ML-driven drug discovery: identifying biomarkers and mechanistic insights of a novel anti-cancer drug (NUC-7738)

Rafael Kollyfas

210017984

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

MSc Artificial Intelligence

at the University of St Andrews, School of Computer Science 16 August 2022

Supervisor: Dr. Ognjen Arandjelović

Abstract

Cancer is a leading cause of death worldwide, causing nearly one in six deaths. Resistance to therapy continues to be a primary cause of cancer treatment failures, resulting in a high number of cancer-related deaths. Recent technological advancements and the availability of a vast amount of omics data have ushered us into a new era of precision medicine, a paradigm shift away from one-size-fits-all cancer treatment toward patient-specific medicine. The main aim of this project was the in-depth statistical analysis of the NUC-7738 drug using multiple omics datasets and the development of machine learning (ML) models that can be used to predict *in vitro* whether an individual would be sensitive or resistant. Through data analysis and visualisation, tumour origins that are likely to be sensitive and resistant to the drug were identified. Extensive correlation analysis identified genes that highly correlate with drug efficacy across different omics datasets (gene expressions, alternative polyadenylation, proteomics, gene effect, mutations). Furthermore, through a range of feature selection and ML methods, predictive models were created that can be used to predict drug response given a patient's molecular characteristics. A multi-omics model that incorporates all the models' predictions was also attempted, with modest success. Clustering cancer cell lines based on their alternative polyadenylation usage yielded two clusters with varied drug sensitivity. Ensemble clustering of the gene effect dataset revealed genes with similar dependency profiles, which is especially beneficial when it comes to identifying alternative druggable targets for genes of interest that are not conventionally druggable. Finally, correlation analysis when the drug was combined with two other promising drugs was performed. This showcased that when the drug was used in combination with Paclitaxel, the HINT1 gene, which is required to activate NUC-7738, had a significantly stronger correlation with drug efficacy.

Acknowledgements

I would foremost like to thank my supervisor Dr Ognjen Arandjelović for always being very supportive and giving me useful advice throughout this project. I would also like to thank Professor David Harrison and Dr Mustafa Elshani for their biological input and willingness to help whenever I had a question. Furthermore, Justyna Orłowska was also a very important emotional support throughout the entirety of this project. Finally, I would like to thank my family without whom I would have not had the opportunity to be here.

Declaration

I declare that the material submitted for assessment is my own work except where credit is explicitly given to others by citation or acknowledgement. This work was performed during the current academic year except where otherwise stated.

The main text of this project report is 14,978 words long.

In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker, and to be made available on the World Wide Web. I retain the copyright in this work.

16th of August 2022.

Rafael Kollyfas

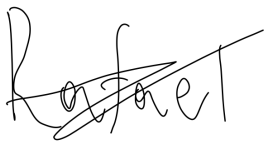


Table of Contents

Κάλλιον το προλαμβάνειν ή το θεραπεύειν, Ιπποκράτης

1 Introduction

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths [7]. The development of a new drug costs up to $2.6 billion and takes up to thirteen years, with only one drug out of every five thousand reaching the stage of market approval [5]. In recent years, machine learning (ML) approaches have been utilized to improve the drug discovery process, providing significant insights, and making the process more efficient by reducing the drug development time and costs. Whole-genome expression (transcriptomic) analysis has become an important tool to identify relevant gene pathways and biomarkers that are deregulated and drive abnormal cellular proliferation and metastatic spread. ML techniques applied to transcriptomic analysis can elucidate the molecular pathways involved in drug efficacy and provide ways to identify patient subgroups who are susceptible or receptive to specific drugs [6]. Furthermore, the integration of numerous other omics datasets (e.g., proteomics, mutations) can be used to obtain a more holistic overview of drug sensitivity biomarkers from different perspectives and further improve cancer prognostics and treatment. These recently emerged possibilities bring us closer to precision medicine, a paradigm shift away from one-size-fits-all cancer treatment and toward patient-specific medicine.

1.1 Aim and Objectives

This project aims to explore multiple omics datasets, leading to the development of a ML pipeline that can be used to predict *in vitro* whether an individual would be sensitive or resistant to the novel NUC-7738 drug.

Primary objectives:

1. Data analysis and visualisation for the cancer cell lines across the three time points to identify tumour origins that appear more sensitive or resistant to the drug. Evaluate what tumour origins might be good targets for the drug.
2. Perform correlation analysis between transcriptome (mRNA) expression data and NUC-7738 sensitivity (*GI50, EC50, AUC, IC50*). Identify gene biomarkers that highly correlate with sensitivity to the drug.
3. Explore clustering of cancer cell lines based on their Alternative Polyadenylation (APA) usage. Examine relationships between the clusters and sensitivity/resistance to the drug.

The primary objectives are critical to the project's success and hence take precedence. The secondary objectives are also important and would considerably enhance the project's outcome; however, given the limited time, their accomplishment is less vital and may be regarded more as prospective extensions that would make the project more intriguing.

Secondary objectives:

1. Gain further biomarker insights by exploring the cancer cell lines' relationship with additional omics datasets.
2. Experiment with the development of a multi-omics ML model that can predict drug sensitivity by combining biomarkers and mechanistic insights identified.
3. Investigate the drug's efficacy and variations in biomarkers detected when the drug is administered in combination with other drugs.

1.2 Structure

The background chapter discusses fundamental concepts and literature. The following chapter describes the techniques' implementation and methodology in detail, while the final chapter evaluates the developed models, discusses the project's success, and briefly discusses potential future work.

2. Background

This chapter provides an overview of traditional cancer treatments, the emergent potential of ML to revolutionize anti-cancer drug therapy, and fundamental background about the NUC-7738 drug. It also explains how ML is used to identify gene biomarkers, why alternative polyadenylation is significant, and how several omics can be integrated to provide a more informed prognosis. Finally, the benefit of combining multiple drugs is briefly discussed.

2.1 Cancer traditional treatments and a new approach

Human cells normally develop and multiply through cell division whenever the body requires them. When cells get old or damaged, they die, and new cells replace them. When this mechanism fails, abnormal or damaged cells grow and reproduce. This may lead to the creation of tumours which are tissue masses [1]. Tumours are divided into two categories: benign, which do not spread to other parts of the body, and cancerous, which are invasive [2]. Cancerous tumours can invade nearby tissues and travel via the blood or lymph system to establish a new tumour in other organs through metastasis [3]. Common cancer treatments include surgery, radiotherapy, and chemotherapy.

Cancer surgery involves cancer removal from the patient's body and is most effective for solid confined tumours. The risks of surgery include infection, bleeding, and damage to nearby tissues [8]. Radiotherapy involves high doses of radiation which to kill cancer cells and shrink tumours. At high doses, it kills or slows the growth of cancer cells by damaging their DNA, causing them to stop dividing or die. Because radiotherapy does not instantly kill cancer cells, this process can take days or weeks [9]. However, radiotherapy can also damage nearby healthy cells, resulting in side effects (e.g., fatigue) which vary depending on the treated part of the body [10]. Chemotherapy is the most common treatment that uses drugs to kill cancer cells by stopping their reproduction and preventing them from growing and spreading [4]. Chemotherapy drugs are discovered through drug discovery methods such as phenotypic and reverse pharmacology, which are often expensive, time-consuming, and complex.

Resistance to therapy continues to be a primary cause of cancer treatment failures, resulting in a high number of cancer-related deaths. The treatment strategy is primarily determined by cancer subtypes and the presence of genetic mutations. However, the existence of a genetic mutation does not always predict therapy response and can vary between cancer subtypes [11]. Recent technological advancements have ushered us into a new era of precision medicine by combining ML and biological research to analyse diseases and extract relevant insights from large amounts of data (multi-omics), improving treatment outcomes. Precision medicine's goal is to provide treatments that raise patients' chances of survival and improve their quality of life by decreasing undesired side effects. This can be accomplished by matching patients with suitable therapies or therapeutic combinations [11].

2.2 NUC-7738

Cordycepin, also known as 3'-deoxyadenosine (3'-dA), is a natural nucleoside analogue and a bioactive component of the fungus *cordyceps sinensi*s, making it a traditional treatment for inflammatory diseases and cancer that has been used in traditional Chinese medicine for hundreds of years [42]. While Cordycepin has great anticancer potential *in vitro*, when administered (*in vivo*), only a small amount of cancer-destroying drug is delivered to the tumour. Cordycepin must be transported into cancer cells by a nucleoside transporter (hENT1), converted to the active anti-cancer metabolite 3'-dATP by a phosphorylating enzyme (ADK), and broken down in the bloodstream by ADA enzyme. These resistance mechanisms linked to transportation, activation, and breakdown result in insufficient anti-cancer metabolite delivery to the tumour [42,73].

ProTides convert a nucleoside prodrug into a phosphorylated nucleotide attached to a phosphonamidite moiety, which makes the molecular more lipophilic and capable of entering cells. The active nucleotide is released during the hydrolysis of the phosphonamidite. NuCana has utilised this technology to design NUC-7738 which is a ProTide transformation of 3'-deoxyadenosine (3'-dA) that activates the anti-cancer metabolite of 3'-dA directly inside cancer cells, thereby bypassing 3'-dA's fundamental limitations of transportation, activation, and breakdown. NUC-7738 has up to 40 times greater potency for killing cancer cells than its parent compound, with limited toxic side effects [42,74].

Pharmacokinetics and tumour samples obtained from the first-in-human phase I clinical trial of the drug evidenced that it is an effective proapoptotic agent in cancer cells with effects on the NF-kB pathway [73]. In the trial, NUC-7738 was given to patients with advanced solid tumours that were resistant to conventional treatment and results showed that it was well-tolerated and has encouraging signs of anti-cancer activity [42].

2.3 Biomarkers

Biomarkers include biological molecules found in blood, other bodily fluids, or tissues that can be used to detect a normal or abnormal process, a condition, or disease [12]. They can be important in establishing treatment methods since their discovery can aid in early cancer prediction and identify patient subgroups who are susceptible or receptive to specific drugs. Using predictive biomarkers to choose appropriate individuals for certain treatments can significantly boost therapeutic efficacy and reduce toxicities, making it a critical process [6]. Most of the research on molecular biomarkers of cancer drug efficacy has been done in pharmacogenomics, which studies genome-level changes as potential biomarkers of drug response [14]. *In vitro* studies, however, show that gene expression variation accounts for even more diversity in drug sensitivity than genomic changes and may provide more insight into clinical therapeutic effectiveness [15]. Recently, genome expression (transcriptomic) analysis has been explored using a variety of ML and DL techniques that can elucidate the molecular pathways involved in drug efficacy and provide potential ways to predict new patients' responses to available therapies. Many of these techniques can account for gene interaction and

joint predictive power, identifying genes that are weak biomarkers individually but have a high joint predictive power. A set of high-quality biomarkers should give a comparable or better prediction performance than the full list of genes since non-related genes can reduce the predictive ability of a classifier.

A popular way of identifying gene biomarkers is through traditional ML feature selection (FS) methods. Pineda et al. [16] used a naïve Bayes classifier using gene expression and DNA methylation data that classified lung adenocarcinoma and lung squamous cell carcinoma. They used ReliefF as their feature selector algorithm to identify the top 30 relevant variables, which were then used to build the classification model. The relevance of the selected genes was then examined using hierarchical clustering and ingenuity pathway analysis (IPA) for gene functional analysis. They discovered 19 genes, four of which (AKR1B10, AQP10, CXCR2, TP73) were linked to specific lung cancer subtypes. Their model had a great classification performance with an 0.89 area under the curve (AUC) score. However, this study did not use another dataset to validate their model's performance. Tarek et al. [17] took a different approach, developing an ensemble model for classifying leukaemia, colon, and breast cancer. Their proposed system utilised three FS methods: singular value decomposition (SVD) entropy, extreme value distribution (EVD), and backward elimination Hilbert-Schmidt independence criterion (BAHSIC). Their ensemble system consists of five KNN (K=3) base classifiers (using majority voting), where each classifier utilises its feature selection parameters to ensure the diversity of the ensemble. Their classification accuracy was 0.92 for leukaemia, 0.80 for colon cancer, and 0.91 for breast cancer. While their method is attractive, their available sample number[1] far outnumbers our own, rendering it challenging to achieve comparable results using a similar approach. Kathad et al. [13] predicted drug sensitivity for the LP-184 compound using a comparable number of samples to us (59 samples). They used multiple FS layers followed by a single XGBoost regression model to derive a 16-gene biomarker signature. The first layer involves Pearson correlation analysis between the genes and the *IC50*, where only the significantly correlated (p<0.05) genes were kept. The subset of genes was further reduced by keeping only biologically relevant genes through pathway enrichment analysis. The Relief algorithm was then used to rank and assign weights to the remaining genes based on drug sensitivity (*IC50*), selecting the top 100 scoring genes. The Boruta[2] feature selection algorithm was then used which iteratively removes features that are statistically less relevant than artificial noise variables introduced. The final step in the process was refining the biomarker genes using a backward sequential feature selector which was trained using XGBoost. They achieved 86% accuracy at the four-fold cut-off, and a strong and significant (r=0.70, p=1.36e06) correlation between actual and predicted LP184 sensitivity. This method is more relevant for our test as we do not classify against different cancer subtypes but predict drug efficacy. A limitation of their approach is the removal of genes that are not linearly correlated (Pearson) with drug sensitivity despite Pearson's correlation being

[1] Their dataset had 6,500 colon, 24,481 breast and 3,571 leukaemia samples. Our dataset contains 95 drug samples.
[2] Used with random forest.

significantly influenced by outliers and gene expression data being heavily skewed [32]. This conceivably led to the un-identification of important biomarker genes.

Deep learning (DL) techniques have also been applied to gene expression data. For example, Danaee et al. [18] used a stacked denoising autoencoder (SDAE) to transform high-dimensional gene expression data into a lower-dimensional, meaningful representation from which functional features could be extracted. The extracted representation was then fed into a support vector machines (SVM) with an RBF kernel, which achieved high accuracy (0.98), and F1-measure (0.983) on the test set. To identify biomarker genes with a strong influence, the SDAE weight matrices of each layer were multiplied to obtain the estimated weights for each gene at the input layer. Then, by fitting with the normal distribution, the top genes were selected. A limitation is that they match the importance of genes to their high-level features. Lyu et al. [14] instead of matching the importance of genes to their high-level features, matched them with their contribution to classification. Their method entailed embedding high-dimensional gene expression data into 2D images, which were then fed into a convolutional neural network (CNN) to classify 33 tumour types. They then generated a significance heatmap using the guided Grad-CAM technique for each tumour type, containing all the genes. By using functional analysis on the genes with high intensities in the heatmap, they validated that these top genes are related to tumour-specific pathways, and some have already been used as biomarkers. They also achieved excellent results having 0.956 accuracy, 0.955 precision, 0.956 recall, and 0.954 f1-score. DL techniques would be difficult to successfully apply in this project as there is a great disparity in data availability between our project and the ones incorporating DL[3]. Furthermore, our goal is to find biomarkers for a drug instead of identifying biomarkers for each cancer type.

2.4 Alternative polyadenylation

Alternative polyadenylation (APA) is a key post-transcriptional regulation mechanism that processes RNA products based on their 3′-untranslated region (3′-UTR) unique sequence signal. Approximately 70% of known human genes have several polyA sites, which produce varying lengths of 3′ untranslated regions (3′ UTR), leading to transcriptomic variation [48]. Furthermore, widespread 3′ UTR shortening has been observed in various forms of cancer [43,44] including cancer cells [45]. Through disruption of the competing endogenous RNA network [46], this shortening activates oncogenes [47] or represses tumour-suppressor genes in trans, promoting carcinogenesis. In human cancers, APA events are critical in oncogenic gene expression, chemotherapy resistance, and tumour microenvironment, and potentially serve as prognostic biomarkers and therefore predict clinical outcomes of human cancers [18].

Zhong et al. [22] used the distal polyA site usage index (PDUI) value to indicate the frequency of APA events. They developed a new metric called APA score by first determining the differential APA events between APA RNA processing patterns. Using the Boruta algorithm

[3] Lyu et al [14] had 10267 tumour samples.

differential APA events with adjusted $p<0.05$ were extracted for FS. The PDUI values of the selected APA events were then selected to perform principal component analysis (PCA). The first two principal components (PC1 and PC2) were extracted and used to calculate the APA score by summing the addition of PC1 with PC2 for all the APA events. By applying the non-negative matrix factorization (NMF) algorithm they clustered cancer patients into two groups based on the top 1,000 APA events with the most variance. Analysis of the two patterns revealed that high APA scores were correlated with undesirable survival outcomes, relatively high response to immunotherapy and low sensitivity to anti-cancer targeted drugs. A limitation of their approach is that NMF cannot capture non-linear interactions between APA events. This may present an opportunity to investigate APA clustering using linear and non-linear dimensionality reduction algorithms.

2.5 Multi-omics

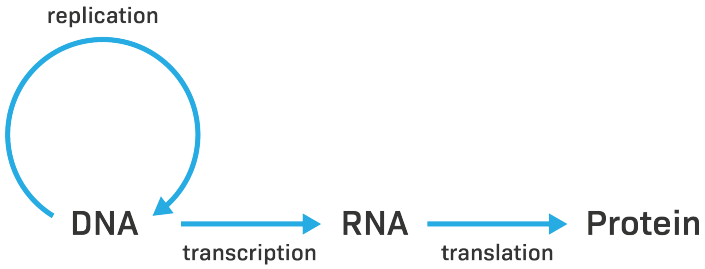


Figure 1: The central dogma of molecular biology.

The central dogma (Figure 1) of molecular biology states that genetic information flows from DNA (deoxyribonucleic acid) to mRNA (messenger ribonucleic acid) to protein [23]. The study of DNA, mRNA, and proteins are known as genomics, transcriptomics, and proteomics respectively. The genetic blueprint of a cell is investigated using genomics, examining an individual's DNA which allows us to investigate the presence of specific genes. Transcriptomics examines the genes that are actively expressed and studies the transcribed genetic material to provide information about what is happening at the cellular level. Proteomics helps in characterizing information flow within cells and organisms through protein pathways and their networks [24]. When multiple omics are used in conjunction, they are known as multi-omics. They can provide information on biomolecules from multiple layers, which is promising for systematically understanding complex biology. [25]. Single omics data can be combined using integrated techniques to study the interaction of molecules [26]. They can aid in analysing the flow of information from one omics level to the next, therefore bridging the genotype-to-phenotype gap. Integrative techniques, by their ability to analyse biological phenomena holistically, can increase disease prognostics and prediction accuracy, and hence help improve treatment and prevention [25-27]. Furthermore, studies have shown that merging omics data sets produces a better knowledge and clearer image of the system under research when compared to single omics data.

Combining different datasets and mining biologically meaningful biomarkers from multi-omics data is a challenging task [27]. The use of ML and DL in data has brought new possibilities for understanding and analysing biological systems' underlying characteristics and complexities, but their integration still presents some unique challenges. Multi-omics datasets are typically heterogeneous as different normalisation and scaling techniques are frequently used for different types of data (e.g., transcriptomics and proteomics), resulting in varying dynamic ranges and data distribution [29]. Furthermore, some omics data (e.g., metabolomics) are more likely to contain null values [30]. Therefore, imputation, outlier detection, and scaling should be considered separately for each dataset before integrating them.

Two popular approaches for analysing and integrating multi-omics data are concatenation-based and model-based. Concatenation-based integration considers developing a model using a joint data matrix and combining multiple omics datasets in a single large matrix. Once all individual omics have been concatenated, the key advantage of using concatenation-based approaches is the ease with which ML may be used to analyse data. These approaches employ all concatenated features equally and identify the most discriminating features for a given phenotype [24]. Fujita et al. [33] employed an unsupervised concatenation-based method that used joint non-negative matrix factorization[4] (JNMF) to discover common clusters (co-modules) from different multi-dimensional omics datasets (RNA expression, microRNA expression, DNA methylation data). They then integrated pathway gene signature analysis through IPA to understand the association of various types of molecules, provide causal networks based on biological relationships curated from the literature, and enable inferences on pathway activation and dependencies. This technique enabled them to identify novel rationale-based and clinically validated biomarkers. One limitation is that JNMF is computationally slow and needs large memory allocation. Furthermore, more modern approaches (e.g., DL models) might be better suited for integrating the data in a matrix. This was explored by Chaudhary et al. [37] where an autoencoder (AE) was used to integrate heterogeneous data from RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), and methylation data to identify robust survival subgroups of hepatocellular carcinoma (HCC). They concluded that their autoencoder transformation tends to aggregate genes sharing similar pathways, making it appealing to interpret biological functions. Their model provides two subgroups of patients with significant survival differences and good model fitness.

Model-based integration methods develop intermediate models for each of the omics data and then construct a final model from the various intermediate models that combines them. Multiple model-based methods were developed for multi-omics data like hierarchical classifiers from Haghighi et al. [34], an ensemble-based approach using XGBoost from Ma et al. [28], and KNN from Shen and Chou [35]. A recent paper published by Xu et al. [36] developed an effective DL method for integrating multi-omics data to classify cancer subtypes. Their method integrated gene expression, miRNA expression and DNA methylation data with stacked autoencoders (SAE) for each of the omics data to learn their representation. All learned representations were then integrated into an autoencoder to learn complex data representations

[4] JNMF is equivalent to a standard NMF with concatenated inputs.

from the three SAEs. The learned complex representation was used as input to a deep flexible neural forest (DFNForest) which classifies the cancer subtypes (Figure 2). Their method facilitates the integration of different omics data and achieves excellent classification performance having an accuracy, precision, recall, and f1-score of more than 0.8 for multi-class classification.

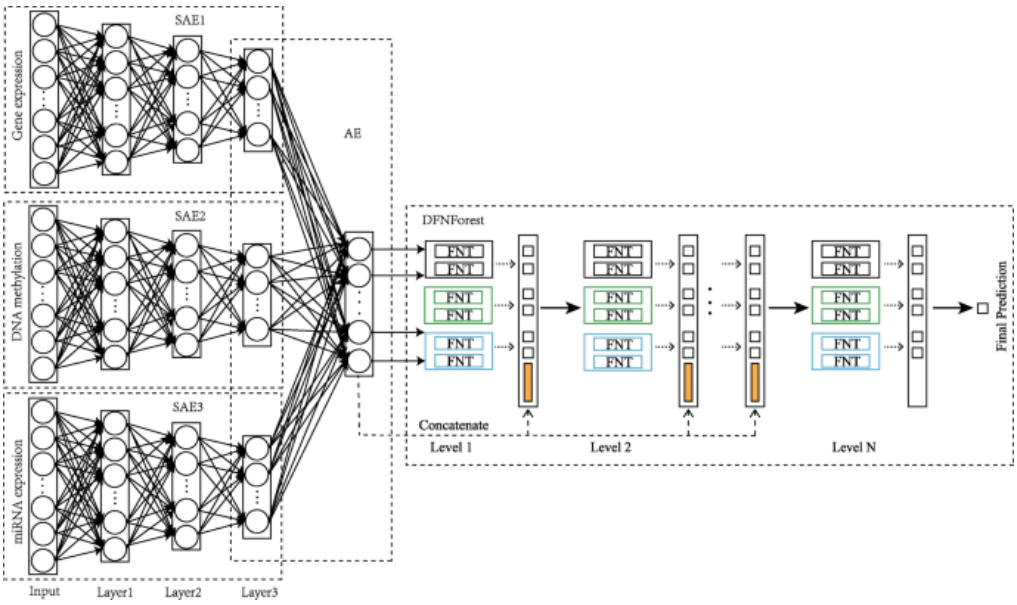


Figure 2: The framework used by Xu et al. [36].

The application of complex DL techniques would again be extremely challenging and almost certainly unsuccessful due to the extremely limited number of samples available, which would impair the ability of complex models to learn effectively. Therefore, the primary objective for the multi-omics aspect of this project is the in-depth exploration of each dataset and identification of potential biomarkers.

2.6 Drugs combinations

Most anti-cancer drug treatments use combinations rather than single agents to achieve a greater spectrum of activity against cancer cells and to minimise side effects if the drugs used in combination have different toxicity profiles. The drug combinations can be classified as synergistic, additive, or antagonistic. When the combined effect[5] of two chemicals exceeds that predicted by their single-agent potencies, they are considered synergistic. If the effect of each drug does not change the aggregate of individual drug effects, it is referred to as additive. In contrast to the synergistic effect, if the aggregate is smaller than the response of individual agents, the combination has an antagonistic effect [31]. Due to the scarcity of cell line availability for drug combinations and time limitations this dissertation only explores gene expression correlation analysis when NUC-7738 is used in combination with Paclitaxel and Erlotinib.

[5] The desired effect in our case is cancer inhibition.

2.7 Summary

The background chapter provided context about cancer, its traditional treatments and current challenges which could be addressed by precision medicine. Section 2.2 introduced the NUC-7738 drug, its mechanistic insights, and its current advancements in human trials. The importance of biomarkers and modern computational approaches which are used were then discussed in section 2.3, followed by a brief discussion on APA and its key role in cancer treatment. The benefit of multi-omics was then discussed, followed by a brief section about the importance of drug combinations.

3. Methodology & implementation

This chapter describes the implementation of all the project's components. An overview of the various datasets and setup used is provided in section 3.1, followed by data analysis of the NUC-7738 drug dataset. Section 3.3 describes the methods developed for identifying gene biomarkers from gene expression data and then clustering cell lines based on their APA events are described. Section 3.5 describes the exploration of additional DepMap datasets and their combination to a model that makes a weighted average prediction considering all the omics datasets.

3.1 Setup

Table 1 presents the datasets used, along with a brief explanation for each.

Dataset	Access	Explanation	Genes	Cell lines	Reference:
NUC-7738	Private	NUC-7738 drug's information; Contains different drug efficacy's metrics for cell lines on different treatment hours (72, 120, 168)	-	95	Provided by NuCana
Gene expressions	Public	Gene expression values	19,161	1,406	DepMap Custom Downloads
Mutations		Information about gene mutations (e.g., variant, somatic)	18,784	1,771	
Gene effect		Gene's significance towards the survival of a cancer cell line	17,386	1,086	
Proteomics		Protein expression values	15,577	365	
APA datasets		Multiple APA datasets with APA event values for genes from various cell lines. Each dataset contains APA events for a particular type of cancer	10,518	735	Zhong et al. study

Table 1: The datasets used.

Several libraries were used, with the most important ones highlighted in Table 2.

Library	Purpose
Pandas	Load datasets, manipulate them, and analysis
NumPy	Numerical and mathematical operations on multi-dimensional matrices

Scikit-learn	Split datasets into training and validation sets, load, and train ML models
Seaborn	Visual analysis and exploration
Mlxtend	Data science and ML algorithms

Table 2: Most important libraries used and their purpose.

The project was mainly written in Python and Jupyter Notebooks, except for a short script for the mutation dataset (section 3.5.1.4). The purpose of each notebook is explained in Appendix A.

3.2 Data exploration and visualisation

This section investigates the distribution of *IC50* at different treatment hours, how the drug's efficacy is affected, and the identification of tumour origins that are more sensitive or resistant to the drug. Appendix B contains additional analysis that delves deeper into the dataset's metrics.

3.2.1 IC50 distribution

IC50 is the most used metric to assess drug efficacy. It measures the concentration of a drug required to inhibit a given biological process or component by 50% in vitro [40].

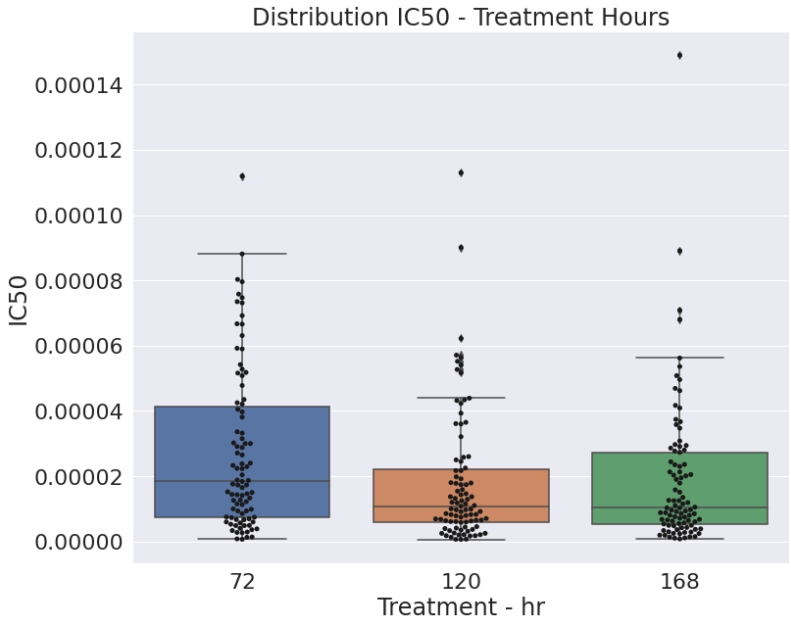


Figure 3: Boxplot of IC50 at different treatment hours.

Figure 3 depicts the dispersion of the *IC50* distribution of the cell lines over the three treatment hours. The box plots show that the median value is the highest at 72 hours while being nearly identical at 120 and 168 hours. Furthermore, the maximum value (top whisker) is at 72 hours, followed by 168 hours, and 120 hours. The over-imposed swarm plot shows that at 120 and 168 hours the distribution is skewed between the minimum point (bottom whisker) and the first

quartile (Q1), while at 72 hours the distribution is more symmetric. The 120 hours had the most outlier points, which were labelled as such by being outside (above or below) 1.5 times the interquartile range (IQR).

$$Q1 - 1.5(IQR) \; or \; Q3 + 1.5(IQR) \qquad (1)$$

When the *IC50* is measured after 120 hours, more cell lines are sensitive to the drug as *IC50* distribution dispersity is lower. At 72 treatment hours, the cell lines seem the most resistant to the drug since their box plot has the greatest dispersion and the most cell lines with higher *IC50* values. Interestingly, at 168 treatment hours, the cell lines appear more resistant than at 120 hours since more cell lines have higher values, as evidenced by the third quartile and the swarm plot.

Based on the above analysis we can conclude that drug efficacy varies considerably between the three treatment hours and therefore each must be analysed separately.

3.2.2 Cell Lines Sensitivity and Resistance

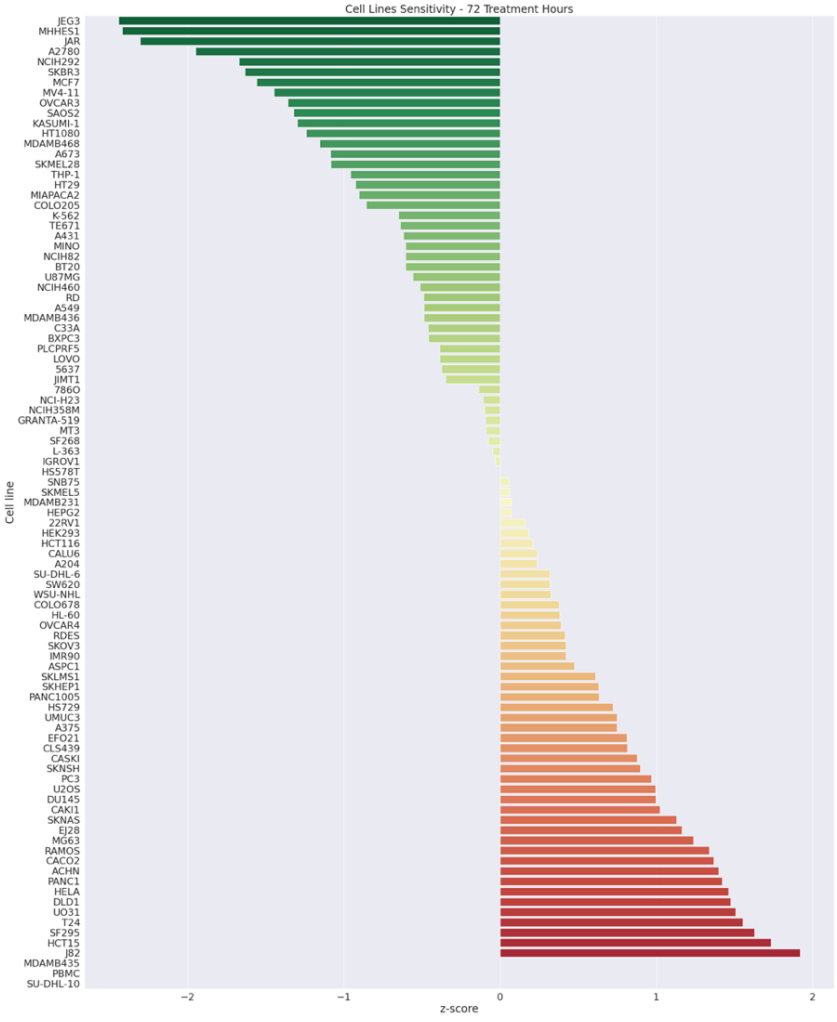


Figure 4: Cell lines sorted by their z-score (72 Treatment Hours).

Figure 4 shows the cell lines sorted by their sensitivity based on the *z-score* at 72 treatment hours[6]. The *z-score* was calculated by NuCana by applying the standard score equation (2) to the cell lines concentration (μM)[7]. It can be used to determine how many standard deviations a given point lies from the mean. Lower *z-score* values indicate more sensitive cell lines, while higher z-score values indicate more resistant ones.

$$Z = \frac{x-\mu}{\sigma} \qquad (2)$$

Table 3 shows the five most sensitive and resistant cell lines at the three treatment hours based on their z-score.

Treatment Hours	5 Most Sensitive Cell Lines	5 Most Resistant Cell Lines
72	JEG3, MHHES1, JAR, A2780, NCIH292	J82, HCT15, SF295, T24, UO31
120	MHHES1, JAR, JEG3, MCF7, SKBR3	UO31, HCT15, J82, DLD1, SU-DHL-10
168	MHHES1, MDAMB468, JEG3, BT20, OVCAR3	UO31, DLD1, ACHN, SF295, T24

Table 3: The 5 most sensitive and resistant cell lines at the three different treatment hours.

Three of the five most sensitive (MHHES1, JAR, JEG3) and resistant (UO31, HCT15, J82) cell lines are common between 72 and 120 hours. At 168 treatment hours, the MHHES1 and JEG3 cell lines are also among the top five most sensitive cell lines, indicating their particular sensitivity to NUC-7738. The UO31 cell line is among the top five most resistant cell lines at all three treatment hours. Interestingly, at 72 and 168 treatment hours, two cell lines (SF295, T24) are present in the top five resistant cell lines, but not at 120 hours. Finally, the DLD1 resistant cell line is common between 120 and 168 treatment hours.

[6] The sensitivity plots for 120 and 168 treatment hours are available in Appendix C.

[7] The three last cell lines in Figure 7 (MDAMB435, PBMC, and SU-DHL-10) do not have a *z-score* value since their concentration value was not successfully read at 72 hours.

3.2.3 Drug's efficacy on tumour origins

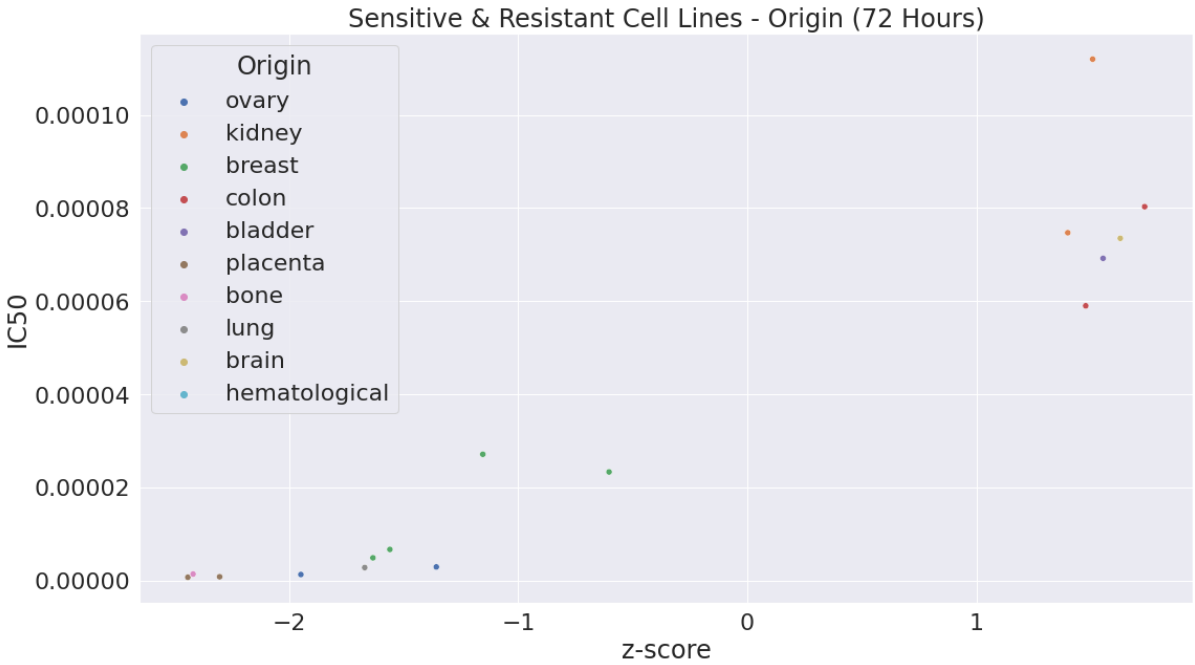


Figure 5: Unique sensitive and resistant cell lines coloured based on their origin.

The cell lines in Table 3 are depicted in Figure 5 and are coloured based on their tissue origin. Noticeably, only the kidney tissue origin appears in both the top sensitive (bottom left) and top resistant (top right) cell lines. The most sensitive cell lines are from the *placenta*, *kidney*, *ovary*, *lung*, and *breast* origins, whilst the most resistant cell lines are from the *colon*, *kidney*, *bladder*, and *brain* origins. In the sensitive group, four out of ten cell lines have a breast tissue origin whilst in the resistant group, four out of six come from *kidney* and *colon* origins.

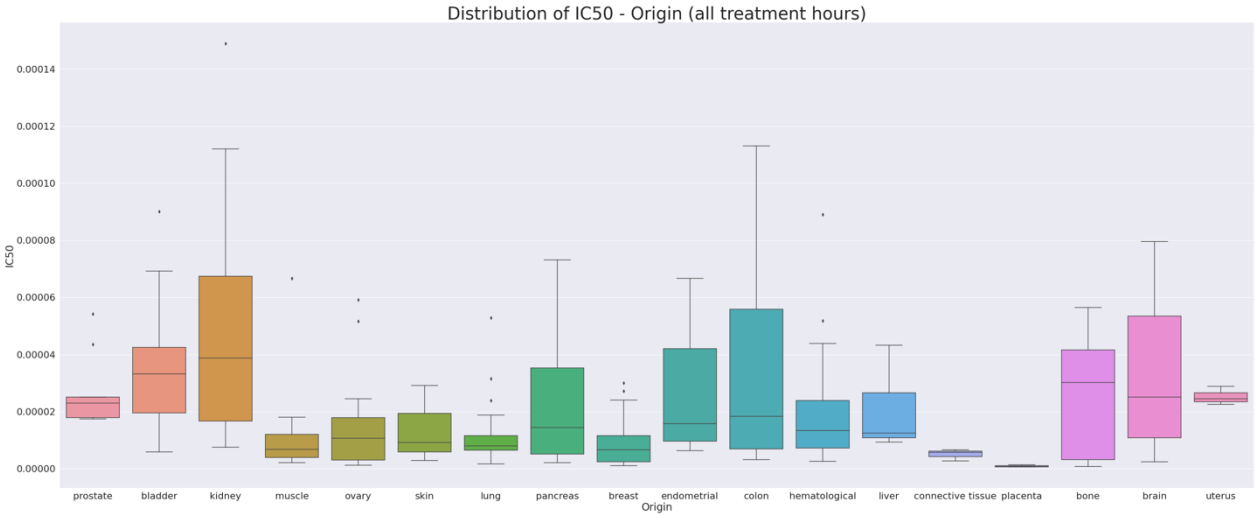


Figure 6: IC50 distribution across different origins.

Figure 6 shows the boxplots of the *IC50* distribution across the different cancer origins. The origins (*placenta*, *ovary*, *lung*, *breast*) identified as sensitive from Figure 5 have relatively low distributions. The *kidney* origin sample that appeared in the sensitive cell lines appears an

outlier. Furthermore, from the plot, we can derive that *muscle, skin,* and *connective tissue* are also sensitive origins. The resistant origins (*colon, kidney, bladder, brain*) from Figure 8 have higher median and *IC50* values. Finally, cell lines originating from *bone* are also resistant to the drug.

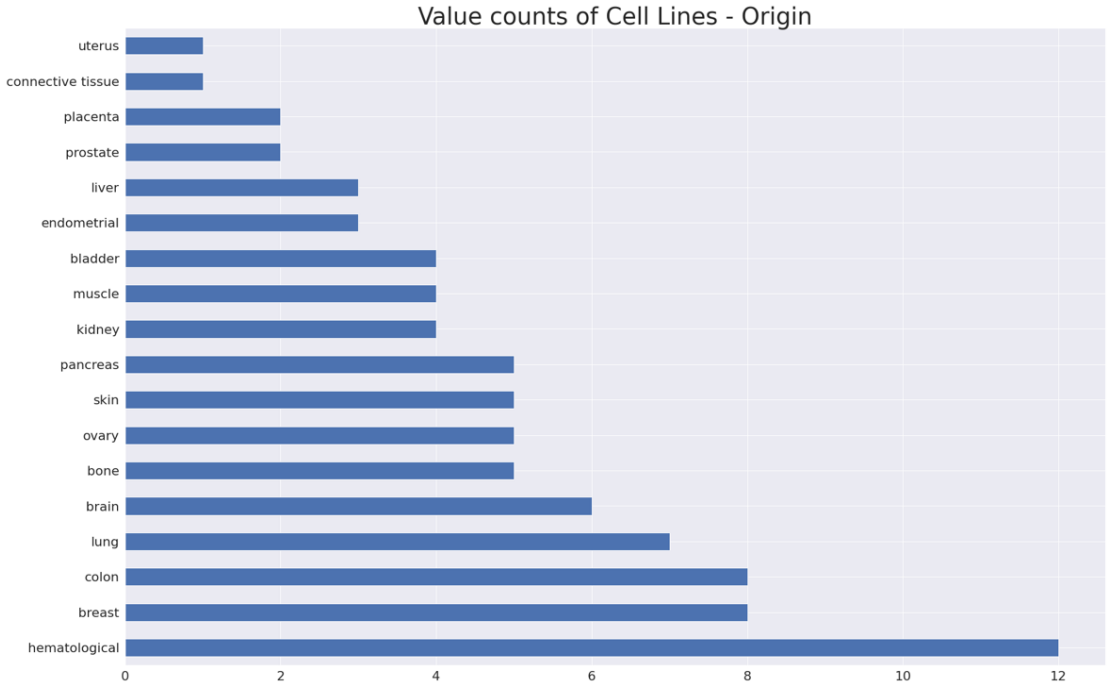


Figure 7: The value counts of cell line origins.

When examining the origins distributions, we must also consider the number of cell lines available for each origin. Figure 7 shows that the *uterus* and *connective tissue* origins have only one cell line sample and therefore their distribution in Figure 9 is simply the difference of one cell line's *IC50* value at the time points. Conclusions about the origin's sensitivity based on such few cell line samples are not representative. Unfortunately, this is a limitation of the dataset which cannot be fixed (e.g., by generating artificial data), since the reliability of medical data is critical. However, inferences can still be drawn from Figure 10. For example, there are the same number of cell lines originating from the *colon* and *breast* but the median and distribution of the *breast* is considerably lower, implying that the probability of the drug being more effective for *breast* is substantial.

Considering the aforementioned, it was decided to only comment about the origins with at least five cell line samples. It emerges that the drug is likely to be more sensitive to cell lines originating from *breast* and *lung*, and more resistant to cell lines originating from *bladder, brain, kidney,* and *bone.* Furthermore, *haematological* originated cell lines seem neither resistant nor sensitive to the drug, despite the relatively large number of samples for that origin.

3.3 Identification of biomarkers - gene expressions

This section discusses the ML approaches used to identify gene expression correlations and biomarkers with a significant impact on NUC-7738 efficacy. The methods were applied for the 72 treatment hours samples, but the same procedure could be used for the additional treatment hours.

Gene expression biomarkers can be composed of a single gene product, or a combination known as a gene expression signature [41]. To incorporate gene expression biomarker identification into our process, DepMap's gene expression dataset was paired with our NUC-7738 dataset based on their cell line ID (*DepMap_ID*). Ten of our NUC-7338 cell lines were not included in the gene expression dataset.

Section 3.1.1 showcases the correlation analysis performed to discover the top correlated genes with *IC50*. Section 3.1.2 highlights the biomarkers obtained and their predictive effectiveness when the approach proposed by Kathad et al. [13] is followed, whereas section 3.1.3 suggests potential alterations in the process.

3.3.1 Correlation Analysis

The correlations between gene expressions and *IC50* were calculated using the Pearson, Spearman, and maximal information coefficient (MIC) techniques. Identifying correlations is crucial as it can help derive mechanistic insights of the drug, help the evaluation of biomarkers, and provide target leads for drug optimisation.

Pearson correlation assigns a value between -1 and 1, where 0 is no correlation, 1 is a total positive[8] correlation, and -1 is a total negative[9] correlation. A higher correlation coefficient indicates a stronger relationship between variables. The formula to calculate the correlation coefficient is given at (3), where x_i is the value of the x-variable in a sample, $\overline{x}$ is the mean of the values of the x variable, y_i is the value of the y-variable in a sample, and $\overline{y}$ is the mean of the values of y.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \qquad (3)$$

Spearman correlation evaluates how effectively an arbitrary monotonic function can explain a relationship between two variables without making any assumptions about the variables' frequency distribution or their linearity [97]. Spearman coefficient is thus able to perfectly (ρ=1) capture relationships like the exponential function, which increases monotonically in the

[8] Both decrease or increase linearly together.
[9] Values go in opposite directions.

same direction but not at a constant rate. The formula[10] is given at (4), where n is the number of observations and d_i is the difference between two ranks of each observation.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \qquad (4)$$

Maximal information coefficient (MIC) assigns a value between 0 and 1, where 0 means statistical independence and 1 a completely noiseless relationship. It does not discern whether the relationship is positive or negative, but rather its strength. It captures both linear and non-linear relationships and is robust to outliers due to its mutual information foundation. In MIC's seminal paper [96] it was proven that the technique works well for gene expression datasets and as such was chosen for experimentation.

Only the statistically significantly correlated genes (p<0.05) were kept, where *p-value* is the probability of obtaining the same results given no correlation among the variables [38]. Before calculating the correlations, columns with over 20% zero values were excluded. Furthermore, three cell lines were dropped since their *IC50* reading was missing.

Table 4 shows the top 10[11] most correlated genes with *IC50* for negative and positive relations identified by the three measures.

Gene Expression – Top 10 most correlated genes with IC50			
Dataset	**Genes**	**Method**	**Relation**
Gene Expression	'MRPL23', 'RPS12', 'PCBP1', 'UQCRH', 'HES6', 'SRD5A3', 'FAM78A', 'PPARGC1B', 'COQ8A', 'CSTA'	Pearson	Negative
Gene Expression	'CSTA', 'HES6', 'SLC5A6', 'ATP5MC3', 'PPARGC1B', 'SNRPA1', 'COQ8A', 'ARID3B', 'NRG4', 'BHMG1'	Spearman	
Gene Expression	'ABCB1', 'LRRN4', 'ARSJ', 'ADAM20', 'ME3', 'ITGA3', 'CCDC122', 'ARHGAP29', 'CRIM1', 'GSTT2B'	Pearson	Positive
Gene Expression	'ARSJ', 'ELFN2', 'SYCE1L', 'ADAM20', 'ITGA3', 'ME3', 'ARHGAP29', 'BIRC3', 'CCN3', 'THBS1'	Spearman	
Gene Expression	'CAMSAP2', 'TNC', 'CASK', 'EPHX2', 'ZBTB45', 'GSTT2B', 'ME3', 'ZNF503',	MIC	Positive and Negative

[10] Assuming no duplicates in the dataset.

[11] For MIC, top 20 genes are displayed since it does not separate between positive and negative.

	'GABBR1', 'DPP7', 'MMP24', 'ZWILCH', 'HES6', 'BORCS6', 'P4HB', 'SEC24B', 'NFIB', 'HADHA', 'FRRS1', 'HSPBAP1'		

Table 4: Most correlated genes with IC50.

Spearman and Pearson commonly identified the HES6, CSTA, PPARGC1B, and COQ8A genes in their top 10 negative correlations. The HES6 gene belongs to the split homolog hairy-enhancer family and has been linked to oncogenesis and cancer progression in several human malignancies, including prostate and breast cancer [98]. Furthermore, in terms of positive correlations, Spearman and Pearson identified ME3, ARHGAP29, ARSJ, ITGA3, and ADAM20, from which ARSJ had the strongest correlation with NUC-7738 sensitivity and has been linked to response in hypoxia to breast cancer [99]. ME3 and HES6 were identified in the top correlations by all three techniques. Interestingly, ME3 is involved in carcinogenesis of pancreatic cancer thus making it a candidate target for diagnosis, treatment, and prognosis [19].

3.3.2 Replicating Kathad et al. process

The initial step was to prepare the merged data for FS and ML-driven biomarker identification. This started by removing all columns that originated from the NUC-7738 dataset except the *IC50*. The columns with a value of zero in over 20% of the samples were also removed, reducing the number of genes from 19,161 to 14,498. The dataset was then split into training and validation sets, with 80% of the data used for training, and the rest for validation. The data were the gene expression values and the label the *pIC50*[12] which was chosen due to it being a common practice in the literature for better-discriminating *IC50* values and because the nature of potency values is logarithmic [20]. The split needed to happen as early as possible to avoid data leakage, which occurs when information from the validation set is leaked, giving the model an unfair advantage to make better predictions in the validation data but failing to generalize to other unseen data. To ensure no data leakage occurs, all FS procedures were performed solely on the training set. The sets were checked for any missing values and three labels were missing. This was addressed by training a K Nearest Neighbour (K=3) imputer using the training set. The imputer found the three nearest cell lines in the training set for each and assigned their averaged *pIC50* value. Figure 8 depicts the process followed to replicate Kathad et al. [13] technique.

[12] *pIC50* is the *IC50* in minus decimal log (base-10) power.

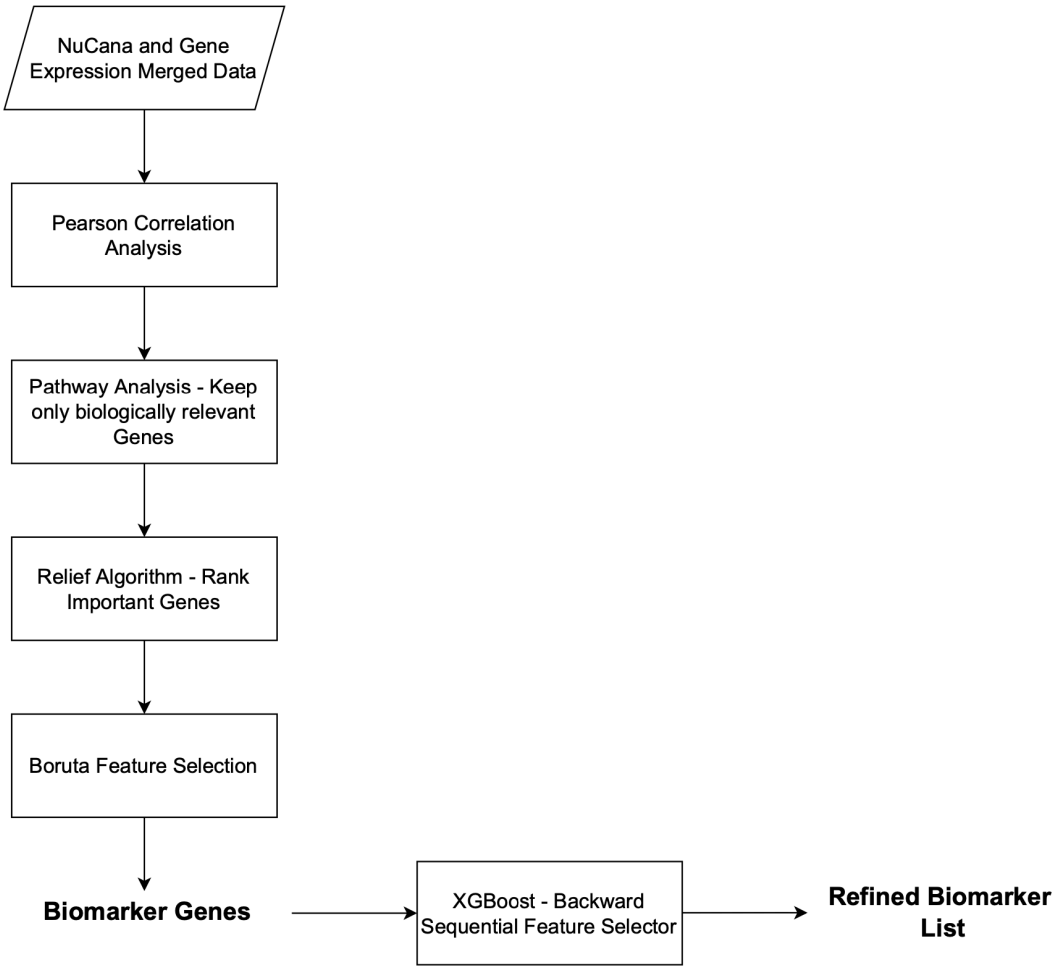


Figure 8: Biomarkers extraction process.

The Pearson correlation analysis kept only genes that were significantly correlated with sensitivity on the training set. This made a big discount to the genes as it reduced them from 14,498 genes to 1,986.

Pathway analysis was then performed using gene set enrichment analysis (GSEA), a powerful tool for interpreting high-throughput expression studies to uncover insights into biological processes or pathways underlying a particular phenotype [39]. GSEA eliminated any genes that are biologically irrelevant. The four pathway-libraries that were selected were biological, molecular, cellular, and KEGG, which are the same used by the paper. A function was created called *union_pathway_genes* which performs three union operations to merge all the genes that were identified by each of the four pathway libraries. Genes not included in the merged list were dropped, causing the genes to reduce from 1,986 to 1,760.

The Relief algorithm was then used to rank and assign weights to the remaining genes based on their relation to drug sensitivity. The algorithm calculates a feature score for each feature which can then be applied to rank and select top-scoring features for FS [56]. A function called *rank_genes_relief* was created which takes in the updated set of genes, and the number

of genes to keep. It then fits and trains the RReliefF algorithm, a regression-specific variant, which selects and returns the 100 top-ranked genes.

The next layer in the FS process was using the Boruta algorithm (used with random forest). Boruta iteratively removes features that are statistically less relevant than artificial noise variables introduced. It does so by creating random shadow (noise) copies of the features and tests the original feature against those copies to determine if it is better than the noise, and therefore worth keeping. In each iteration, rejected variables are removed from consideration for the next iteration [57]. The results of this process and the number of biomarkers returned highly depend on the max-depth of the random forest (RF), and the number of importance ranks kept. To apply this technique, a function was created called *subset_genes_boruta* which runs the Boruta feature selector on the genes matrix from the previous step utilising the RF model and keeping only the genes that are below the *ranking_threshold* specified. This gives an initial biomarker signature.

The next step was the integration of a sequential feature selector (SFS) method using extreme gradient boosting (XGBoost), a tree-based supervised learning approach that integrates estimates from a series of simpler and weaker models into an ensemble by computing their weighted sum. Each of the weaker models uses gradient descent optimization to update the model's weights to reach a local minimum of the cost function [68]. The SFS further refines the biomarker gene list, keeping a signature of biomarkers with the most predictive power by using backward elimination to sequentially delete features that do not significantly contribute to the prediction of the dependent variable. The best feature to be deleted is based on 10 times repeated 5-fold cross-validation (CV) score, where CV is a model evaluation technique that separates the data into K subsets, with one of the K subsets serving as the test set and the remaining K-1 subsets forming the training set [70]. The average error for all K trials is then computed and based on that the best feature is selected. Once the best feature subset is found, a new XGBoost regressor gets trained on the new subset of features, using the training set. The model then predicts the *pIC50* value of the validation set and computes the Pearson correlation coefficient (PCC) between the predicted and actual values.

Table 5 summarises the most-predictive biomarker subsets and their performance on the validation set, across different values for the *max_depth* parameter. The Boruta ranking threshold was set to 3, to keep only relevant features.

Max Depth	**Genes Subset**	**Validation Set - Pearson Coefficient**
1	DKK3, CSTA, SERPINF1, ME3	0.3504 (p=0.1679)
3	ANXA2, SLC38A5, SALL4, KMT5C, ME3, CRABP2	0.5263 (p=0.0299)
5	ANXA2, SLC38A5, SALL4, KMT5C, ME3, CRABP2	0.5263 (p=0.0299)
7	HIPK2, ANXA2, SLC38A5, SALL4, KMT5C, ME3, CRABP2	0.5975 (p=0.0113)

8	CSTA, SNX19, HIPK2, ANXA2, MBNL2, SLC38A5, GLIS3, SALL4, KMT5C, ME3, CRABP2, CCN1	0.5272 (p=0.0296)
10	HIPK2, ANXA2, SLC38A5, SALL4, KMT5C, ME3, CRABP2	0.5975 (p=0.0113)

Table 5: Performance on validation set.

The same subset of genes was selected when *max_depth* was 3 & 5, and 7 & 10. The best models produced have a PCC of 0.5975 (p=0.0113) on the validation set.

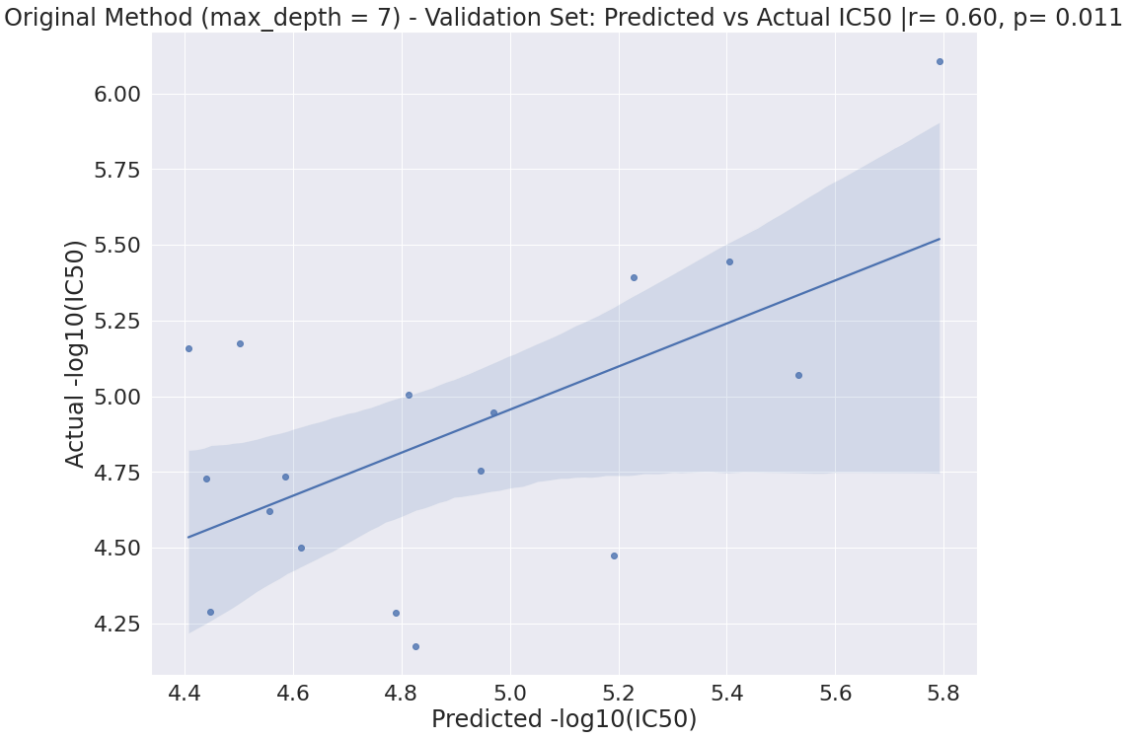


Figure 9: Predicted vs actual pIC50 value.

3.3.2 A proposed alteration

This section proposes an alternative approach (Figure 10) to the technique developed in the previous section, with the differences highlighted with a grey background.

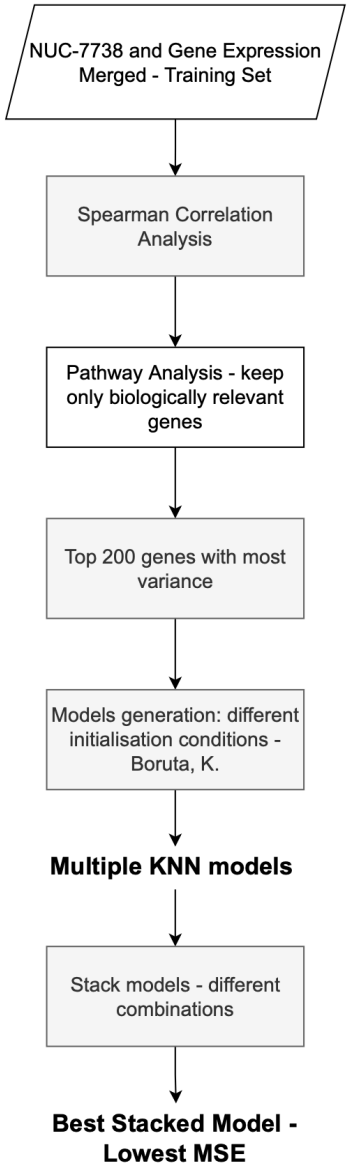


Figure 10: The new approach developed.

To also capture monotonic correlations between gene expression values and drug sensitivity, the Pearson correlation analysis at the commencement of their workflow was replaced by a Spearman correlation analysis. Furthermore, the Relief algorithm was replaced by a variance threshold algorithm, retaining the top 200 genes with the highest variance. This was done as genes whose expressions have low variance should not differ significantly between the sensitive and resistant cell lines, hence resulting in a model with poorer predictive power [69]. We hypothesized that because of the extremely limited samples, sophisticated models' predictive ability would be hampered and that a simpler model would be a better fit. The next element in our proposed method is the generation of multiple K-Nearest Neighbour (KNN) models, where each was created using different initial values for the *max-depth* parameter of the Boruta algorithm and the number of neighbours (K). Before deciding on KNN, experimentation with multiple models was performed, including ridge regression, decision trees, and RF (see Appendix D). KNN is a non-parametric algorithm that can handle non-linear

data by approximating the relationship between independent variables and a continuous target value [67]. Rationally, the use of KNN is meaningful since cell lines with similar gene expression levels to biomarker genes should respond similarly to the drug. The KNN implementation used computes a weighted average of its K neighbours, with closer neighbours having a greater influence. The experimentation with different K values is consequently essential as it yields different models. The *max-depth* parameter for Boruta is also important as different values lead to a marginally varied initial signature. Models for all the possible permutations of the values listed in Table 6 were thus generated. Before passing the training samples to KNN, their values were standardised[13].

Parameter	Values
Max_depth	1,2,3,4,5,6,7,8,9,10
K	2,3,4,5

Table 6: Values tried for the model generation process.

Using the SFS procedure, the best-performing model on the validation set for each max depth value was discovered and preserved. This left us with ten promising KNN models, that share some features. The final step in the pipeline is the combination of different models based on stacking, a powerful ensembled process that learns how to optimally aggregate several ML models to generate a more powerful predictor that benefits from the differences between the models. The idea to ensemble multiple KNN models was inspired by the success of Tarek et al. [17], as discussed in section 2.3.1. The stacked model (meta-model) was chosen to be a support vector regression (SVR) model after experimentation with XGBoost, SVR, Ridge, and RF, as it was shown that it syndicates the KNN models the best (lowest MSE error). SVR trains with a symmetrical loss function that penalizes both high and low misestimates equally and seeks the hyperplane with the most points [112]. Multiple stacked models were developed greedily, with the first stacked model combining the best[14] two performing KNN models, the second combining the first three best KNN models, etc. The stacked model that was saved is the one that combined the best two KNN models as it performed the best on the validation set, having the lowest MSE score. The final model performs better than the one developed in the previous section, having a more significant stronger correlation coefficient of 0.65 (p=0.0051) with the experimentally derived *pIC50* values on the validation set (Figure 11).

[13] Mean=0 and standard deviation=1.

[14] Best models are selected based on their validation-set performance.

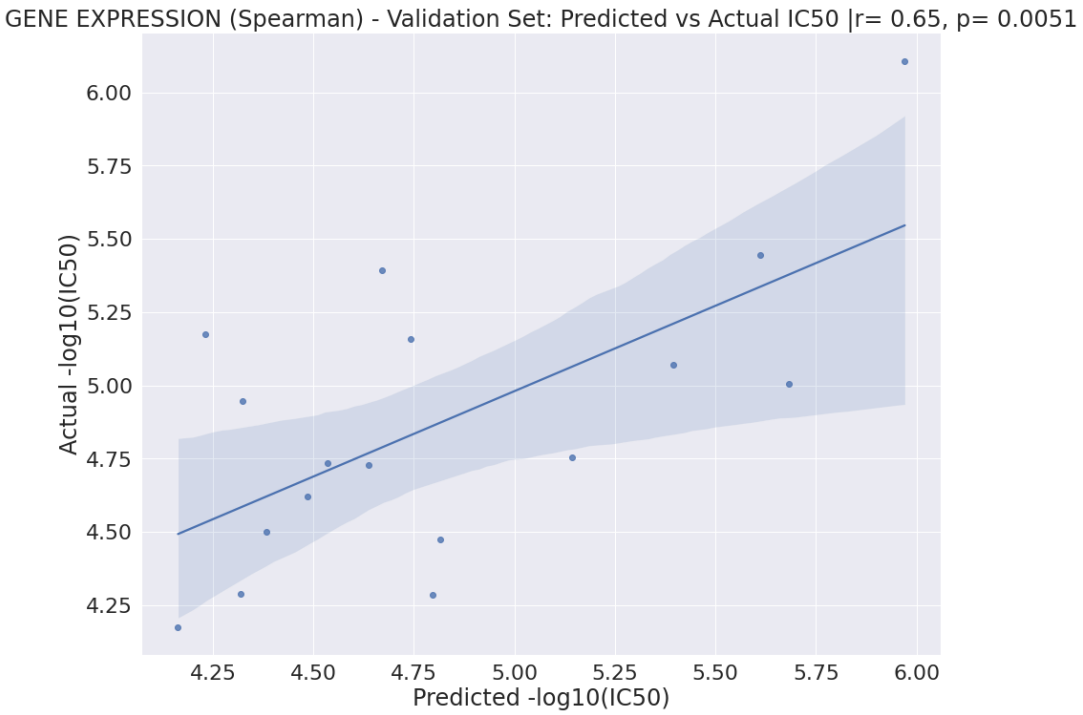


Figure 11: Predicted vs actual pIC50 values (validation set).

The resulting 10-gene biomarker signature is the following:

ITGA3, RARRES2, CTHRC1, MST1R, ARHGAP29, ABCC3, DCBLD2, ANXA2, BIRC3, HLA-A

3.4 Clustering Alternative Polyadenylation

As elucidated in 2.4, APA events can act as important biomarkers for predicting drug resistance and clinical outcomes of cancer [18]. Clustering of cell lines based on APA events allows to better understand how APA influences drug sensitivity while allowing for further exploration of important biomarkers and mechanistic insights that distinguish the clusters. This section details the clustering techniques explored. An overview of the pre-processing procedure used to prepare the data for clustering is available at appendix L. Pre-processing yielded the1000 genes with the most variance for 53 cell lines.

3.4.1 Clustering APA events

Clustering is an unsupervised ML task which involves the automatic discovery of similar groups in data by considering the feature space. Clustering high-dimensional data has an inherent problem in the "curse of dimensionality". This term defines the finding that many algorithmic problems get more complex as the number of dimensions increases [49]. Further studying of this problem showed that neighbourhoods and distances among data objects are known to become increasingly meaningless in high-dimensional spaces. Beyer et al. [50] proved that the distance of an object *o* to its nearest neighbour *dmin(o)*, and the distance to its farthest object *dmax(o)* eventually become indistinguishable in high dimensions (5).

$$\forall \varepsilon \lim_{|D|\to\infty} p\big(\mathrm{dmax}(o)\big) < (1+\varepsilon)\mathrm{dmin}(o)) = 1 \quad (5)$$

This is a major problem for many traditional clustering algorithms which rely on similarity distance-based metrics. This is commonly addressed using dimensionality reduction methods to project high-dimensional data to a lower dimensional subspace. Dimension reduction is the discovery of lower-dimensional representations of high-dimensional data while retaining some significant structure, like sample distance. The study of dimension reduction techniques originates in Pearson's seminal paper on finding curves and surfaces of best-fit given a set of noisy measurements [60]. Dimension reduction techniques generally fall into two categories, linear which include the most popular techniques like principal component analysis (PCA) and non-negative matrix factorisation (NMF), and non-linear techniques which include t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP). To experiment with the clustering of APA events, both linear (PCA) and non-linear techniques (UMAP) were examined.

3.4.1.1 Clustering with PCA and Spherical K-Means

PCA creates new uncorrelated variables that successively maximize variance to lower the dimensionality of large datasets while avoiding information loss. Identifying the principal components abridges to solving an eigenvalue/eigenvector problem [58]. To determine the number of principal components (PCs) needed to retain the essence of our dataset, the cumulative explained variance preserved over different numbers of PCs was calculated.

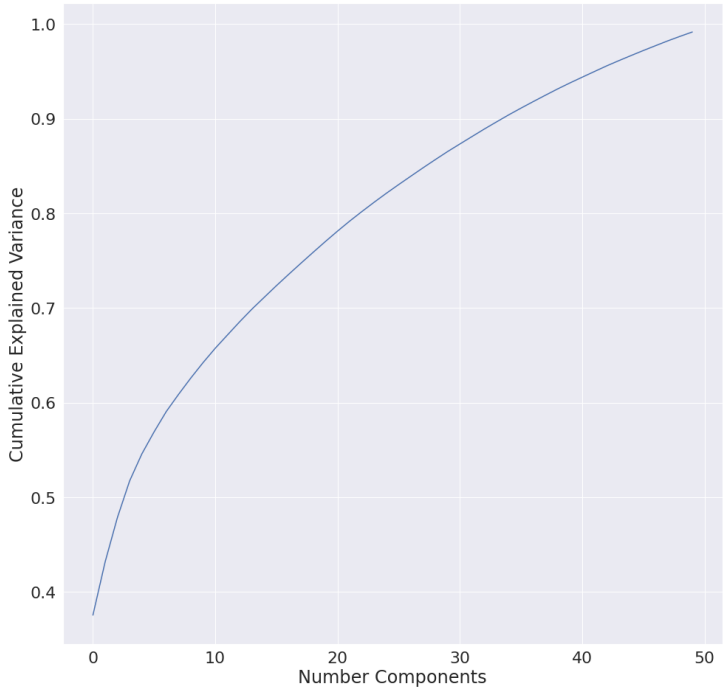


Figure 12: Cumulative explained variance against different number of components.

Reducing the data to 13 PCs would keep 70% of cumulative explained variance of the original data, 22 PCs would keep 80%, 34 PCs 90%, and 42 PCs 95%. Keeping more PCs would lead

to less information loss but also to a greater chance of suffering from the curse of dimensionality due to the increased number of dimensions.

Once the dimensions are reduced, the data can be clustered more easily. A popular clustering algorithm is K-Means, an iterative algorithm that tries to partition the dataset into K distinct non-overlapping clusters where each data point belongs to only one group. The algorithm begins by guessing K centroids at random, and then it repeatedly 1) assigns each data point to the closest cluster, and 2) updates the centroids by taking the average of all the data in each cluster until the centroids no longer change (converge) [51]. The algorithm is simple to implement, generalizes to clusters of different shapes and sizes, and guarantees convergence. However, it requires knowing the number of clusters in advance, and it also uses Euclidean distance which leads to the curse of dimensionality on high-dimensional data. A variant of K-Means that was proven to perform better in high-dimensional space is spherical K-Means [52]. Its main alteration is that it uses cosine distance rather than Euclidean. An implementation of spherical K-Means was developed (*sphericalKMeans*) and used.

To evaluate the trade-off between the different number of PCs and select the most appropriate number of clusters in a principled fashion, the minimum description length (MDL) was used. MDL states that the best description of the data is given by the model which compresses it the best, by having the lowest description length (DL) which is the number of bits required to encode the parameters of a model and the data, given the model. The principle is based on the idea that a simpler explanation is better than a more complex one (Occam's razor) and that the more the data can be compressed, the more the algorithm has learned about it and therefore better it can predict it [59, 60]. MDL for statistical models can be formally written as in (6), where d_i are individual data points from the data (D), N is their count, and N_M is the number of free model parameters [60].

$$DL(M,D) = \frac{1}{2} N_M \log_2 N - \sum_{i=0}^{N-1} \log_2 p(d_i) \qquad (6)$$

The number of free parameters for a K-Means model is the number of clusters (K) multiplied by the number of dimensions. In terms of the likelihood term, K-Means is a special form of a Gaussian mixture model (GMM) where its probability density function is given by:

$$p(x) = \sum_{k=1}^{K} \frac{1}{K} N(x \mid \mu_k, I) \qquad (7)$$

With the prior distributed uniformly for the number of clusters and the covariances fixed to the identity matrix. Due to the high number of dimensions, the multivariate Gaussian distribution is needed to represent the model's variables. Therefore, when expanded, the DL equation becomes:

$$DL(M,D) = \frac{1}{2} N_M \log_2 N - \sum_{i=0}^{N-1} \log_2 \sum_{k=1}^{K} \frac{1}{K} \left(\frac{1}{\sqrt{2\pi^d \, |\Sigma|}} e^{\left\{ -\frac{1}{2} (x-\mu)^{\mathrm{T}} \, \Sigma^{-1} (x-\mu) \right\}} \right) \qquad (8)$$

To calculate the DL of the clustering, a function called *calculate_DL* was developed which takes as parameters the data, the number of clusters, and the predicted labels and calculates the DL for a model using the formula above. Appendix L encloses the results when the PCs needed to retain 70%, 80%, 90%, and 95% of the cumulative explained variance, and K between 2 and 5 were tried.

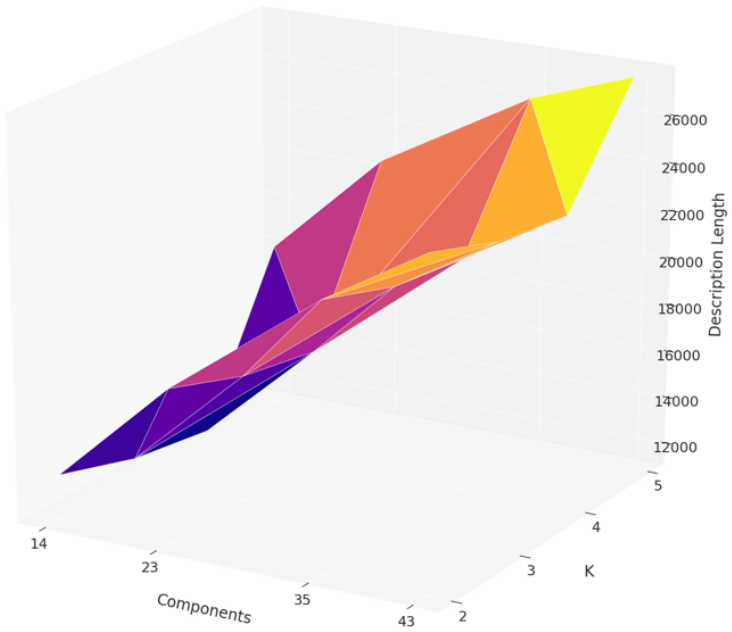


Figure 13: Visualisation MDL of PCA with Spherical K-Means.

The combination with the minimum DL is fourteen principal components and four clusters (Figure 13), which suggests that keeping 70% of explained variance captures the general structure well and clusters the samples correctly. MDL shows the information gained by keeping 95% of the variance does not justify the increase in the model's complexity.

It should be noted that while the visualization in Figure 14 appears to show poor cluster separation, what is displayed is a projection of the first two PCs which does not accurately reflect the distribution in the 14-dimensional space.

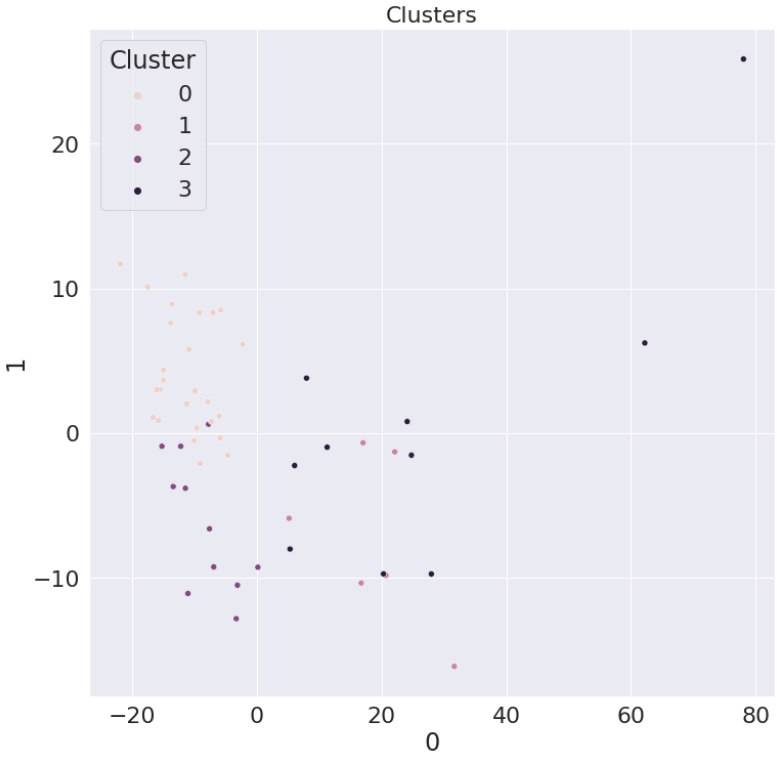


Figure 14: Clusters created by K-Means with 13 PCs and K=4.

3.4.1.2 Clustering with UMAP and HDBSCAN

The problem with using linear dimensionality reduction techniques is their frequent failure to account for potential non-linear interactions. The most popular modern non-linear dimension reduction approaches are t-SNE and UMAP. It was shown that UMAP outperforms t-SNE in terms of global structure preservation and run-time performance [61]. Since preserving global structure is crucial in obtaining a more accurate representation of the clusters, UMAP was chosen. It mainly builds upon mathematical foundations related to the work of Belkin and Niyogi [62] on Laplacian eigenmaps. It creates high-dimensional graph representation of the data and then optimizes a low-dimensional graph to maximise structural similarity [61,63]. UMAP is commonly used for clustering in conjunction with hierarchical density-based spatial clustering of applications with noise (HDBSCAN), a non-parametric method that uses a density-based approach to search for a cluster hierarchy shaped by the multivariate modes of the underlying distribution. K-Means are effective when the clusters are spherical, equally sized, equally dense, most dense in the centre of the sphere, and do not contain any noise or outliers. In practice, however, the clusters often have arbitrary shapes, different sizes, different densities, and noise. Intuitively, HDBSCAN instead of seeking clusters of a specific shape looks for data regions that are denser than the surrounding space, resulting in a superior performance than K-Means [64].

The most principal UMAP parameter is the number of neighbours (*n_neighbors*) which controls the number of nearest neighbours[15] of each point. This is important for the balance between preserving local (low values) versus global (high values) structure. The trade-off of

[15] This number includes its own sample.

preserving a more global structure is that fine-grained details are lost [63]. The other parameter of UMAP is *min_dist* which controls how closely the data are together, with lower values putting the points close together and higher values more loosely. The most important HDBSCAN parameters are *min_cluster_size* and *min_samples,* with the first setting the smallest size grouping that would be considered a cluster, and the second controls how conservative the clustering is. The larger the value for *min_samples*, the more data points are declared as noise, and clusters will be restricted to only the densest areas [65]. Since the clusters should be as undisputable as possible, the *min_samples* remained fixed to its default value (i.e., conservative).

After experimentation with multiple clustering parameters (see Appendix E), it was found that smaller values for *n_neighbors* (3-5) and *min_dist* (0-0.05) in combination with small values for *min_cluster_size* (3-7) created the most distinct clusters. Furthermore, dimensionality reduction on the entire dataset identified more dissimilar groups when compared to clustering using the 1,000 genes with the most variance and was thus preferred. Figure 15 shows a promising cluster identified during this experimentation with *n_neighbors*=5, *min_dist*=0.01, and *min_cluster_size*=5.

Figure 15: Promising cluster with UMAP and HDBSCAN.

Profiling of the two clusters (Figure 16) revealed that their sensitivity varies, with the "0" cluster having a lower median and maximum value. The "1" cluster has more spread out *IC50* values with a larger IQR range. However, the disparity in sample size between the two clusters, with the "0" cluster having only 13 samples and the "1" having 40, must be considered.

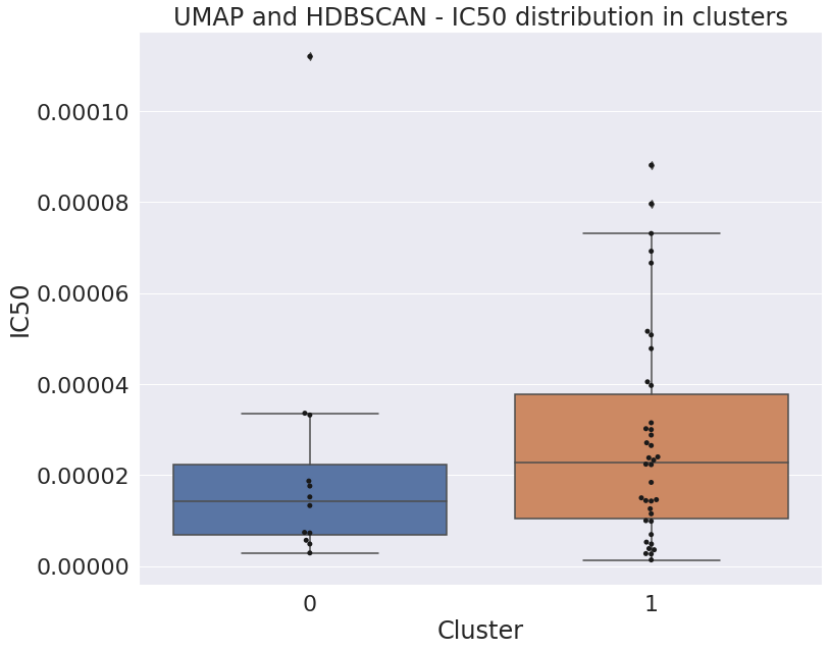


Figure 16: IC50 distribution between clusters.

Figure 17 shows a cluster map of the ten genes with the most significant difference in their average APA event frequency between the clusters. The identification of these is important since it helps gain further mechanistic insights between the two clusters.

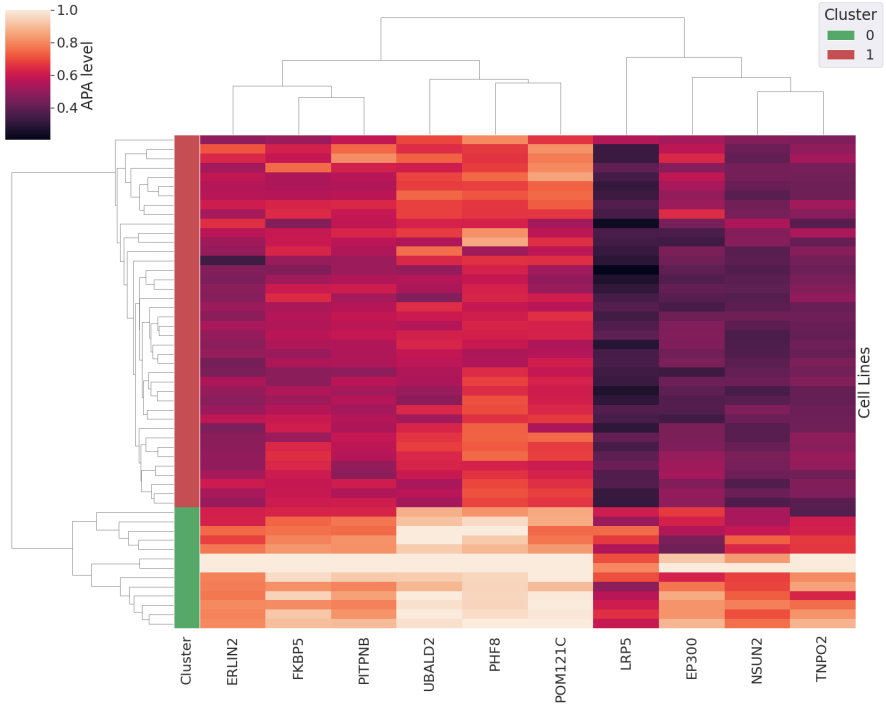


Figure 17: Ten genes with significant differences in their APA event frequency between clusters.

3.5 Exploration of additional DepMap datasets

Proteomics, gene effect, and mutations DepMap datasets were also explored. The incorporation of the proteomics dataset [72] into the pipeline is critical as proteins are essential to life, performing structural, metabolic, transport, immune, and regulatory functions [109]. The gene

effect dataset contains each gene's significance towards the survival of a cancer cell line. Negative scores[16] indicate that cell growth has been inhibited and/or death has occurred because of gene knockout (KO), a procedure that disrupts a cell's genomic DNA so that the expression of a specific gene is permanently stopped [66]. The mutations dataset included information about mutations that occurred on each cell line. It has been further explored as identifying mutation markers and selecting appropriate treatment for patients with specific genome mutations are key steps in the development of targeted therapies [53].

Their exploration initially involved the in-depth correlation analysis and identification of potential biomarkers for each (section 3.5.1) and then their integration to facilitate the development of a multi-omics predictive model (section 3.5.2). Missing data from public datasets was a major issue, with some datasets (e.g., APA) containing nearly half of the NUC-7738 cell lines, severely limiting experimentation possibilities.

3.5.1 Correlation Analysis and potential biomarkers

Before correlation analysis, all datasets underwent the same pre-processing steps, in which columns with more than 20% zero values were discarded and cell lines with missing *IC50* readings were dropped. The top 10 most correlated genes with *IC50* for negative and positive relationships identified by the three measures are discussed, with a focus on genes that were identified consistently across the techniques[17]. Biomarkers were discovered using the proposed approach in section 3.3.3. Appendix F includes the correlation tables for each.

3.5.1.1 Alternative Polyadenylation

Five top positively correlated genes (PRR12, CLUH, CAMK2D, PRPF8, DCTN5, NPM3) with *IC50* were common between Pearson and Spearman[18]. PRPF8 is a known significant factor for splicing in breast cancer progression [100], while the most highly common positively correlated gene, CAMK2D, affects various metabolic and cancer-related signaling pathways, which are important for cisplatin response and drug resistance development [101]. CCDC97 is in the top 10 positively correlated identified genes with *IC50* by both Pearson and Spearman. MIC identified IMPA2 as its top correlated gene, an oncogene involved in the proliferation and migration of cervical cancer which regulated the MAPK signalling pathway [111]. None of the top correlated genes identified were common in all three techniques.

The biomarker discovery procedure identified an eight-biomarker APA level gene signature: FAM3C, FANCC, SLC5A3, TAF4, XPA, SUPT16H, CBX3, and TRA2A. The predictive model has a 0.43 PCC (p=0.18) with the actual sensitivity values on the validation set (Figure 18). The poor performance may be attributed to the limited samples available[19].

[16] DepMap standardised the dataset's scores such that non-essential genes have a median score of 0 and independently identified common essentials have a median score of -1.
[17] The same techniques were used as in Section 3.3.1 (Pearson, Spearman, and MIC).
[18] See Appendix F part a.
[19] Almost half of our cell lines were missing from the APA datasets, leading to only 53 samples.

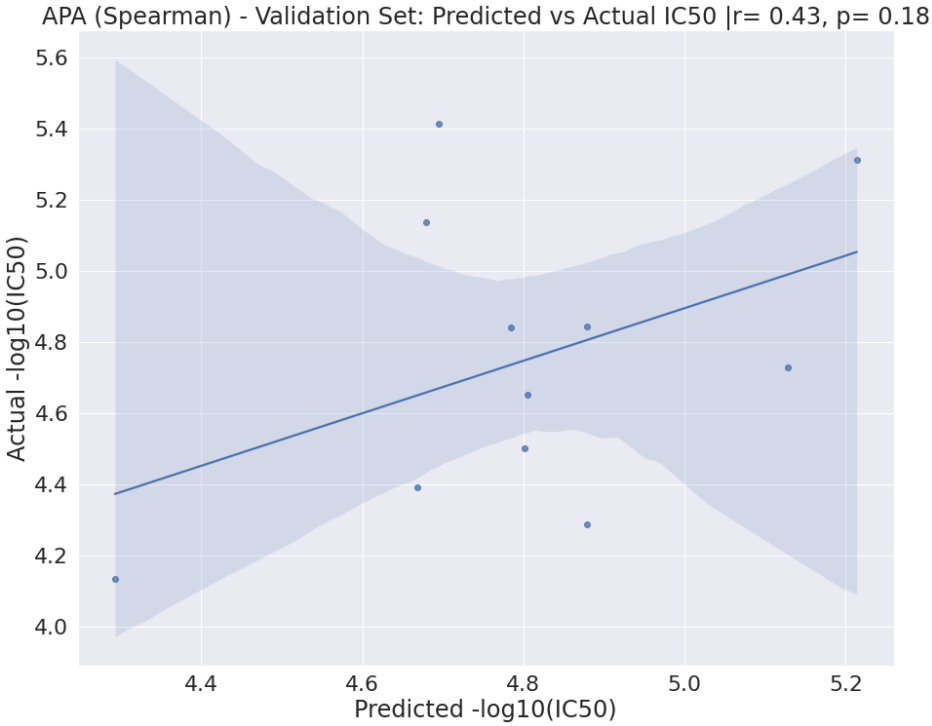


Figure 18: Predicted vs actual pIC50 values – APA.

3.5.1.2 Proteomics

The ORC3, ORC4, TMED3 and WDR18 proteins were common in the top 10 negative correlations between Pearson and Spearman[20]. Interestingly, seven (APBB2, BCAR1, KIF3B, LRRC20, NUDCD3, RANBP3, TNS3) out of the top 10 positively correlated proteins, were common between Pearson and Spearman. KIF3B has been shown to be over-expressed in multiple human cancers (gastric, oral, pancreatic, prostate, seminoma, hepatocellular), and Wang et al. [104] showed it could serve as a potential therapeutic target for breast cancer treatment. NUDCD3 was identified in the top three positively correlated proteins by all three measures.

An 18-gene signature was derived by applying the biomarker predictive pipeline: AHNAK, PEX19, HBD, GLG1, GCA, GMFG, DOCK1, SIRT1, GID8, RTL8A, NUDCD3, APBB2, CALCOCO2, ACSL3, PBRM1, HMGN4, C18orf25, EPG5. The developed predictive model attained a 0.8 correlation coefficient (p=0.0032) on the validation set (Figure 19).

[20] See Appendix F part b.

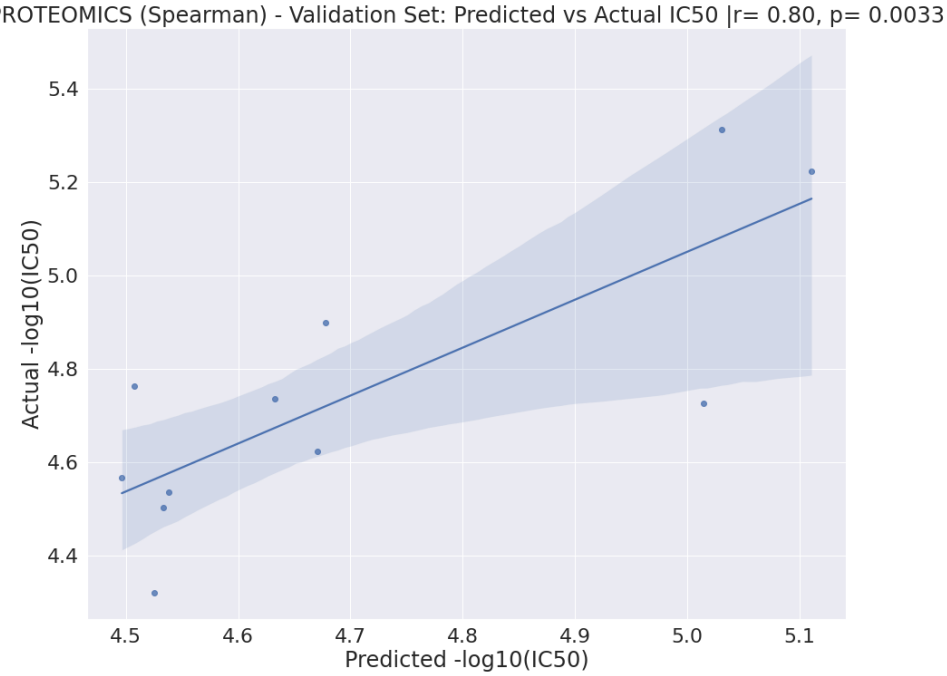


Figure 19: Predicted vs actual pIC50 values – Proteomics.

3.5.1.3 Gene Effect

Aside from correlation analysis, clustering was also performed on the gene effect dataset.

3.5.1.3.1 Clustering dependencies

Following suggestions from Dr Mustafa Elshani, the possible targets for our drug include: CSTF(n)[21], CPSF(n), PCF11, SRSF(n), CLP1, and PAPOLA. The primary goal in clustering using the gene effect dataset is to discover genes with similar dependency profiles[22], which may be useful in identifying alternative druggable targets for genes of interest that are not conventionally druggable [91]. To produce reliable dependency profiles, clustering was done on the entire (1,086 cell lines) gene effect dataset.

The developed technique (Figure 20) begins by identifying the top 100 positively and negatively correlated genes whose effects are highly correlated (Spearman) with each potential target gene. It then compiles all the correlated genes into a single list and updates the gene effect dataset to include them. To determine the coordinates of each data point in a 2D plane, a pairwise Spearman distance matrix is generated and passed into UMAP. The data points are then given to HDBSCAN which clusters the genes. Ensemble clustering was used to ensure robust clustering and discover genes that were consistently assigned to the same cluster across 100 runs with different seeds. This technique was influenced by ECHODOTS, developed by Shimada et al. [91]. The main differences include tailoring the technique for the NUC-7738 drug by finding similar dependencies to its potential targets and replacing t-SNE with UMAP and DBSCAN with HDBSCAN based on the arguments presented in section 3.4.2.2.

[21] All genes that start with CSTF and end with a number.
[22] Similar effect on cell lines.

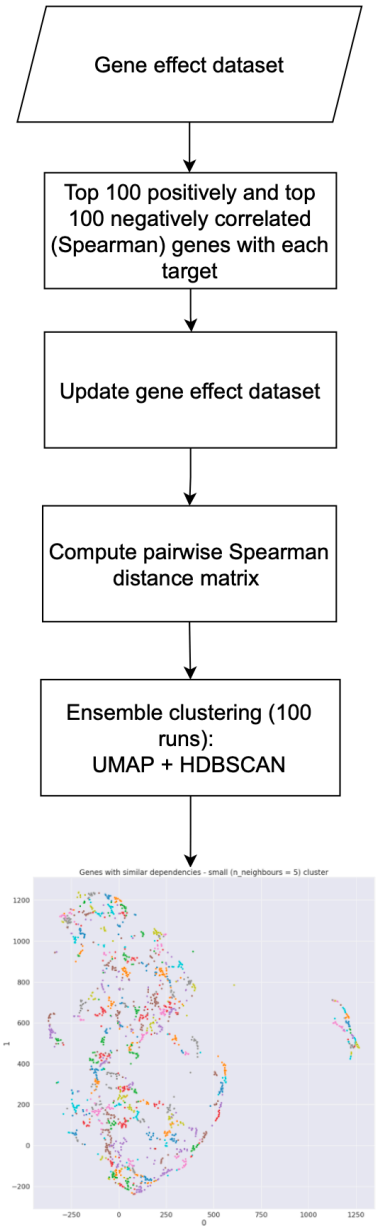


Figure 20: Technique developed to identify genes with similar dependency profiles.

The size and number of clusters are determined by the *n_neighbours* parameter. Figure 21 depicts the resulting cluster when *n_neighbours*=5. Appendix G contains clusters generated when *n_neighbours* is 10 and 15.

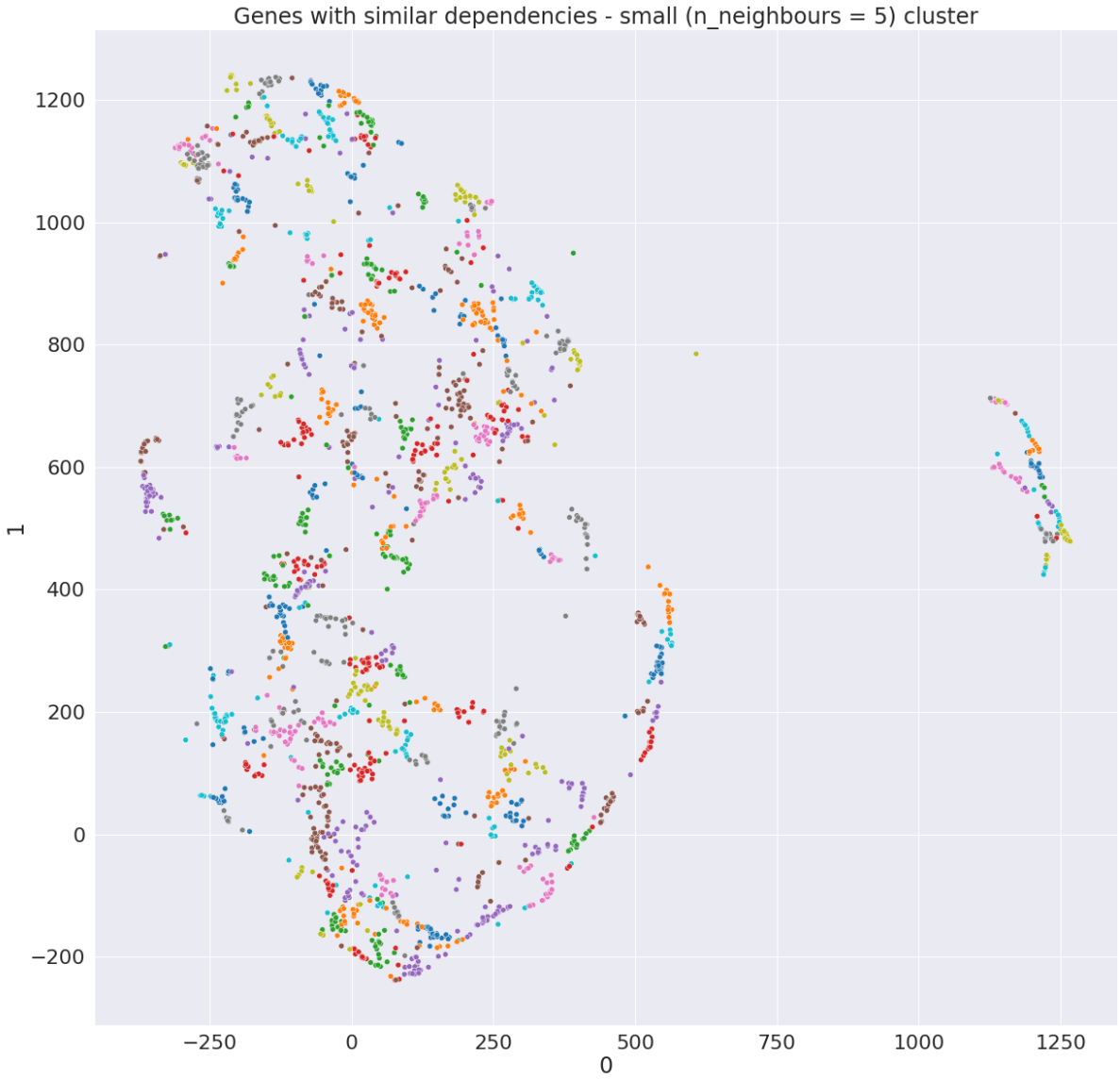


Figure 21: Cluster of genes with similar dependencies.

3.5.1.3.2 Correlation analysis

CADM1 and ATG2A are in the top 10 negatively correlated gene effects with *IC50* in both Pearson and Spearman[23]. On the other hand, RAD1, RARS2, DTNBP1, MRPS31 were commonly identified by the techniques in the top 10 positive correlations. CADM1 has the highest negative linear correlation (Pearson) and second highest monotonic (Spearman). Its expression reduces melanoma cell invasion and migration, and it was recently shown that a knockdown of the gene increased melanoma cell invasion [106]. MRPS31 is also interesting, as its knockdown enhanced hepatoma cell invasiveness in hepatocellular carcinoma [107]. The top correlated gene effect identified by MIC is in the SPRR1A gene, which is a biomarker for the prognosis of colon cancer [108].

[23] See Appendix F part c.

Ten genes were identified as biomarkers: LSM3, EIF2S3, CWC22, TCF3, CLNS1A, AMD1, TYMS, PSMD4, RIOK1, CDK8, with the predictive model having a 0.54 PCC (p=0.036) on the validation set (Figure 22).

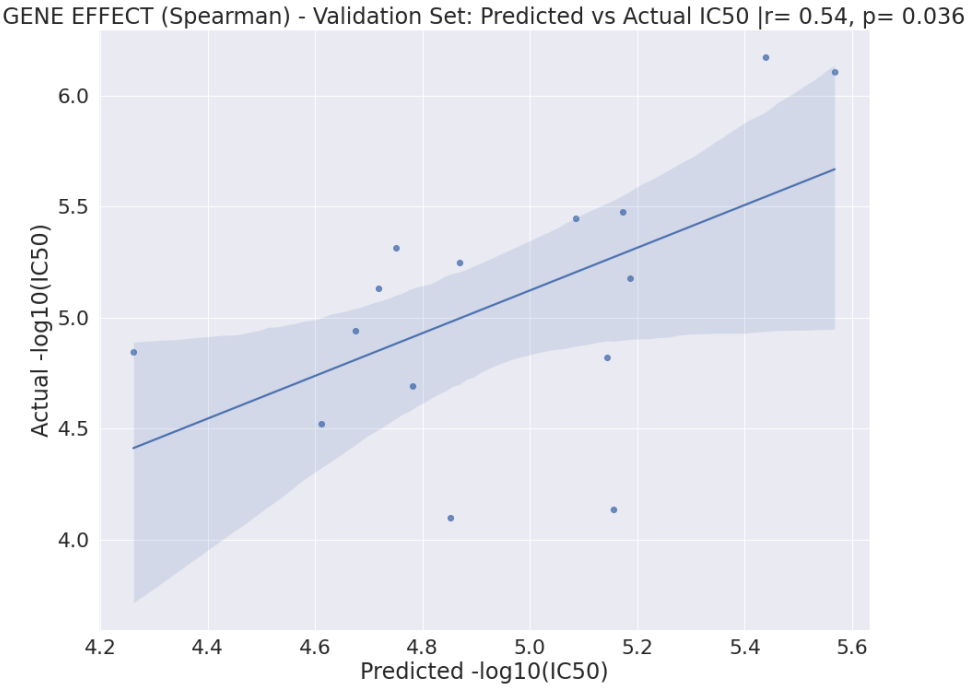


Figure 22: Predicted vs actual pIC50 values – Gene Effect.

3.5.1.4 Mutations

Aside from correlation analysis and biomarker discovery, the mutations dataset was explored further to examine the number of mutations at different sensitivity level bands and identify the gene mutations which best separate sensitive and resistant cell lines.

3.5.1.4.1 Mutation numbers at different sensitivity level bands

The cell lines were separated into *sensitive*, *middle*, and *resistant* categories based on their *z-score*, where cell lines with a *z-score* less than -0.5 were classified as *sensitive*, cell lines with *z-score* between -0.5 and 0.5 as *middle*, and the ones with a value greater than 0.5 as *resistant*. The *middle* class was created to avoid the introduction of a harsh threshold as the cell lines close to 0 z-score have similar sensitive levels and it would therefore be incongruous for one to be *sensitive* and another to be *resistant*. The average and standard deviation were calculated for each sensitivity level (Table 7).

Sensitivity Level (classification)	Average number - Gene Mutations	Standard Deviation - Gene Mutations
Sensitive	450.4	160.9
Middle	976.8	1136.1
Resistant	1237.6	2692.4

Table 7: Average number of Gene Mutations in three sensitivity levels.

By looking at the average one can conclude that the number of mutations is highly correlated with drug resistance. However, once the standard deviation is considered, the results are not as clear. The resistant cell lines have a higher standard deviation, indicating that the number of mutated genes is more spread out in resistant cell lines than in sensitive cell lines. As a result, the assumption that more resistant cell lines have more gene mutations is less reliable.

3.5.1.4.2 Mutations with the most impact on z-score

To identify the gene mutation with the most impact on the sensitivity category of the cell line, a function was created (*create_one_hot_all_genes*) that transformed the dataset into a binary representation, with the rows representing the cell lines, and the columns the genes. If a cell line had a mutation on the specific gene, a value of 1 was assigned; otherwise, 0. The Pearson correlation between the mutated genes and *z-score* was then calculated, from which only the significant correlations (p<0.05) were kept. The results were confirmed using a point biserial correlation. To evaluate the validity of these findings, the top 50 positively and negatively correlated mutated genes were plotted on a heatmap (Figure 23) against the three sensitivity level categories.

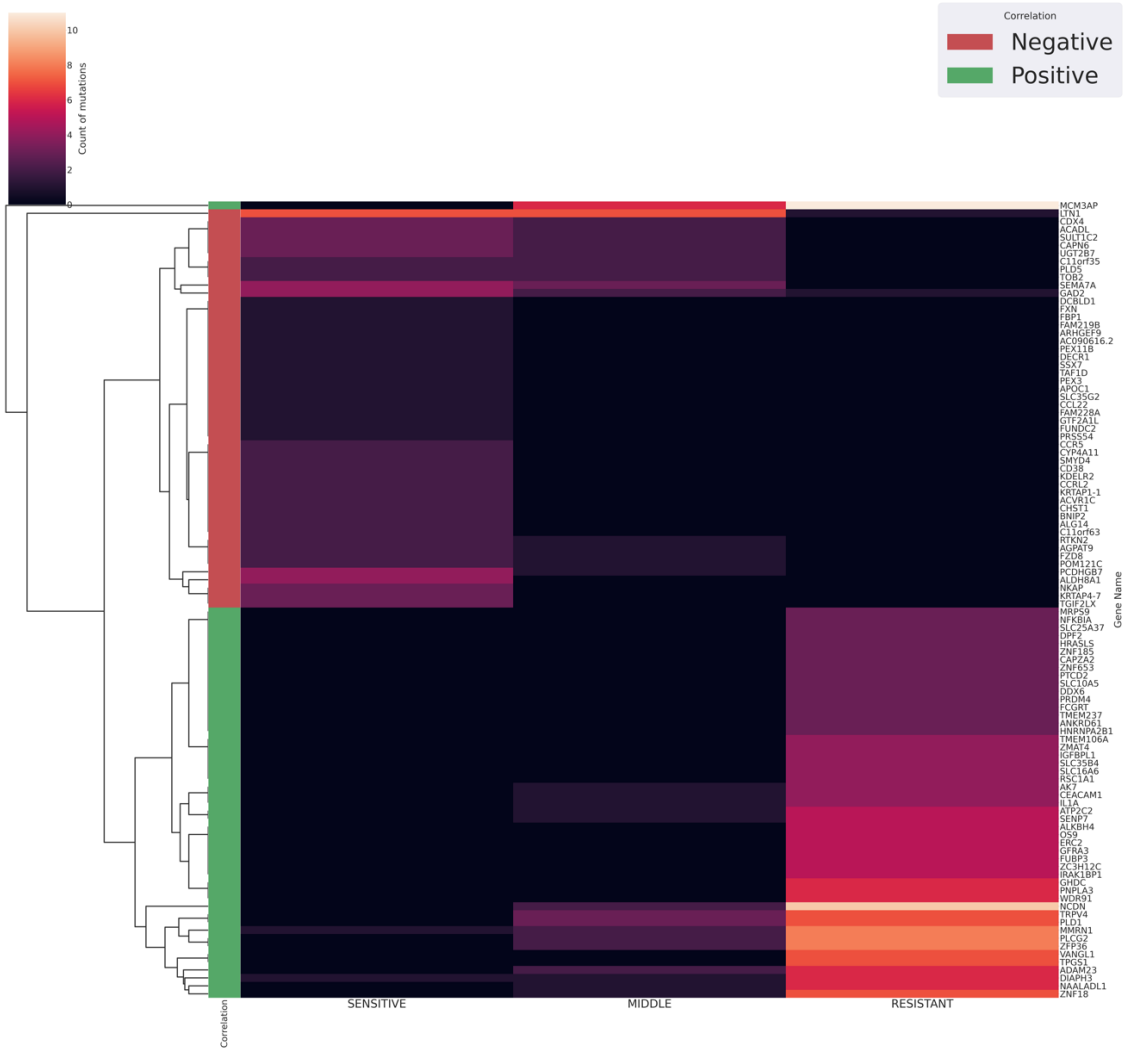


Figure 23: Heatmap of top 50 positively and negatively correlated gene mutations and the class to which they appear.

Mutations in negatively correlated genes are more prevalent in sensitive cell lines, whereas mutations in positively correlated genes are substantially more common in resistant cell lines. From the negatively correlated genes mutations only for the GAD2 and LTN1 genes appear in the resistant cell lines. In the sensitive cell lines, only MMRN1 and DIAPH3 are found from the positively correlated genes. Some of these mutated genes, as expected, also occur in the middle sensitivity level. However, since middle is difficult to define, it is intriguing that it appears considerably different from high and low. The findings confirm that mutations in those genes are associated with drug sensitivity and resistance, respectively.

3.5.1.4.3 Correlation analysis and biomarker discovery

Since the mutations dataset contains binary data and a continuous label (*IC50*), point-biserial correlation is a more appropriate correlation measure. A positive correlation indicates that a mutation is linked to an increase in drug resistance. A negative correlation, however, indicates that a mutation in a particular gene is associated with higher drug sensitivity. An interesting gene that was in the ten highest negative correlations with *IC50*, is BRAF. Mutation in this gene has been linked to cancer growth in a variety of cancer types, including melanoma, ovarian, and colorectal cancer [90].

Due to the binary nature of the mutation dataset, the biomarker discovery approach had to be modified (Figure 24). Spearman correlation analysis was replaced with point-biserial correlation analysis, and the steps following the variance threshold algorithm with LOBICO, a technique developed by Knijnenburg et al. [92] that infers small and easily interpretable logic models of binary input features that explain a continuous output variable.

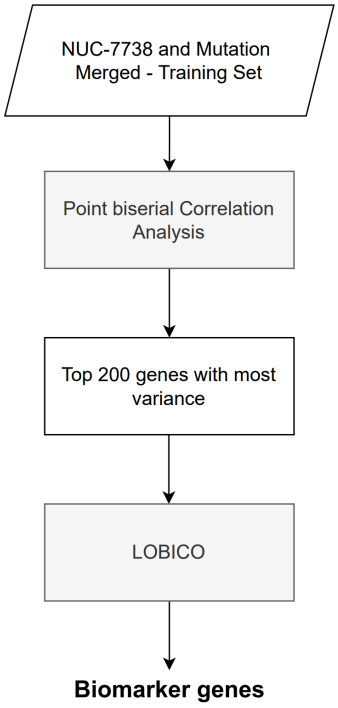


Figure 24: The process followed to identify biomarkers in the mutation dataset.

LOBICO differentiates sensitive and resistant cell lines using a binarization threshold technique they developed. It determines an optimal logic function of binary mutation features

that minimizes the sum of the weighted misclassified cell lines, with the weight being proportional to the distance to the binarization threshold. LOBICO minimizes the total weight of misclassified samples (9), where y′ is a binary vector with the predicted binary labels, and y is the actual labels vector.

$$\hat{\theta} = \arg\min_{\Theta}(|y' - y|) \quad (9)$$

LOBICO's logic functions are in disjunctive normal form (DNF), and their complexity is governed by two parameters: *K*, the number of disjunctive terms, and *M*, the maximum number of chosen features per disjunctive term [92]. The authors provided LOBICO's source code[24], which was applied to our data. The binarization technique to decide the binary threshold could not be replicated as there was no data about the lower and upper bounds of our *IC50* confidence intervals. Therefore, the cell lines in the training set were sorted in ascending order based on their *pIC50* and separated on resistant and sensitive cell lines based on a threshold[25]. Once trained, LOBICO can be used to make a binary prediction (sensitive or resistant) based on the optimal logic function it identified during training.

A major argument by Knijnenburg et al. was that in every drug most cell lines are resistant. Therefore, we experimented with three separate thresholds: 50%, 65%, and 80%, representing the percentage of cell lines that are classified as resistant (class=0). Multiple logic models were generated by experimenting with different *K* and *M* values, for each of the thresholds (see Appendix H). The best performing model on the validation set is at the 50% threshold (*pIC50*= 4.73), with *K*=1 and *M*=3. The optimal logic function identified is:

~INSRR & ~ANO4 & ~INSR

Its performance was relatively poor with 0.47 accuracy, 0.54 precision, 0.88 recall, and 0.67 F1-score.

3.5.1 Multi-omics model

As discussed in Chapter 2, the integration of multi-omics data can be useful in analysing a patient's biological data holistically and providing a more accurate diagnosis. The fact that each dataset had different cell lines available, precluded the potential of investigating a concatenation-based model as we would have an extremely small (~30 samples) high-dimensional dataset. Furthermore, developing a complex model-based approach would also be very challenging as the extremely scarce data available would be a major drawback in incorporating DL techniques such as SAE, which would overfit the training data. Considering these constraints, a simple model-based technique was devised in which the intermediate models for each omics data set are combined to produce the final multi-omics prediction.

[24] Written in MATLAB.

[25] For example, if the percentage threshold was 75%, then 75% of the cell lines with the lowest *pIC50* value were classified as resistant cell lines.

When the multi-omics model is given a new sample, each of the omics models makes an individual prediction. The weighted average (10) of these predictions is then computed, with models with lower MSE scores on the validation set receiving greater weight (inverse proportional). The weights are represented by *w*, and the model predictions by *x*.

$$\frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i} \qquad (10)$$

To calculate the weights, the reciprocal of each MSE score is calculated and the weights are normalised such that they sum to one. Finally, after each of the models makes a prediction, the normalised weights are used to compute the weighted average following the formula above.

3.6 Drug combinations analysis

Following consultation with Professor David Harrison[26], it was decided to investigate only gene expression correlations with drug sensitivity when NUC-7738 is administered in combination with 1) Paclitaxel and 2) Erlotinib. Instead of providing the *IC50* values for these combinations, *IC* values[27] for each cell line at distinct points were recorded. Appendix I explains how one-dimensional piecewise linear interpolation was used to calculate *IC50* and includes the correlation results. Noteworthy, the combination *IC50* values are for the 120 treatment hours and therefore the appendix also includes gene expressions correlation analysis with NUC-7738[28] for that hour.

The most interesting insight derived from the correlation analysis is that the gene expression of the HINT1 gene is considerably more correlated with *IC50* (both Pearson and Spearman) when the drug is used in combination with Paclitaxel (Table 8). This is an especially important finding as the NUC-7738 paper [73] stated that HINT1 is essential to activate NUC-7738. Furthermore, they asserted that HINT1 protein is prevalent across multiple cancer types and its depletion was not dose-limiting, showing that even low levels of HINT1 are adequate for NUC-7738 activation.

Drug(s)	HINT1 Pearson	HINT1 Spearman
NUC-7738	r=0.13, *p*=0.23	ρ=0.12, *p*=0.27
Paclitaxel	r=-0.02, *p*=0.93	ρ=0.53, *p*=0.004
Erlotinib	r=0.14, *p*=0.58	ρ=0.32, *p*=0.21
NUC-7738+Paclitaxel	r=0.52, *p*=0.006	ρ=0.78, *p*=2.73e-06
NUC-7738+ Erlotinib	r=0.11, *p*=0.45	ρ=0.21, *p*=0.45

Table 8: HINT1 correlation with IC50 on different drug combinations.

[26] Medicine professor at the University of St Andrews.
[27] Each cell line consisted of approximately four to five *IC* points.
[28] Correlation analysis table in section 3.3.1 is for 72 treatment hours.

3.7 Summary

This chapter started by giving context to the datasets and packages utilized. Data analysis was then used to show the distribution of *IC50* across the three treatment hours and to identify lineages that are more susceptible and resistant to the drug. The gene expression dataset was then subjected to correlation analysis, which was followed by two ML-driven biomarker discovery methods. Unsupervised clustering on the APA dataset helped reveal clusters with differences in their drug sensitivity and genes with substantial APA frequency discrepancies. The exploration of the proteomics, gene effect, and mutation datasets was then discussed, with special attention on their correlation analysis and the biomarkers revealed. Following that, a technique for producing multi-omics predictions was discussed, and finally, the correlation analysis of two drug combinations was investigated.

4. Discussion

With only 95 cell lines in our dataset, many of which were not publicly available, the sophistication of techniques we could employ was severely limited. Regardless, we experimented with the development of a biomarker discovery technique by experimenting with multiple methods (see Appendix D) and were able to develop a biomarker discovery pipeline that appears promising based on validation set performance. This chapter begins by critically evaluating the performance of the proposed method. Section 4.2 delves deeper into the biomarkers discovered, and section 4.3 highlights the potential of the ensemble clustering technique to identify genes with similar dependency profiles. Section 4.4 assesses the project's success concerning its objectives and the final section delivers closing remarks.

4.1 Critical evaluation of proposed biomarker discovery method

The predictive models are evaluated by assessing their performance on a holdout set, followed by a calculation of their 2-fold and 4-fold accuracy.

4.1.1 Holdout-set performance

To properly assess the models' performance and generalisability, a holdout set was used. The holdout set consists of 10 cell lines that were withheld until the model development process was completed.

Method	Model	Cell lines	Correlation Coefficient	p-value
Kathad	Gene expressions	'ACH-000411', 'ACH-000555', 'ACH-000429', 'ACH-000234', 'ACH-000336', 'ACH-000406', 'ACH-001687', 'ACH-001688'	0.84	0.0085
Proposed			0.77	0.0256
Proposed	APA	'ACH-000234', 'ACH-000555', 'ACH-000429', 'ACH-000336'	-0.3	0.7029
	Proteomics	'ACH-000411', 'ACH-000555', 'ACH-000429', 'ACH-000234', 'ACH-000406'	0.91	0.0296
	Gene effect	'ACH-000234', 'ACH-000336', 'ACH-000406', 'ACH-000411', 'ACH-001687', 'ACH-001688'	0.76	0.077
	Combined (without APA)	'ACH-000406', 'ACH-000411', 'ACH-000234'	0.89	0.2948

Table 9: Holdout set performance - omics data.

Both methods produced models with strong and significant correlations between predicted and actual drug efficacy for the gene expression dataset. However, the model developed following

Kathad et al. outperforms our approach on the holdout set, despite having a worse performance on the validation set. The model developed for the proteomics dataset appear to generalise effectively on the holdout set as their predictions have significantly (*p-value*<0.05) high correlations with the experimentally derived *IC50* values. Whilst the gene effect and combined models also have high correlation coefficients, the *p-value* shows that these correlations are not statistically significant and therefore should be not relied on. The worst performance is for the APA dataset, which was expected given the model's poor performance on the validation set. This could be due to the extremely low number[29] of samples available, but also as a result that it is the only dataset that was not present on DepMap, and thus its creators may have used different pre-processing[30] steps, rendering our pipeline ineffective. Excluding the APA dataset, the technique's performance appears promising, however, with such limited testing samples[31], definitive conclusions cannot be drawn. The correlation plots for each model's performance are available in Appendix J.

4.1.2 Accuracy

4.1.2.1 Proposed method models

To further validate the performance of the models, their prediction accuracy (Table 10) was calculated by dividing the number of cell lines with predicted *IC50*[32] within two or four-fold from actual *IC50*, by the total number of cell lines in the holdout set.

Method	Model	Dataset	Accuracy 2-fold	Accuracy 4-fold
Kathad	Gene expressions	Validation Set	58.8%	76.5%
Proposed	Gene expressions	Validation Set	47.1%	76.5%
Proposed	APA	Validation Set	63.6%	90.9%
	Proteomics	Validation Set	90.9%	100%
	Gene effect	Validation Set	33.3%	80%
Kathad	Gene expressions	Holdout Set	75%	87.5%
Proposed	Gene expressions	Holdout Set	75%	87.5%
Proposed	APA	Holdout Set	50%	75%
	Proteomics	Holdout Set	80%	80%
	Gene effect	Holdout Set	50%	100%

Table 10: 2-fold and 4-fold accuracy between the predicted and actual IC50 values.

The best performing model on the validation set was the one for proteomics, having accuracy of 90.9% on 2-fold accuracy, and 100% for 4-fold. The worst performing model was the gene effect model, where its predictions were 33.3% within 2-folds of the actual *IC50* values and

[29] 53 out of our 95 cell lines.
[30] All the other datasets were initially pre-processed by DepMap using their pre-processing pipeline.
[31] Each dataset contained a subset of these 10 cell lines, rendering our task even more challenging.
[32] The *pIC50* predictions were converted to their original scale (Molar).

80% within 4-folds. As expected, the APA model has poor 2-fold performance on the holdout set. Interestingly, on the holdout set, the proteomics model had an accuracy of 80% on both 2-fold and 4-fold, indicating that the 20% of its predictions were more than four times off. The gene effect model appears to perform considerably better on the holdout set, with all its predictions being within 4-fold of the actual values. The performance of the models on this metric is encouraging since most predictions appear reasonably close to the actual experimentally derived *IC50* values.

4.1.2.1 LOBICO – Mutation model

Figure 25 shows the performance of the mutation model on the holdout set. By looking at the confusion matrix a bias towards predicting the sensitive class (class=1) is observable as six out of its eight predictions were of that class.

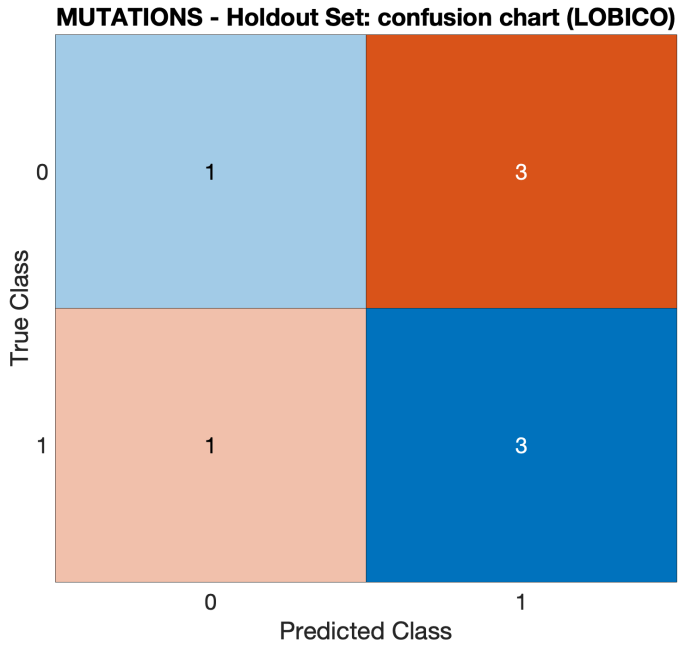


Figure 25: Confusion matrix for the mutations model - holdout set.

The accuracy, recall, precision, and F1-score of the mutations model is shown in Table 11.

Accuracy	Precision	Recall	F1
0.5	0.5	0.75	0.6

Table 11: Mutations model performance across classification metrics.

For our given task, the LOBICO model does not perform well. However, the fact that we were unable to calculate the binarization threshold (see section 3.5.1.4.3) in the manner described in the paper may have hampered its performance. The low number of samples available in comparison to the amount used by the LOBICO authors may have also played a critical role.

4.2 Biomarkers evaluation

This section discusses the biomarkers discovered for each of the different datasets using the proposed method and then demonstrates the increase in predictive ability when only the biomarkers are used instead of the whole dataset.

4.2.1 Biomarkers Discussion

The developed approach has merit since a significant number of biomarkers identified have been linked to cancer in respected journals, either as prognostic biomarkers, therapeutic targets, or tumour suppressors.

4.2.1.1 Gene Expressions

The most correlated[33] biomarker gene with drug sensitivity is ARHGAP29 (Figure 26), with a correlation coefficient of 0.42 (p=8.6e-05). This is a gene that was recently demonstrated to be associated with cancer cell growth in multiple studies. Shimizu et al. [75] identified it as a prognostic biomarker and therapeutic target for prostate cancer, while Kolb et al. [76] showed that it could be an important factor in the invasion of aggressive and mesenchymal-transformed breast cancer cells.

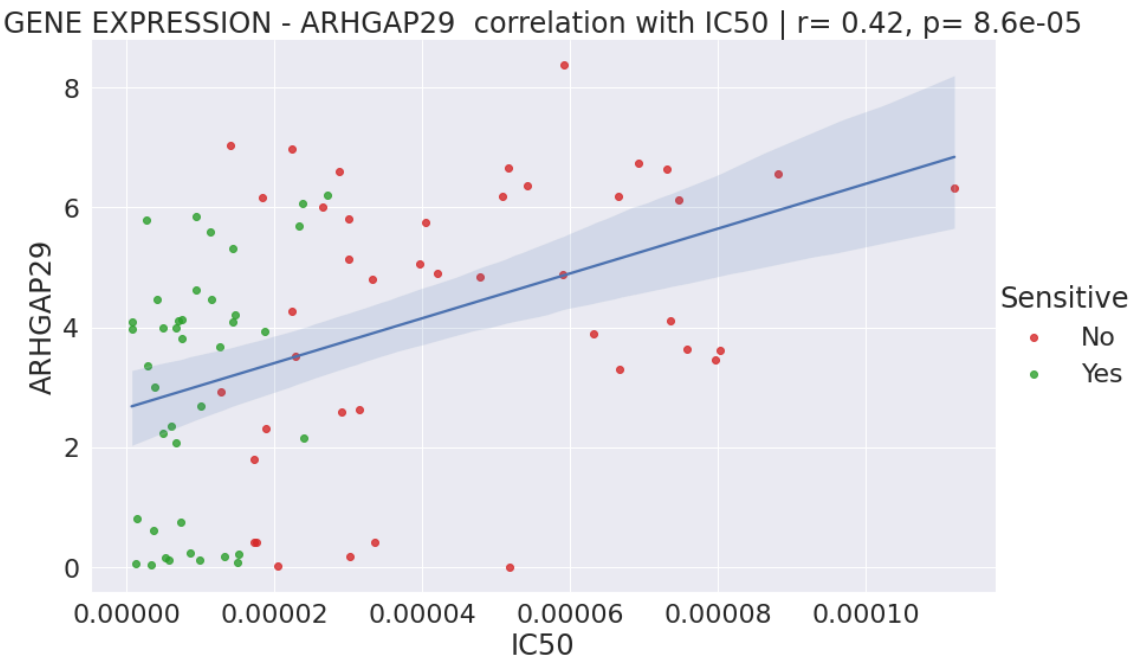


Figure 26: ARHGAP expression correlation with IC50.

Importantly, Appendix K contains a selection of publications for each of the other biomarker genes discovered.

A heatmap of biomarker gene expression values for the top 5 sensitive (green) and top 5 resistant (red) cell lines was produced to investigate the relationship between biomarker gene

[33] Based on Spearman correlation coefficient.

expression values and drug sensitivity. There is a considerable difference in gene expression values between the sensitive and resistant cell lines, further verifying their predictive validity. The genes are less expressed in sensitive cell lines and more in resistant ones. ANXA2 appears to be an exception, with high expression across all cell lines. However, the model uses the joint predictive power of the genes rather than each individually, thus although ANXA2 expression does not differentiate between groups, its combination with other biomarker values may be vital.

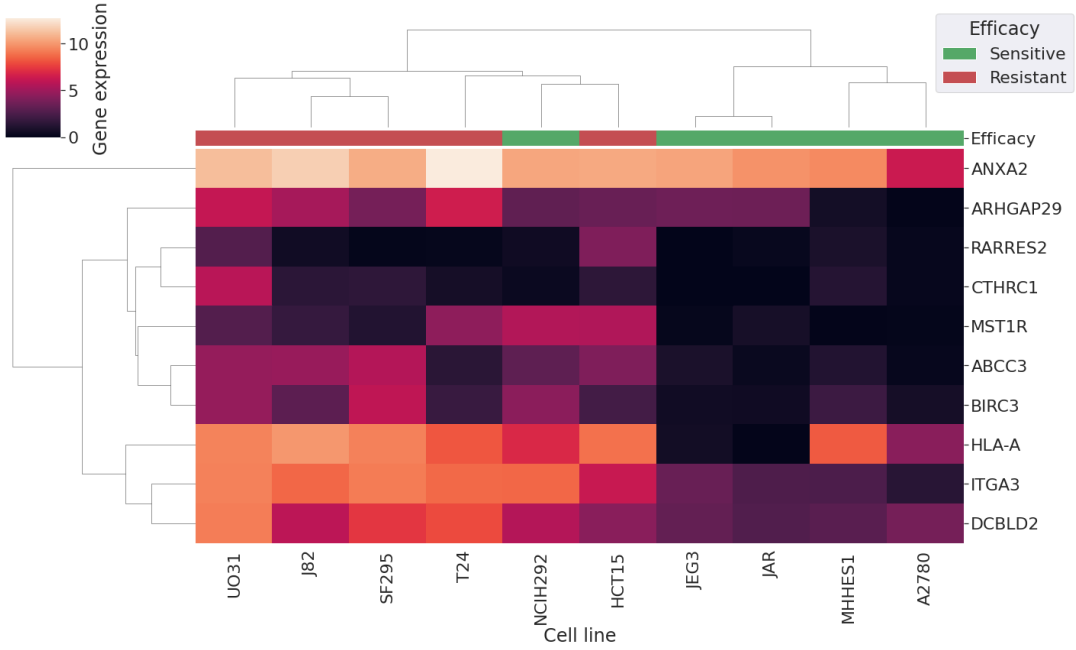


Figure 27: gene expression biomarkers heatmap values against top sensitive and resistant cell lines.

4.2.1.2 APA

The FAM3C gene has a critical role in tumour formation, invasion, metastasis, and survival, making it a biomarker and potential therapeutic target for various cancers [102]. Kazantseva et al. [103] discovered that TAF4 alternative splicing is helpful for cell reprogramming and provides a unique strategy for generation of melanocyte-like cells and recapitulating stages of melanoma progression. The APA frequency of biomarker genes varied considerably between the top sensitive and resistant cell lines, with lower APA levels in most of the resistant cell lines. This is confounding given the poor predictive performance of the model. The results could imply that the FS extraction was effective but there were simply not enough data to build a strong predictor, or the experimental data was noisy, a plausible scenario given that it is the only dataset downloaded from a source other than DepMap.

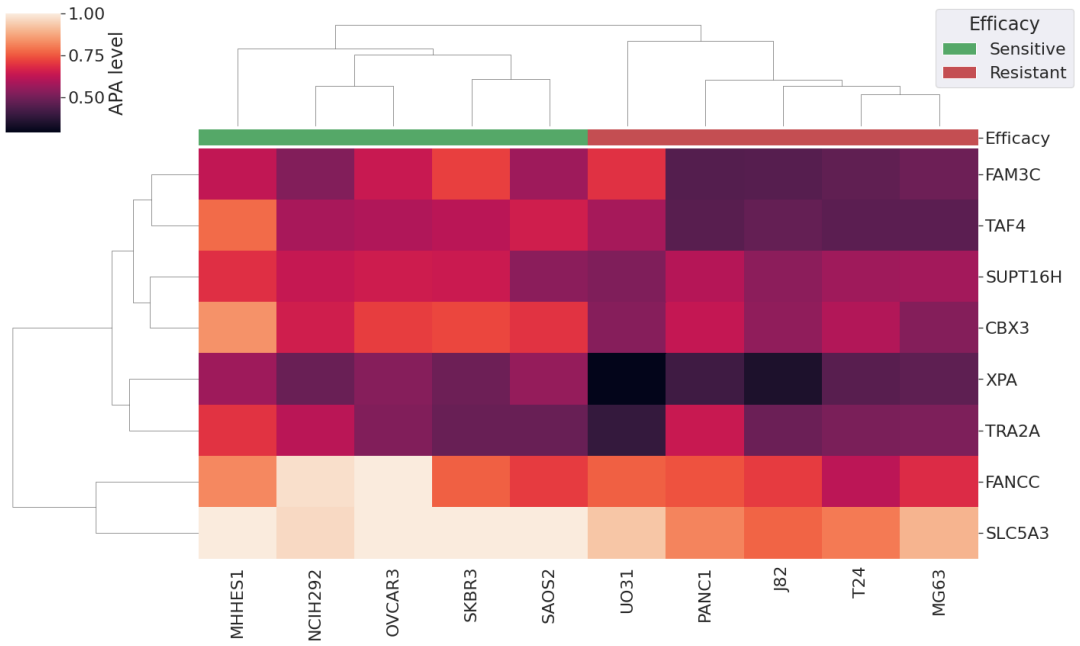


Figure 28: APA biomarkers heatmap values against top sensitive and resistant cell lines.

4.2.1.3 Proteomics

From the top highly correlated proteins with *IC50* identified in correlation analysis, NUDCD3 was included in the signature. AHNAK is associated with increased invasion and metastasis of human lung cancer cells and its expression is correlated with cell migration and invasive ability in malignant mesothelioma cells [105]. The heatmap (Figure 29) shows that the expressions for GLG1, PBRM1, and HBD are higher in resistant cell lines. On the contrary, expressions of RTL8A, PEX19, SIRT1, GID8, and NUDCD3 are higher in sensitive cell lines.

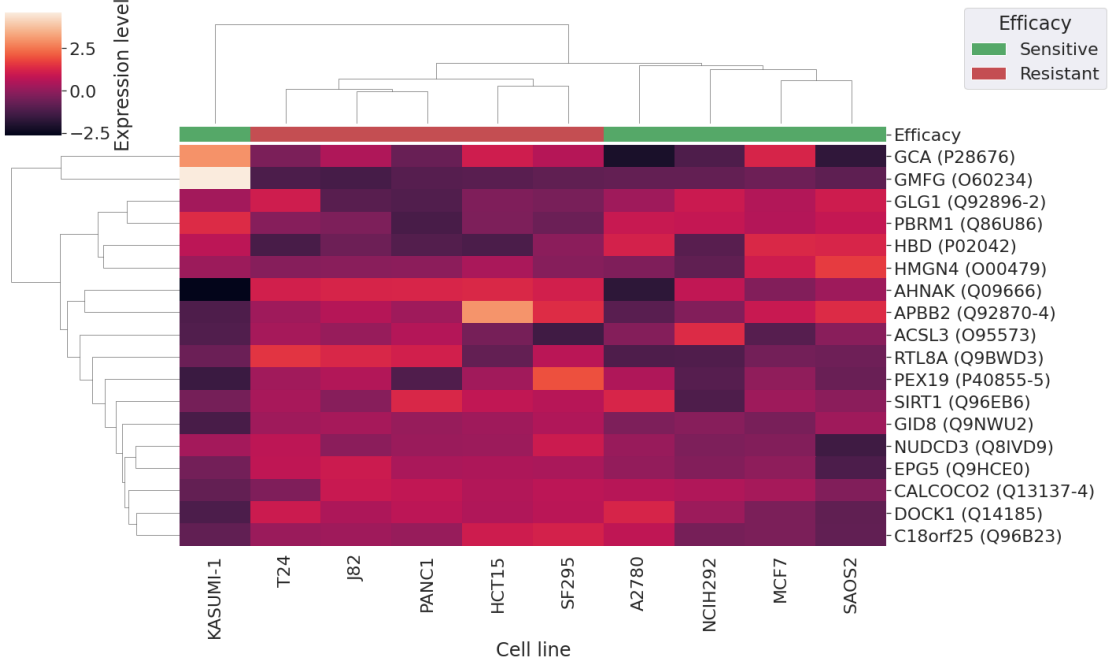


Figure 29: Proteomics biomarkers heatmap values against top sensitive and resistant cell lines.

4.2.1.4 Gene Effect

The biomarker genes come from the positive and negative correlations with *IC50*. CDK8 most significantly positively correlates with *IC50* (ρ=0.44) and its knockout suppresses both the transcriptional and the mitogenic effects of oestrogen in oestrogen receptor positive breast cancer cells [110]. The differences of the biomarkers' values between the sensitive and resistant cell lines are not as distinct as in the other datasets (Figure 30).

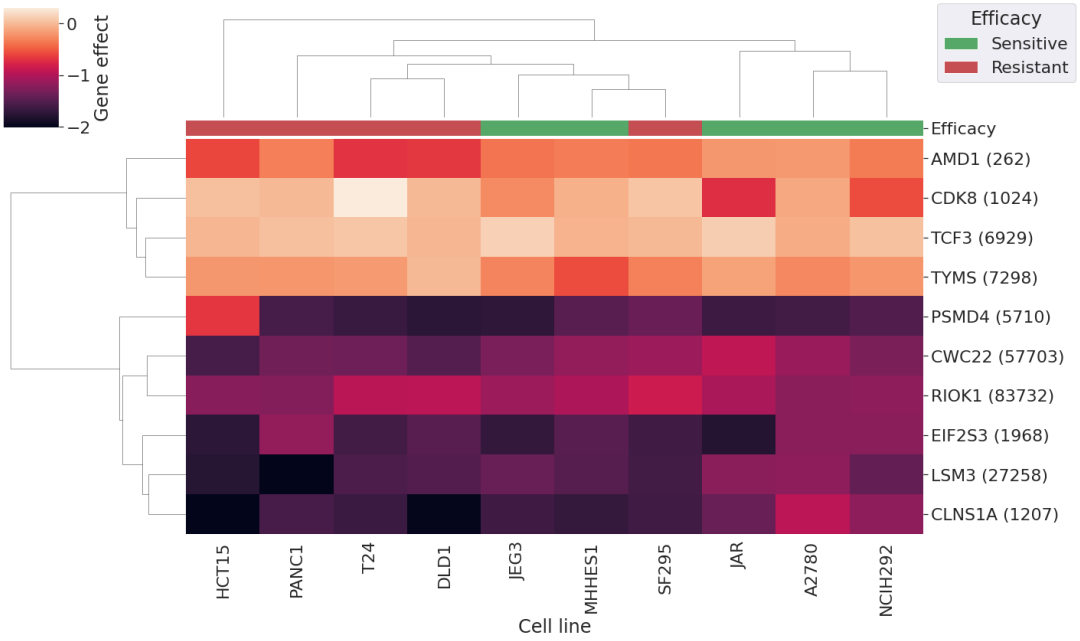


Figure 30: Gene Effect biomarkers heatmap values against top sensitive and resistant cell lines.

4.2.2 Predictive Performance

As discussed in section 2.3, a set of high-quality biomarkers should enhance the predictive performance of a model over the full list of genes, since non-related genes are eliminated. To examine this, multiple KNN (K=2,3,4,5) and XGBoost models[34] were developed with all the genes being kept (without FS). Their performance is depicted in Figure 31.

[34] Trained using the same training & validation split as the models that incorporated feature selection.

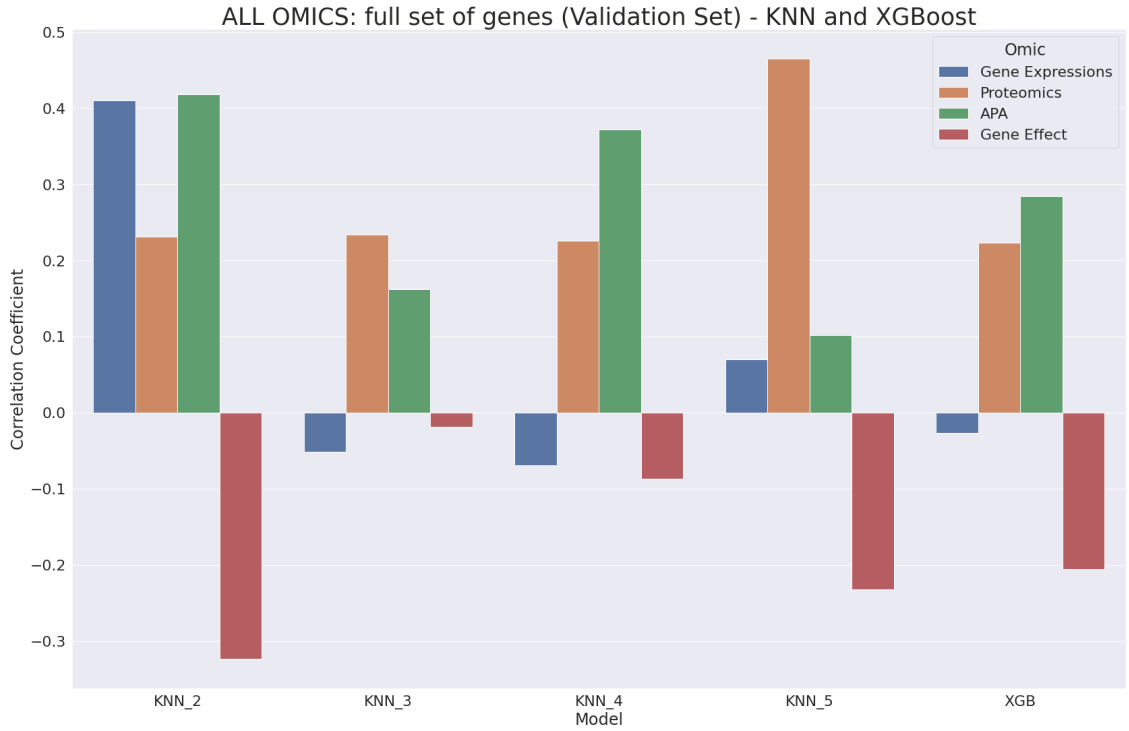


Figure 31: Performance on KNN and XGBoost models when all the genes are kept.

The models perform worse than when the biomarker discovery process is used which confirms the identified biomarkers' ability to improve drug efficacy prediction.

4.3 Dependency clustering

This section explores a potential use case of the dependency clusters developed (section 4.2.1). As previously stated, the CPSF genes may be of interest to our drug. For example, if the CPSF4 gene is targeted but is revealed to be undruggable, the technique can be used to focus attention on other genes with similar dependency profiles.

Table 12 shows the genes that were clustered with CPSF4 (*n_neighbours*=5).

Target	Clustered genes
CPSF4	RPP40, RAD51D, ROMO1, BUB3, RPP30, ILF2, PUS1, DHX33, CIAO2B, TP53RK, CDAN1, CHORDC1, EXOSC4, MVD, WDR82, UTP23, CPSF4, GFER, POP5, CENPA, TAF1C

Table 12: Clustered genes with CPSF4 target.

Examining the clusters from various *n_neighbours* (5,10,15) values demonstrated that the technique consistently identified the same genes, proving its stability and robustness. The main difference was that runs with a higher *n_neighbours* value contained more genes in their clusters.

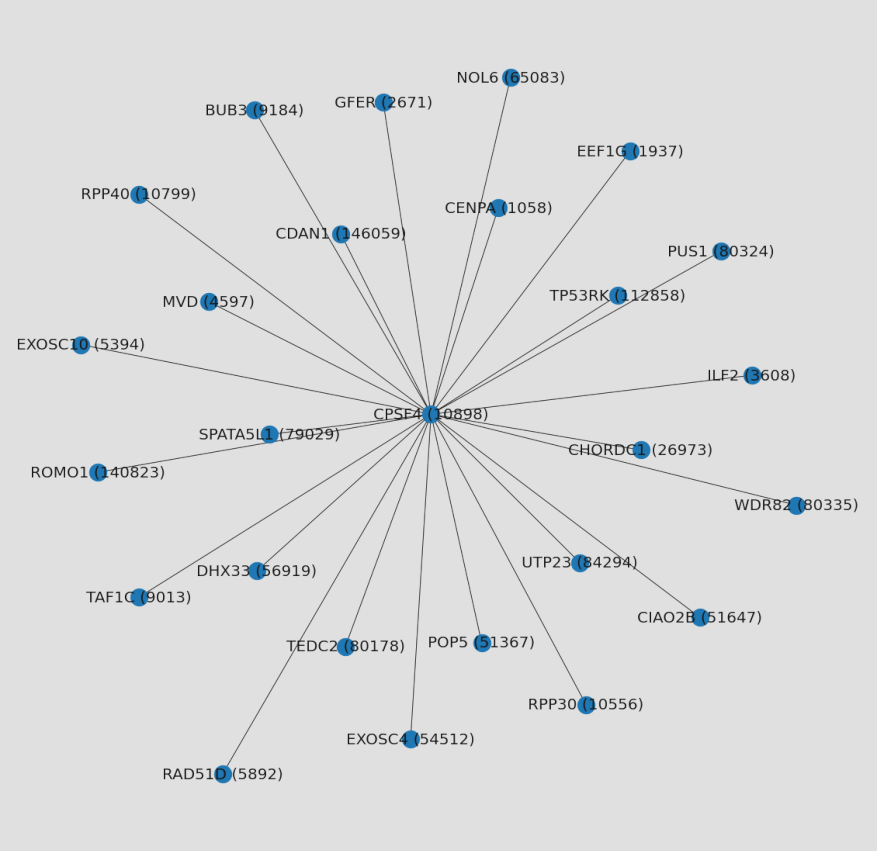


Figure 32: Genes with shortest distance have higher co-dependency with CPSF4.

Figure 32 displays the CPSF4 dependency network, with genes that their effect is more highly correlated (Spearman) with the effect of the target gene being closer. This could lead to a systematic redirection to alternative targets which could be beneficial in elucidating alternative options for every possible targeted gene. The clusters of all target genes (for all runs) are available in the supplemental material.

4.4 Evaluation based on objectives

This section evaluates the success of the initial objectives and discusses the challenges faced where applicable.

4.4.1 Primary

Primary 1: Successful. The data analysis performed (section 3.2) checked the distribution of sensitivity of the cell lines across the three treatment hours and plotted the distributions of the sensitivity in different origins. Through critical evaluation, I identified that the drug is more likely to be sensitive in *breast* and *lung* origins and more resistant to *bladder, brain, kidney,* and *bone.*

Primary 2: Successful. In-depth correlation analysis (Pearson, Spearman, MIC) was performed for all the datasets for *IC50*. When starting the project, I set as objective to perform correlation analysis using all sensitivity metrics (GI50, AUC, EC50), however data analysis and further

research elucidated that the metrics are highly correlated between them and that *IC50* is the principal metric, with the others rarely being used. Potential biomarker genes with the most predictive ability were identified using the ML predictive models developed. The development of the models was particularly challenging as any sophisticated model overfitted the data but failed to generalise effectively on the validation set. Other challenges included missing values both in terms of values for the features (gene expressions) but also sensitivity (IC50). Imputing the data (based on the training set) could damage the validity of the dataset but the alternative, dropping the samples with missing values, would further reduce the number of samples. The first approach was preferred as it is the most common approach in the literature for similar problems. Despite these challenges, extensive efforts were made to develop a pipeline that is primarily methodologically correct and can predict sensitivity values with a high correlation with the actual values. The workflow developed appears promising but more rigorous testing on additional cell lines is required to derive safe conclusions. For the mutation dataset, the LOBICO approach was used which performed poorly on the validation and holdout sets. This could be attributed to the small number of samples used to derive the optimal logic functions and the divergence made in the binarization threshold calculation.

Primary 3: Successful. The risk for the curse of dimensionality was initially identified and therefore dimensionality reduction techniques were used before clustering. Two separate techniques, PCA with Spherical K-Means, and UMAP with HDBSCAN, were used to identify clusters. Profiling the second approach's clusters revealed differences in the two clusters' sensitivity levels and genes with significant differences in their APA levels. The APA dataset contained data only for 53 out of our 95 cell lines, making the process more challenging and less reliable.

4.4.2 Secondary

Secondary 1: Successful. Besides gene expressions and APA, the mutations, gene effect, and proteomics datasets were also explored. This involved in-depth correlation analysis and biomarker discovery. A key component was also the gene effect clustering, which can be utilised to discover alternative druggable targets. A major challenge was that all the datasets were missing a considerable amount of our cell lines. This showcases how even large-scale initiatives with multi-million pounds investments, such as DepMap, still contain incomplete data.

Secondary 2: Successful. Following a discussion with Dr Arandjelović, he suggested using the separate models for each omics individually due to the limited number of features being a disadvantage for learning any kind of relationship between the multiple omics' datasets. Considering the advice, I decided that the best approach to satisfy this objective would be to average the predictions made by each of the methods and derive to a final weighted average value. This solution can be improved as the different ML models cannot share knowledge or leverage complementarity information between them at any stage throughout the learning

process and therefore cannot capture inter-omics interactions. However, given the limitations of our dataset, I believe that this was the best possible solution.

Secondary 3: Partially successful. Dr Arandjelović and Professor David Harrison advice not exploring this objective in detail due to the scarcity of data and limited time. Nonetheless, correlation analysis of two interesting drug combinations with NUC-7738 was performed. This identified that when the NUC-7738 drug is used in combination with Paclitaxel, the HINT1 gene is more significantly positively correlated to *IC50.* This is an important discovery as HINT1 is essential to activate NUC-7738, thus suggesting that the Paclitaxel+NUC-7738 combination might be more effective than NUC-7738 as a single agent.

4.5 Conclusion

This section outlines final remarks about the significance of this project, learning outcomes, and a discussion of potential future enhancements.

4.5.1 Significance

Through statistical and ML methods I identified origins that are more sensitive to the drug, genes that are highly correlated with drug efficacy, and potential biomarkers in different omics datasets. This is critical since it provides mechanistic insights into how the drug works and in which scenarios it performs best. Furthermore, I developed predictive models which can predict sensitivity or resistance to the drug, in isolation, and combination (multi-omics). A review of the literature revealed that most of the biomarkers obtained are associated with cancer, demonstrating the effectiveness of this approach. Correlation analysis with two other drugs identified that the key HINT1 gene is more highly expressed when NUC-7738 is used in combination with Paclitaxel. Finally, genes that share similar dependency profiles to the drug's possible target genes were also identified. This project can be the steppingstone in utilising precision medicine for the NUC-7738 drug, enabling the treatment of patients according to their molecular characteristics. The effect of successful deployment of such a system can be invaluable, as it can increase the chances of successful treatment, saving lives.

4.5.2 Final reflections and future work

Due to lack of prior experience with such high-dimensional datasets, most of the feature selection and dimensionality reduction techniques used were novel to me. Another challenge posed by the project was the requirement for a thorough understanding of advanced biomedical concepts that far exceeded the level of knowledge expected of an artificial intelligence master's student. Due to its large scale, the project was also extremely programming demanding, necessitating me to be systematic and methodical. The project's scale also caused challenges in the write up of this report, as I could not expand on every topic as much as I would like. For example, If I had more words available, I would also discuss the biomarkers discovered using the Kathad et al. [13] approach. Reflecting on my learning experience, I significantly increased my knowledge of omics and ML, while gaining valuable research experience.

Despite the project's success, there is still room for improvement. For example, if more data becomes accessible, more complex techniques could be used for developing and integrating the models. If I had more time, I would also build predictive models and identify biomarkers for the other treatment hours available. Furthermore, I would also explore the drug combination section in greater detail by finding correlation with more promising drugs. My vision for this project is the creation of a web platform that would enable clinicians to utilise the predictive ability of the developed models. I made an initial attempt to develop this, but owing to time constraints, I was unable to complete it (see Appendix M). Concluding, this project was a great learning odyssey that enabled me to apply my skills in a meaningful domain. I truly enjoyed it.

References

[1] National Cancer Institute. n.d. What Is Cancer?. [online] Available at: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] National Cancer Institute. n.d. NCI Dictionary of Cancer Terms. [online] Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/benign

[3] National Cancer Institute. n.d. NCI Dictionary of Cancer Terms. [online] Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis

[4] Mayoclinic.org. n.d. Cancer treatment - Mayo Clinic. [online] Available at: https://www.mayoclinic.org/tests-procedures/cancer-treatment/about/pac-20393344

[5] Biobide. n.d. The Drug Discovery Process: What Is It and Its Major Steps. [online] Available at: https://blog.biobide.com/the-drug-discovery-process

[6] Collins, F. and Varmus, H., 2015. A New Initiative on Precision Medicine. New England Journal of Medicine, 372(9):793-795.

[7] World Health Organisation. 2022. Cancer. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/cancer

[8] National Cancer Institute. 2015. Surgery for Cancer. [online] Available at: https://www.cancer.gov/about-cancer/treatment/types/surgery

[9] National Cancer Institute. 2019. Radiation Therapy for Cancer. [online] Available at: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy#RTCCSE

[10] National Cancer Institute. 2022. Radiation Therapy Side Effects. [online] Available at: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy/side-effects

[11] Rafique, R., Islam, S. and Kazi, J., 2021. Machine learning in the prediction of cancer therapy. Computational and Structural Biotechnology Journal, 19:4003-4017.

[12] National Cancer Institute. n.d. NCI Dictionary of Cancer Terms. [online] Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker

[13] Kathad, U., Kulkarni, A., McDermott, J., Wegner, J., Carr, P., Biyani, N., Modali, R., Richard, J., Sharma, P. and Bhatia, K., 2021. A machine learning-based gene signature of response to the novel alkylating agent LP-184 distinguishes its potential tumor indications. BMC Bioinformatics, 22(1):102.

[14] Lauschke, V., Milani, L. and Ingelman-Sundberg, M., 2017. Pharmacogenomic Biomarkers for Improved Drug Therapy—Recent Progress and Future Developments. The AAPS Journal, 20(1):4.

[15] Costello, J., Heiser, L., Georgii, E., Gönen, M., Menden, M., Wang, N., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S., Mpindi, J., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J. and Stolovitzky, G., 2014. A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology, 32(12):1202-1212.

[16] Pineda, A., Ogoe, H., Balasubramanian, J., Rangel Escareño, C., Visweswaran, S., Herman, J. and Gopalakrishnan, V., 2016. On Predicting lung cancer subtypes using 'omic' data from tumor and tumor-adjacent histologically-normal tissue. BMC Cancer, 16(1):184.

[17] Tarek, S., Abd Elwahab, R. and Shoman, M., 2017. Gene expression based cancer classification. Egyptian Informatics Journal, 18(3):151-159.

[18] Danaee, P., Ghaeini, R. and Hendrix, D., 2016. A deep learning approach for cancer detection and relevant gene identification. Biocomputing 2017:219-229.

[19] Zhang, Q., Li, J., Tan, X. and Zhao, Q., 2019. Effects of ME3 on the proliferation, invasion and metastasis of pancreatic cancer cells through epithelial-mesenchymal transition. Neoplasma, 66(06):896-907

[20] Collaborative Drug Discovery Inc. (CDD). 2018. What is pIC50?. [online] Available at: https://www.collaborativedrug.com/what-is-pic50-2/

[21] Ren, F., Zhang, N., Zhang, L., Miller, E. and Pu, J., 2020. Alternative Polyadenylation: a new frontier in post transcriptional regulation. Biomarker Research, 8(1):67.

[22] Zhong, W., Wu, Y., Zhu, M., Zhong, H., Huang, C., Lin, Y. and Huang, J., 2022. Alternative splicing and alternative polyadenylation define tumor immune microenvironment and pharmacogenomic landscape in clear cell renal carcinoma. Molecular Therapy - Nucleic Acids, 27:927-946.

[23] Li, G. and Xie, X., 2011. Central dogma at the single-molecule level in living cells. Nature, 475(7356):308-315.

[24] Reel, P., Reel, S., Pearson, E., Trucco, E. and Jefferson, E., 2021. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49:107739.

[25] Yan, J., Risacher, S., Shen, L. and Saykin, A., 2017. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Briefings in Bioinformatics, 19(6):1370-1381

[26] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. and Milanesi, L., 2016. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17:15

[27] Alberts, B., 2007. Molecular biology of the cell. Enskede: TPB.

[28] Ma, B., Meng, F., Yan, G., Yan, H., Chai, B. and Song, F., 2020. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in Biology and Medicine, 121:103761.

[29] Antonelli, Claggett, Henglin, Kim, Ovsak, Kim, Deng, Rao, Tyagi, Watrous, Lagerborg, Hushcha, Demler, Mora, Niiranen, Pereira, Jain and Cheng, 2019. Statistical Workflow for Feature Selection in Human Metabolomics Data. Metabolites, 9(7):143

[30] Liew, A., Law, N. and Yan, H., 2010. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics, 12(5):498-513.

[31] Güvenç Paltun, B., Kaski, S. and Mamitsuka, H., 2021. Machine learning approaches for drug combination therapies. Briefings in Bioinformatics, 22(6):1-16.

[32] Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. and Shyr, Y., 2013. Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. PLoS ONE, 8(8):e71462.

[33] Fujita, N., Mizuarai, S., Murakami, K. and Nakai, K., 2018. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Scientific Reports, 8(1);9743.

[34] Bavafaye Haghighi, E., Knudsen, M., Elmedal Laursen, B. and Besenbacher, S., 2019. Hierarchical Classification of Cancers of Unknown Primary Using Multi-Omics Data, 18:1176935119872163.

[35] Shen, H. and Chou, K., 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics, 22(14):1717-1722

[36] Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H. and Dawood, H., 2019. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinformatics, 20:527.

[37] Chaudhary, K., Poirion, O., Lu, L. and Garmire, L., 2018. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clinical Cancer Research, 24(6):1248-1259

[38] The Data School. 2021. Correlation and P value. [online] Available at: https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/

[39] Zito, A., Lualdi, M., Granata, P., Cocciadiferro, D., Novelli, A., Alberio, T., Casalone, R. and Fasano, M., 2021. Gene Set Enrichment Analysis of Interaction Networks Weighted by Node Centrality. Frontiers in Genetics, 12:577623.

[40] Web.archive.org. n.d. IC50 versus EC50. [online] Available at: https://web.archive.org/web/20170528053210/https://www.fda.gov/ohrms/dockets/ac/00/slides/3621s1d/sld036.htm

[41] Stransky, B. and de Souza, S., 2013. Gene Expression Biomarkers. Encyclopedia of Systems Biology.

[42] Ox.ac.uk. 2021. Anti-cancer drug derived from fungus shows promise in clinical trials | University of Oxford. [online] Available at: https://www.ox.ac.uk/news/2021-10-08-anti-cancer-drug-derived-fungus-shows-promise-clinical-trials

[43] Xia, Z., Donehower, L., Cooper, T., Neilson, J., Wheeler, D., Wagner, E. and Li, W., 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nature Communications, 5:5274.

[44] Xiang, Y., Ye, Y., Lou, Y., Yang, Y., Cai, C., Zhang, Z., Mills, T., Chen, N., Kim, Y., Muge Ozguc, F., Diao, L., Karmouty-Quintana, H., Xia, Y., Kellems, R., Chen, Z., Blackburn, M., Yoo, S., Shyu, A., Mills, G. and Han, L., 2017. Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. JNCI: Journal of the National Cancer Institute, 110(4):379-389.

[45] Mayr, C. and Bartel, D., 2009. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells, 138(4):673-84.

[46] Park, H., Ji, P., Kim, S., Xia, Z., Rodriguez, B., Li, L., Su, J., Chen, K., Masamha, C., Baillat, D., Fontes-Garfias, C., Shyu, A., Neilson, J., Wagner, E. and Li, W., 2018. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk, 50:783-789.

[47] Masamha, C., Xia, Z., Yang, J., Albrecht, T., Li, M., Shyu, A., Li, W. and Wagner, E., 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature, 510:412-416.

[48] Derti, A., Garrett-Engele, P., MacIsaac, K., Stevens, R., Sriram, S., Chen, R., Rohl, C., Johnson, J. and Babak, T., 2012. A quantitative atlas of polyadenylation in five mammals. Genome Research, 22(6):1173-83

[49] Kuo, F. and Sloan, I., 2005. Lifting the curse of dimensionality. Notices Am Math Soc.

[50] Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U., 1999. When Is "Nearest Neighbor" Meaningful?. Lecture Notes in Computer Science, 1540:217-35.

[51] Dabbura, I., 2018. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. [online] Medium. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[52] Strehl, A., Ghosh, J., Mooney, R., 2001. Impact of Similarity Measures on Web-page Clustering. In Workshop on Artificial Intelligence for Web Search (AAAI), 2000:58-64.

[53] Shen, Y., Zhang, Y., Xue, W. and Yue, Z., 2021. dbMCS: A Database for Exploring the Mutation Markers of Anti-Cancer Drug Sensitivity. IEEE Journal of Biomedical and Health Informatics, 25(11):4229-4237.

[54] Tang, Y. and Gottlieb, A., 2021. Explainable drug sensitivity prediction through cancer pathway enrichment. Scientific Reports, 11(1):3128.

[55] Sepia2.unil.ch. n.d. Area under the Curve – Pharmacokinetics. [online] Available at: https://sepia2.unil.ch/pharmacology/parameters/areaunderthecurve/

[56] Kira, K. and Rendell, L., 1992. The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings, 129-134.

[57] Kursa, M. and Rudnicki, W., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11):1-13.

[58] Jolliffe, I. and Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202.

[59] Singh, A., 2012. Lecture 13: Minimum Description Length. [online] Cs.cmu.edu. Available at: https://www.cs.cmu.edu/~aarti/Class/10704/lec13-MDL.pdf

[60] Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11): 559-572.

[61] McInnes, L., Healy, J., Saul, N. and Großberger, L., 2018. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29):861.

[62] Belkin, M. and Niyogi, P., 2002. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Advances in Neural Information Processing Systems 14, 14(6):585–591.

[63] Coenen, A. and Pearce, A., n.d. Understanding UMAP. [online] Pair-code. Available at: https://pair-code.github.io/understanding-umap/index.html

[64] Berba, P., 2020. Understanding HDBSCAN and Density-Based Clustering. [online] pepe berba. Available at: https://pberba.github.io/stats/2020/01/17/hdbscan/

[65] HDBSCAN Documentation. n.d. Parameter Selection for HDBSCAN* — hdbscan 0.8.1 documentation. [online] Available at: https://hdbscan.readthedocs.io/en/latest/parameter_selection.html

[66] Integrated DNA technologies. n.d. Gene knockout. [online] Available at: https://eu.idtdna.com/pages/applications/gene-knockout

[67] Pandey, A. and Jain, A., 2017. Comparative Analysis of KNN Algorithm using Various Normalization Techniques. International Journal of Computer Network and Information Security, 9(11):36-42.

[68] Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[69] Hackstadt, A. and Hess, A., 2009. Filtering for increased power for microarray data analysis. BMC Bioinformatics, 10:11

[70] Schneider, J., 1997. Cross Validation. [online] Cs.cmu.edu. Available at: https://www.cs.cmu.edu/~schneide/tut5/node42.html

[71] Picard, M., Scott-Boyer, M., Bodein, A., Périn, O. and Droit, A., 2021. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal, 19:3735-3746

[72] Nusinow, D., Szpyt, J., Ghandi, M., Rose, C., McDonald, E., Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D., Jedrychowski, M., Golji, J., Porter, D., Rejtar, T., Wang, Y., Kryukov, G., Stegmeier, F., Erickson, B., Garraway, L., Sellers, W. and Gygi, S., 2020. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell, 180(2), pp.387-402.e16.

[73] Schwenzer, H., De Zan, E., Elshani, M., van Stiphout, R., Kudsy, M., Morris, J., Ferrari, V., Um, I., Chettle, J., Kazmi, F., Campo, L., Easton, A., Nijman, S., Serpi, M., Symeonides, S., Plummer, R., Harrison, D., Bond, G. and Blagden, S., 2021. The Novel Nucleoside Analogue ProTide NUC-7738 Overcomes Cancer Resistance Mechanisms In Vitro and in a First-In-Human Phase I Clinical Trial. Clinical Cancer Research, 27(23):6500-6513.

[74] Nucana.com. n.d. NUC-7738 - A transformation of 3'-deoxyadenosine (3'-dA). [online] Available at: http://www.nucana.com/nuc7738.html

[75] Shimizu, K., Matsumoto, H., Hirata, H., Ueno, K., Samoto, M., Mori, J., Fujii, N., Kawai, Y., Inoue, R., Yamamoto, Y., Yano, S., Shimabukuro, T., Furutani-Seiki, M. and Matsuyama, H., 2020. ARHGAP29 expression may be a novel prognostic factor of cell proliferation and invasion in prostate cancer, 44(6):2735-2745.

[76] Kolb, K., Hellinger, J., Kansy, M., Wegwitz, F., Bauerschmitz, G., Emons, G. and Gründker, C., 2020. Influence of ARHGAP29 on the Invasion of Mesenchymal-Transformed Breast Cancer Cells. Cells, 9(12):2616.

[77] Jiao, Y., Li, Y., Liu, S., Chen, Q. and Liu, Y., 2019. ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer. OncoTargets and Therapy, 12:4141-4152.

[78] Zhang, G., Li, B. and Lin, Y., 2022. Evaluation of ITGA3 as a Biomarker of Progression and Recurrence in Papillary Thyroid Carcinoma. Frontiers in Oncology, 11:614955.

[79] Zhou, F., Shen, D., Xiong, Y., Cheng, S., Xu, H., Wang, G., Qian, K., Ju, L. and Zhang, X., 2021. CTHRC1 Is a Prognostic Biomarker and Correlated With Immune Infiltrates in Kidney Renal Papillary Cell Carcinoma and Kidney Renal Clear Cell Carcinoma. Frontiers in Oncology, 10:570819.

[80] Zhang, X., Hu, L., Yang, Q., Qin, W., Wang, X., Xu, C., Tian, G., Yang, X., Yao, L., Zhu, L., Nie, H., Li, Q., Xu, Q., Zhang, Z., Zhang, Y., Li, J., Wang, Y. and Jiang, S., 2021. CTHRC1 promotes liver metastasis by reshaping infiltrated macrophages through physical interactions with TGF-β receptors in colorectal cancer. Oncogene, 40:3959-3973.

[81Meng, C., Zhang, Y., Jiang, D. and Wang, J., 2022. CTHRC1 is a prognosis-related biomarker correlated with immune infiltrates in colon adenocarcinoma. World Journal of Surgical Oncology, 20:89.

[82] Feng, Z., Li, K., Wu, Y. and Peng, C., 2021. Transcriptomic Profiling Identifies DCBLD2 as a Diagnostic and Prognostic Biomarker in Pancreatic Ductal Adenocarcinoma. Frontiers in Molecular Biosciences, 8:659168.

[83] Chen, X., Lv, Y., Xu, K., Wang, X., Zhao, Y., Li, J., Qin, X., Shi, Y., Wang, L., Chang, A., Huang, C. and Xiang, R., 2021. DCBLD2 Mediates Epithelial-Mesenchymal Transition-Induced Metastasis by Cisplatin in Lung Adenocarcinoma. Cancers, 13(6):1403.

[84] Liu-Chittenden, Y., Jain, M., Gaskins, K., Wang, S., Merino, M., Kotian, S., Kumar Gara, S., Davis, S., Zhang, L. and Kebebew, E., 2017. RARRES2 functions as a tumor suppressor by promoting β-catenin phosphorylation/degradation and inhibiting p38 phosphorylation in adrenocortical carcinoma. Oncogene, 36(25):3541-3552.

[85] Chen, C., Lin, Y., Chen, C. and Chen, Y., 2018. Annexin A2-mediated cancer progression and therapeutic resistance in nasopharyngeal carcinoma. Journal of Biomedical Science, 25(1):30.

[86] Wang, T., Wang, Z., Niu, R. and Wang, L., 2019. Crucial role of Anxa2 in cancer progression: highlights on its novel regulatory mechanism. Cancer Biology & Medicine, 16(4):671-687.

[87] Frazzi, R., 2021. BIRC3 and BIRC5: multi-faceted inhibitors in cancer. Cell, Bioscience, 11(1):8.

[88] National Cancer Institute. 2022. HLA Gene May Predict if Cancer Immunotherapy Will Work. [online] Available at: https://www.cancer.gov/news-events/cancer-currents-blog/2022/immunotherapy-cancer-biomarker-hla-gene

[89] Michelakos, T., Kontos, F., Kurokawa, T., Cai, L., Sadagopan, A., Krijgsman, D., Weichert, W., Durrant, L., Kuppen, P., R Ferrone, C. and Ferrone, S., 2022. Differential role of HLA-A and HLA-B, C expression levels as prognostic markers in colon and rectal cancer. Immunotherapy Cancer, 10(3): 4115.

[90] Hopkinsmedicine.org. n.d. BRAF Mutation and Cancer. [online] Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/braf-mutation-and-cancer

[91] Shimada, K., Bachman, J., Muhlich, J. and Mitchison, T., 2021. shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. eLife, 10:e57116

[92] Knijnenburg, T., Klau, G., Iorio, F., Garnett, M., McDermott, U., Shmulevich, I. and Wessels, L., 2016. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Scientific Reports, 6(1):36812.

[93] Li, P., Xiao, J., Zhou, B., Wei, J., Luo, J. and Chen, W., 2020. SYNE1 mutation may enhance the response to immune checkpoint blockade therapy in clear cell renal cell carcinoma patients. Aging, 12(19):19316-19324.

[94] Adamska, A., Ferro, R., Lattanzio, R., Capone, E., Domenichini, A., Damiani, V., Chiorino, G., Akkaya, B., Linton, K., De Laurenzi, V., Sala, G. and Falasca, M., 2019. ABCC3 is a novel target for the treatment of pancreatic cancer. Advances in Biological Regulation, 73:100634.

[95] Hunt, B., Wicker, C., Bourn, J., Lower, E., Takiar, V. and Waltz, S., 2020. MST1R (RON) expression is a novel prognostic biomarker for metastatic progression in breast cancer patients. Breast Cancer Research and Treatment, 181(3):529-540.

[96] Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M. and Sabeti, P., 2011. Detecting Novel Associations in Large Data Sets. Science, 334(6062):1518-1524.

[97] Hauke, J. and Kossowski, T., 2011. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. QUAGEO, 30(2):87-93.

[98] Xu, Y., Liu, X., Zhang, H., Zhu, Z., Wu, X., Wu, X., Li, S., Song, L. and Xu, X., 2018. Overexpression of HES6 has prognostic value and promotes metastasis via the Wnt/β-catenin signaling pathway in colorectal cancer. Oncology Reports, 40(3):1261-1274.

[99] Rodriguez, E., Maroufy, V., Zheng, W., Wu, H. and Soltanalizadeh, B., 2020. Modelling of hypoxia gene expression for three different cancer cell lines. International Journal of Computational Biology and Drug Design, 13(1):124-143

[100] Cao, D., Xue, J., Huang, G., An, J. and An, W., 2022. The role of splicing factor PRPF8 in breast cancer. Technology and Health Care, 30(S1):293-301.

[101] Xu, X., Zheng, Z., Jia, L., Suo, S., Liu, B., Shao, T., Tu, Q., Hua, Y. and Xu, H., 2018. Overexpression of SMARCA2 or CAMK2D is associated with cisplatin resistance in human epithelial ovarian cancer. Oncology Letters, 16(3):3796-3804.

[102] Zhu, Y., Pu, Z., Wang, G., Li, Y., Wang, Y., Li, N. and Peng, F., 2021. FAM3C: an emerging biomarker and potential therapeutic target for cancer. Biomarkers in Medicine, 15(5):373-384.

[103] Kazantseva, J., Sadam, H., Neuman, T. and Palm, K., 2016. Targeted alternative splicing of TAF4: a new strategy for cell reprogramming. Scientific Reports, 6(1):30852.

[104] Wang, C., Zhang, R., Wang, X., Zheng, Y., Jia, H., Li, H., Wang, J., Wang, N., Xiang, F. and Li, Y., 2021. Silencing of KIF3B Suppresses Breast Cancer Progression by Regulating EMT and Wnt/β-Catenin Signaling. Frontiers in Oncology, 10:597464.

[105] Sohn, M., Shin, S., Yoo, J., Goh, Y., Lee, I. and Bae, Y., 2018. Ahnak promotes tumor metastasis through transforming growth factor-β-mediated epithelial-mesenchymal transition. Scientific Reports, 8(1):14379.

[106] Hartsough, E., Weiss, M., Heilman, S., Purwin, T., Kugel, C., Rosenbaum, S., Erkes, D., Tiago, M., HooKim, K., Chervoneva, I. and Aplin, A., 2019. CADM1 is a TWIST1-regulated suppressor of invasion and survival. Cell Death & Disease, 10(4):281.

[107] Min, S., Lee, Y., Hong, J., Park, T., Woo, H., Kwon, S. and Yoon, G., 2021. MRPS31 loss is a key driver of mitochondrial deregulation and hepatocellular carcinoma aggressiveness. Cell Death & Disease, 12(11):1076.

[108] Deng, Y., Zheng, X., Zhang, Y., Xu, M., Ye, C., Lin, M., Pan, J., Xu, Z., Lu, X. and Chi, P., 2020. High SPRR1A expression is associated with poor survival in patients with colon cancer. Oncology Letters, 19(5):3417-3424.

[109] Timp, W. and Timp, G., 2020. Beyond mass spectrometry, the next step in proteomics. Science Advances, 6(2):8978.

[110] Crown, J., 2017. CDK8: a new breast cancer target. Oncotarget, 8(9):14269-14270

[111] Zhang, K., Liu, L., Wang, M., Yang, M., Li, X., Xia, X., Tian, J., Tan, S. and Luo, L., 2020. A novel function of IMPA2, plays a tumor-promoting role in cervical cancer. Cell Death & Disease, 11(5):371.

[112] Awad, M. and Khanna, R., 2015. Support Vector Regression. Efficient Learning Machines, 67-80.

Appendices

Appendix A – Developed Jupyter Notebooks overview

Multiple notebooks were developed to separate concerns. The developed software is available in the *software & supporting material* folder.

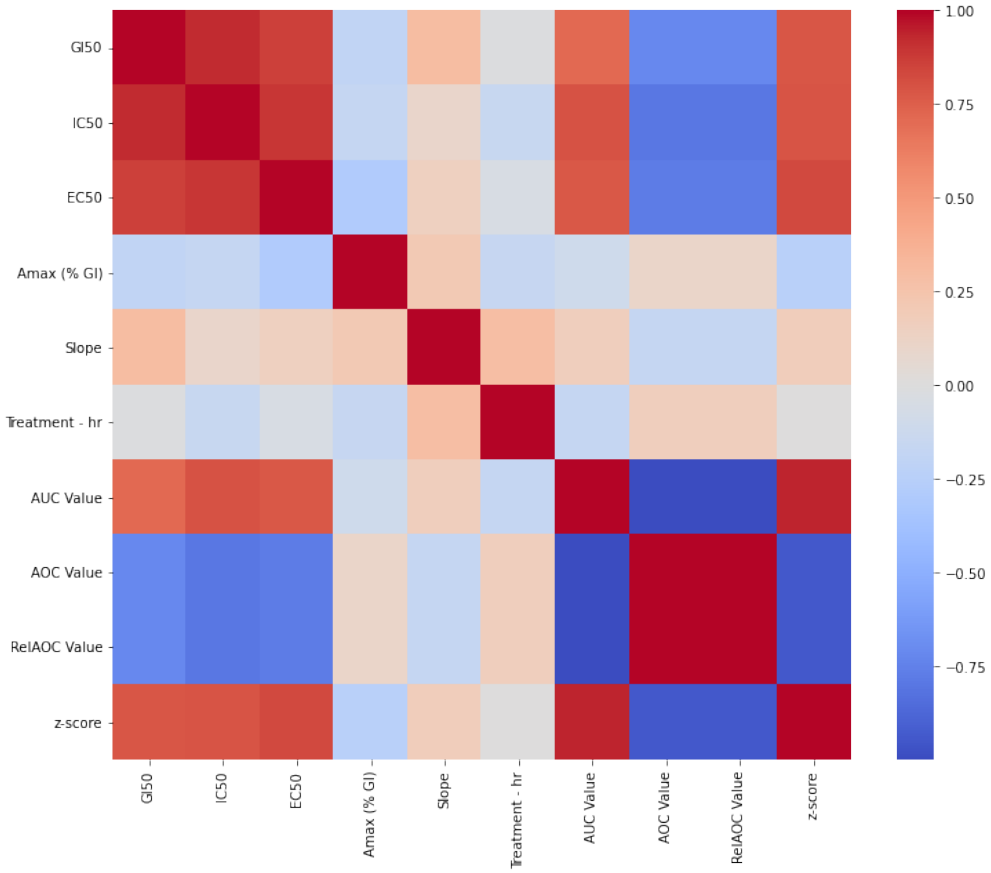
Language	Notebook	Content
Python	Data analysis - NUC-7738.ipynb	Analysis performed on the NUC-7738 dataset (Section 3.2).
	Correlation Analysis Pipeline.ipynb	Correlation analysis for all the datasets, for all three treatment hours (Section 3.3.1 & Section 3.5).
	Identifying Biomarkers Gene Expressions using Kathad et al approach.ipynb	Development of the method proposed by Kathad et al. (Section 3.3.2)
	Development of a Biomarkers Pipeline.ipynb	Implementation of the alteration technique proposed for biomarker discovery (Section 3.3.3). Biomarker discovery for all omics datasets.
	Clustering IC50 with PDUI - APA.ipynb	Cluster analysis performed for the APA dataset (Section 3.4).
	Exploring Mutations.ipynb	Exploration of the mutations dataset (Section 3.5.1.4)
	Clustering dependent genes.ipynb	Clustering genes that are dependent between them based on their gene effect (Section 3.5.1.3)
	Helper Functions.py	Functions that are commonly/sharely used by other notebooks.
Matlab	Mutations_NUC7738.m	Code to run LOBICO on the exported (pre-processed) mutations dataset from the Development of a Biomarkers Pipeline notebook (Section 3.5.1.4).

Appendix B – NUC-7738 Metrics Examination

The following table describes the most important metrics from the NUC-7738 dataset.

Metric	Description
GI50	The dose that inhibits the growth of cells by 50%.
EC50	The plasma concentration required for obtaining 50% of the maximum effect *in vivo* [40].
Treatment - hr	The number of hours after the dose was administered that the other metrics were calculated.
AUC Value	The area under the drug concentration-time curve in plasma (AUC). It is expressed in mg*h/L and reflects the actual body exposure to the drug after administration of a dose of the drug [55].
z-score	Where a cell line lies in terms of drug sensitivity in the normal distribution curve.

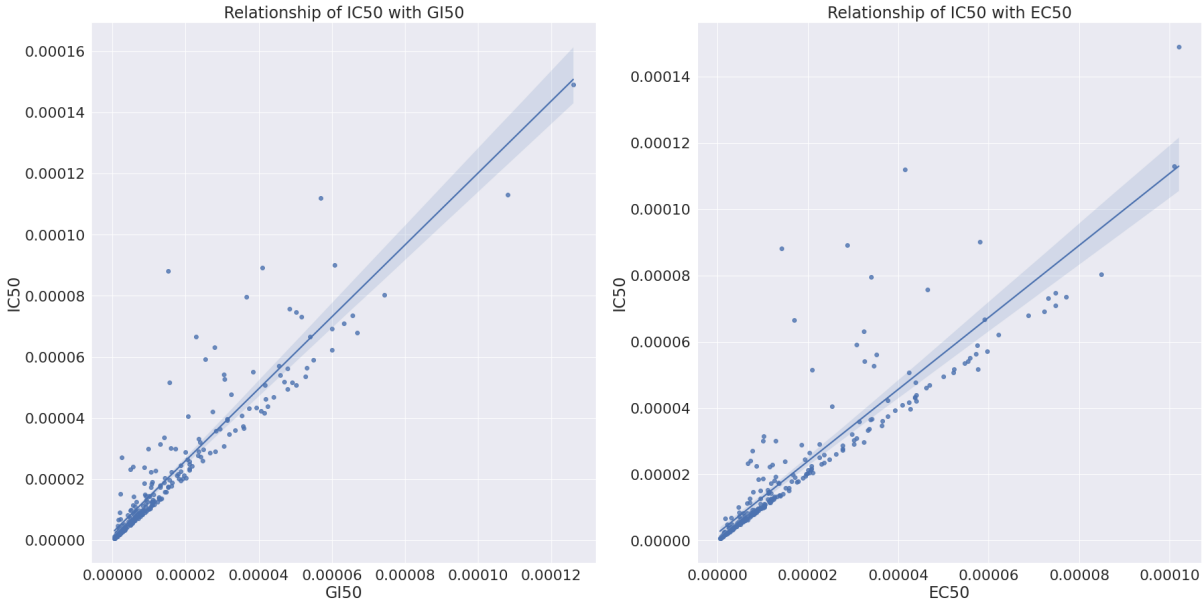
The correlation matrix in the Figure below investigates the strength of correlations between the metrics.



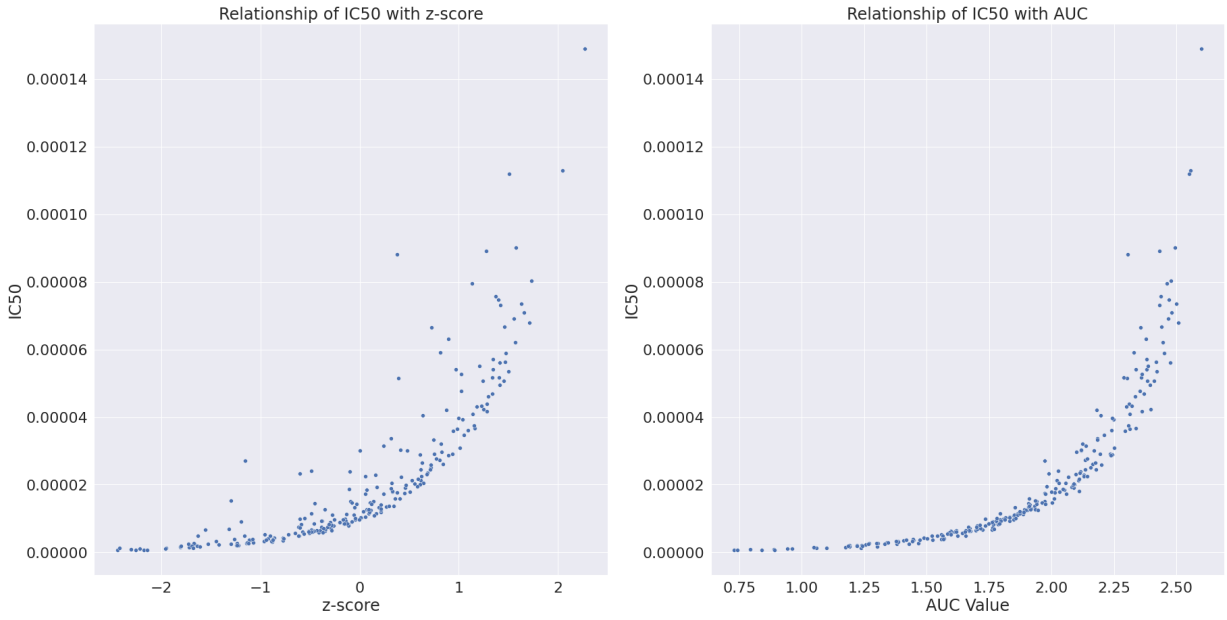
There are some significant correlations. *IC50*, *GI50*, and *EC50* are all highly positively correlated between them. Furthermore, they also have a strong positive correlation with *AUC Value* and *z-score*. *Treatment–hr* has a weak negative correlation with *IC50*, suggesting that

on higher *Treatment–hr* values, the *IC50* value is lower. The *z-score* metric is highly correlated with the *GI50*, *IC50*, *EC50*, *AUC*, *AOC*, and *RelAOC* values, indicating that it might be an appropriate metric for distinguishing sensitive and resistant cell lines.

Since *IC50* is the primary metric used to assess drug efficacy, its relationships to other metrics were further investigated (using all cell lines from the three different time points).

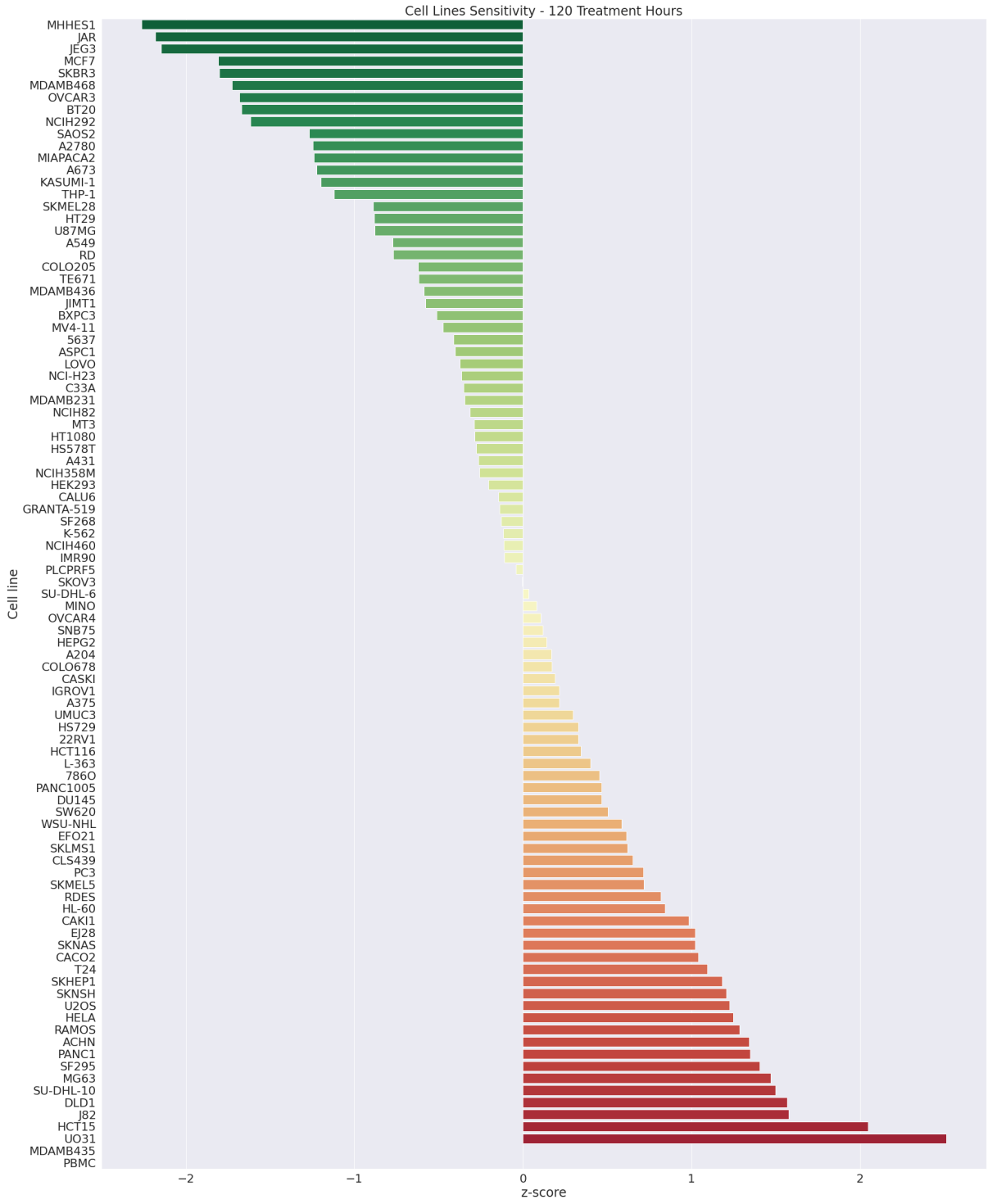


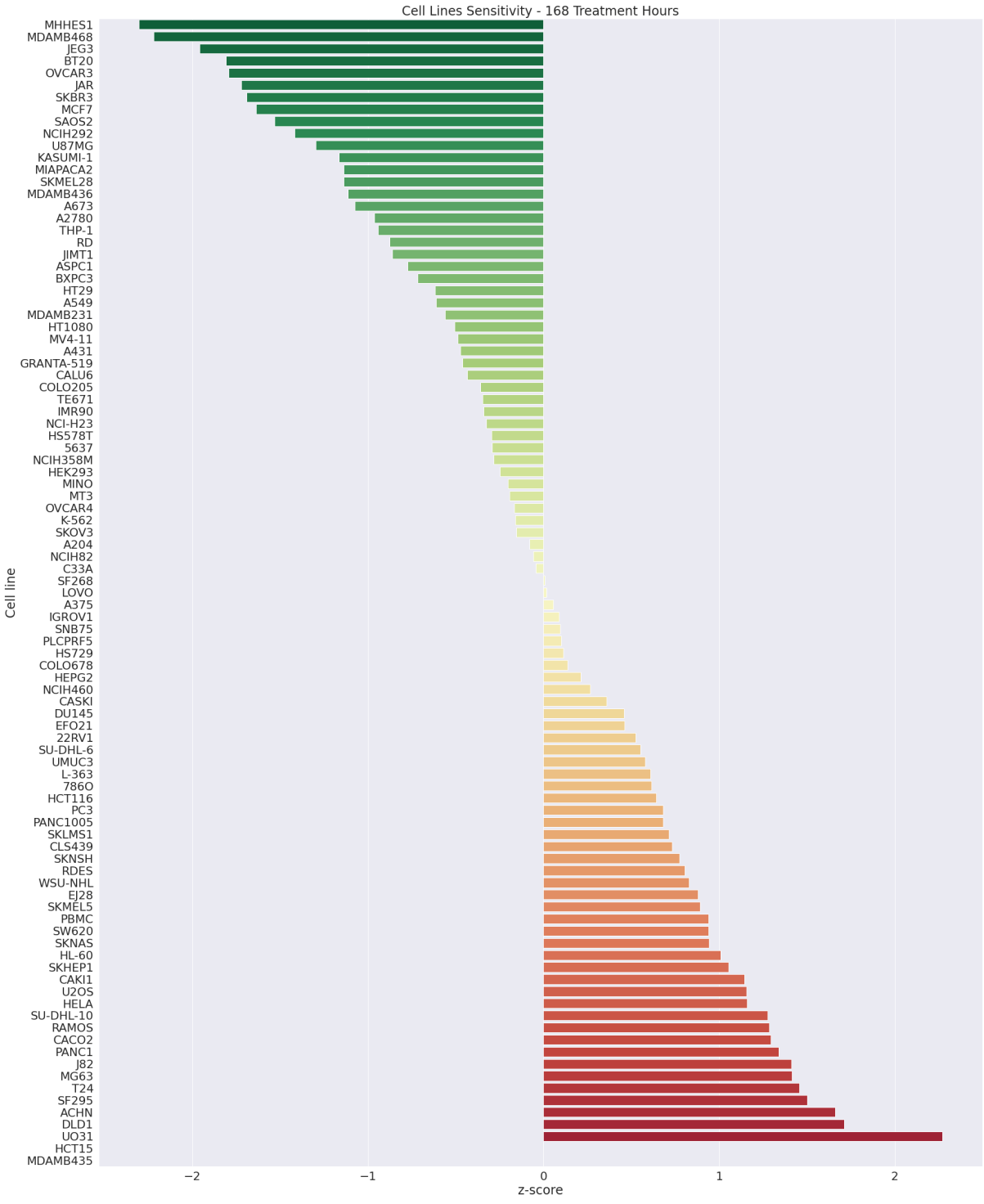
There is a linear relationship between *IC50* and both *GI50* and *EC50*, whereas *IC50* increases or decreases, so do the other two metrics.



There is an exponential growth relationship between *z-score* and *AUC* with *IC50*. As the *z-score* or *AUC* value increases, the *IC50* value increases exponentially (and vice versa).

Appendix C: Cell lines sensitivity using Z-score – 120 and 168 hours




Cell Lines Sensitivity - 168 Treatment Hours
Cell line
MHHES1
MDAMB468
JEG3
BT20
OVCAR3
JAR
SKBR3
MCF7
SAOS2
NCIH292
U87MG
KASUMI-1
MIAPACA2
SKMEL28
MDAMB436
A673
A2780
THP-1
RD
JIMT1
ASPC1
BXPC3
HT29
A549
MDAMB231
HT1080
MV4-11
A431
GRANTA-519
CALU6
COLO205
TE671
IMR90
NCI-H23
HS578T
5637
NCIH358M
HEK293
MINO
MT3
OVCAR4
K-562
SKOV3
A204
NCIH82
C33A
SF268
LOVO
A375
IGROV1
SNB75
PLCPRF5
HS729
COLO678
HEPG2
NCIH460
CASKI
DU145
EFO21
22RV1
SU-DHL-6
UMUC3
L-363
786O
HCT116
PC3
PANC1005
SKLMS1
CLS439
SKNSH
RDES
WSU-NHL
EJ28
SKMEL5
PBMC
SW620
SKNAS
HL-60
SKHEP1
CAKI1
U2OS
HELA
SU-DHL-10
RAMOS
CACO2
PANC1
J82
MG63
T24
SF295
ACHN
DLD1
UO31
HCT15
MDAMB435
−2
−1
0
1
2
z-score

Appendix D – Model Experimentation Performance (Validation Set)

Model Name	Model	Correlation Coefficient	p-value
dt_1	Decision Trees	0.213445051	0.41075473
dt_2	Decision Trees	0.481039054	0.05060532
dt_3	Decision Trees	-0.036472934	0.88947191
dt_4	Decision Trees	-0.366707113	0.14766872
dt_5	Decision Trees	-0.000645371	0.99803862
dt_6	Decision Trees	-0.036472934	0.88947191
dt_7	Decision Trees	-0.000645371	0.99803862
dt_8	Decision Trees	-0.000645371	0.99803862
dt_9	Decision Trees	-0.005915549	0.9820231
dt_10	Decision Trees	-0.005915549	0.9820231
knn_1	KNN	0.636906673	0.00596832
knn_2	KNN	0.627506941	0.00700477
knn_3	KNN	0.646145418	0.00507361
knn_4	KNN	0.526834608	0.02979073
knn_5	KNN	0.616049923	0.00845828
knn_6	KNN	0.586736284	0.01329169
knn_7	KNN	0.616049923	0.00845828
knn_8	KNN	0.485124783	0.04839809
knn_9	KNN	0.585103469	0.01361423
knn_10	KNN	0.581843773	0.01427659
rf_1	Random Forest	0.368387791	0.14568013
rf_2	Random Forest	0.357432551	0.15897664
rf_3	Random Forest	0.217201757	0.40236892
rf_4	Random Forest	0.149447697	0.5669937
rf_5	Random Forest	0.259701405	0.31411502
rf_6	Random Forest	0.251448315	0.33028387
rf_7	Random Forest	0.259701405	0.31411502
rf_8	Random Forest	0.236553029	0.36065828

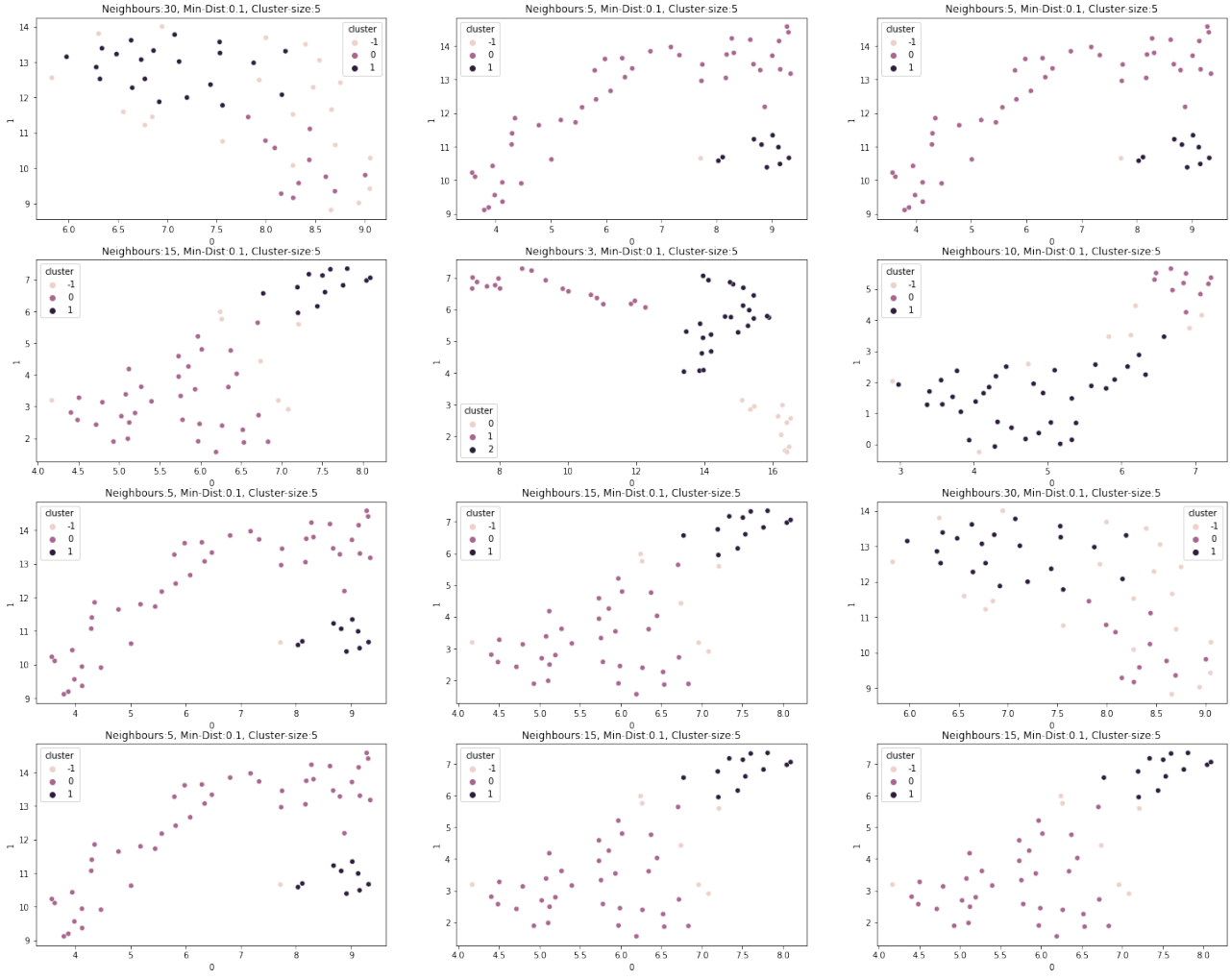
rf_9	Random Forest	0.221208463	0.39352739
rf_10	Random Forest	0.216823636	0.40320878
ridge_1	Ridge Regression	0.459118592	0.06375749
ridge_2	Ridge Regression	0.476839102	0.05295204
ridge_3	Ridge Regression	0.41846686	0.09458238
ridge_4	Ridge Regression	0.41468569	0.09791159
ridge_5	Ridge Regression	0.452384024	0.06826397
ridge_6	Ridge Regression	0.489043973	0.04634956
ridge_7	Ridge Regression	0.452384024	0.06826397
ridge_8	Ridge Regression	0.455300798	0.06628431
ridge_9	Ridge Regression	0.365867241	0.14866939
ridge_10	Ridge Regression	0.40091594	0.11074379
XGBoost_1	XGBoost	0.420474845	0.09284781
XGBoost_2	XGBoost	0.631412181	0.00655793
XGBoost_3	XGBoost	0.310319129	0.2254242
XGBoost_4	XGBoost	0.316927753	0.21517753
XGBoost_5	XGBoost	0.334452246	0.18948048
XGBoost_6	XGBoost	0.465954417	0.05941205
XGBoost_7	XGBoost	0.334452246	0.18948048
XGBoost_8	XGBoost	0.300639001	0.2409873
XGBoost_9	XGBoost	0.371044358	0.14257443
XGBoost_10	XGBoost	0.369668559	0.14417709

Appendix E: How UMAP and HDBSCAN parameters affect clustering

To investigate and visualise how different parameter values affect the clustering of the data, two functions (plot_multiple and plot_multiple_structured) were developed, with the first selecting various combinations at random, whilst the second using combinations in the order they were given in the lists.

Experiment 1 – values tried (random method)		
n_neighbours List	**min_dist list**	**min_cluster_size**
[3,5,10,15,30]	[0.1]	[5]

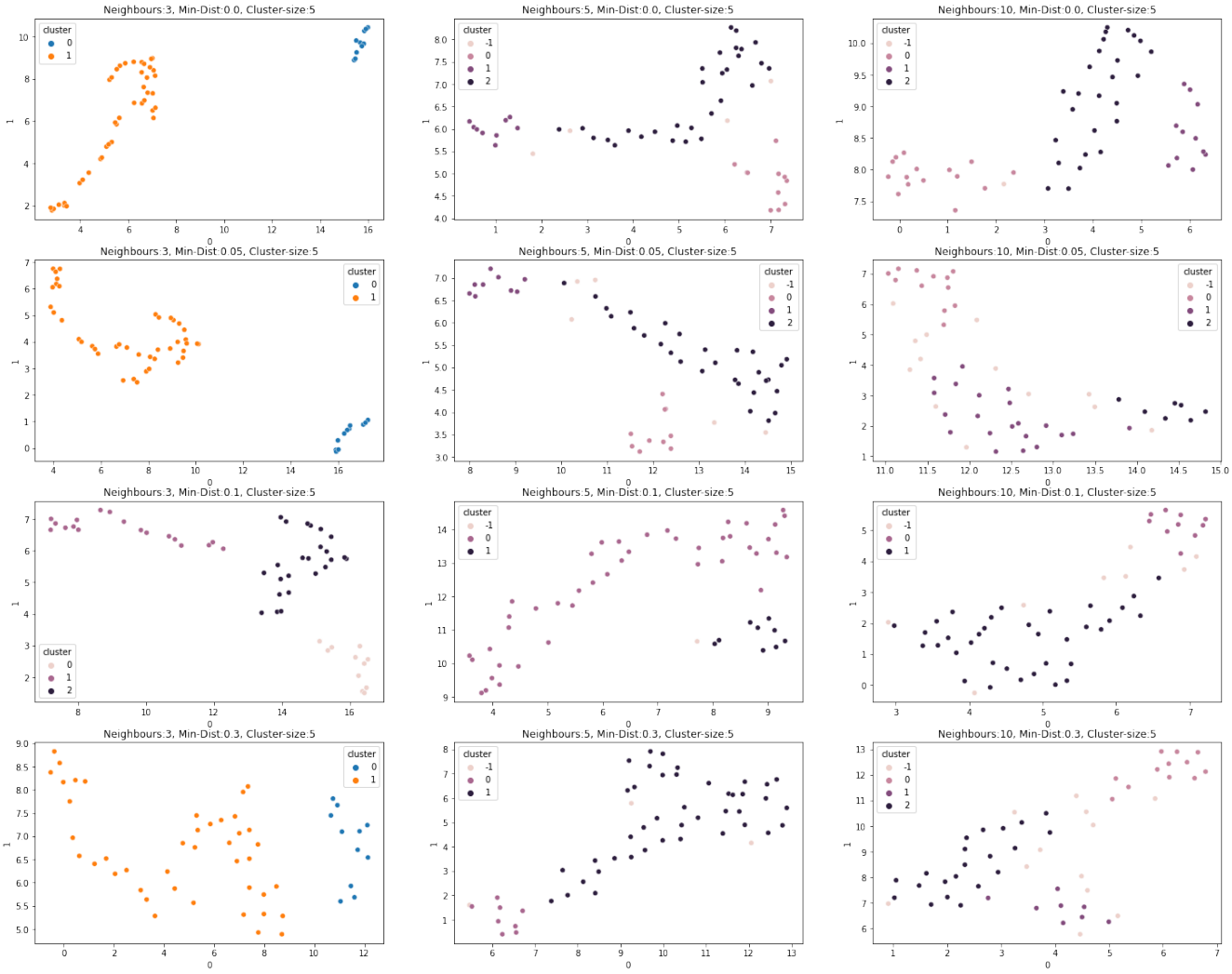
Randomly select different combinations from the given values and plot them.



Smaller values for n_neighbors seem to work better. Therefore, in next attempt I tried to find min_dist, while keeping n_neighbours to small values - [3,5,10].

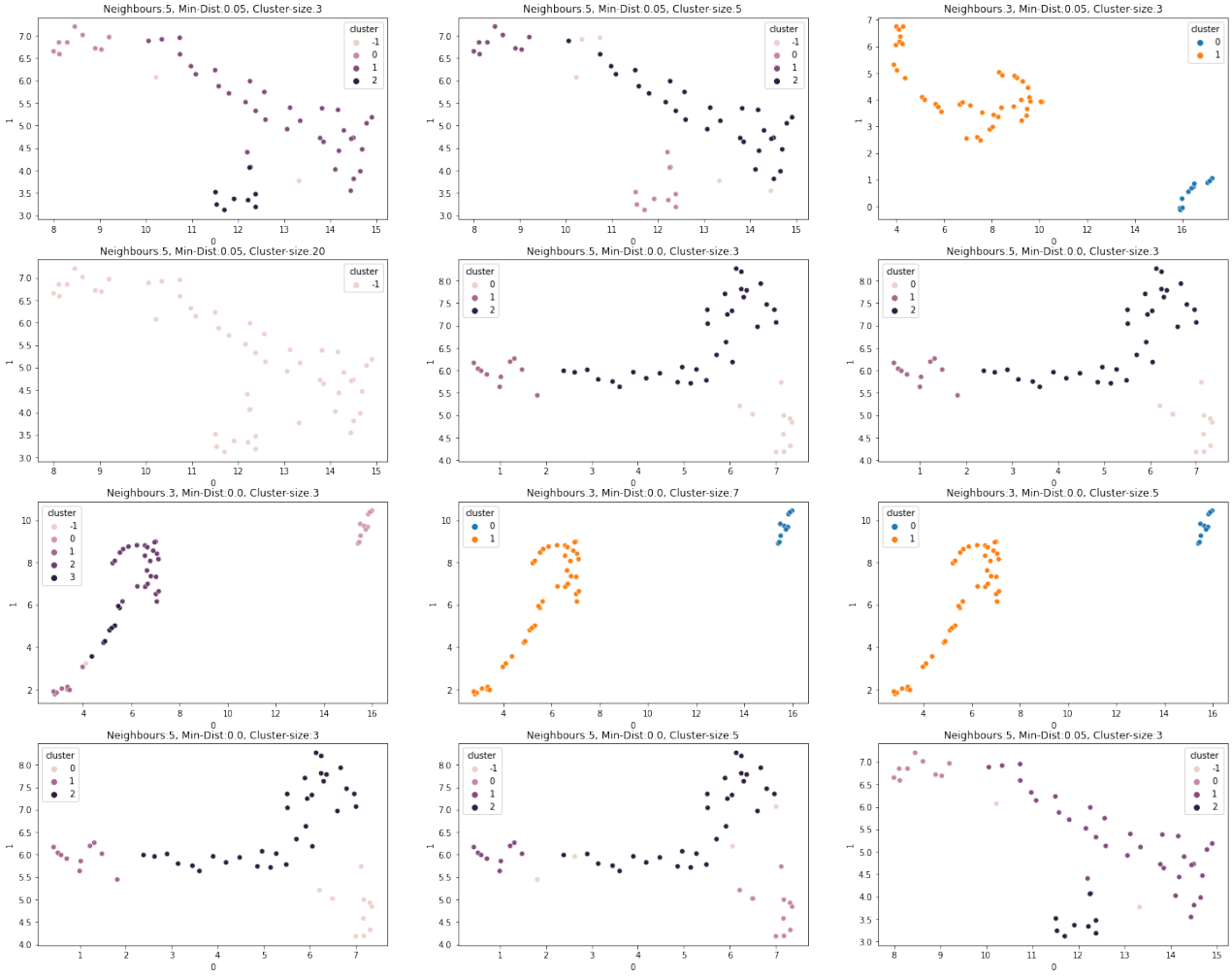
Experiment 2 – values tried (structured method)		
n_neighbours List	**min_dist list**	**min_cluster_size**
[3,5,10]	[0, 0.05, 0.1, 0.3]	[5]

Plot them in the order given in the lists (view the results vertically).



This experiment shows how the minimum distance affects the clusters (see the results vertically). Smaller values in min_dist create denser clusters. Neighbours = 3 seems to create the most distinct clusters.

Experiment 3 – values tried (random method)		
n_neighbours List	**min_dist list**	**min_cluster_size**
[3,5]	[0, 0.05]	[3, 5, 7, 10, 20]



Randomly select different combinations from the given values and plot them.

Cluster size 3,5,7 appears to work best.

Appendix F: Correlation Analysis – Additional Datasets (72 Hours)

A) APA

Alternative Polyadenylation - Top 10 most correlated genes with IC50 (72 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Alternative Polyadenylation	'ANKRD54', 'TMEM260', 'CCDC97', 'EDRF1', 'ULK3', 'USPL1', 'SNRK', 'ARRB2', 'BCOR', 'IPPK'	Pearson	Negative
Alternative Polyadenylation	'BUD13', 'CCDC97', 'CHAC1', 'SNX11', 'IMPA2', 'BORCS7', 'TCFL5', 'FOPNL', 'ARPP19', 'CBX3'	Spearman	Negative
Alternative Polyadenylation	'CRLS1', 'CAMK2D', 'AP5Z1', 'PRPF8', 'ITM2B', 'PRR12', 'NPM3', 'CLUH', 'DCTN5', 'DPP3'	Pearson	Positive
Alternative Polyadenylation	'CLUH', 'CAMK2D', 'NPM3', 'MAZ', 'TRAF7', 'DCTN5', 'PRPF8', 'COPG2', 'C6orf47', 'PRR12'	Spearman	Positive
Alternative Polyadenylation	'IMPA2', 'UBE2A', 'CCDC59', 'DUSP3', 'PDE8A', 'PATL1', 'HNRNPDL', 'WDR11', 'FAM3C', 'C1QTNF3-AMACR', 'ISCU', 'FAM8A1', 'BET1L', 'ASF1A', 'TNFRSF12A', 'HSD17B12', 'TMEM2', 'HSD17B10', 'TMEM41B', 'CBLL1'	MIC	Positive and Negative

B) Gene Effect

Gene Effect - Top 10 most correlated genes with IC50 (72 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Gene Effect	'CADM1 (23705)', 'ITGB3 (3690)', 'USP39 (10713)', 'SMG5 (23381)', 'ATG2A (23130)', 'PHOX2A (401)', 'EDC3 (80153)', 'RNF41 (10193)', 'HOXD12 (3238)', 'SRP54 (6729)'	Pearson	Negative
Gene Effect	'CGN (57530)', 'CADM1 (23705)', 'SKP1 (6500)', 'CENPL (91687)', 'XAB2 (56949)', 'CCKBR (887)', 'AKTIP (64400)', 'ATG2A (23130)', 'F3 (2152)', 'PSMD11 (5717)'	Spearman	Negative

Gene Effect	'MRPS31 (10240)', 'DTNBP1 (84062)', 'PDE4C (5143)', 'CHRNA7 (1139)', 'HUS1 (3364)', 'RARS2 (57038)', 'TCIM (56892)', 'BTBD11 (121551)', 'RAD1 (5810)', 'VPS41 (27072)'	Pearson	Positive
Gene Effect	'MRPS31 (10240)', 'RARS2 (57038)', 'RAD1 (5810)', 'EIF3G (8666)', 'DTNBP1 (84062)', 'AKT2 (208)', 'CDK8 (1024)', 'NUB1 (51667)', 'EIF3F (8665)', 'HACD1 (9200)'	Spearman	Positive
Gene Effect	'SPRR1A (6698)', 'TMEM259 (91304)', 'CCDC115 (84317)', 'KRT38 (8687)', 'NBR1 (4077)', 'TBC1D7 (51256)', 'KLHL30 (377007)', 'EPS8L2 (64787)', 'RGS10 (6001)', 'VPS36 (51028)', 'BVES (11149)', 'MORF4L2 (9643)', 'STAM (8027)', 'CCKBR (887)', 'UBE2I (7329)', 'DHX9 (1660)', 'TEX47 (219557)', 'KAT6A (7994)', 'UFD1 (7353)', 'ZBTB20 (26137)'	MIC	Positive and Negative

C) Proteomics

Proteomics - Top 10 most correlated genes with IC50 (72 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Proteomics	'TMED3 (Q9Y3Q3)', 'WDR18 (Q9BV38)', 'ORC3 (Q9UBD5-2)', 'ORC4 (O43929)', 'DHRS4 (Q9BTZ2)', 'LAS1L (Q9Y4W2)', 'TEX10 (Q9NXF1)', 'TRABD (Q9H4I3)', 'DHRS4L1 (P0CG22)', 'TAF1C (Q15572-6)'	Pearson	Negative
Proteomics	'WDR18 (Q9BV38)', 'ORC3 (Q9UBD5-2)', 'DHX34 (Q14147)', 'HIRA (P54198)', 'TMED3 (Q9Y3Q3)', 'CDK5RAP3 (Q96JB5)', 'ORC4 (O43929)', 'UFL1 (O94874)', 'PBRM1 (Q86U86)', 'EXOSC10 (Q01780)'	Spearman	Negative
Proteomics	'SH2B1 (Q9NRF2)', 'NUDCD3 (Q8IVD9)', 'LRRC20 (Q8TCA0)', 'KIF3B (O15066)', 'TENM3 (Q9P273)', 'SCRN1 (Q12765-2)',	Pearson	Positive

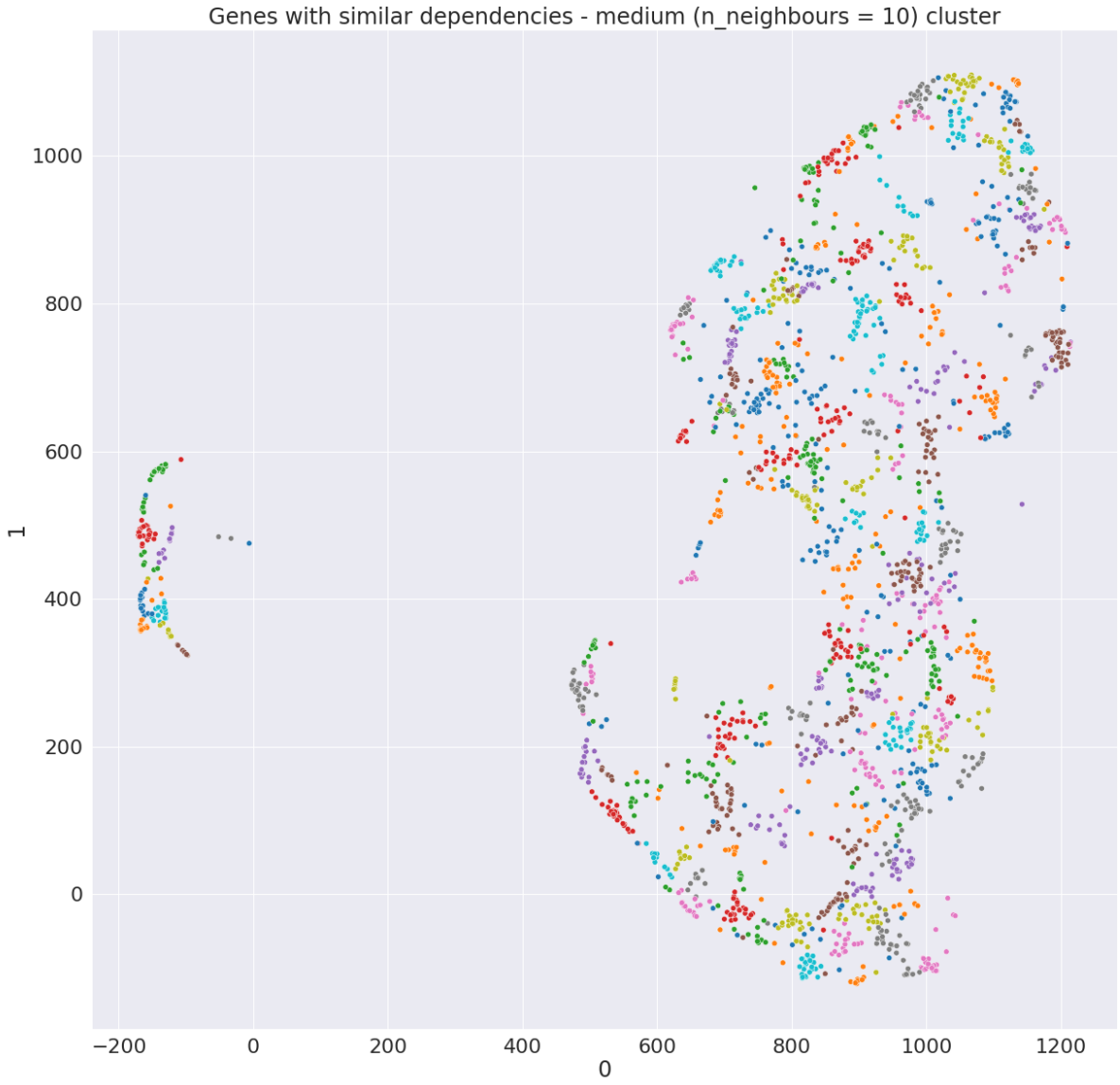
	'BCAR1 (P56945-6)', 'APBB2 (Q92870-4)', 'RANBP3 (Q9H6Z4)', 'TNS3 (Q68CZ2)'		
Proteomics	'NUDCD3 (Q8IVD9)', 'KIF3B (O15066)', 'LRRC20 (Q8TCA0)', 'DCTN2 (Q13561)', 'SIRT1 (Q96EB6)', 'BCAR1 (P56945-6)', 'RANBP3 (Q9H6Z4)', 'TNS3 (Q68CZ2)', 'APBB2 (Q92870-4)', 'RTL8A (Q9BWD3)'	Spearman	Positive
Proteomics	'P4HB (P07237)', 'ETV6 (P41212)', 'NUDCD3 (Q8IVD9)', 'NSUN5 (Q96P11-2)', 'LRRFIP1 (Q32MZ4)', 'BCAM (P50895)', 'OSBPL9 (Q96SU4-2)', 'PIBF1 (Q8WXW3)', 'METTL26 (Q96S19)', 'ITCH (Q96J02)', 'SERINC1 (Q9NRX5)', 'LRRC59 (Q96AG4)', 'PFKFB2 (O60825)', 'STBD1 (O95210)', 'C7orf25 (Q9BPX7-2)', 'MAPK10 (P53779)', 'SETD7 (Q8WTS6)', 'THUMPD1 (Q9NXG2)', 'FUNDC2 (Q9BWH2)', 'KIF22 (Q14807)'	MIC	Positive

D) Mutation

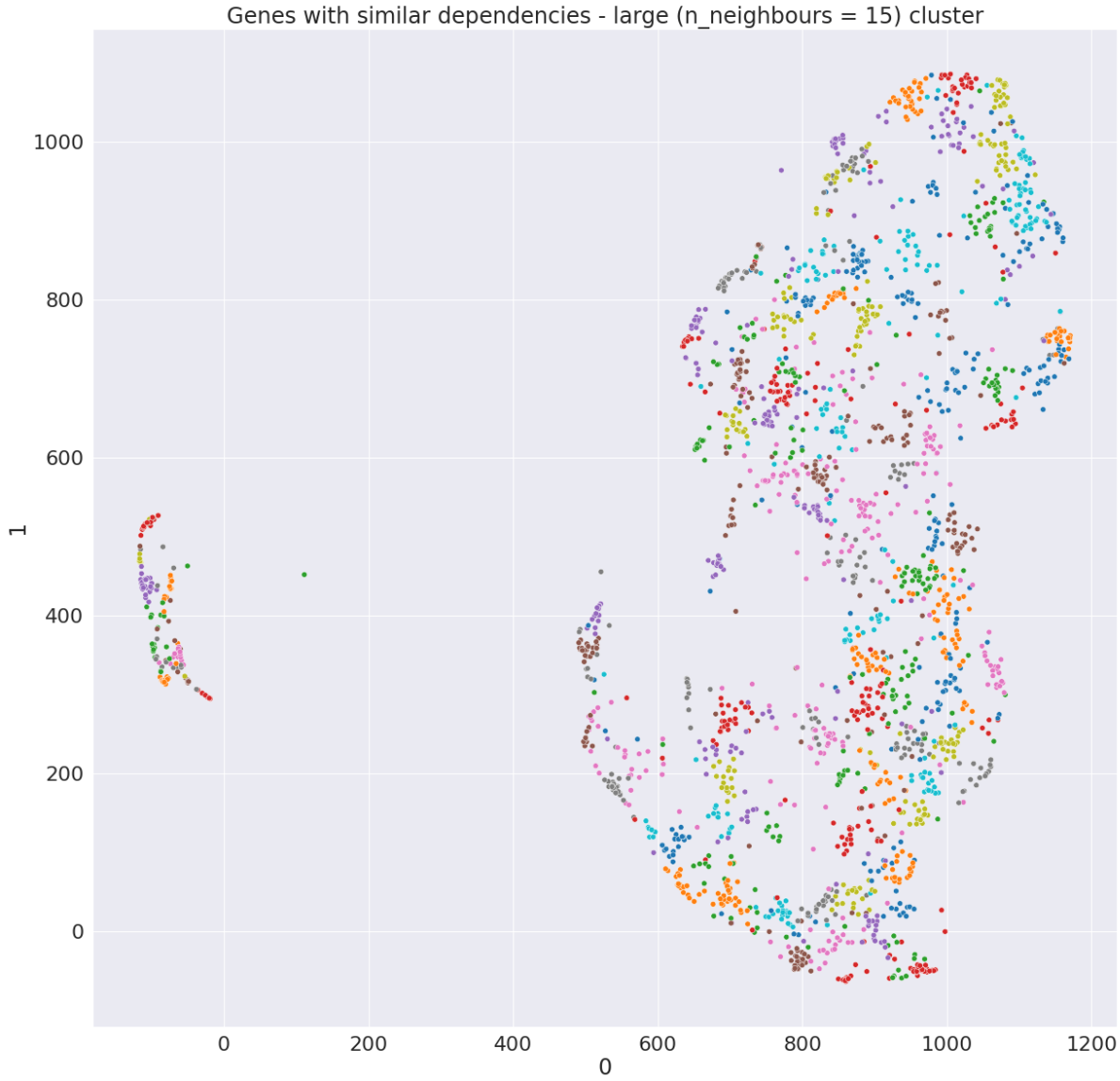
Mutations - Top 10 most correlated genes with IC50 (72 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Mutations	'LTN1', 'SEMA7A', 'ITPKC', 'CEP170B', 'BRAF', 'CAPN6', 'MAGEL2', 'ARHGEF16', 'PCDHGB7', 'MAGEC1'	Point-biserial	Negative
Mutations	'NAALADL1', 'ZC3H12C', 'VANGL1', 'SLC10A5', 'RBM27', 'DDX6', 'ATP2C2', 'RSC1A1', 'PNPLA3', 'IRAK1BP1'	Point-biserial	Positive

Appendix G – Ensemble clustering – gene dependencies.

Medium cluster: *n-neighbours*=10



Large cluster: *n-neighbours*=15



The genes present in each cluster are present in the supporting data, on the *Ensemble Clustering* folder.

Appendix H – Mutations performance

Best performance for each threshold is highlighted with a blue colour. NaN occurs when the divisor is 0.

50% resistant - binary threshold: *pIC50* = 4.727

K	M	Description	Optimal Logic Model	Accuracy	Precision	Recall	F1
1	1	Single predictor	~SYNE1	0.41	0.5	0.86	0.63
1	2	2-input AND	~SYNE1 & ~TRIP6	0.41	0.5	0.71	0.59
1	3	3-input AND	~INSRR & ~ANO4 & ~INSR	0.47	0.54	0.88	0.67
1	4	4-input AND	~SYNE2 & ~PLD1 & ~INSR & ~CASP10	0.29	0.42	1	0.59
2	1	2-input OR	~SYNE1 | ITPKC	0.41	0.5	0.86	0.63
3	1	3-input OR	GAD2 | FAM115A | SMARCC1	0.47	1	0.13	0.22
4	1	4-input OR	GAD2 | ITPKC | FAM115A | AFF4	0.47	1	0.13	0.22
2	2	2-by-2	~ANO4 & PCDHGB7 | ~SYNE1 & ~ADCY1	0.35	0.45	0.83	0.59

65% resistant - binary threshold: *pIC50* = 4.846

K	M	Description	Optimal Logic Model	Accuracy	Precision	Recall	F1
1	1	Single predictor	~SYNE1	0.47	0.5	0.75	0.6
1	2	2-input AND	~SYNE1 & ~TRIP6	0.47	0.5	0.63	0.56
1	3	3-input AND	~SYNE2 & ~PKD1L1 & ~PTPN13	0.53	0.55	0.67	0.6
1	4	4-input AND	~SYNE2 & ~PLD1 & ~INSR & ~CASP10	0.3529	0.42	0.83	0.56
2	1	2-input OR	~SYNE1 | ITPKC	0.47	0.5	0.75	0.6
3	1	3-input OR	GAD2 | FAM115A | SMARCC1	0.41	0	0	NaN
4	1	4-input OR	GAD2 | ITPKC | FAM115A | AFF4	0.53	1	0.11	0.2

2	2	2-by-2	~ANO4 & PCDHGB7 | ~SYNE1 & ~ADCY1	0.41	0.45	0.71	0.56

80% resistant - binary threshold: *pIC50* = 5.20

K	M	Description	Optimal Logic Model	Accuracy	Precision	Recall	F1
1	1	Single predictor	~SYNE1	0.29	0.08	0.2	0.12
1	2	2-input AND	~SYNE1 & ~INSRR	0.35	0.09	0.17	0.12
1	3	3-input AND	~SYNE1 & ~PKD1L1 & ~CSPG4	0.35	0.09	0.17	0.12
1	4	4-input AND	~SYNE2 & ~PKD1L1 & ~PTPN13 & ~HSPA12A	0.41	0.1	0.14	0.12
2	1	2-input OR	GAD2 | AFF4	0.88	NaN	0	NaN
3	1	3-input OR	GAD2 | FAM115A | AFF4	0.88	NaN	0	NaN
4	1	4-input OR	GAD2 | ITPKC | FAM115A | AFF4	0.88	NaN	0	NaN
2	2	2-by-2	~PTK2B & UGT2B7 | ~PKD1L1 & GAD2	0.88	NaN	0	NaN

Best performing model is at the 50% threshold, when K=1 and M=3. Its predictive biomarkers are INSRR, ANO4, and INSR.

Appendix I –NUC-7738+Paclitaxel and NUC-7738+Erlotinib combinations

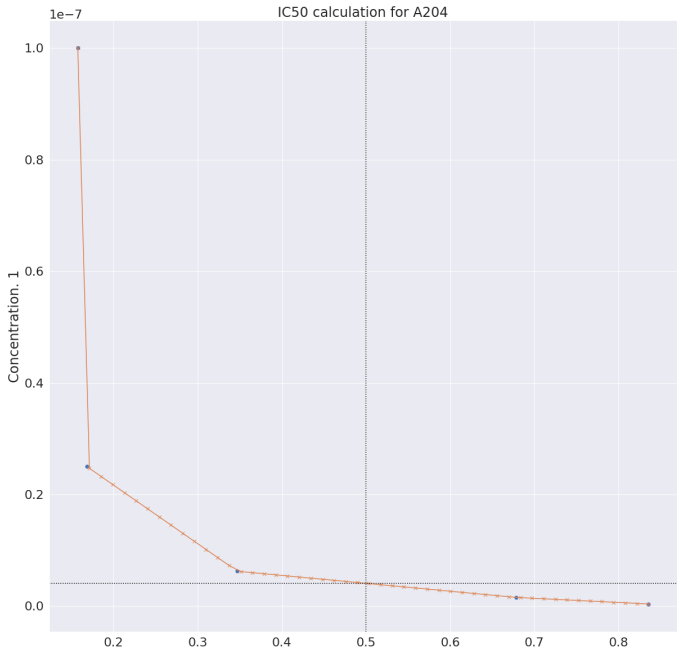
A) IC50 Interpolation

This appendix shows an example of how the *IC50* was interpolated for each cell line for a drug combination. The data available for each cell line of a combination looked like the table below.

Cell line	Compound 1	Compound 2	Origin	Conc. 1	Conc. 2	IC
A204	Paclitaxel	Nuc-7738	muscle	3.910000e-10	0.000001	0.8356
A204	Paclitaxel	Nuc-7738	muscle	1.560000e-09	0.000001	0.6781
A204	Paclitaxel	Nuc-7738	muscle	6.250000e-09	0.000001	0.3471
A204	Paclitaxel	Nuc-7738	muscle	2.500000e-08	0.000001	0.1684
A204	Paclitaxel	Nuc-7738	muscle	1.000000e-07	0.000001	0.1575

The concentration (Conc. 2) for the NUC-7738 compound was fixed at 0.000001, while the concentration (Conc. 1) of the other drug (Paclitaxel in this case) varied. To get the *IC50* value we need to interpolate the concentration of Paclitaxel when *IC* is 50% (since the concentration for NUC-7738 is fixed).[35] Important to note, the IC50 of a cell line was calculated if and only if there is at least one *IC* value for that cell line that is less than 0.5.

A new function was created called *get_IC50_for_drug_combination* which passes the IC values as *xp* and Conc.1 as *fp* in NumPy's *interp* function, which returns the one-dimensional piecewise linear interpolant at a given point *x*. When 50 different data points are requested (across a linear space) the following plot is generated.



[35] This was confirmed by Dr Mustafa Elshani, a researcher at the medical department.

This enables us to find the *IC50* values for all the cell lines in the combinations. The interpolated IC50 values and generated plots are available in the combinations folder (Supplementary Material).

B) Correlation Analysis Combinations

NUC-7738: Gene expressions - Top 10 most correlated genes with IC50 (120 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Gene expressions	'MTERF1', 'FCHO1', 'MRI1', 'ERMP1', 'KDM5B', 'NOTCH1', 'ANKRD6', 'ATP6V1B1', 'ATAD2B', 'ICAM5'	Pearson	Negative
Gene expressions	'ATP6V1B1', 'CSTA', 'KMT5C', 'LGR4', 'ERMP1', 'VAX2', 'NUDT14', 'B4GALNT3', 'KDM5B', 'PPP1R3D'	Spearman	Negative
Gene expressions	'ABCB1', 'ERBIN', 'SNF8', 'WNT5B', 'GSTT2B', 'CLDN2', 'ABCC4', 'C4BPB', 'TMEM14B', 'RHOBTB3'	Pearson	Positive
Gene expressions	'ABCB1', 'CWC25', 'ERBIN', 'GLYCTK', 'ARL2BP', 'C12orf75', 'ABCC4', 'MMP24', 'CNTNAP1', 'IL7'	Spearman	Positive
Gene expressions	'PITX1', 'ZBTB14', 'KDM5A', 'GDF7', 'NDUFA10', 'RGL2', 'PARP4', 'SMAD7', 'RHOBTB1', 'POGLUT2', 'ZCRB1', 'BECN1', 'LGR4', 'RAB8A', 'ATP5F1C', 'RASGRP4', 'CEP55', 'ARL16', 'HERC5', 'CSNK1A1'	MIC	Positive

NUC-7738+Paclitaxel: Gene expressions - Top 10 most correlated genes with IC50 (120 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Gene expressions	'HIP1R', 'OAT', 'ZFP14', 'ZBTB10', 'CCDC121', 'DIPK1B', 'PTEN', 'SIGIRR', 'GAMT', 'EIF4EBP2'	Pearson	Negative
Gene expressions	'IGSF9', 'SBK1', 'SLC29A2', 'EPS8L1', 'EPN3', 'DLG3', 'ENDOU', 'ARHGAP8', 'RHOV', 'OVOL1'	Spearman	Negative
Gene expressions	'RUNDC3B', 'DIO2', 'GPAT2', 'ABCB1', 'MGAM', 'IL6', 'TMC3', 'KCNE5', 'ELAVL2', 'RGS4'	Pearson	Positive

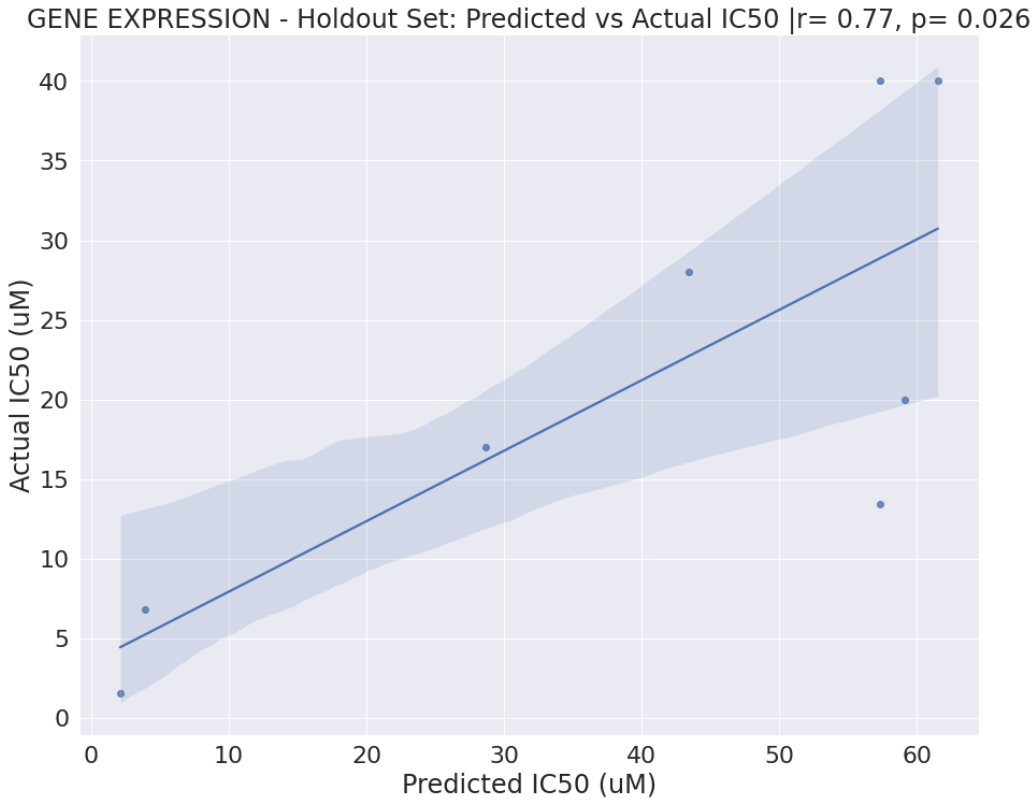
Gene expressions	'HINT1', 'SEC13', 'CUTA', 'MTRF1L', 'GOPC', 'CENPW', 'LYSMD3', 'SLC25A20', 'FAM114A1', 'QKI'	Spearman	Positive
Gene expressions	'IGSF9', 'CKMT1A', 'RRAGA', 'LANCL1', 'MTRF1L', 'MAPRE2', 'SEC13', 'ARL2BP', 'TPM4', 'RIPK1', 'CENPQ', 'GNG12', 'MTM1', 'DIP2B', 'GRHL1', 'CNKSR1', 'HRK', 'ZCRB1', 'DLG3', 'ECHDC1'	MIC	Positive

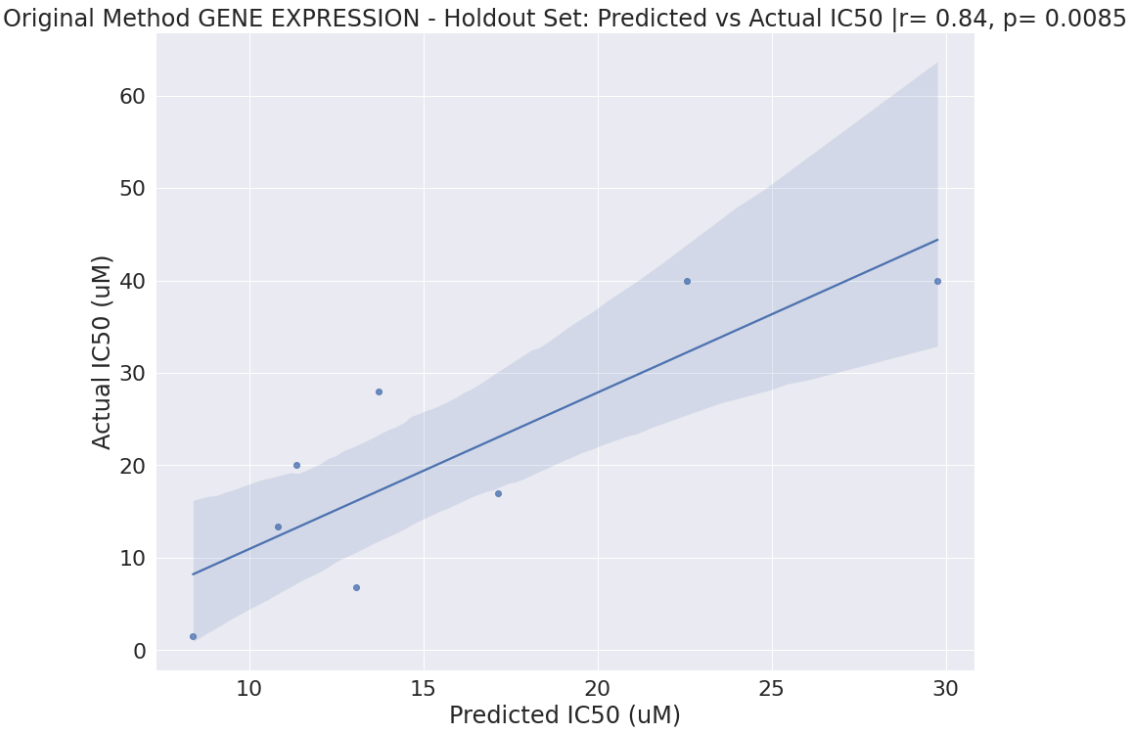
NUC-7738+Erlotinib: Gene expressions - Top 10 most correlated genes with IC50 (120 Hours)			
Dataset	**Genes**	**Method**	**Relation**
Gene expressions	'P2RY6', 'GLDC', 'SMPD2', 'SYTL3', 'LRRC8D', 'ANKRD2', 'NPNT', 'CORO6', 'PPP2R2C', 'ADGRL2'	Pearson	Negative
Gene expressions	'TDRD9', 'GLDC', 'SMPD2', 'TIMM17A', 'ECT2L', 'CTSL', 'HK1', 'GLUL', 'TNF', 'TRAF2'	Spearman	Negative
Gene expressions	'BTF3', 'DOK1', 'DGUOK', 'PCLAF', 'BTBD1', 'FAM227B', 'ZMYM5', 'TMEM107', 'CUL4B', 'NDUFAF6'	Pearson	Positive
Gene expressions	'CST7', 'NDUFAF6', 'ZMYM5', 'IGBP1', 'BTF3', 'COMMD6', 'GPALPP1', 'CUL4B', 'DPH6', 'PIGV'	Spearman	Positive
Gene expressions	'ZMYM5', 'IGBP1', 'TDRD9', 'BORCS6', 'GPALPP1', 'CCDC115', 'TMEM107', 'SKOR1', 'RPL9', 'PIGV', 'C11orf1', 'RAB4B', 'ST3GAL5', 'ST20-MTHFS', 'DENND4A', 'ALKBH8', 'VAMP2', 'CELF2', 'NCAPG2', 'JADE2'	MIC	Positive

Appendix J: Biomarkers summary and correlation plots on the holdout set

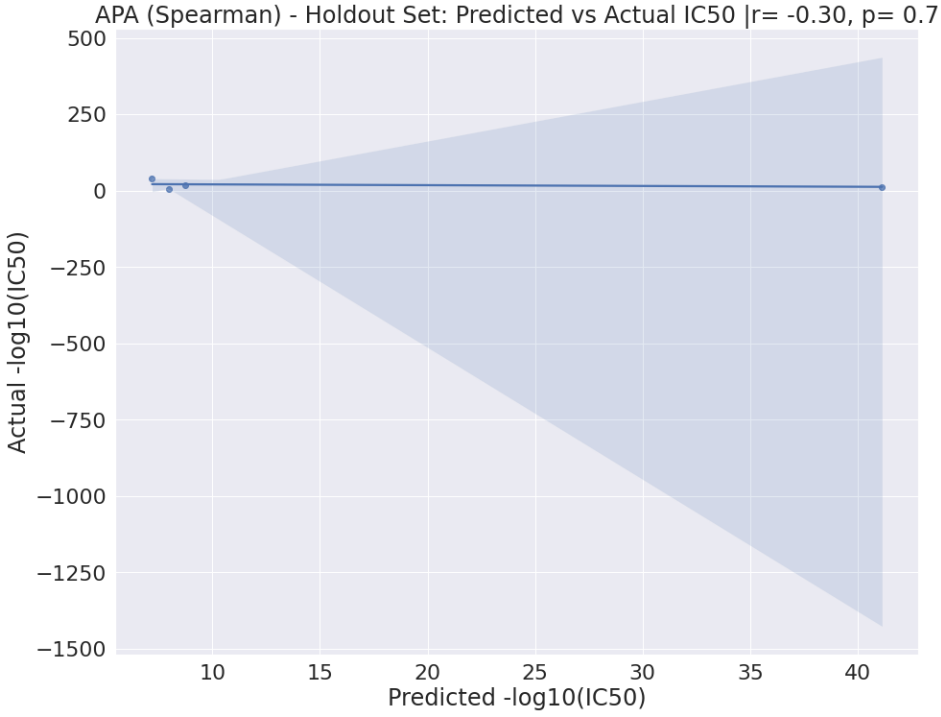
Model	Biomarkers
Gene expressions (Kathad)	HIPK2, ANXA2, SLC38A5, SALL4, KMT5C, ME3, CRABP2
Gene expressions	ITGA3, RARRES2, CTHRC1, MST1R, ARHGAP29, ABCC3, DCBLD2, ANXA2, BIRC3, HLA-A
APA	FAM3C, FANCC, SLC5A3, TAF4, XPA, SUPT16H, CBX3, TRA2A
Proteomics	AHNAK, PEX19, HBD, GLG1, GCA, GMFG, DOCK1, SIRT1, GID8, RTL8A, NUDCD3, APBB2, CALCOCO2, ACSL3, PBRM1, HMGN4, C18orf25, EPG5
Gene effect	LSM3, EIF2S3, CWC22, TCF3, CLNS1A, AMD1, TYMS, PSMD4, RIOK1, CDK8
Mutations	FAM220A, FAM3C, FANCC, XPA

A) Gene Expressions

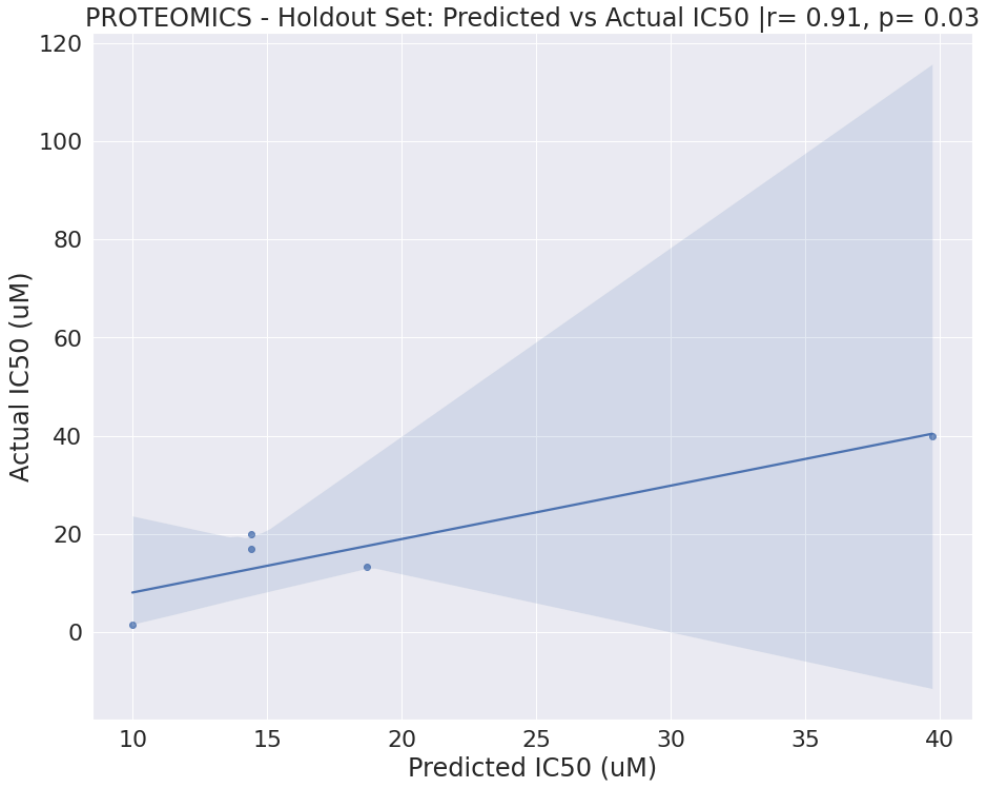



Original Method GENE EXPRESSION - Holdout Set: Predicted vs Actual IC50 |r= 0.84, p= 0.0085
Actual IC50 (uM)
Predicted IC50 (uM)
0
10
20
30
40
50
60
10
15
20
25
30

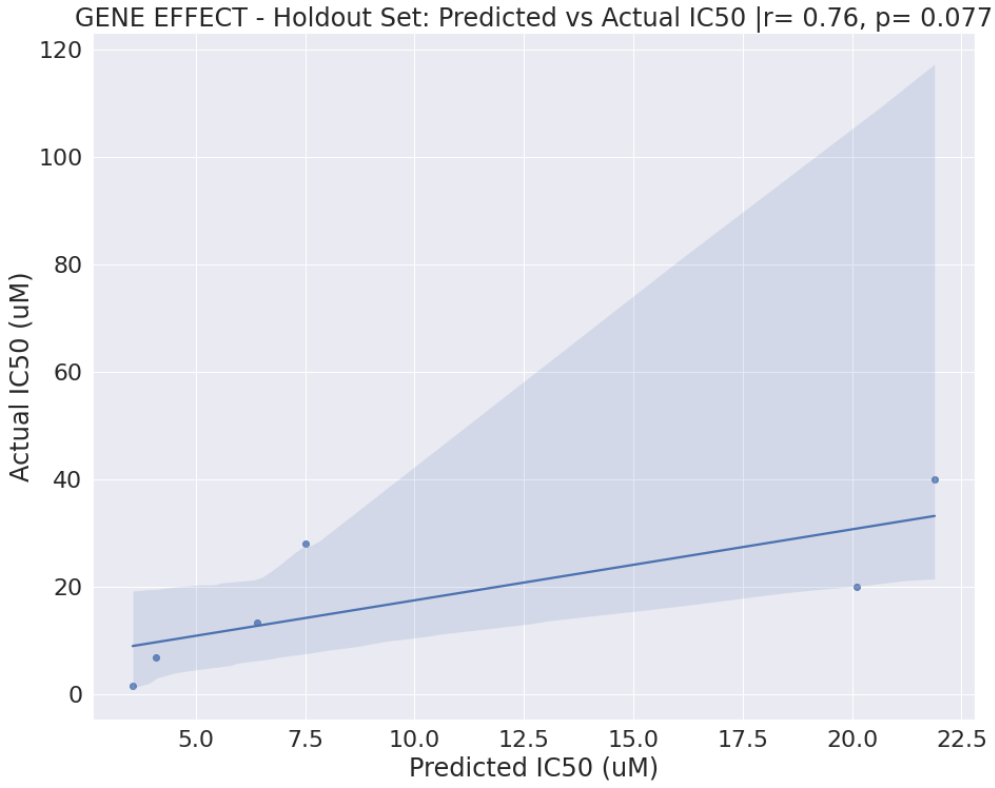
B) APA


APA (Spearman) - Holdout Set: Predicted vs Actual IC50 |r= -0.30, p= 0.7
Actual -log10(IC50)
Predicted -log10(IC50)
500
250
0
−250
−500
−750
−1000
−1250
−1500
10
15
20
25
30
35
40

C) Proteomics



D) Gene Effect



Appendix K – Gene expression biomarkers cancer relation

The table below contains a selection of publications for each of the other biomarker genes discovered.

Biomarker gene	Selected Paper(s)
ITGA3	• ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer [77]. • Evaluation of ITGA3 as a Biomarker of Progression and Recurrence in Papillary Thyroid Carcinoma [78].
CTHRC1	• CTHRC1 Is a Prognostic Biomarker and Correlated with Immune Infiltrates in Kidney Renal Papillary Cell Carcinoma and Kidney Renal Clear Cell Carcinoma [79] • CTHRC1 promotes liver metastasis by reshaping infiltrated macrophages through physical interactions with TGF-β receptors in colorectal cancer [80] • CTHRC1 is a prognosis-related biomarker correlated with immune infiltrates in colon adenocarcinoma [81]
DCBLD2	• Transcriptomic Profiling Identifies DCBLD2 as a Diagnostic and Prognostic Biomarker in Pancreatic Ductal Adenocarcinoma [82]. • DCBLD2 Mediates Epithelial-Mesenchymal Transition-Induced Metastasis by Cisplatin in Lung Adenocarcinoma [83].
RARRES2	• RARRES2 functions as a tumor suppressor by promoting β-catenin phosphorylation/degradation and inhibiting p38 phosphorylation in adrenocortical carcinoma [84].
ANXA2	• Annexin A2-mediated cancer progression and therapeutic resistance in nasopharyngeal carcinoma, Chen et al. [85]. • Crucial role of Anxa2 in cancer progression: highlights on its novel regulatory mechanism [86].
BIRC3	• Frazzi, R., 2021. BIRC3 and BIRC5: multi-faceted inhibitors in cancer. Cell & Bioscience, 11(1):8 [87]
HLA-A	• HLA Gene May Predict if Cancer Immunotherapy Will Work [88]. • Differential role of HLA-A and HLA-B, C expression levels as prognostic markers in colon and rectal cancer [89]
ABCC3	• ABCC3 is a novel target for the treatment of pancreatic cancer [94]
MST1R	• MST1R (RON) expression is a novel prognostic biomarker for metastatic progression in breast cancer patients [95].

Appendix L – Clustering APA

A) Pre-processing APA datasets

Clustering of the cell lines based on their APA events was used by incorporating multiple APA datasets, where each had APA events and cell lines for one specific type of cancer (e.g., bladder). To use our cancer cell lines which consist of multiple cancer types, they had to be combined. The *PDUI* value, which ranged from 0 to 1, was used to calculate the frequency of APA events and was proportional to the transcript's distal polyadenylation site.

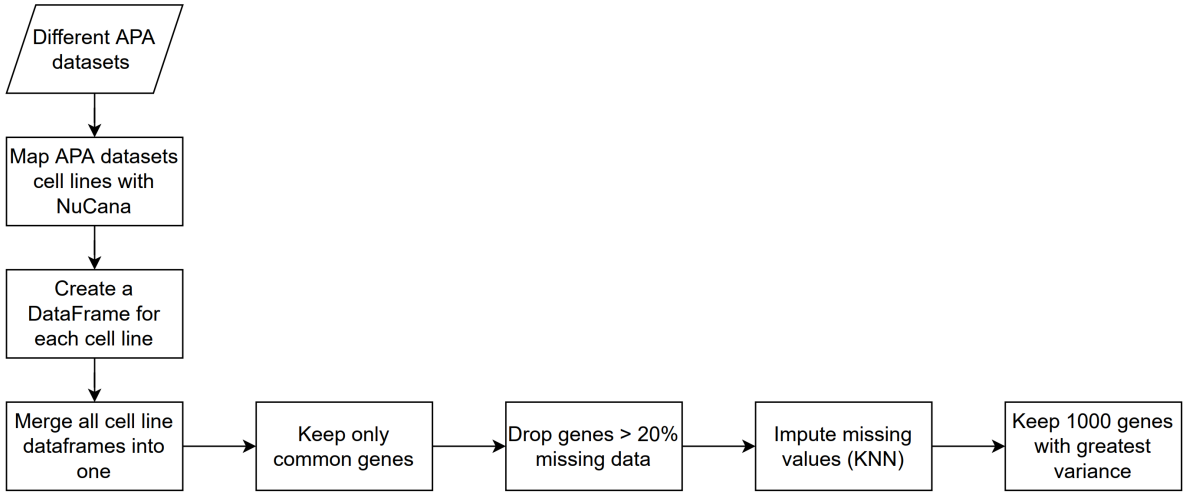


Figure above depicts the pre-processing steps taken to integrate the APA datasets, clean the data, and retain only useful genes. The first step was to manually identify the names for the cell lines used in the APA datasets and map them with our NUC-7738 dataset names. This was a manual process, from which only 53 out of 95 cell lines were matched between the two datasets. Each APA dataset was then loaded into a DataFrame and the *IC50* value of each cell line was retrieved. For every cell line, a DataFrame was created storing the *PDUI* scores for each gene. This was done by identifying in which APA dataset the cell line was located and then retrieving its values. All the cell lines DataFrames were then merged into a single DataFrame from which only the common genes were kept. This was done to address the issue that some datasets contain unique genes that were only found in them. The genes with more than 20% missing data were then dropped since imputing them would make the dataset unreliable due to the volume of artificial data. The remaining missing data were imputed using a KNN imputer (K=3). Finally, following an important pre-processing step from Zhong et al. [22], only the 1,000 genes with the highest variance were retained.

B) MDL – PCA with Spherical K-Means

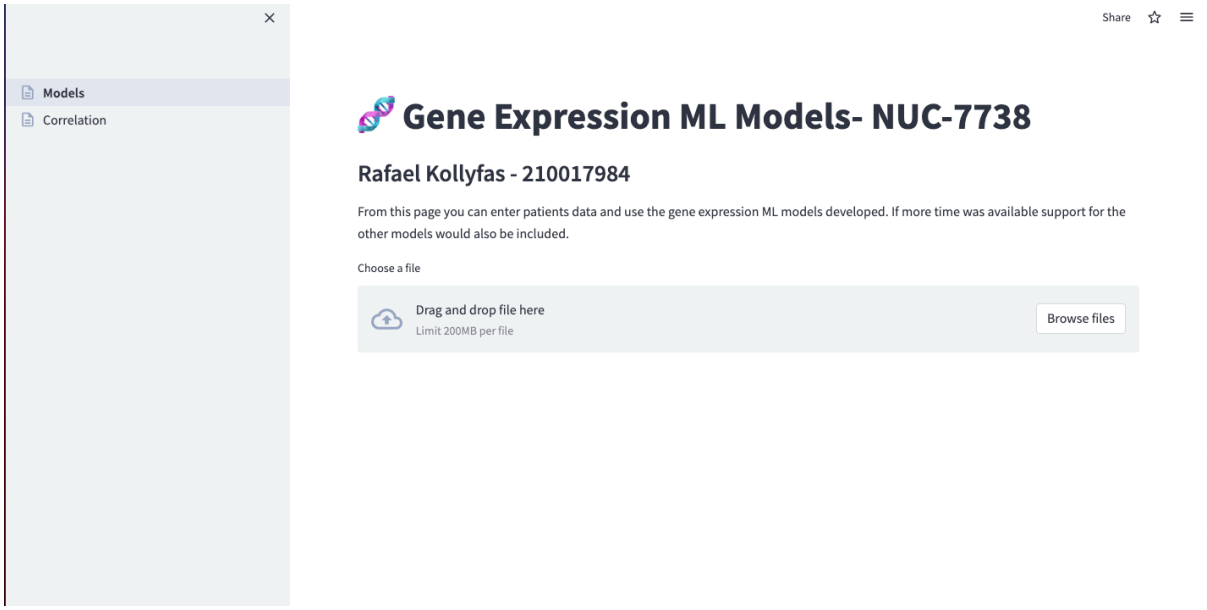
Principal Components	K	DL
14	2	12,961.7

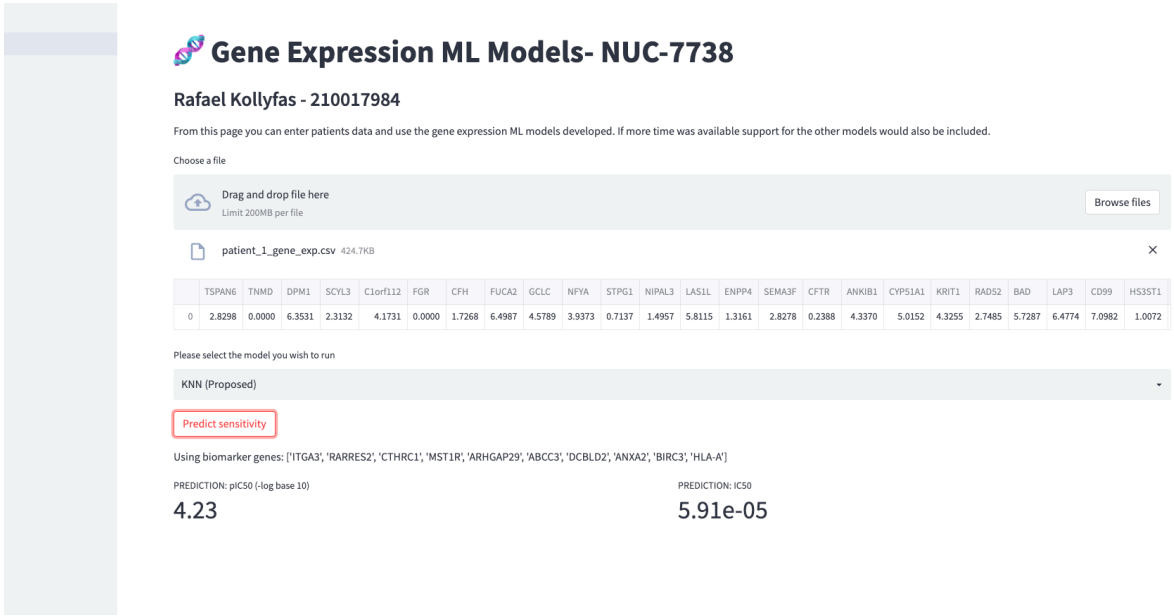
23	2	17,506.4
35	2	22,306.9
43	2	24,886
14	3	11,921.2
23	3	16,437.6
35	3	21,280.9
43	3	23,837.3
14	4	11,450
23	4	15,745.7
35	4	20,859.4
43	4	23,394.8
14	5	18,040.3
23	5	22,547.3
35	5	26,184.7
43	5	27,699.3

Appendix M - Platform

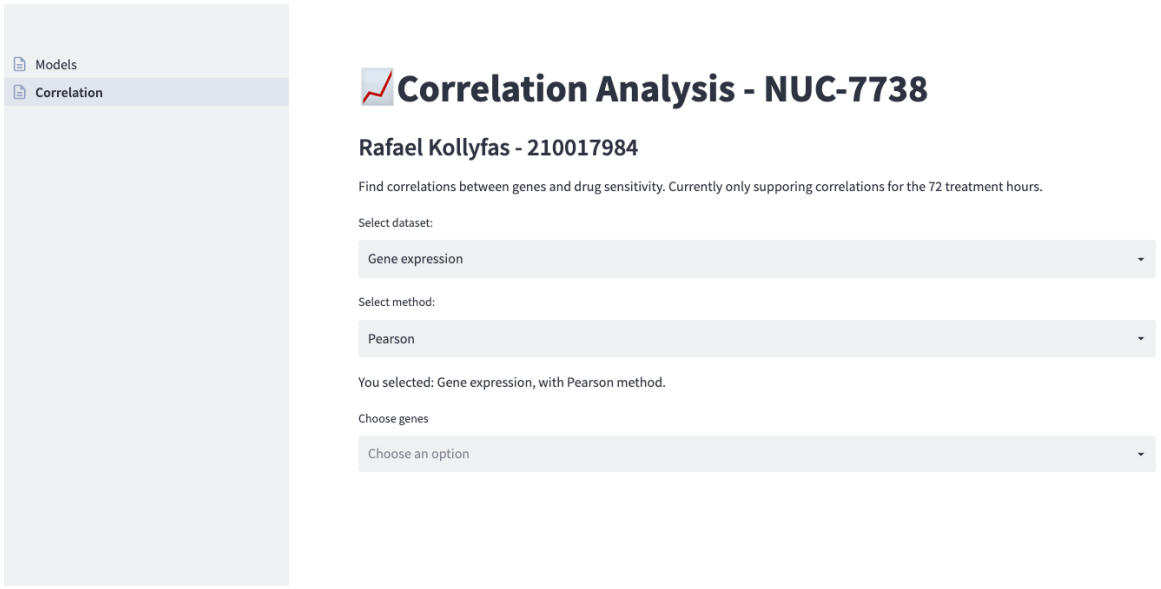
A simple platform interface that was developed using Streamlit, a python-based web library. The platform is at its very early stages as I could not fully develop it due to timing constraints.

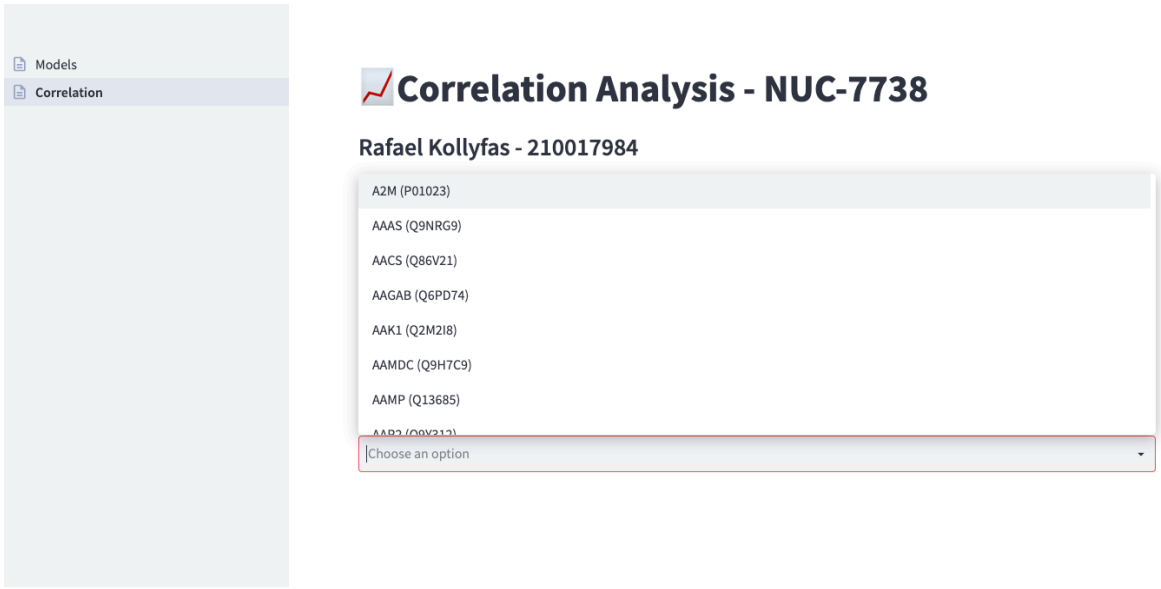
It currently allows:
1) Predict drug efficacy based on a new patient's gene expression data using both models developed (KNN/Proposed and XGBoost/Kathad).

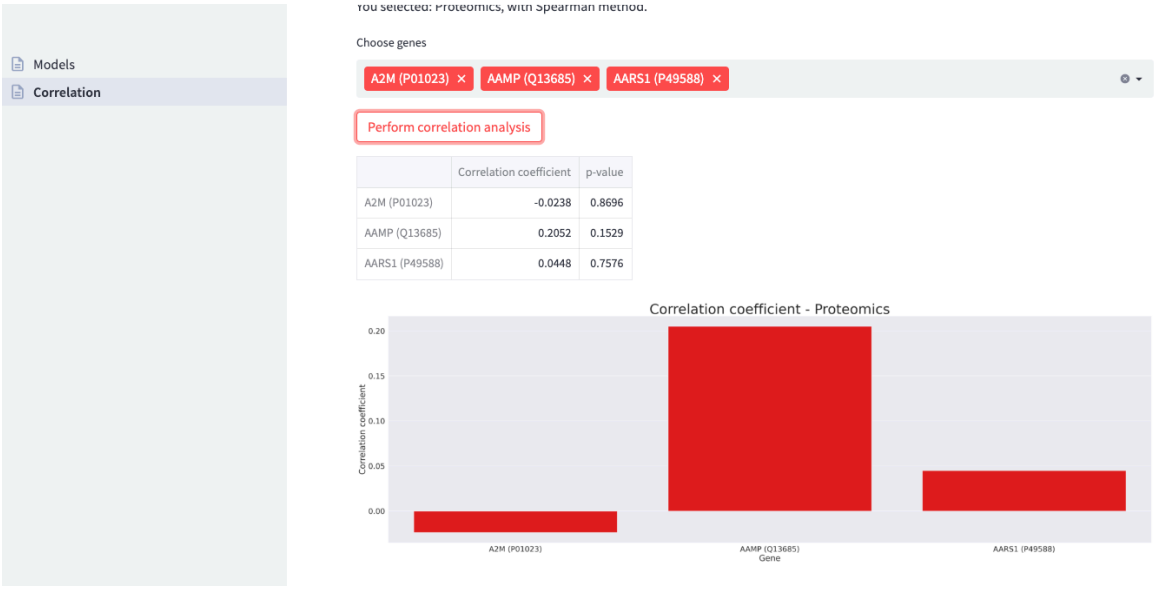




2) Get the correlation coefficient between *IC50* and chosen gene(s). This works for all the datasets and allows the user to choose between Pearson, Spearman, and point-biserial (for mutations) correlation analysis. It returns the results in a table and as a visualisation.







If more time was available, the following would have been implemented:

1. Support for other omics models developed.
2. Ability to search for a targeted gene's cluster to identify genes with similar dependency profiles. Colour-based visualisation to show location of cluster.

The platform is hosted at:
https://rafkol98-ml-driven-omics-platform-models-ipez2r.streamlitapp.com/

Appendix N - Ethics

UNIVERSITY OF ST ANDREWS
TEACHING AND RESEARCH ETHICS COMMITTEE (UTREC)
SCHOOL OF COMPUTER SCIENCE
PRELIMINARY ETHICS SELF-ASSESSMENT FORM

This Preliminary Ethics Self-Assessment Form is to be conducted by the researcher, and completed in conjunction with the Guidelines for Ethical Research Practice. All staff and students of the School of Computer Science must complete it prior to commencing research.

This Form will act as a formal record of your ethical considerations.
Tick one box

☐ **Staff Project**
☒ **Postgraduate Project**
☐ **Undergraduate Project**

Title of project

ML–Driven Drug Discovery: Identifying biomarkers and mechanistic insights of a novel anti cancer drug

Name of researcher(s)

Rafael Kollyfas

Name of supervisor (for student research)

Ognjen Arandelovic

OVERALL ASSESSMENT (to be signed after questions, overleaf, have been completed)

Self audit has been conducted **YES** ☒ **NO** ☐

There are no ethical issues raised by this project

Signature Student or Researcher

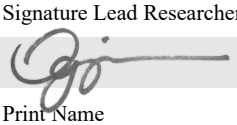
Print Name

Rafael Kollyfas

Date

01/06/2022

Signature Lead Researcher or Supervisor



Print Name

Ognjen Arandelovic

Date

01/06/2022

This form must be date stamped and held in the files of the Lead Researcher or Supervisor. If fieldwork is required, a copy must also be lodged with appropriate Risk Assessment forms.

Computer Science Preliminary Ethics Self-Assessment Form

Research with human subjects

Does your research involve human subjects or have potential adverse consequences for human welfare and wellbeing?

YES ☐ **NO** ☒

If YES, full ethics review required
For example:
Will you be surveying, observing or interviewing human subjects?
Will you be analysing secondary data that could significantly affect human subjects?
Does your research have the potential to have a significant negative effect on people in the study area?

Potential physical or psychological harm, discomfort or stress

Are there any foreseeable risks to the researcher, or to any participants in this research?

YES ☐ **NO** ☒

If YES, full ethics review required
For example:
Is there any potential that there could be physical harm for anyone involved in the research?
Is there any potential for psychological harm, discomfort or stress for anyone involved in the research?

Conflicts of interest

Do any conflicts of interest arise?

YES ☐ **NO** ☒

If YES, full ethics review required
For example:
Might research objectivity be compromised by sponsorship?
Might any issues of intellectual property or roles in research be raised?

Funding

Is your research funded externally?

YES ☐ **NO** ☒

If YES, does the funder appear on the 'currently automatically approved' list on the UTREC website?

YES ☐ **NO** ☒

If NO, you will need to submit a Funding Approval Application as per instructions on the UTREC website.

Research with animals

Does your research involve the use of living animals?

YES ☐ **NO** ☒

If YES, your proposal must be referred to the University's Animal Welfare and Ethics Committee (AWEC)

University Teaching and Research Ethics Committee (UTREC) pages
http://www.st-andrews.ac.uk/utrec/