position, reference allele, variant allele, variant type, and functional impact. This mutation data serves multiple purposes, including exploring the genetic heterogeneity and diversity of cancer cell lines. It also aids in identifying potential biomarkers or drivers of drug sensitivity or resistance. For instance, researchers have employed CCLE mutation data in combination with machine learning methods to predict the response of tumors to drugs using integrated genomic profiles. Furthermore, by integrating mutation data with other CCLE datasets, such as copy number variations, gene expression, DNA methylation, proteomics, and metabolomics, a comprehensive understanding of the molecular mechanisms underlying cancer biology and therapeutics can be achieved. CCLE mutation data is accessible through various sources, including the Broad DepMap Portal, the R package depmapAnalysis, and the Harmonizome database. Researchers can employ different tools and libraries, such as R, Python, Bioconductor, or cBioPortal, to analyze and interpret the data for their specific research needs.

2.2.2 SIFT scores and is-deleterious mutations

SIFT scores serve as a predictive measure for assessing the impact of amino acid alterations on protein function. The basis of SIFT scores lies in the notion that amino acid changes conserved across diverse species tend to be crucial for protein function, while variable changes across species are more likely to be tolerated. By comparing amino acid frequencies in a collection of related sequences, SIFT scores estimate the probability of an amino acid change being tolerated. SIFT scores ranges from 0.0 to 1.0, with lower scores indicating reduced tolerance and higher scores indicating greater tolerance. These scores enable the classification of amino acid changes into two categories: deleterious or tolerated. Deleterious changes are expected to impair protein function, while tolerated changes are predicted to have minimal or no effect. SIFT scores aid in the identification of mutations that might be linked to diseases, phenotypes, or drug responses.

SIFT score equation

The SIFT score equation is:

$$SIFT\ score = p_{\text{not tolerated}} = \prod(1 - p_i)$$

where p_i represents the frequencies of amino acid changes at each position in the alignment.

The threshold for defining deleterious changes can vary depending on the intended application and the sensitivity of the prediction. A common threshold is 0.05, meaning that amino acid changes with SIFT scores below 0.05 are predicted to be deleterious. Various sources provide SIFT scores, including public databases, web servers, and software tools. The SIFT 4G annotator stands out as one of the most comprehensive sources, offering precomputed SIFT scores for over 200 organisms based on their reference genomes. Users also have the option to utilize the SIFT 4G algorithm to compute SIFT scores for their own sequences or variants. Other sources of SIFT scores encompass the dbNSFP database, the Ensembl Variant Effect Predictor, and the ANNOVAR tool. Although SIFT scores are widely employed and valuable for predicting the functional consequences of amino acid changes, they do possess limitations and encounter

challenges. Some of these include: SIFT scores heavily rely on the availability and quality of homologous sequences for each protein. Insufficient or unrepresentative homologous sequences can result in inaccurate or unreliable SIFT scores. SIFT scores do not account for additional factors that can impact protein function, such as protein structure, interactions, post-translational modifications, or environmental conditions. Consequently, SIFT scores may not fully capture the intricate nature and context-dependent aspects of protein function.SIFT scores do not provide mechanistic explanations for the deleterious or tolerated nature of an amino acid change. Therefore, they may not reveal the underlying molecular pathways or targets involved in mediating the functional consequences. It is crucial to interpret SIFT scores cautiously and supplement them with other sources of information, such as experimental validation, structural analysis, or functional annotation, to gain a more comprehensive understanding and improve predictions of the functional impact of amino acid changes.

2.3 Drug Response Data

2.3.1 Creammist Database

Introduction to Creammist database

CREAMMIST database is a comprehensive probabilistic database for predicting the response of cancer drugs Precision oncology strives to customize the most effective treatment for each patient based on their unique genomic characteristics, and one of the primary objectives is to predict drug response accurately. However, this task is challenging due to the intricate and diverse nature of cancer, limited data availability, and imperfect methodologies employed. In vitro cancer drug screening datasets, which evaluate the sensitivity of numerous cancer cell lines to different drugs, serve as a common data source for predicting drug response. These datasets offer valuable insights for identifying biomarkers and developing machine learning models for drug response prediction. Nonetheless, they do come with certain limitations, including inconsistencies and variabilities in the drug sensitivity scores across various sources, owing to differences in experimental setups and preprocessing techniques. To address these limitations, an integrative probabilistic database called CREAMMIST (https://creammist.mtms.dev)

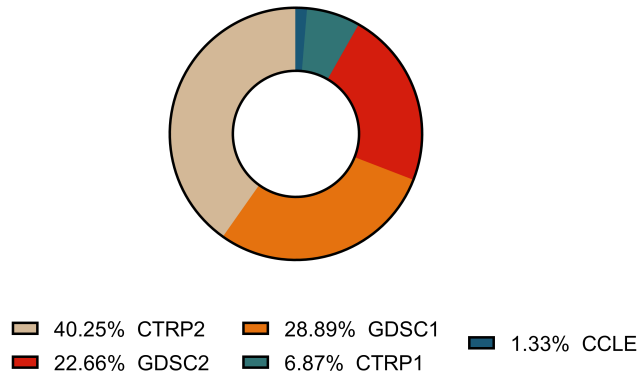

Figure 2.2: Sources of different drugs

has been developed specifically for cancer drug response prediction. CREAMMIST consolidates five widely utilized datasets on cancer cell-line drug response, comprising over 14 million dose-response data points. It employs a Bayesian framework to generate integrative dose-response curves, taking into account the uncertainty associated with different sources (or high certainty when multiple datasets are well aligned). The database also offers user-friendly statistics derived from the integrative dose-response curves.

Measuring the drug response

When measuring drug response, specific metrics such as IC50 or AUC can be used. The equation for calculating IC50, for example, can be expressed as:

$$IC50 = \frac{1}{2} \times \text{maximum response}$$

where the maximum response is determined from the dose-response curve.

Dose-Response Curve

The dose-response curve is a fundamental tool used in pharmacology and toxicology to describe the relationship between the dose or concentration of a compound and the biological response it produces. It provides valuable information about the compound's efficacy, potency, and potential toxicity.

The mathematical equation commonly used to model the dose-response relationship

is the Hill equation:

$$E = \frac{E_{\text{max}} \cdot C^n}{C_{50}^n + C^n} \tag{2.1}$$

In this equation, E represents the observed response, E_{max} is the maximum attainable response, C is the concentration or dose of the compound, C_{50} is the concentration or dose that produces a 50% response, and n is the Hill coefficient that describes the steepness of the curve. The Hill equation allows for the characterization of various dose-response curve shapes. When $n > 1$, the curve exhibits positive cooperativity, indicating a sigmoidal shape. Conversely, when $n = 1$, the curve is linear. The value of C_{50} represents the concentration or dose at which the response is half-maximal and is often used as a measure of potency. By fitting experimental data to the Hill equation, researchers can estimate the parameters E_{max}, C_{50}, and n, providing quantitative insights into compound activity and allowing for comparisons between different compounds and biological systems.

Potential of Creammsit database

The CREAMMIST data holds immense potential for diverse downstream analyses, such as biomarker discovery, determination of optimal drug concentrations for experiments, and training robust machine learning models. Additionally, it serves as a valuable resource for conducting pharmacogenomics analysis, which seeks to elucidate the molecular mechanisms underlying drug action and resistance in cancer. Researchers and clinicians interested in precision oncology and cancer drug response prediction can greatly benefit from the wealth of information provided by the CREAMMIST data.

CHAPTER 3

Methodology

3.1 Featurization in Machine Learning

The process of data representation holds significant importance in the field of machine learning, enabling models to gain insights and make informed predictions or decisions based on available data. However, not all data types are initially compatible with machine learning algorithms, especially if they lack a numerical format. Different data formats, such as text, graphs, time-series, and images, require conversion into numerical vectors before they can be effectively utilized for machine learning tasks. Data representation involves the conversion of non-numerical data into numerical vectors that encapsulate the pertinent information and characteristics of the original data. It is a distinct process from feature engineering, which focuses on modifying numerical features to improve their suitability for machine learning. Data representation is also known as feature extraction or feature representation. Featurization techniques vary depending on the type and structure of the data being transformed. For instance, text data can be processed using techniques like bag-of-words, term frequency-inverse document frequency (TF-IDF), word embeddings, or topic modeling, which convert text into numerical vectors representing words, frequencies, meanings, or topics. Graph data, on the other hand, can utilize techniques like node degree, clustering coefficient, PageRank, shortest path, or graph neural networks to convert nodes and edges into numerical vectors representing properties, centrality, similarity, or embeddings. Time-series data can be transformed using methods such as Fourier transform, wavelet transform, autocorrelation, or recurrent neural networks, capturing frequency, amplitude, trend, seasonality, or patterns. Similarly, image data can be processed using techniques like pixel values, color histograms, edge detection, or convolutional neural networks to generate numerical vectors representing intensity, color, shape, or features in images. Featurization techniques are crucial in machine learning as they enable the utilization of non-numerical data for tasks like classification, regression, clustering, or recommendation. These techniques

help reduce data dimensionality and complexity, enhance model performance and accuracy, and facilitate result interpretation and visualization. In the domain of omics data, which encompasses biological data related to genomics, transcriptomics, proteomics, metabolomics, and phenomics, featurization techniques are especially important due to the high-dimensionality, heterogeneity, and noise inherent in such data. Featurization in omics data involves preprocessing steps such as cleaning, normalizing, scaling, and filtering to remove noise and outliers and ensure data consistency. Additionally, dimensionality reduction techniques like principal component analysis can be used to reduce the aberrant number of features and keep the relevant ones. Featurization in omics data enables machine learning applications in biomarker discovery, disease diagnosis, drug response prediction, and personalized medicine. Moreover, it aids in uncovering the underlying molecular mechanisms and interactions that drive the biological phenomena of interest. Featurization techniques are vital in machine learning as they enable the utilization of non-numerical data for tasks such as classification, regression, clustering, and recommendation. These techniques also help in reducing data complexity, improving model performance and accuracy, and facilitating result interpretation and visualization. In the field of omics data, which encompasses genomics, transcriptomics, proteomics, metabolomics, and phenomics, featurization techniques are particularly significant due to the high-dimensionality, heterogeneity, and noise inherent in such data. The process involves preprocessing steps like cleaning, normalizing, scaling, and filtering, followed by dimensionality reduction techniques such as feature selection or extraction. Integration of different omics data types is also performed to gain a comprehensive understanding of the biological system. Featurization in omics data allows for machine learning applications in biomarker discovery, disease diagnosis, drug response prediction, and personalized medicine. Moreover, these techniques aid in uncovering underlying molecular mechanisms and interactions that drive biological phenomena. Overall, featurization techniques are instrumental in enabling machine learning to handle diverse data types, enhance model performance, and extract valuable insights from complex datasets.

3.2 Featurization of Drug SMILES

3.2.1 Approaches for featurization of SMILES

Converting drug SMILES (Simplified Molecular Input Line Entry System) into machine learning features is crucial for utilizing machine learning models in drug discovery applications. However, since SMILES strings are categorical, variable-length, and non-unique, they are unsuitable for machine learning models. To address this, several approaches have been developed to transform SMILES into numerical representations that capture the structural and functional characteristics of molecules. One common method is the use of molecular fingerprints, which are binary vectors indicating the presence or absence of specific substructures or patterns in a molecule. Examples include Morgan fingerprints, generated using a circular hashing algorithm, and MACCS keys, which employ predefined yes/no questions about a molecule's properties and structure. Molecular fingerprints find applications in similarity search, clustering, and classification tasks. Another approach involves calculating molecular descriptors, which are numeric values reflecting various physicochemical and pharmacological properties of a molecule, such as molecular weight, logP, polar surface area, and hydrogen bond donors/acceptors. Software tools like RDKit or OpenBabel can be used to derive molecular descriptors from SMILES strings. These descriptors are useful for regression and quantitative structure-activity relationship (QSAR) modeling. Molecular embeddings represent another method, employing representation learning techniques like autoencoders, variational autoencoders, or graph neural networks to learn continuous vectors that capture latent semantic information about molecules. These embeddings aim to capture aspects like chemical diversity, biological activity, or synthetic feasibility and can be employed in generative, transfer learning, and multi-task learning tasks. Recent studies have explored the combination of molecular images, 2D representations showing the spatial arrangement of atoms and bonds, with vector representations for drug classification. These studies propose augmenting molecular images with additional binary vectors encoding information not readily apparent from the images alone, such as Morgan fingerprints or MACCS keys. The combination of molecular images and vector representations has shown promising results in tasks like HIV inhibition classification. Another novel approach, Affinity2Vec, addresses drug-target binding affinity (DTBA)

prediction as a graph-based problem. It constructs a weighted heterogeneous graph incorporating drug-drug similarity, target-target similarity, and drug-target binding affinities. By employing feature representation learning, graph mining, and machine learning techniques, Affinity2Vec achieves superior performance compared to existing methods for predicting binding affinities without relying on 3D structural data. Selecting the appropriate method for transforming drug SMILES into machine learning features is crucial for effective drug discovery applications. Each method has its advantages and limitations, dependent on the specific task and available data. Careful evaluation of performance and interpretability is essential to choose the most suitable approach for each problem.

3.2.2 Comparison of different approaches

To enable machine learning models to predict drug properties and interactions based on their structures, it is necessary to convert drugs, typically represented by SMILES (Simplified Molecular Input Line Entry System), into numerical features. SMILES strings are alphanumeric notations describing the connectivity and stereochemistry of atoms in a drug. For example, the SMILES string for aspirin is CC(=O)Oc1ccccc1C(=O)O. However, SMILES strings are unsuitable for machine learning models due to their categorical nature, variable lengths, and lack of uniqueness. Consequently, several methods have been developed to transform SMILES strings into numerical vectors or matrices capable of capturing the structural and functional information of drugs. Here are some methods for converting drugs using SMILES:

Mordred: This software computes over 1800 molecular descriptors from SMILES strings. These descriptors represent numerical values quantifying various physicochemical and pharmacological properties, such as molecular weight, logP, polar surface area, hydrogen bond donors and acceptors, and more. Mordred prioritizes ease of installation and use, supports extensive molecular descriptors, delivers rapid calculations, and includes automated tests.

Graph2vec: This method learns continuous embeddings for graphs derived from SMILES strings. Treating each drug as a graph, with atoms as nodes and bonds as edges, Graph2vec employs a graph kernel to measure graph similarity. Subsequently, a

neural network is employed to learn low-dimensional vector representations for graphs. Graph2vec effectively captures structural and functional information within a compact space.

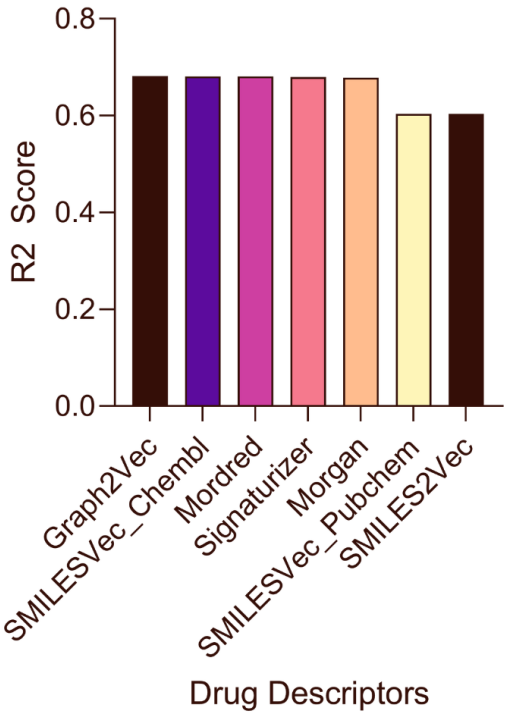

Figure 3.1: Comparison of different drug descriptors

SmilesVec: This method learns continuous embeddings for SMILES strings by treating each drug as a character sequence. Utilizing a recurrent neural network (RNN), SmilesVec encodes the sequence into a vector, thereby capturing syntactic and semantic information in a lower-dimensional space.

Smiles2Vec: Similar to SmilesVec, this method learns continuous embeddings for SMILES strings. It treats each drug as a token sequence, where tokens represent atoms or bond symbols. Employing an RNN, Smiles2Vec encodes the sequence into a vector, capturing latent semantic information in a lower-dimensional space.

Morgan: This method generates binary fingerprints for drugs based on their SMILES strings. Morgan fingerprints employ a circular hashing algorithm to create fixed-length fingerprints, with each bit indicating the presence or absence of specific substructures or patterns in the drug. Widely used for similarity search, clustering, and classification, Morgan fingerprints provide valuable insights.

Signaturizer: This method generates binary fingerprints for drugs using SMILES strings. Signaturizer utilizes a signature hashing algorithm to generate fixed-length fingerprints, with each bit indicating the presence or absence of specific signatures within the drug. Signatures consist of sets of atoms and bonds forming paths within the drug, allowing for the capture of higher-level structural information compared to Morgan fingerprints.

Transforming drugs utilizing SMILES is crucial for leveraging machine learning models in drug discovery. Different methods offer distinct advantages and limitations depending on the specific task and available data. It is therefore important to choose the most appropriate method for each problem and evaluate its performance and interpretability carefully.

3.3 Featurization of Cell Line Mutations

3.3.1 Examples of featurization techniques for mutations

The CCLE project is an extensive endeavor with the objective of characterizing the genetic and molecular attributes of over 1000 cancer cell lines, widely utilized models for studying cancer biology and drug response. Mutation data involves identifying changes in the DNA sequence of the cell lines. These mutations can impact the functionality of genes and pathways involved in cancer development and progression. Mutation data also enables the inference of mutational signatures, which reflect exposure to different mutagenic processes.

Some examples of featurization techniques for mutation data include: Mutation count: This simple feature denotes the total number of mutations in a cell line. It can be calculated for different mutation types or genomic regions, providing insights into mutational load and potential response to immunotherapy. Mutation frequency: This feature represents the proportion of mutations in a cell line relative to a reference genome or normal sample. It can be computed for different mutation types or genomic regions, reflecting genomic instability and heterogeneity. Mutation spectrum: This feature describes the distribution of mutation types in a cell line, considering nucleotide contexts or functional consequences. It unveils mutational signatures and underlying

mutagenic processes. Mutation annotation: This feature captures the functional impact of mutations in a cell line, utilizing tools and databases to annotate mutations based on their effects on genes, pathways, protein domains, structure, and interactions. It aids in identifying driver mutations and potential therapeutic targets. Mutation clustering: This feature highlights the spatial distribution of mutations in a cell line, identifying regions of high mutation density or recurrence across multiple cell lines. It indicates genomic fragility, hotspots of oncogenic activity, and potential therapeutic vulnerabilities.

These examples illustrate featurization techniques applicable to CCLE mutation data, although other methods may also be utilized. The choice of featurization techniques depends on the research question, available data, and machine learning model. These techniques assist in extracting meaningful features from CCLE mutation data, enabling exploration of the genetic diversity, complexity, and relationships with other molecular and phenotypic features of cancer cell lines.

3.3.2 Binary Representations

What is binary representation?

The featurization techniques discussed above are naïve and we need better representations. The binary representation method for featurizing mutations involves breaking down a genomic sequence into smaller units, such as nucleotides, codons, amino acids, genes, or regions. Each unit is assigned a unique index in a binary vector of predetermined length. Initially, all indices in the binary vector are set to zero. For each mutation in the genomic sequence, the corresponding index in the binary vector is switched to one. This resulting binary vector effectively represents the mutation pattern of the genomic sequence. By employing the binary representation method, mutation patterns of genomic sequences can be effectively captured in binary vectors. These binary vectors can subsequently be utilized for various analyses, including similarity assessments, clustering, and classification tasks.

Challenges Associated with Binary Representation

Loss of information: Binary representation only captures the presence or absence of a mutation, failing to consider crucial details such as mutation type, frequency, consequence, or location. For example, both a synonymous mutation and a nonsense mutation are encoded as 1, despite having vastly different effects on protein function. Similarly, a common mutation and a rare or unique mutation are both represented as 1, despite their distinct implications for genetic diversity and disease susceptibility.

High dimensionality: Binary representation requires a large vector length to accommodate all possible subunits of a genomic sequence. For instance, when using nucleotides as subunits, the vector length needed to represent a sequence of length n is 4^n. If codons are used, the vector length becomes 64^n for a sequence of length 3n. This high dimensionality poses challenges in terms of memory and computational costs, as well as issues related to sparsity and noise. Lack of robustness: Binary representation is sensitive to minor changes or errors in the genomic sequence or mutation data. Even a single nucleotide insertion or deletion can completely alter the binary vector by shifting the alignment of subunits. Additionally, a missed or misidentified mutation in the data can significantly impact the binary vector. Lack of generalization: Binary representation is specific to individual genomic sequences or sets of mutations and fails to capture shared features among different sequences or mutations. For example, sequences with similar mutation patterns but different subunit compositions yield vastly different binary vectors. Similarly, mutations with similar functional effects occurring in different subunits result in disparate binary indices.

To address the limitations of binary representation, various solutions and alternatives can be explored: Incorporating more information: Instead of using a single bit to encode the presence or absence of a mutation, multiple bits can be employed to represent additional information such as mutation type, frequency, consequence, or location. This expanded representation, such as using 4 bits to encode different SNV types or 3 bits to denote SNV consequences, increases the information content and enhances the discriminative power of the binary vector. Reducing dimensionality: Rather than using all possible subunits as indices in the binary vector, a subset of relevant or informative subunits can be selected based on the specific task at hand. For instance, considering only genes or regions known to be associated with a particular disease or phenotype as

indices in the binary vector can alleviate memory, computational, sparsity, and noise concerns. Improving robustness: Instead of relying on exact matching, approximate matching or alignment methods can be employed to tolerate small changes or errors in the genomic sequence or mutation data. Measures such as edit distance or Hamming distance can be utilized to assess similarity between subunits and assign them to the nearest index in the binary vector, thereby enhancing the robustness and stability of the representation. Moving beyond specific subunits, more general features capturing common characteristics among different sequences or mutations can serve as indices in the binary vector. Molecular descriptors or fingerprints describing chemical structure and properties can be employed, or machine learning techniques like autoencoders or neural networks can learn latent features from raw data for use as indices. This approach enhances the generalization and transferability of the binary vector.

3.3.3 Featurization of mutations by PPR algorithm

What is PageRank Algorithm?

PageRank is an algorithm used by search engines to rank web pages based on their importance. It assigns a numerical score to each page in a network, representing the probability that a random surfer will visit that page. The core idea behind PageRank is that important pages are likely to receive more links from other important pages.

Let's consider a directed graph $G = (V, E)$, where V is the set of vertices (web pages) and E is the set of edges (links between pages). We represent the graph using an adjacency matrix A, where A_{ij} is 1 if there is an edge from page i to page j, and 0 otherwise.

The PageRank score of a page i is calculated iteratively using the following equation:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{PR(j)}{L(j)} \tag{3.1}$$

where $PR(i)$ is the PageRank score of page i, d is the damping factor (typically set to 0.85), N is the total number of pages, $M(i)$ is the set of pages that link to page i, and

$L(j)$ is the number of outgoing links from page j.

This equation represents a random surfer model. The surfer starts on a random page, follows links on the current page with a probability of $1 - d$, and jumps to a random page with a probability of $\frac{d}{N}$. The PageRank score of a page is the sum of the PageRank scores of the pages that link to it, weighted by the inverse of their outgoing link counts.

Random Walk with Restarts (RWR)

Random Walk with Restarts (RWR) is a variant of PageRank that introduces a restart mechanism. Instead of starting from a random page, the surfer restarts from a specified set of seed nodes with a certain probability at each step.

Let's denote the set of seed nodes as S. The RWR score of a page i after k steps is calculated as:

$$RWR(i,k) = (1-r) \sum_{j \in M(i)} \frac{RWR(j,k-1)}{L(j)} + r \sum_{s \in S} \frac{1}{|S|} \tag{3.2}$$

where $RWR(i,k)$ is the RWR score of page i after k steps, r is the restart probability, $M(i)$ is the set of pages that link to page i, $L(j)$ is the number of outgoing links from page j, and $|S|$ is the number of seed nodes.

This equation represents a random surfer who can either follow links on the current page with a probability of $1 - r$, or restart from a seed node with a probability of $\frac{r}{|S|}$. The RWR score of a page is the sum of the RWR scores of the pages that link to it, weighted by the inverse of their outgoing link counts, plus a restart term that distributes the restart probability equally among the seed nodes.

Personalized PageRank (PPR)

Personalized PageRank (PPR) extends RWR by allowing the surfer to have a personalized preference towards specific pages. It calculates the probability that the surfer will reach each page from the seed nodes, taking into account both the global importance of the page (as in PageRank) and the personalized preferences.

Let's denote the personalized preference vector as p, where p_i is the preference score

for page i. The PPR score of a page i after k steps is calculated as:

$$PPR(i,k) = (1-r) \sum_{j \in M(i)} \frac{PPR(j,k-1)}{L(j)} + r \sum_{s \in S} \frac{p_s}{|p|_1} \tag{3.3}$$

where $PPR(i,k)$ is the PPR score of page i after k steps, r is the restart probability, $M(i)$ is the set of pages that link to page i, $L(j)$ is the number of outgoing links from page j, S is the set of seed nodes, p_s is the preference score for seed node s, and $|p|_1$ is the ℓ_1 norm of the preference vector.

This equation combines the RWR score with a personalized preference term. The preference scores are normalized to sum up to 1 using the ℓ_1 norm, and the preference term is scaled by the restart probability and added to the RWR score. This allows the surfer to prefer certain pages based on the personalized preferences.

PPR can be used in various applications, such as recommendation systems, personalized search, and network analysis, to incorporate both global importance and personalized preferences in the ranking of pages.

3.3.4 PPR algorithm applied to gene interaction network

Analyzing the gene interaction network with SIFT deleterious genes as seed nodes using the PPR algorithm enables us to comprehend the impact of mutation-related deleteriousness signals on various genes within the network. Gene interaction networks are visual depictions of the functional associations between genes, encompassing protein-protein interactions, co-expression, and co-regulation. These networks provide insights into the underlying molecular mechanisms and pathways involved in complex phenomena like diseases, drug responses, and evolution. Nevertheless, gene interaction networks often suffer from incompleteness, noise, and biases due to limitations in experimental methods and data sources. Hence, computational techniques become necessary for inferring or expanding gene interaction networks by leveraging prior knowledge and additional information.

The Personalized PageRank (PPR) algorithm constitutes one such computational approach for inferring or expanding gene interaction networks. Derived from the PageRank algorithm, which assesses the importance and relevance of web pages to a specific

query, the PPR algorithm ranks genes based on their significance and relevance to a given set of seed genes. By evaluating the connectivity of genes to multiple seed genes through short paths in the gene interaction network, the PPR algorithm assigns higher scores to such genes. To better comprehend the impact of mutation-related deleteriousness signals, we can employ the PPR algorithm to analyze the gene interaction network, utilizing SIFT deleterious genes as seed nodes. SIFT deleterious genes are genes containing amino acid changes predicted to be detrimental according to the SIFT algorithm. This algorithm predicts the impact of amino acid changes on protein function by considering amino acid conservation across various species. SIFT deleterious genes may hold relevance to diseases, phenotypes, or drug responses. By utilizing the PPR algorithm to scrutinize the gene interaction network with SIFT deleterious genes as seed nodes, we can identify functionally related genes that interact with these deleterious genes. Such genes are also susceptible to the influence of mutation-related deleteriousness signals and may possess crucial roles in biological processes or pathways. Additionally, comparing PPR scores across different sets of seed nodes corresponding to distinct phenotypes or conditions allows us to explore how mutation-related deleteriousness signals propagate and impact various genes in the gene interaction network across diverse scenarios.

Example

Let's consider a gene interaction network represented as a directed graph $G = (V, E)$, where V is the set of genes and E is the set of interactions between genes. We can represent the graph using an adjacency matrix A, where A_{ij} is 1 if there is an interaction from gene i to gene j, and 0 otherwise. In the context of our gene interaction network, we can select a set of seed genes known to be deleterious according to SIFT predictions. These seed genes will serve as the personalized preference nodes for the PPR algorithm.

PPR score calculation for the gene interaction network

Let's denote the set of seed genes as S, where $|S|$ is the number of seed genes. The PPR algorithm assigns importance scores to genes based on the global network structure and personalized preferences. The PPR score of a gene i is calculated iteratively as follows:

$$PPR(i,k) = (1-r) \sum_{j \in M(i)} \frac{PPR(j,k-1)}{L(j)} + r \sum_{s \in S} \frac{p_s}{|p|_1} \tag{3.4}$$

Here, $PPR(i,k)$ represents the PPR score of gene i after k iterations, r is the restart probability, $M(i)$ is the set of genes that interact with gene i, $L(j)$ is the number of outgoing interactions from gene j, S is the set of seed genes, p_s represents the preference score for seed gene s, and $|p|_1$ is the ℓ_1 norm of the preference vector. The first term in the equation represents the contribution from the neighboring genes in the network. It calculates the PPR scores of genes j that interact with gene i in the previous iteration $(k-1)$, weights them by the inverse of the outgoing interactions of gene j ($L(j)$), and sums them up. The second term represents the personalized preference component. It calculates the preference score of each seed gene s (p_s), scales it by the restart probability r and the inverse of the ℓ_1 norm of the preference vector $|p|_1$, and sums them up. This term ensures that the personalized preferences influence the importance scores of genes. The algorithm iteratively updates the PPR scores until convergence or a specified number of iterations.

3.4 Modeling

To build accurate drug response prediction models, we need to incorporate both drug features and cell line mutation features as inputs for a machine learning model. The output of the model will be the predicted drug response, typically represented by metrics like IC50 (half-maximal inhibitory concentration). In this section, we will explain the process of modeling in detail.

3.4.1 Cell Line Mutation Features

For the cell line mutation features, we utilize Personalized PageRank (PPR) scores for all 20,000 genes in the gene interaction network. These scores quantify the importance and relevance of each gene to a set of seed genes known to contain deleterious mutations. To reduce dimensionality and noise, we focus on genes that are annotated by the OncoKB database, a curated collection of cancer-related genes and variants. This se-

lection ensures that we concentrate on genes relevant to cancer biology. By filtering the features, we can effectively capture the impact of genetic mutations on drug response prediction.

3.4.2 Drug Features

In addition to cell line mutation features, we incorporate drug features into our prediction models. We employ Mordred, a software tool that calculates molecular descriptors from drug structures. Molecular descriptors are numerical values that characterize various chemical properties and similarities of drugs. To ensure compatibility with the machine learning model, we exclude features that contain non-numerical values, such as strings or missing data. This process guarantees that the drug features are suitable for subsequent analysis and modeling.

3.4.3 Feature Matrix

To combine the drug and cell line mutation features, we construct a feature matrix. Each row of the matrix represents a drug-cell line pair, while each column corresponds to a specific feature. The resulting feature matrix consists of approximately 1,400 features and 600,000 drug-cell line pairs. By merging the two sets of features, we create a comprehensive representation of the drug-cell line interaction space. This matrix serves as the input for our subsequent machine learning algorithms.

3.4.4 Data Splitting

To assess the performance and generalizability of our models, we split the data into training and testing sets based on the cell lines. We ensure that the cell lines present in the testing set are not included in the training set. This separation guarantees that our model can effectively generalize to unseen cell lines and accurately predict drug response for new samples. The training set is used to train the models, while the testing set is employed to evaluate their performance.

3.4.5 Model Selection

We perform model selection by evaluating various machine learning models, such as linear regression, random forest, extra trees, and catboost. The evaluation is based on metrics such as the Pearson correlation coefficient, which measures the linear correlation between the predicted and actual drug response values.

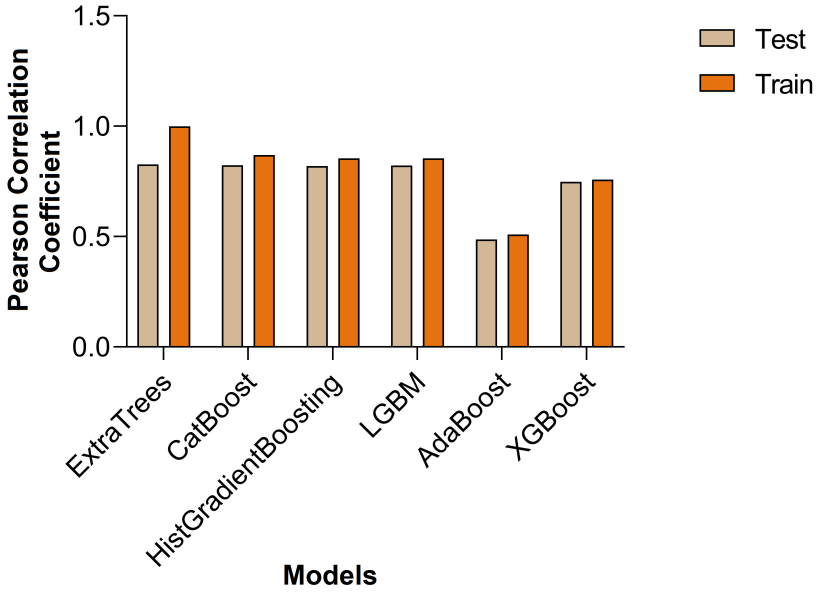

Figure 3.2: Results of Model Selection

In Figure 3.2, we present the results of model selection, showing the performance of each model. After thorough evaluation, we choose the catboost model as it demonstrates superior performance on the testing set without overfitting to the training set, unlike the extra trees model. Catboost is a gradient-boosting model capable of handling complex and high-dimensional data, making it suitable for our drug response prediction task.

3.4.6 Hyperparameter Tuning

To optimize the catboost model's performance, we conduct hyperparameter tuning. Hyperparameters are configuration settings that control the learning process of the model. We explore different hyperparameters, such as learning rate, depth, and iterations. To ensure robustness and prevent overfitting, we employ 5-fold cross-validation. This technique divides the training data into five subsets, using four subsets for training and the

remaining subset for validation. We repeat this process five times, ensuring that each subset serves as the validation set once. By evaluating the model's performance on the validation set for different hyperparameter configurations, we can select the optimal values that minimize the prediction error.

Through meticulous modeling, we can construct accurate drug response prediction models that incorporate both drug and cell line mutation features. This approach enables us to make informed predictions about drug efficacy, contributing to the advancement of personalized medicine and precision oncology.

CHAPTER 4

Results and Conclusion

4.1 Results

The study aimed to tackle the intricate nature of tumor heterogeneity and personalized responses to drug treatments in the field of oncology. We developed a predictive modeling approach, named Precily 2.0, was developed to forecast drug responses in cancer cell lines by utilizing a combination of drug and somatic mutation-based features. To construct a comprehensive dataset for each drug-cell-line pair, drug descriptors derived from SMILES representations were concatenated with mutation features obtained through the Personalized PageRank (PPR) algorithm applied on a Protein-Protein Interaction Network (PPIN). This approach provided valuable insights into gene deleteriousness propagation within the PPIN. After careful model selection and hyperparameter tuning, Precily 2.0 emerged as the most effective model for the prediction task. During evaluation, the performance of Precily 2.0 was found to be promising, generating reliable results.

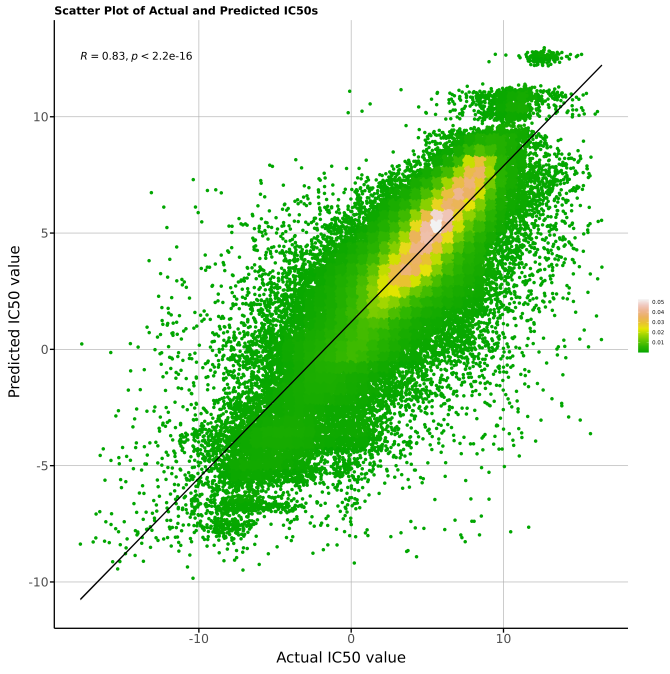

Figure 4.1: Test set performance of Precily 2.0

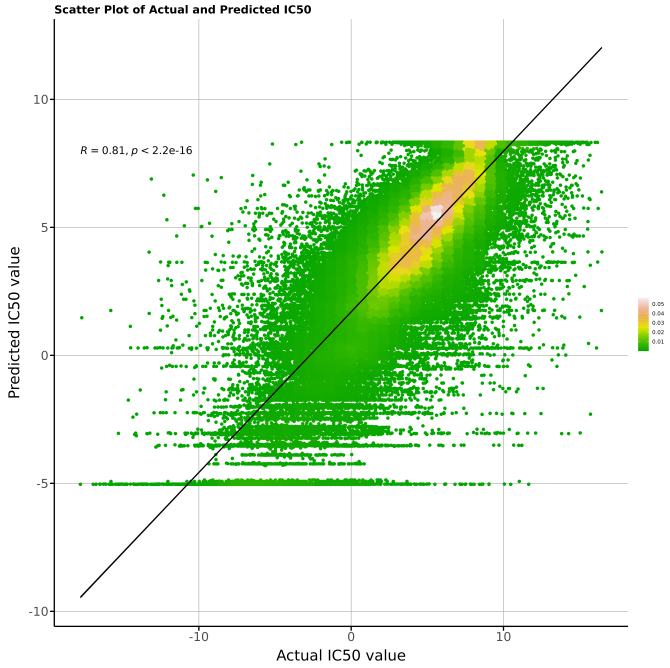

Figure 4.2: Test set performance of DrugCell

Precily 2.0 achieved a Pearson Correlation Coefficient of 0.83, surpassing other models with similar methodologies. The successful implementation of Precily 2.0 highlighted the potential of incorporating drug descriptors and network-based mutation propagation as predictive features for determining drug responses in cancer cell lines. DrugCell on the other hand trained and tested on the same data scored a Pearson Correlation Coefficient of 0.81. From the scatter plots of both, we can see a clear distinction between not just the performance of the two models but also the stability of them. It is clear that Precily 2.0 is more stable than DrugCell.

4.2 Conclusion

The evaluation of Precily 2.0 demonstrated its promising performance, as evidenced by a Pearson Correlation Coefficient of 0.83. This outperformed other models with similar methodologies like DrigCell, indicating the superiority of the developed approach. The integration of drug descriptors and network-based mutation propagation as predictive features proved effective in determining drug responses in cancer cell lines. The findings of this study have significant implications for precision oncology. Precily 2.0 can empower oncologists to make informed treatment decisions for individual patients, contributing to the advancement of personalized medicine. By incorporating multiple

data sources and advanced modeling techniques, Precily 2.0 has the potential to improve patient outcomes and enhance the effectiveness of cancer therapies. In summary, this research demonstrates the value of integrating drug descriptors and network-based mutation propagation to address the challenges of tumor heterogeneity. The results highlight the potential of Precily 2.0 in facilitating precision oncology and advancing the field towards more targeted and effective cancer treatments. DRP modeling is a problem of paramount importance in precision oncology. Modeling using chemo-genomic data, the model performance is not as good as chemo-transcriptomic data, but it overcomes the problem with the chemo-transcriptomic data-based model, which is the problem of RNA degradation in FFPE blocks, hence for translational purposes we can rely on Precily 2.0 to some extent.

4.3 Future directions

Moving forward, there are several promising avenues for future research that can expand our understanding of drug responses in cancer cell lines, particularly regarding the mechanisms of action (MOA). One intriguing direction involves leveraging multi-omics data to gain insights into these complex processes.

With the advent of advanced technologies, the generation of large-scale multi-omics datasets has become increasingly feasible. These datasets encompass a wide range of molecular information, including genomics, transcriptomics, epigenomics, proteomics, and metabolomics. Integrating and analyzing these diverse layers of data can provide a comprehensive perspective on the intricate biological mechanisms underlying drug responses. By incorporating multi-omics data into predictive modeling approaches, we can uncover valuable insights into the MOA of drugs and potentially identify relevant biomarkers associated with treatment outcomes.

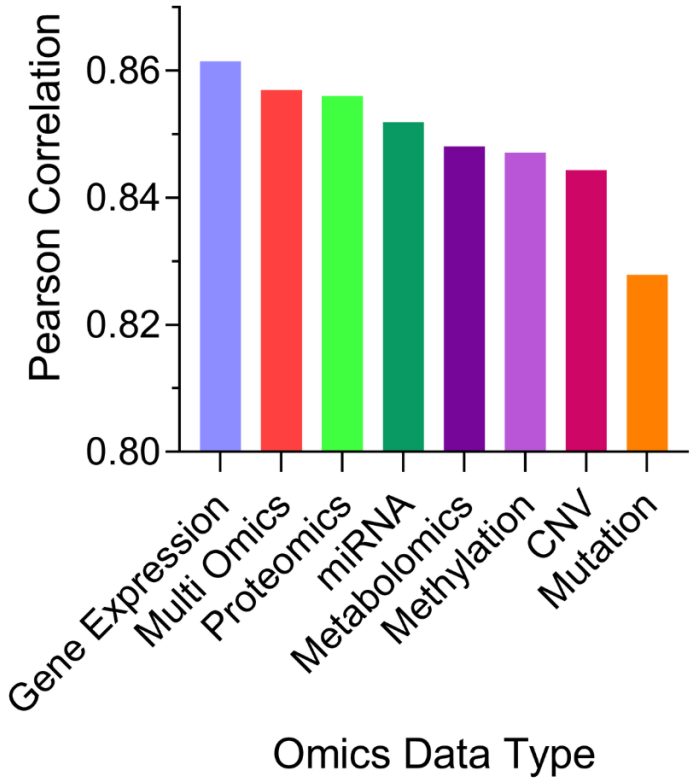

Figure 4.3: The potential of MultiOmics

The integration of multi-omics data offers an opportunity to explore the interplay between various molecular features, such as genetic alterations, gene expression patterns, epigenetic modifications, protein-protein interactions, and metabolic pathways. By examining these interactions, researchers can uncover hidden relationships and unravel the complex mechanisms driving drug efficacy. Furthermore, the analysis of multi-omics data can aid in identifying novel therapeutic targets, understanding drug resistance mechanisms, and facilitating the development of personalized treatment strategies.

Another promising aspect of incorporating multi-omics data is the potential to discover synergistic drug combinations. By examining the molecular characteristics and interactions within multi-omics datasets, researchers can identify specific features that interact synergistically or antagonistically with certain drugs. This knowledge can then be utilized to optimize treatment regimens, overcome drug resistance, and enhance patient outcomes.

To explore these future directions, collaborative efforts and the establishment of comprehensive multi-omics datasets will be essential. Additionally, the development

and utilization of advanced computational techniques, such as machine learning and network analysis, will play a crucial role in effectively integrating and analyzing the complex multi-omics data.

In summary, the integration of multi-omics data represents a promising avenue for future research in understanding drug responses in cancer cell lines. By leveraging diverse molecular information, we can deepen our understanding of the MOA, identify potential biomarkers, and enhance treatment strategies within the realm of precision oncology. These advancements have the potential to revolutionize cancer treatment by enabling personalized approaches that consider the unique molecular characteristics of individual patients.

REFERENCES

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

Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, Ma J, Ideker T. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell. 2020 Nov 9;38(5):672-684.e6. doi: 10.1016/j.ccell.2020.09.014. Epub 2020 Oct 22. PMID: 33096023; PMCID: PMC7737474.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. Journal of cheminformatics, 3(1), 1-14.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), 31-36.

Garraway, L. A., Verweij, J., Ballman, K. V. (2013). Precision oncology: an overview. J Clin Oncol, 31(15), 1803-5.

Yingtaweesittikul, H., Wu, J., Mongia, A., Peres, R., Ko, K., Nagarajan, N., Suphavilai, C. (2023). CREAMMIST: an integrative probabilistic database for cancer drug response prediction. Nucleic Acids Research, 51(D1), D1242-D1248. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.

Basu A., Bodycombe N.E., Cheah J.H., Price E.V.(2013) Cancer Therapeutics Response Portal (CTRP): linking resources to discover candidate therapeutic compounds

Yang W., Soares J., Grellier G.(2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells

Liu Yifan et al.(2021) PPR: A Pathway-Based Predictive Model for Drug Response Prediction in Cancer Cell Lines Frontiers in Genetics