Final Year Project Report

Predicting Drug Sensitivity from Genetic Screens in Cancer Cell Lines

Orla Cullen

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Colm Ryan

UCD School of Computer Science

University College Dublin

September 2017

Project Specification

One of the goals of precision medicine in cancer is to identify treatments that will work in molecularly defined cohorts of patients (e.g. patients whose tumour harbours a mutation in gene X will respond well to drug Y). One approach to identifying these associations is to screen existing drugs in panels of cancer cell lines to see which drugs kill which cell lines. By connecting this data with genome sequencing in the cell lines it is possible to associate a particular mutation with increased sensitivity to particular drugs. However, the majority of human genes cannot be inhibited with existing drugs and consequently this approach may miss targets that would be very effective if we developed new drugs for them. An alternative approach is to use loss of function genetic screening to identify genes that different cancer cell lines depend upon for survival. Various experimental techniques (RNA interference, CRISPR screening) can be used to inhibit individual genes and measure how this impacts cell line growth. In recent years very large scale efforts have been performed to measure the sensitivity of hundreds of cancer cell lines to either collections of drugs or to collections of gene targeting reagents. A comprehensive evaluation of the relationship between the two data types has yet to be performed. Theoretically loss-of-function genetic screens should provide sufficient information to predict the results of drug sensitivity screens. If a drug targets the function of a specific gene, and we know that inhibiting that gene kills a specific cell line, then we could predict that the drug should also kill the cell line. However, things are rarely that straightforward - many drugs will inhibit the function of multiple genes, while for some drugs we do not know which genes they inhibit. The goal of this project is to assess how accurately we can predict drug sensitivity from genetic screens.

Core:

Loss of function screen datasets and drug sensitivity datasets will be provided

- For drugs with known target genes, quantify how well the gene inhibition measurements predict drug sensitivity measurements
- For all drugs, apply a machine learning approach (e.g. random forest regression) to predict the sensitivity of cell lines to that drug by using the loss-of-function screens as input features
- Analyse the relationship between feature importance (as determined by the machine learning model) and reported drug targets (e.g. is the reported target gene of a drug always the most important feature, if two genes are targeted by a drug are both identified as important features)

Advanced:

- Compare the predictive power derived from loss-of-function screens to that derived from other sources (e.g. gene expression)
- Assess the robustness of the predictive models by testing predictions on additional resources.

Abstract

This study encapsulates the fields of machine learning, statistical analysis, data mining and bioinformatics. The purpose of this study was to apply these computer science concepts to biological datasets to identify if we can accurately predict drug sensitivity by using genetic screens as input features. In cancer cell lines, it has been noticed that gene mutations provide an identifiable marker in the detection of its relationship with drug sensitivity or resistance. Predicting drug responses from these screens provides knowledge of patterns in the gene mutations that provoke a sensitive or resistant drug response. Multiple machine learning models were implemented with varying results. This shows that the selection of the machine learning model is an important aspect of gaining accurate prediction. The random forest model worked, more effectively on the overall dataset, when using a cross validated approach, achieving a >70% accuracy. As a result by applying a machine learning approach, we can say that it is possible to use genetic screens as input features in training a model to predict this sensitivity. However, this approach works best where the mappings from drug to gene are clearer. This approach would enable us to quickly pair drug responses with gene alterations which could potentially guide the development and design of cancer treatments and trials.

Table of Contents

1 Introduction

The following report intends to provide the reader with relevant knowledge and insight into the application of data mining and machine learning techniques, to extract and predict drug sensitivity from genetic screens in cancer cell lines. This will be achieved by utilising loss of function screens and subsequently genetic expression data. A comparative analysis will then be undertook on the results, of the implementation between the machine learning models. The project is based upon applied research so an important part of this project is to assimilate and implement theories, knowledge and techniques which have been previously defined in academic arenas.

1.1 Background & Justification

According to the World Health Organization [1] cancer is one of the top leading causes of death around the world with *"8.8 million deaths in 2015".* An integral part of addressing this issue, is developing treatments that are highly responsive to the cancer cells. Clinical trials can be biased because they divide the data into clusters dependent upon some preclinical the data. This can lead to results of success where failure was inevitable and vice versa. The discovery of new and effective cancer treatments is a lengthy process, where a patient's progress and results are analyzed throughout their treatment. However, the data generated from human cancer cell lines and drug sensitivity over the years, can be analyzed for common traits in genes, by mapping the similarities of cancer cell lines. Subsequently, by identifying these common traits, we can discover similarities between both types of cancer. This also promotes, the discovery of the patterns of drugs that respond, most positively to the cancer. Thus individualized cancer treatments can be offered to patients based upon the patterns discovered [2].

For these reasons, applying machine learning and data mining principles to the data sets extracted from the genetic screens it is anticipated that we could potentially predict drug sensitivity from loss of function screens with some level of accuracy. This would allow the predictive models created to guide the production of new treatments and perhaps assist medical staff on the selection of a cancer treatment based on their genetic expression on a case by case basis.

1.2 Project Goals

There are two distinct parts to completing this project. These are the essential or core elements of the project. This element of the project will be focused around the loss of function screens. Later, the project will compare the predictive power of the loss of function screens to the genetic expression. Finally, the robustness of the models will be assessed by validating the results

using Genomics of Drug Sensitivity in Cancer (GDSC) which consists of the shared compounds and cell lines with the CCLE datasets.

The core area of this project which will be focused on in the early part of the project development will draw attention to drugs with known target genes. This data will be used to measure the predicted drug sensitivity, given the loss of function screens and the relationship or correlation to gene inhibition. Subsequently, all drug data will be measured and evaluated using a machine learning approach and the loss of function screens as input features to forecast the expected gene inhibition for each drug. Afterward,s the results will be analyzed to determine which machine learning model is better suited to the data. A key aspect of this analysis, will be to pay particular attention to feature importance, observing if the target gene is always the most important feature.

Subsequently, the advanced part of this project on the completion of the core elements, as described in the aforementioned paragraph will delve deeper into the comparative analysis of the projections derived from the genetic screens, to that of the projections made from gene expression. by observing the predictive power of both datasets. As a final task, the intention is to assess the models robustness by verifying the project findings using GDSC.

1.3 Structure of the Coming Report

After, briefly introducing the reader to this report in the last section by concisely giving an overview of the project. Followed by, indicating some reasons that bioinformatics has become an attractive field of study, that branches off from computer science, merging the fields of pharmacology, genetics, biological medicine, and computational fields, to develop more individualized and effective cancer treatments. The report will progress to Chapter 2, where the report will go further to focus on research in the area of predicting drug sensitivity from genetic screens, delving deeper into some methods and techniques previously applied, for instance the random forest and elastic net algorithms. These models have uncovered useful predictive patterns, that can be exploited to provide learning models, which can better predict the most effective treatment for a given gene mutation. Chapter 3, I will then provide the reader with an account of the datasets used in the project along with an overview of the implementation. Subsequently , the focus will move on towards the analysis. This will encompass, the comparative analysis of the predictive models. Finally, the report will conclude by making reference to what has yet to be achieved and the future work, in advance of bringing the report to a close with a brief summary of the project findings.

2 Background Research

Cancer is a leading cause of death in Ireland with the statistics depicting that 1 out of every four deaths being attributed to cancer [3]. Thus, it is imperative, that we can develop cancer treatments that are effective in stopping the progress of gene mutations. This can be achieved by providing more precise treatments that target certain features of the genetic makeup of the cancer cell. This field which combines medicine, statistics and machine learning concepts, is commonly referred to as Precision Medicine. These techniques when combined, can be applied to biomedical data to help us discover relevant data and uncover patterns that appear frequently. Hence, the hope is that the patterns explored and produced can be exploited to give cancer research and drug development more accurate information[4].

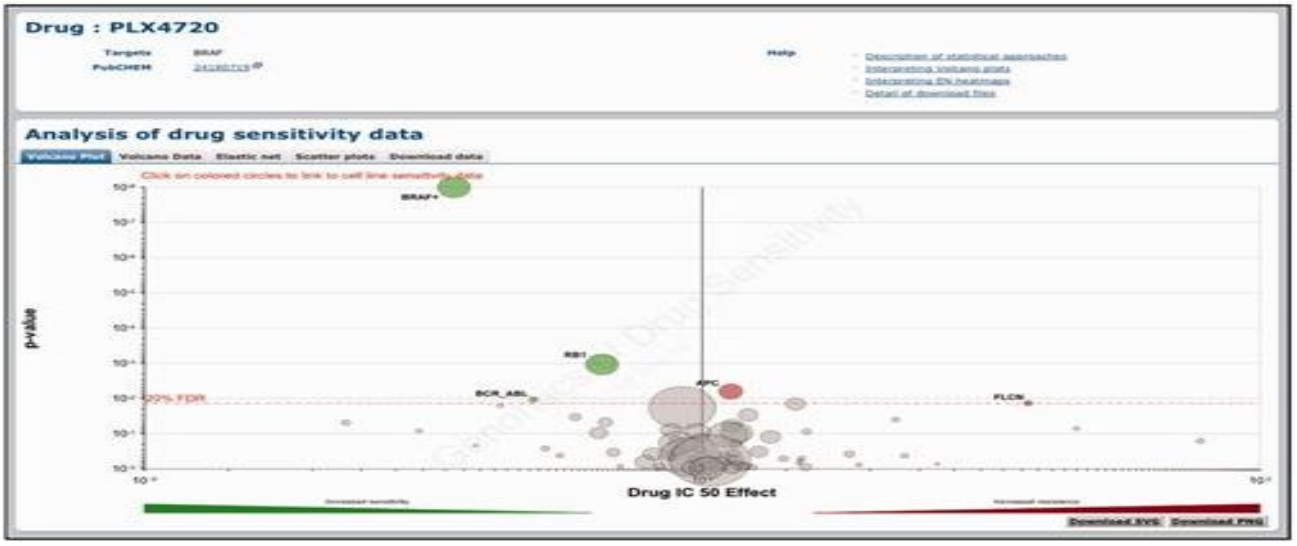
There have been noticeable advances in predicting cancer cell sensitivity to drugs by combining the fields of medicine and machine learning. It provides methods that can be used in research to predict drug sensitivity given loss of function screens. Thus, developing more efficient, effective and individualized treatment programs, based upon the patient's genetic makeup and particular strain of cancer. Cancer clinical trials and the development of cancer treatments can be a complicated process, that extends over a long period by exploiting the patterns that the machine learning models produce, links between cancer cells and their reaction to a host of drugs can be identified[5], [6]. Subsequently, it provides us with indication of which drugs provide some recovery from a cancer. These drugs once identified can then be refined to produce more potent treatments. Precision medicine, tries to address the problem of diversity in the effects seen on the loss of function screens by analyzing the relationship between the drug resistance scores in terms of the IC50 EC50 and ActArea scores. This allows discovery of minor differences in the genomic patterns where, the loss of function screens show resistance instead of sensitivity to the treatment. This depicts that a one treatment fits all approach cannot be applied to the cancer field adequately, and a more individualised approach could be taken to address differences in mutation of the genes and other differences in the cancer being treated. The journals that have been researched over the project all predict the drug sensitivity from genetic screens by using either different models or a combination of different datasets. As a result of the previous research in this area, it has shown that using either loss of function scores or gene expression is both possible and accurate in the prediction of drug sensitivity [2], [5], [7], [8]. Furthermore, by using the gene expression dataset as input features, the results show increased sensitivity. This is due to DNA being unique for every person and the gene mutation can be targeted in many locations of the DNA by using the copy number of the gene affected [5], [9].

According to the research paper *"Genomics of Drug Sensitivity in Cancer(GDSC) : a resource for therapeutic biomarker discovery*" the manner in which the patient will respond to treatment, is heavily dependent on the changes observed in the patient's genetic material [9]. As part of the GDSC project. they created a database that hosted three diverse datasets. This database is updated every few months, which ensures that new researchers are always working on the most recent data. The data sets hosted here are:

1. A data set that contains both endorsed and drugs in their infancy of development-Cell Line drug sensitivity.
2. A data set that represents the portfolio of cancer types which includes gene information and mutations
3. The third data set consolidates both data sets to determine genetic markers to drug response [9]–[11].

It can be clearly seen throughout the research, that the visualization of the data is an important aspect of the data mining process. It can give us clarity by producing plots that appear to express some information about the data set and the patterns can become more definitive. Below, we can see the visualization of the drug sensitivity. In this representation we can observe drug resistance and drug sensitivity delineated by red and green respectively in the *figure 1*. The dimensions of each circle are larger, which signifies the amount of cell variants screened for each drug. This particular figure shows that the BRAF gene commonly found in breast cancer is highly responsive to the PLX4720. Whilst, the APC gene which plays a vital role in colorectal cancer is highly resistant to the same drug. These studies show the importance of applying these techniques to biological data because from this pattern it can be said that if a patient has colorectal cancer, then it would be absurd to treat that patient with PLX4720, due to the fact the gene displays high resistance to that particular drug. Subsequently, having noticed this pattern the cancer treatment development, can proceed to discover drugs that will show sensitivity in the APC gene [9].

figure 1:[9]



As seen from the above illustration cancer cells show that the are distinctly sensitive or resistant to particular drugs, dependant on the particular strand of mutation caused in cancer cells. This allows us to eliminate drugs that show high resistance in the treatment of certain cancer cell lines and concentrate on the creation of treatments, that pinpoint genes that display high sensitivity to a drug. In this project the idea is to replicate the elements produced in previous research papers, then run experiments and analyse the findings of this project.

2.1 Correlation Coefficient

The correlation coefficient is also known as the R value. It can be used to determine if there is any influence in the model. There are many correlation models however the pearson correlation is a common model implemented on linear models. It measures the degree to which two variables show commonality. The pearson correlation coefficient return a value into the range of -1 to +1. If there is no linear relationship then a value of 0 is returned whilst a negative correlation is returned usually between the ranges of - 0.1 to -1. Similarly a positive correlation will have 0.1 to 1 with values closer to one being highly similar. The r value can be squared to allow ease of evaluation which would then return a percentage value [12]. This value is used to depict the explanatory power of the model with the value closest to 1 illustrating how closely the regression line regression line fits the real data.

2.2 Feature Importance

When applying a machine learning algorithm, the selection of the features utilized in the model is highly important. In the decision trees algorithm, the features are selected based upon the amount of information gained or entropy. Entropy is a calculation of spontaneity in the dataset. The more arbitrary or the higher the entropy to 1 results in the most information gained from the dataset. The formula for information gain is split into two parts[13]. In the first part (see *figure 2)* calculates the entropy of the training set S when p represents the positive samples and q the negative samples [13].

figure 2

$$H(S) = -p\ log_2\ (p) - q\ log_2\ (q)$$

In part two (see figure 3) of the equation the reduction of uncertainty by assessing the predictive class of a feature f. This part of the equation uses part 1 of the equation and takes away part two of the equation from part 1 to get the information gain.

figure 3

$$IG(S,f) = H(S) - \sum_{v \in values(f)} \frac{|Sv|}{(s)} H(Sv)$$

In the random forest regression model we can use standard deviation and standard deviation reduction to calculate the amount of information gained. The standard deviation reduction results in the reduction in standard deviation when slit on an attribute. The split with the highest standard deviation reduction gives use the most alike attribute to the previous.

2.3 Regression

Regression is a technique used when the data cannot be classified as the values a continuous[14]. This means they cannot be clearly defined into a class. It is a technique which is commonly used in predicting future outcomes and identifying relationships between a predictor and the target. In this case it tries to identify the drug sensitivity form the genetic screens. It takes into consideration the variation between two data points and to what extent the points ly away from the curve or line. It is a useful tool in the analysis of data and when it comes to using a regression model it can be implemented in many ways which will be addressed in the following sections[15].

2.4 Linear Regression

Linear regression is the simplest form of regression. The goal of the linear regression is to examine the relationship between the predictor and the outcome[16]–[18]. Linear Regression is normally fitted to the line, using a method known as the ordinary least squares which acts as the estimator in the model. This is basically the sum of the values achieved, when we subtract the predicted value from the true value. These are also called residuals in regression. It is particularly useful where the variables are dependant on each other, and can produce results that show significant relationships, which are indicated by the beta estimate. The extent of the estimate quantifies, how dependant or close the relationship is between two variables. The simplest linear regression formulae (*see figure 4*)where y is the estimated dependant variable score c represents some constant the b value is the regression coefficient and x is the score of the independent variable.

figure 4

$$y = c + b * x$$

The estimates produced provide an explanation of the relationship between a dependent variable and the independent values .These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

2.5 Elastic Net Regularization

Elastic net is an algorithm that is particularly useful in the adjustment of linear and logistic regression models. It applies regularization which identifies badly formed functions and tries to devise changes in these issues during the refinement of the algorithm. It provides the regularization in the algorithm by accessing the combinatorial consequences on L1 and L2 [19].

Elastic net, is in simple terms a blended algorithm, that merges the best aspect of lasso regression and ridge regression. Both the lasso and ridge regression focus on the cost function, but apply different methods of evaluation. Both algorithms make use of the calculation of the residual sum of the squared errors(RSS) [11], [20]. This is calculated by calculating the squared errors of the predicted outcome versus the actual outcome. Subsequently, ridge regression attempts to mitigate the impact of the weight of the sum of the square magnitude, plus the RSS[1], which results in suppressing the L2 norm. This relates to the euclidean distance or the distance between two points based on the path between the points as the crow flies.

Alternatively, the lasso regression algorithm focuses on suppressing the L1 norm. Thus, in contrast the L1 norm is associated with manhattan distance or the distance between the points given the horizontal and vertical paths. The lasso algorithm calculates the cost as the RSS[1] plus the sum of the absolute values of the weights. The result of using the L1 will cause some of the coefficients to be reduced to zero. The advantage of this property, is that it bodes well in feature selection, where once the coefficient has reduced to zero it is excluded from the model. Ridge regression and Lasso algorithm each have their distinct advantages at opposite ends of the scale where Lasso performs badly and Ridge regression well. The Elastic net algorithm combines the Lasso and Ridge regression therefore optimising the algorithm to perform in well in either situation by balancing the L1 and L2 norms effectively [20].

2.6 Random Forest

Random forest is an ensemble method in machine learning. Ensemble methods have been used in oncogenic data research and have shown peak functionality on a continual basis [21] This type of algorithm allows the combination of different learning models for instance bagging. Consequently, the model gains' accuracy in its classification by its use of bagging which normalizes noisy and impartial models. The basic concept in the random forest algorithm is that, it creates multiple small decision trees by breaking the main data set into subsets, normally called stumps. These are used to decide on a classification, this type of algorithm is generally used on large collection of data [17], [22]. Random forests are part of the CART [2] Model which means they can be used to both classify and predict. In the classification task the output is generally categorized into some class. In contrast, regression predicts a numerical value by assessing the variable with

[1] Residual Sum of the squared errors
[2] CART- Classification and Regression Tree

respect to everything already known in the dataset. The random forest regressor model is usually implemented on continuous variables.

The algorithm for classification and regression for random forests is applied in the same way. As, a first step we select with replacement n-trees from the original dataset. Next, for every sample the tree, allow the tree to develop without pruning any branches from the classification or regression tree. However, at every node instead of allowing the branches to be chosen by the best split, a random sample of the predictors is used. Subsequently, we can anticipate a classification or regression for new data based on the majority vote for n - trees [23], [24].

The error rate which relates to the performance of the model can also be approximated from the training data. This gives the analyst an idea of the accuracy of the model. For this, in every sample chosen with replacement, we observe the data not selected. This is known in machine learning as the out of bag data. This out of bag data is added together on the premise of its average of every data point. According to the estimated value for the error rate of the out of bag data is a good measure of the accuracy and performance of the model as long as a sufficient amount of trees have been utilized in the model [24] .

2.7 Decision Trees

The decision tree builds an unpruned regression tree of maximum depth unless otherwise specified[25]. The key element of this algorithm is that it uses the gini index to split the branch until the purity of is closest to zero. In this way, it is the opposite of the entropy function which measures information gain where the best value is equal to 1. The advantages of using the gini index is that it is better suited to bigger datasets because it also reduces the computational power as it doesnt use the logarithmic function like the entropy. This means that where the node that shows the most purity will be at the top of the tree.Yet, this can be very ambiguous when we are dealing with continuous variables because two branches could have identical or extremely close to identical distribution[26].

2.8 Multiple drug-gene interactions

Significantly, projects and investigations in the area of drug sensitivity, providing a methodology by which cancer treatments can be improved objectively by assessing biomarkers and their sensitivity only taking into consideration the patterns contained in the results. By analysing the data in a neutral clinical way breakthroughs can be achieved by uncovering connections present that may appear unconventional and atypical. However, these connections can provide groundbreaking patterns that may have been overlooked when analysing this data without the application of machine learning techniques. A Landscape of Pharmacogenomic Interactions in Cancer [5] attempts to further delve deeper into oncogenesis by considering that, there may be interplay between cancer functional events. Therefore, assessing multiple drug gene relationships in comparison with single exchanges has shown to offer more precise drug sensitivity measures. This can be seen in " *Improved Survival with Vemurafenib in Melanoma with BRAF*

V600E Mutation" in which the BRAF gene mutations had been several molecular relationships [27]. The results of this study emphasise that better treatment of cancer is dependant upon accessing multiple drug gene relationships. The body of research work has shown that for every type of cancer, we can create models, that are accurate in identifying cancer driver mutations and copy numbers that provide a more robust predictive model. Consequently, by locating changes in the levels of activity, we can refine the predictive model. It has been portrayed that the cancer treatment field, could benefit considerably if the collection of data concerning changes in the activity area were seen as a key objective[5]. While this process would take time and resources, over time it could prove to be more cost effective in the development of new cancer treatments, because applying the machine learning algorithms to the data, the outcome would indicate the genes and the mutations that would best treat the type of cancer.

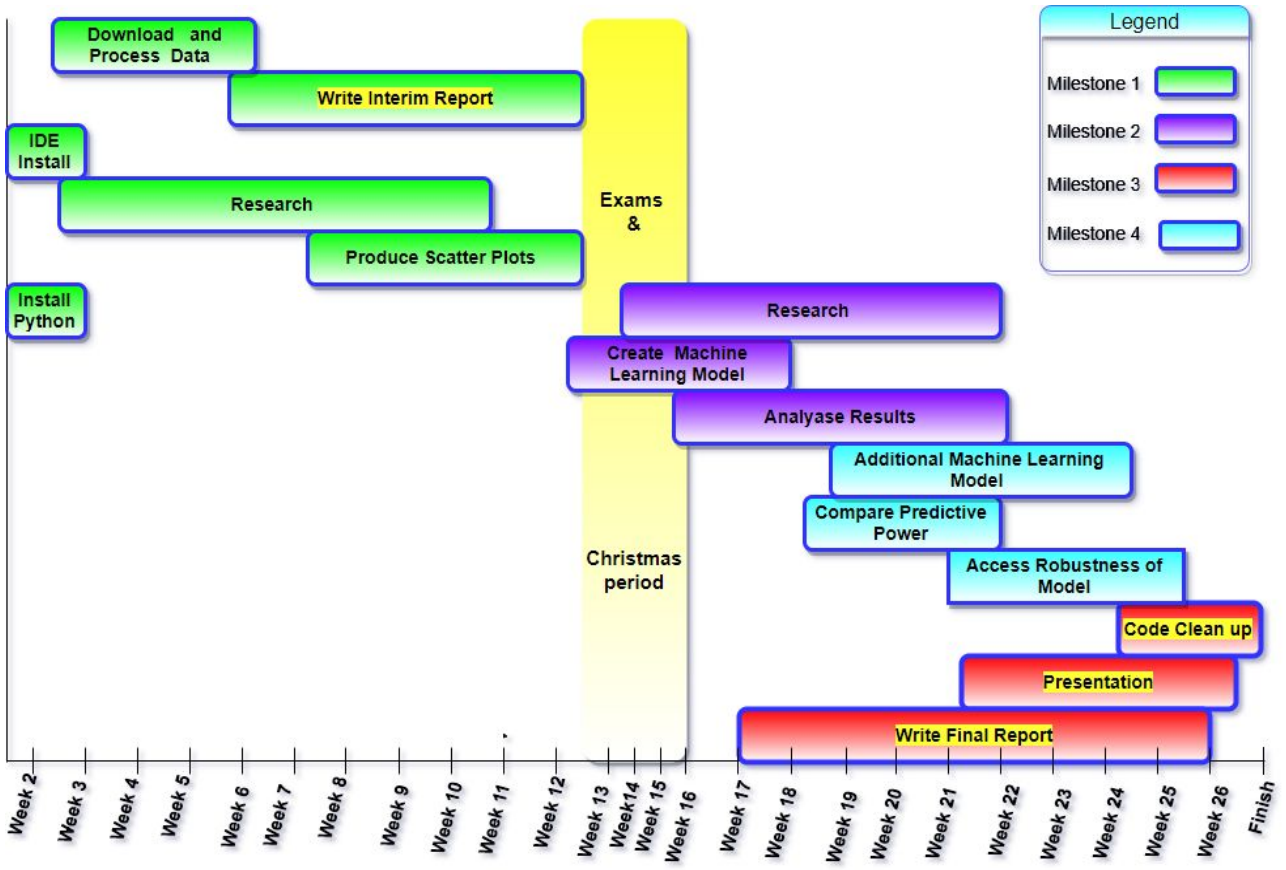
According to the previous research, in this area of predicting drug sensitivity from genetic screens, the majority of the results have shown that its is possible to predict drug sensitivity in in a systematic way[2], [6], [8]. By applying these techniques to biological data especially the CCLE data, has resulted in being able to identify biomarkers of gene mutations. Thus by learning about the patterns and identifying the genetic sequences, more complex and predictive models can be created which in the future will assist in the enhancement and advancement of developing better clinical trial more individualized treatments and identifying whether a drug might be effective for a particular cancer[9].

3 Detailed Design and Implementation

3.1 Project Management

The most important aspect of the project at the beginning was to identify the main goals of the project and the tasks that were involved in achieving the outcome of the project. As a first task I identified 4 key milestones as seen in *figure 5* below. I have coloured the tasks related to the time frame to achieve the task and what milestone the task belongs to. This diagram will serve to provide a rough outline of the project timeline and the accomplishments that need to be achieved in order to complete the project on time. In the following sections I will make reference to some of the design and implementation concerns giving a brief overview of the some of the key element. Subsequently the next chapter will focus on the analysis aspects of the project leading towards the future work and conclusions of the project. The tasks with the highlighted text represent the items that need to be submitted over the course of the project.

figure 5



3.2 Programming Language Selection

One of the key considerations that arose during the planning and preparation of the project, was that a programming language would need to be decided upon, which would be used to produce scatter plots and machine learning models within the project. Python and R were the two languages accessed during this process, as they maintain good libraries for statistical analysis and machine learning. Python seemed like a good option due to its ability to also accommodate the use of R within python. This provides some flexibility by allowing for code to be written like R. A key element, that was played a massive part in the decision to use Python is that Python is highly used within industry. Hence, learning Python as part of this project, would help develop skills in this area. It also provides extensive machine learning libraries like scikit, numpy, scipy and matplotlib, that would be useful in implementing aspects of the project. Taking into account previous experience with R versus no experience in Python, I decided to use python, due to its prominent use in the bioinformatics and the promise of developing a new skill over the course of project, that is highly used in industry. Following the installation of Python 3.6, there was issues with installing some of the libraries needed for this project on Windows after some research around this, it was clear that this was an issue for a lot of Windows users. After further investigation, all the packages needed were installed through Anaconda which is is a package manager that allows you to set up environments run notebooks and download packages[28].

3.3 The Data Sets

The data sets that will be utilised in this project can be downloaded through the Broad Institute website portal for this a user account must be created.

Drug Sensitivity Dataset

The first data set contains pharmacological profiles of 24 anti cancer drugs with 504 cancer cell lines. This dataset also contains key information about the mean the standard deviation and how the curve would be depicted over a period of doses. However, the values that will be most important in providing some knowledge about drug and gene interactions after doing some background research will be IC50, EC50, Activity Area (Act Area) and the Amax values. These values represent quantitative values that characterize the level of response between the cell line and the drug with the Amax value showing the maximal level response, the IC50 is the concentration of inhibitor where the response is reduced by half and the EC50 characterizes the potency of the drug where the concentration gives a half maximal response.

Loss of Function Dataset

The second data set also can be found on the Broad institute site as part of Project Achilles. This data set contains DEMETER inferred gene knock-down effect in cell lines. The DEMETER algorithm [21] forecasts the criticality of a gene by providing us with numerical data with describes by using standard deviation how far away from the mean the data point is, represented on the bell curve the values were most interested in are the values at the upper and lower ends of the bell

curve . These values will refer to sensitivity or resistance of the gene. The values that fall into the center of the bell curve does not provide any useful knowledge . The values provided in this dataset are known as a z scores which can help in understanding the function of a gene and pinpoint genes that are characteristically similar.

Gene Expression Dataset

The third data set can be downloaded from the Broad institute portal. The dataset would have been formulated using microarray analysis techniques. This contain mRNA which carry the genetic information which translates DNA to a sequence of protein products [7].

3.4 Data Preprocessing

A fundamental concept in any machine learning or data mining project is to ensure the data that will be utilise has been processed for any irregularities. Thus using a data set that contains discrepancies or consists of redundant values that provide no useful information creates noise. This can skew the results of the dataset and important knowledge will be lost particularly when we progressing to the machine learning aspects of the project.

As we have seen this project contains three data sets. As a first task, on the drug sensity dataset part of the process involved creating 4 pivot tables of IC50, EC50, Activity area and Amax scores. The creation of the pivot tables allows us the select an index and a columns and transform the data into the respective cell. It simple terms it allows us to extract key information from our original dataset by creating a smaller dataset based on the data identified. This was done 4 times due having four drug sensitivity scores in the dataset which are mapped 24 drugs to the 504 cell lines. After completing this process I was able to reduce the dimension of the data frames to only contain the same column headers by utilise the intersection and difference functions in python. This produced data frames that were (24,260) for the drug sensitivity values and (~17,000, 260) for the loss of function dataset. Next , the nan values needed to be addressed so where a column contained all nan values it was dropped. On the other hand where there was some nan values which could not be dropped, the missing data was filled with the mean of the row. During this step the reordering of the columns took place so that both datasets had the same shape and order . This would be an important aspect especially when moving into applying cross validation in the predictive models where the datasets are split into test and train sets. these must be divided so that the train and test of the drug sensitivity matches the ordering and distribution of the loss of function dataset. This process was repeated for the gene expression dataset which had a shape of (~18500,258). Due to the shape changing the drug sensitivity scores had to be preprocessed again. Since, the gene expression was only assessed in applying the machine learning ,models and not the scatterplot, the activity area scores were the only drug sensitivity scores that needed to be reprocessed.

3.5 Plotting the Scatter Plots

This task involved taking the data which had been preprocessed and turning the raw data into scatter plots. For this task the plots are of only the drugs with known target genes. The total number of scatter plots produced was 200. Each plot contains an R value was also calculated that tells how closely the drug sensitivity and gene inhibition are related. The R value is calculated using the pearson's correlation which returns a value between -1 and 1 with values closer to 1 representing a positive relationship and likewise values closer to -1 a negative relationship and values nearer to zero contain no relationship. For every relationship there is four scatter plots each representing a different drug response, the IC50, EC50, Amax and Act Area values.In hindsight, it has been difficult to access the scatter plots as I haven't labelled the data points. This is something I will try to address later in the project. However on analysing the scatter plots the quantification that appears to achieve better separation in the results is the Act Area. Where the gene inhibition decreases the sensitivity to the drug increases creating a negative relationship [8]. The plots were produced using matplotlib which is a python library for producing 2D plots in conjunction using the backend module a pdf of all the plots was produced [29].

3.6 Machine Learning Models

A requirement of predicting the drug sensitivity from genetic screens was to apply and compare the accuracy of machine learning models. During the basic part of the project elastic net was selected as the main model that be would apply. The selection of elastic net was based on the fact that it combines the best aspects of the LASSO and Ridge methods to create a more optimized model. This has also shown that due to its use of regularization, it can generalize to the date more effectively, which can assist in the prevention of overfitting . All of the machine learning algorithms were implemented using the scikit python library[30].

Elastic Net

The default values were used in creation of the elastic net model with the exception of alpha which is normally defaulted to 1 . The alpha value was reduced to 0.7 which provided the feature importance with more values to analysis. The alpha value in the case of the elastic net is the constant that multiplies the penalty. The L1 ratio was then set to 0.2 instead of 0.5. The L1 ratio is usually a value between 0 and 1 this weights the penalty applied towards L2 or L1 values closer to 0 result in an L2 penalty where values closer to one result in an L1 penalty.

Linear Regression

The default values were used in this model.

Decision Tree Regressor

Again in the case of the Decision tree regressor the default values were chosen, as the max dept was not specified, this would result in an unpruned fully developed tree which could be problematic for memory consumption with large datasets.

Random Forest

All the tuning parameters used in the random forest were again mostly the defaults with the exception of the n estimator which is normally defaulted to 10. This value relates to how many trees are allowed in the forest by increasing the integer to 24 the accuracy of the model also increased.

3.7 Predicting the drug sensitivity

In order to find out the accuracy of a model it is standard to divide the data set into three data sets the training the test and the validation set. As a result the testing capability and the model itself can be affected, especially when the datasets are small. The datasets in this project were only representative of the intersection of the drug sensitivity data set and the loss of function dataset because of this the data that was being used in this project was on the smaller side. For this reason, a cross validation approach was used in predicting the drug sensitivity from loss of function screens. Subsequently, we could eliminate the need for the validation set by using the train set to train the model and the test set to predict the values. This works as it allows the partitioning of the data into a k amount of subsets and completes the prediction on each subset combines the predictions then divides the result by k. This return the average prediction of the results of multiple rounds and minimizes inconsistency in the predictive model(*see figure 6*).

figure 6

4 Analysis and Evaluation of predictive models.

This chapter will focus on the evaluation and analysis of the results achieved by the predictive models implemented after which we will proceed to the future work and the conclusion of the report.

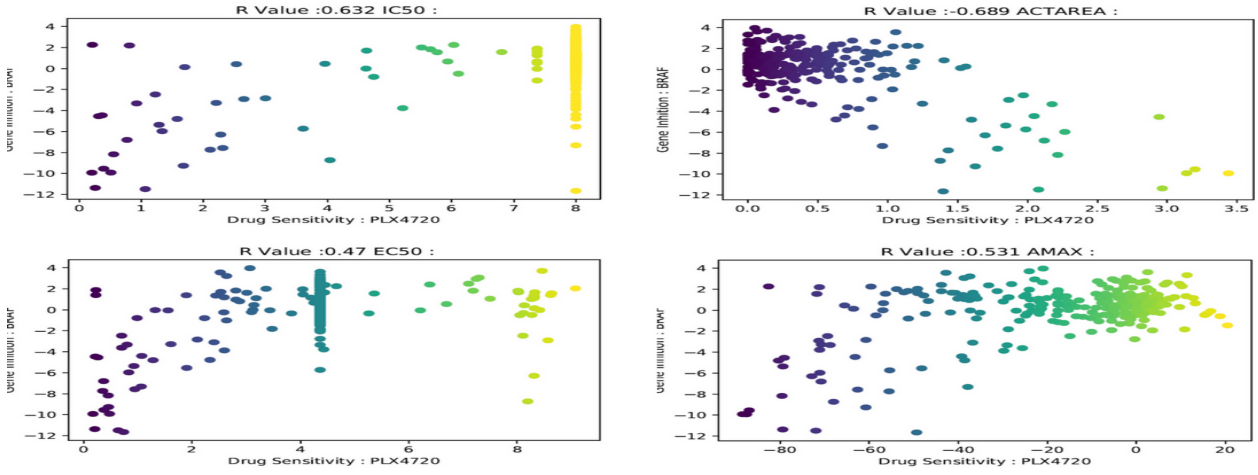
4.1 Quantification of gene inhibition measurements predicting drug sensitivity.

Gene inhibition measurements can be utilised to predict the drug sensitivity measurements by computing the correlation coefficient, it is possible to assess the correctness of the quantification. The visualization of a known genes relationship with the drug sensitivity allows us to illustrate, the results more clearly. The average correlation across all drugs for each drug measurement are as follows :

	IC50	EC50	Activity Area	Amax
Average Correlation	0.148000	0.028000	0.209000	0.224000

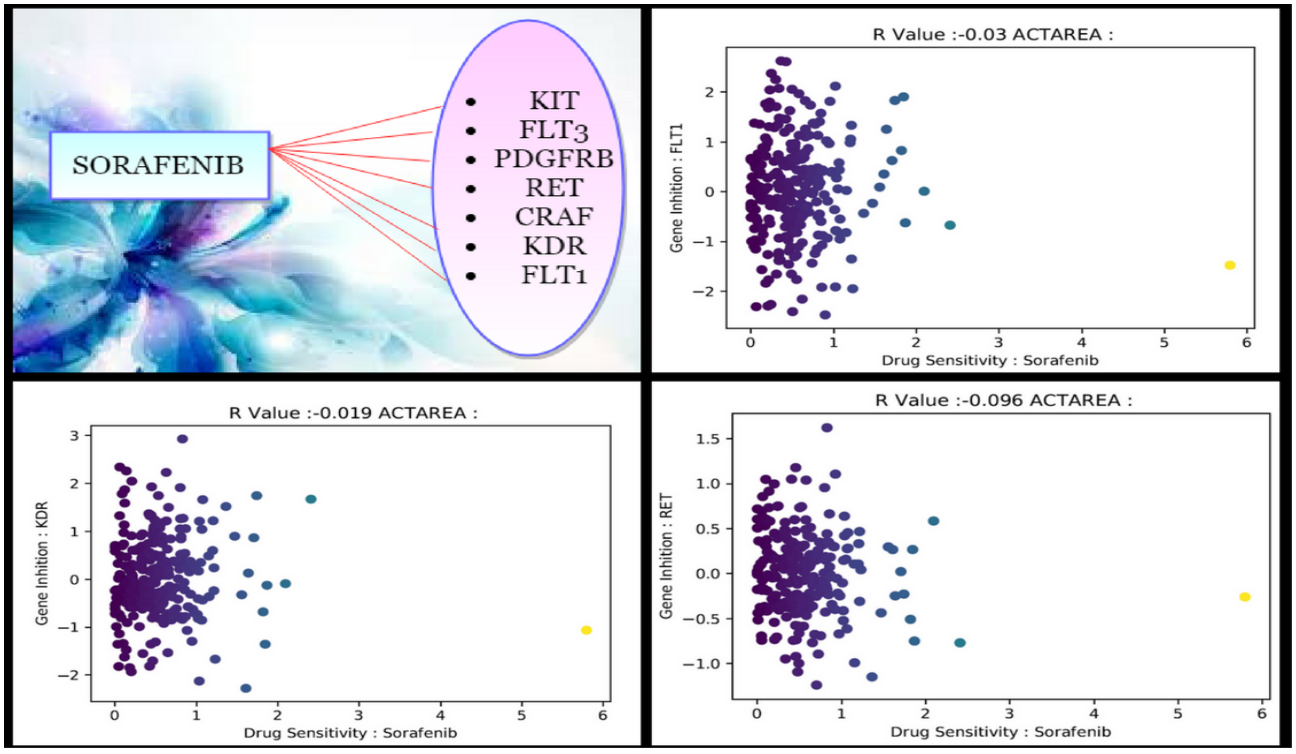
The average correlation above allows us to assess the scatterplots of known gene inhibition measurements in how well they can predict drug sensitivity. Thus assessing the below combination of scatterplots for a known drug PLX4720 which we know targets BRAF we can see that IC50 EC50 and the Amax scores all show a more positive relationship than that of the average correlation. It would be expected that this would be true as in the scatterplot we are just producing the correlation for known drugs to gene targets. Likewise the activity area shows a much more negative relationship than the other 3 measurements. This seems to work well for drugs that have single known targets like PLX4720 (*see figure 7*) more m scatter plots can be found at the link [All scatter plots.]

figure 7



However the drugs mapped to the gene targets do not have a one to one relationship in most cases like for instance Sorafenib which has a one to many relationship. In this case scatter plots are less conclusive, for instance just observing some of the drug sensitivity prediction for known gene targets, where there are multiple targets, we can see that compared to PLX4720, there appears to be little to no relationship. Given that the correlation is closer to 0 (*see figure 8*). The below figure consists of an example of the one to many relationship for the drug sorafenib which we know interacts with the list of genes in the top left of the image. All of the 3 other images show the scatter plots for sorafenib with respect to the genes FLT1 ,RET and KDR in which we can clearly see the lack of relationship where there are multiple targets.

figure 8



The above analysis would suggest that while the scatterplots quantify that the gene inhibition can predict drug sensitivity effectively for drugs with single targets , this is not the case when we address the same question for multiple targets. In the event of multiple targets scatter plots do not accurately represent the data and is not a good quantification method when multiple targets are involved.

4.2 Predicting the sensitivity of cell lines to a drug (loss of function).

After, predicting drug sensitivity using known target genes and quantifying the results by depicting the data on scatter plots. The next part of the project was to apply the machine learning approach just using the the loss of function data set. In this i choose to apply both the random forest and the elastic net.

We can see from the resulting figures 9 and 10, a contradiction in the R^2 score when the model is applied to all the drugs see figure compared to when it's applied to just one see figure . When the predictions are made using just one drug PLX4720 the elastic net outscores the random forest however when applied to the full model the random forest is better. In this case when applied to just 1 drug the performance could be increased in the elastic net due to the dispersion of the data and the noise also the nodel may pick out the negative linear pattern as seen in the scatterplots. However when it comes to the overall model the random forest is more predictive and suppresses the error better this is indicative of the ensemble model where multiple techniques are combined which results in the reduction of overfitting.

figure 9

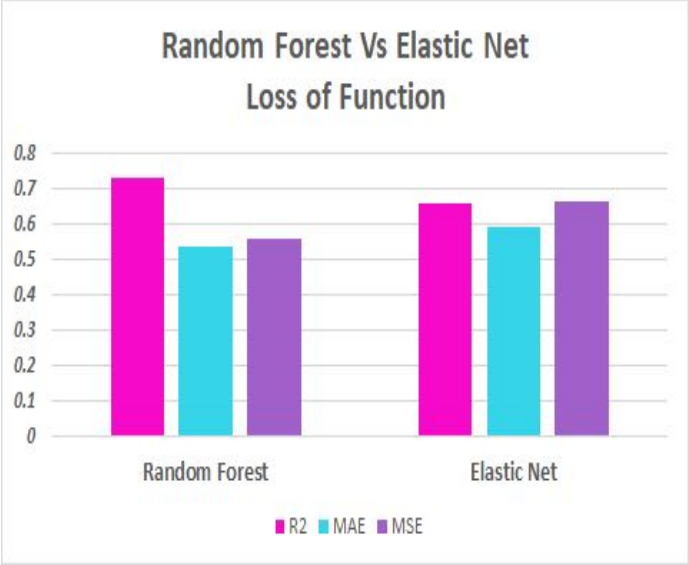
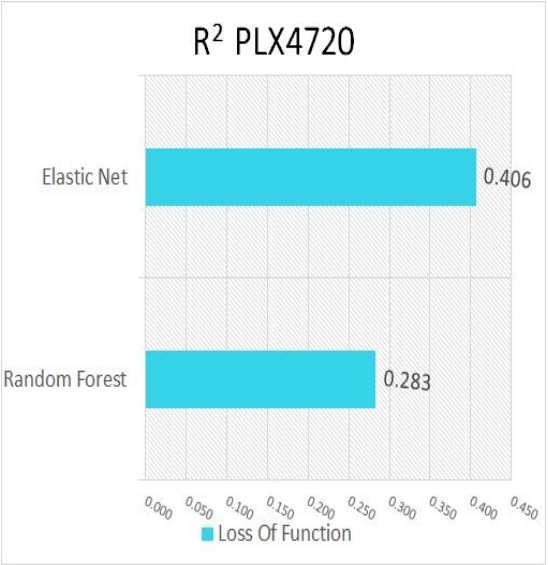


figure 10



The snapshots of the results of the R^2 values for the elastic net (*see figure 11*) and the random forest *(see figure 12*) show the elastic net while a good predictor of PLX4720 the model does not generalize as well to the overall model. In comparison, the random forest appears to balance the variance and bias by using bagging or bootstrap aggregation which uses a random subset of the train set to train the model. The models are then voted on with equal weight. It is clear for the illustrations that the random forest produces scores that are all in the same plane where the elastic net it is clear the R^2 score is highly biased in particular to the PLX4720 drug, making the model a good predictor for some but not all drugs.

figure 11 (elastic net)

```
Results for 17-AAG  is :0.020711
Results for AEW541  is :-0.162355
Results for AZD0530  is :0.016263
Results for AZD6244  is :0.018219
Results for Erlotinib  is :0.106668
Results for Irinotecan  is :0.116015
Results for L-685458  is :-0.166551
Results for LBW242  is :-0.241120
Results for Lapatinib  is :0.087067
Results for Nilotinib  is :0.013985
Results for Nutlin-3  is :-0.233790
Results for PD-0325901  is :0.121917
Results for PD-0332991  is :-0.218005
Results for PF2341066  is :-0.135675
Results for PHA-665752  is :-0.125694
Results for PLX4720  is :0.545472
Results for Paclitaxel  is :-0.143110
Results for Panobinostat  is :-0.066476
Results for RAF265  is :0.079778
Results for Sorafenib  is :-0.135012
Results for TAE684  is :-0.196066
Results for TKI258  is :-0.091494
Results for Topotecan  is :0.068276
Results for ZD-6474  is :-0.285388
```

figure 12(random forest)

```
Results for 17-AAG  is :0.036835
Results for AEW541  is :0.030563
Results for AZD0530  is :-0.209550
Results for AZD6244  is :-0.026401
Results for Erlotinib  is :-0.084457
Results for Irinotecan  is :0.085566
Results for L-685458  is :0.041012
Results for LBW242  is :-0.038975
Results for Lapatinib  is :-0.017089
Results for Nilotinib  is :-0.051380
Results for Nutlin-3  is :-0.122167
Results for PD-0325901  is :0.048569
Results for PD-0332991  is :-0.109737
Results for PF2341066  is :0.012106
Results for PHA-665752  is :-0.042589
Results for PLX4720  is :0.175984
Results for Paclitaxel  is :0.016791
Results for Panobinostat  is :-0.017432
Results for RAF265  is :0.020512
Results for Sorafenib  is :-0.082282
Results for TAE684  is :-0.103828
Results for TKI258  is :-0.033861
Results for Topotecan  is :-0.131189
Results for ZD-6474  is :-0.079679
```

4.3 Relationship between feature importance and reported drug targets

After applying the elastic net regularization, to predict the sensitivity of cell lines to a drug using the loss of function screens as input features, we can analyse the relationship between the feature importance and the reported drug targets. Similarly, to the scatter plots we are looking to establish if the reported target gene or known target gene is always the most important feature. In the event two genes are targeted are they both identified as important features? Likewise, what happens when there are multiple targets?

Thus, taking these questions into consideration when we are analysing the results of the feature importance *(see figure 13),* like the scatter plots shows a lack of consistency when there are multiple know target genes. Sorafenib, as we can see from the figure below shows that while we expected KIT ,FLT3, RET, KDR to be important features, due to the fact that we know these genes show high sensitivity to the drug, it is interesting to see, none of these genes appear in the top 6, when the feature importance is assessed by our elastic net model. Moreover, after delving a little deeper into the top features these known target genes don't even appear within the top 40. However, where we have a drug like PLX4720 where BRAF is the expected know target gene on seeing the results we can see that BRAF preserves its feature importance as the number one feature importance for the PLX4720 drug. This may be due to the data, for instance we know much more about BRAF and PLX4720. Thus, because it has been researched in depth it could simply have a much higher scores within the dataset that shows its definitive relationship and interaction. When assessing the top 10, PLX4720 compared to other drugs known target gene did have substantially higher value for the BRAF gene compared to the next top gene.

figure 13

Feature Importance

Expected

	LAPATINIB	PLX4720	SORAFENIB
1	EGFR	BRAF	KIT
2	ERBB2		FLT3
3			RET
4			KDR

Results

	LAPATINIB	PLX4720	SORAFENIB
1	SRPR	BRAF	RCCD1
2	COPS6	CPA6	SNRPF
3	CDC37	PPP6C	SRRT
4	SNRPC	TAF1	HDAC11
5	ERBB3	ZNF19	FLNA
6	EGFR	SSRP1	TACR2

Where a less obvious relationship exists between a drug and a gene, the feature importance becomes less important especially where a drug targets multiple genes with equal or close to equal values. As the values become more separated or further away from each other the feature importance works better. Hence, as cancer research develops and drugs are linked more effectively to genomic changes the effectiveness of the drug like in the case of PLX4720, would increase for certain genes, which in the case of multiple target genes. It would be expected that the feature importance may show some improvement in predicting drug sensitivity, when there are multiple targets.

4.4 Comparison of the predictive power of the models

This was part of the advanced aspects of the project. The function of this part of the analysis is to compare the models ability to predict the drug sensitivity from the loss of function screens to its ability to make drug sensitivity predictions based upon the gene expression data set.

In this step, the comparison was completed with two models the random forest and the elastic net for each of the datasets. The below figures give an illustration of the predictive power of the models using the following measurements R^2 , mean absolute error(MAE) and mean squared deviation(MSE). These values measure the accuracy of the model by assessing the error and the goodness of the fit of the model. The R^2 is based upon the total variation of the model it tells us how accurate the model is and the best possible value of R^2 is 1 . The MAE[3] and The MSE[4] both have a best possible value of zero they give us an idea about the variance and bias tradeoff which allows us to assess whether the model is a good predictor of the data.

First, the results of the model can be validated by predicting the drug sensitivity on both datasets by running the prediction on the train then by running the prediction on the test. The idea is to analyse the difference in the error. This tells us if the model is adequate for predicting

[3] Mean absolute error
[4] Mean squared error

the drug sensitivity . We can see from the table which is only based on the elastic net that the training and test error remain relatively stable. This would indicate that the model is a good fit and has equal amounts of variance and bias. It means that roughly 50% of the time it is of the time the predictions are not precise . However the model is accuracy is around 72%. If there was a high error in both the test and train it would indicate that the model was highly biased on the other hand if there is a low train error and high test error it would suggest that the model is overfit or contains high variance. However from the table below we can see the models error rate is similar in both the test and train . This tells us the data is well fit and the model is accurate. If we could decrease the error and increase the R^2 this would make the model more precise and more accurate.

Elastic Net	Loss of Function		Gene Expression	
	Train	Test	Train	Test
R^2	0.724	0.656	0.734	0.702
MAE[3]	0.531	0.59	0.517	0.542
MSE[4]	0.533	0.683	0.518	0.577

We can also observe from the table that the gene expression dataset appears to yield a more accurate model, showing a slight increase in the R^2 value. However, the error especially when you consider the MSE[4], shows more stability in the error compared to the loss of function.

Again looking at the figures below we can see from the the results in the below illustrations of the loss of function *(see figure 14)* and the gene expression*(see figure 15)* when we apply a different model like the random forest. In this illustration, we can see that the Random forest is the better predictive model as the R^2 value is higher in the random forest than the elastic net but also it shows a lower error rate which suggests it is slightly more precise with less risk than the elastic net for the loss of function. In comparison, by using the gene expression we can again increase the accuracy of the model and the error when we use the gene expression is decreased in both models.

figure 14

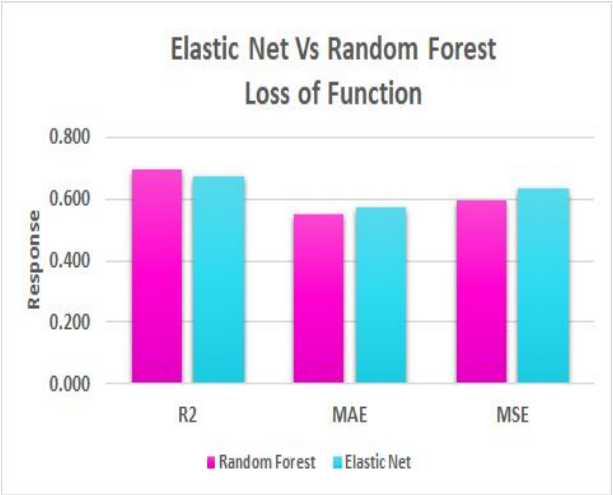
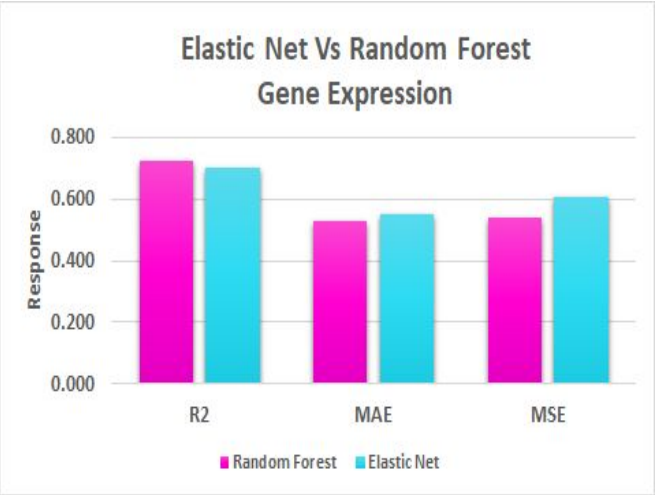


figure 15



After assessing the outcomes of the predictive models on both the loss of function and the gene expression, we can see how both the elastic net and the random forest work well as a predictive model. However, the best predictive model appears to be the random forest with the gene expression data set as it reduces the risk increases the R^2 values and fits the model more accurately.

4.5 Comparison of the robustness of the predictive models

The initial objective of this part of the analysis was to assess the robustness of the predictive models by testing the predictiveness of the models with other resources like the Genomics of Drug Sensitivity in Cancer (GDSC) datastes. However, due to the amount of preprocessing of the datasets needed to complete this part of the project and the time constraints it was not possible to complete this part as specified. Insted a comparison of more models using both of the data sets was undertook.

Subsequently, the analysis of this piece of the project will look at the predictiveness of decision trees and linear regression, alongwith the previously selected models i.e random forest and elastic net. First, we will the address predictiveness of the models by just looking at a known drug PLX4700 which we already know shows good sensitivity to BRAF. We will look at the accuracy score of PLX4720 over the whole model.

The charts below depict the accuracy of each of the models for the drug PLX4720 with respect to the loss of function *(see figure 16)* and the gene expression *(see figure 17)*. We can see that if we just take into consideration this one drug, the loss of function provides the best model. It is surprising from the results that linear regression appears to work better than the random forest , even though this is not a linear problem. The decision tree as we can see from the results is by far the worst model. The decision tree regressor uses the gini impurity as this data contains continuous variables. The negative R value indicates in this case that thread of the data isn't followed by the predictive model. This means that the Decision tree is particularly bad in this case.

If we are just to observe the PLX4720, the reason the PLX4720 may appear better, could be because of the data distribution, together with the fact that PLX4720, even though we know it targets BRAF particularly well, gene expression is not calculated in the same way so a different drug may be more predictive due to targeting copy numbers instead of sole genes.

figure 16

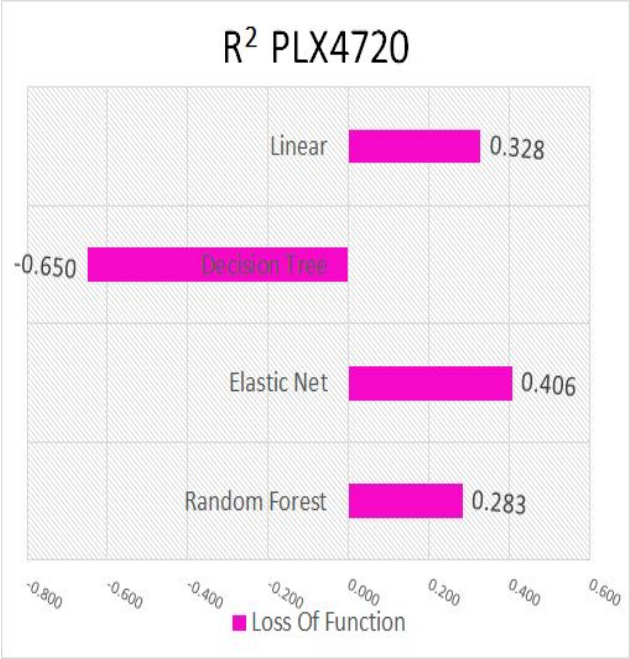
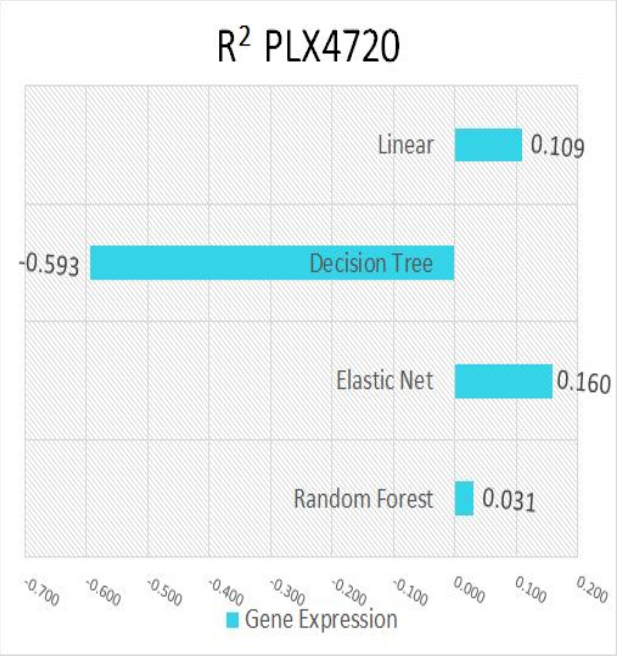


figure 17



Subsequently, when we observe the overall robustness of the models again we can see that the random forest with the gene expression is the best model followed closely by the elastic net with the gene expression. Interestingly, even though the loss of function is better when looking at one drug this observation is not true when applied to the overall model *(see figure 18-19)*.

figure 18

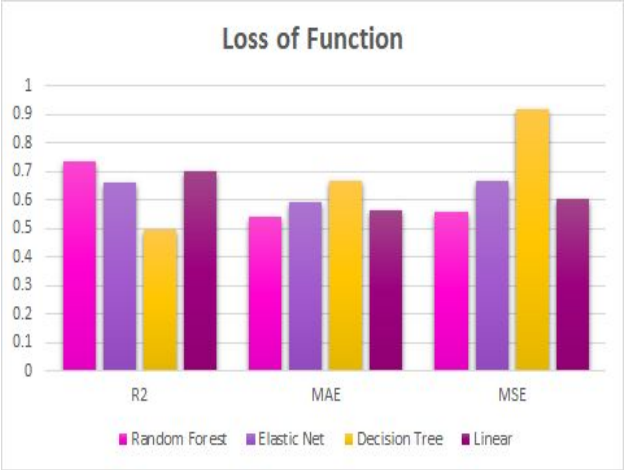
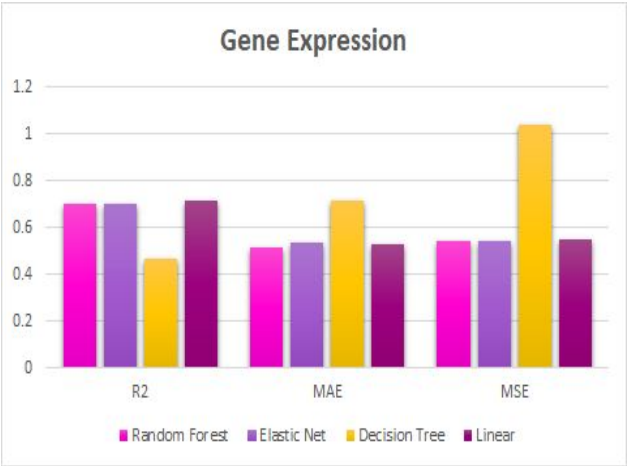


figure 19



A reason the gene expression may work better is that instead of targeting a gene like the loss of function it target the gene expression. A gene can reside in many chromosomes or strands of a human DNA and the quantity and location of these gene mutations are unique in every individual. The gene expression information uin very basic terms contains something similar to a unique identifier that when targeted doesn't just target the mutant gene itself but the gene in every strand on DNA . Hence , it would be expected that the gene expression would work better due to the predictive model being able to predict the drug sensitivity for the gene expression. This would also suggest that it is possible with more research and development in this area that the personalization of cancer treatments may be possible in the future. In terms of robustness both the elastic net and random forest provide a robust model especially when paired with the gene expression data set. These models are good, however further tuning could improve the fit, making them precise. Again, we can see overall the decision tree is the worst predictive model, in this case it grows the tree without pruning. This can also lead to high memory consumption. Linear regression, in this case also appears to work well even though the data is represented was fitted to a sigmoid function. This may be because the general line of a sigmoid function is s shaped, but there is a generalised linear representation, given the data split and the individual data points may follow a linearized pattern in the case of these data sets. These models would need to be assessed on unseen data and an additional dataset, in order to provide a more accurate estimate of the robustness of the model. It would be fair to say that the elastic net and the random forest provide the better predictive models. Random forest probably has the edge, due to it being an ensemble that helps in reducing the possibility of overfitting the data. However in terms of time to run elastic net would be more useful as it takes significantly less time to run. The data sets contained in this project are quite small, running the same experiments on bigger datasets would provide further information on the fit accuracy and robustness of the models.

5 Conclusions & Future Work

As we have seen from the previous chapters we were able to identify that gene inhibition measurements can be used to predict drug sensitivity especially in cases where there is one definitive target gene. However as the target gene becomes more tenetative with multiple target genes showing similar responses the scatter plots dont clearly depict the relationship between the drug and gene. We have established that when a machine learning model is applied the feature importance can be predicted more effectively where a single conclusive gene is the main target. Whils, as more targets are involved and the results become less definitive so does the feature importance. This suggests that as cancer research is an ever developing area, in which there vast amounts of information that remain unknown. Hence, as the field links the interactions between target genes and a drug is linked more successfully, we would expect the feature importance to also improve.

Similarly, we have seen that we can predict drug sensitivity from genetic screens and the accuracy and predictive power is dependant on the machine learning model selected but also on the dataset used as input features. It is clear from the findings that using gene expression yields a better predictive model than when the loss of function data set is applied. The robustness of the model was not assessed in line with the project specifications. Insted as we have seen from the analysis the robustness of the models was assessed by comparing multiple models. The motive behind assessing multiple models using first the loss of function dataset as input features and then the gene expression data set as input features, was that there was a significant amount of data preparation involved in cleaning and organising the GDSC data into a usable dataset. Subsequently, additional time would have been needed to achieve this advanced requirement.

Furthermore, to achieve a more rounded observation on the robustness of the models it would be ideal to evaluate the validity of the models on the GDSC data or another independent resource in the future. This would further enhance our knowledge of the stability of the model given that the testing and training error remain consistent.

Additionally, the machine learning models used in this research were applied with minute amounts of tuning of the model parameters. In some cases the defaults of the model were used. For this reason, the researcher would suggest the further experiments with different model parameters would be beneficial. This could provide further insights and better accuracy by tuning of the model parameters to the most optimal constraints. It would also be worth considering the time complexity of the models when deciding upon the predictive model to apply when investigation drug sensitivity from genetic screens . Whilst we have seen that both elastic net and random forest models performed reasonably well on the data. As we have seen the time complexity difference of the models is highly evident with the random forest taking almost three times the amount of time it takes to run the elastic net regression. Hence, it would be worthwhile

considering if the extra risk associated with the elastic net would drastically deteriorate you machine learning model, when deciding on which machine learning model to apply.

In conclusion, this field of study is in its infancy and it is highly dependant on the biological variables, the transcription of previous patients genomic responses to a drug. It is somewhat restricted as the experiments behind the datasets provided are usually performed outside of the natural environment using petri dishes and test tubes. However, this study would suggest that utilizing machine learning techniques could assist in the discovery of drug sensitivity patterns with gene alterations. Furthermore, it could provide insights that could potentially change the landscape of pharmacogenomics, from expediting new interactions it could help guide the drug development process and assist in the design of clinical trials. It could also provide an avenue to create a much more individualised treatment plan in the future given your specific gene mutations. It is clear that combining biological principles, pharmacogenomics and machine learning could provide useful knowledge that may assist in the area of cancer treatment development. Nonetheless, it still remains a significant task.

6 References

[1] "WHO | Cancer," *World Health Organization*, 23-Mar-2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed: 21-Oct-2017].

[2] J. Licinio and M.-L. Wong, *Pharmacogenomics: The Search for Individualized Therapies*. John Wiley & Sons, 2009.

[3] "Cancer statistics," 12-Oct-2011. [Online]. Available: https://www.cancer.ie/about-us/media-centre/cancer-statistics. [Accessed: 20-Nov-2017].

[4] "All | Systems Biology Ireland." [Online]. Available: http://www.ucd.ie/sbi/blog/all/blogbodycontent,330785,en.html. [Accessed: 07-Dec-2017].

[5] F. Iorio *et al.*, "A Landscape of Pharmacogenomic Interactions in Cancer," *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016.

[6] J. Barretina *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012.

[7] A. Tsherniak *et al.*, "Defining a Cancer Dependency Map," *Cell*, vol. 170, no. 3, pp. 564–576.e16, Jul. 2017.

[8] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," *Pac. Symp. Biocomput.*, pp. 63–74, 2014.

[9] W. Yang *et al.*, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D955–61, Jan. 2013.

[10] S. A. Forbes *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D945–50, Jan. 2011.

[11] M. J. Garnett *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, Mar. 2012.

[12] "Correlation - Statistical Techniques, Rating Scales, Correlation Coefficients, and More - Creative Research Systems." [Online]. Available: https://www.surveysystem.com/correlation.htm. [Accessed: 20-Nov-2017].

[13] D. Greene, "COMP 47490 Feature Selection." Autumn-2017.

[14] A. Gallo, "A Refresher on Regression Analysis," *Harvard Business Review*, 04-Nov-2015.

[15] "Regression Analysis: Step by Step Articles, Videos, Simple Definitions," *Statistics How To*. [Online]. Available: http://www.statisticshowto.com/probability-and-statistics/regression-analysis/. [Accessed: 04-Apr-2018].

[16] "What is Linear Regression? - Statistics Solutions," *Statistics Solutions*. [Online]. Available: https://www.statisticssolutions.com/what-is-linear-regression/. [Accessed: 03-Apr-2018].

[17] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Biometrics*, 2002.

[18] "How to run Linear regression in Python scikit-Learn," *Big Data Made Simple - One source. Many perspectives.*, 05-Mar-2018. [Online]. Available:

http://bigdata-madesimple.com/how-to-run-linear-regression-in-python-scikit-learn/. [Accessed: 03-Apr-2018].
[19] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.
[20] A. Jain, S. Jain, F. Shaikh, G. Singh, and NSS, “A Complete Tutorial on Ridge and Lasso Regression in Python,” *Analytics Vidhya*, 28-Jan-2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/. [Accessed: 05-Dec-2017].
[21] M. Gönen *et al.*, “A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell Lines,” *Cell Syst*, Oct. 2017.
[22] D. Greene, “COMP 47490 Ensembles.” Autumn-2017.
[23] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R news*, 2002.
[24] T. Bylander, “Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates,” *Mach. Learn.*, vol. 48, no. 1–3, pp. 287–297, Jul. 2002.
[25] “sklearn.tree.DecisionTreeRegressor — scikit-learn 0.19.1 documentation.” [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html. [Accessed: 03-Apr-2018].
[26] C. Spark, “Getting started with regression and decision trees.” [Online]. Available: http://cambridgespark.com/content/tutorials/getting-started-with-regression-and-decision-trees/index.html. [Accessed: 03-Apr-2018].
[27] P. B. Chapman *et al.*, “Improved survival with vemurafenib in melanoma with BRAF V600E mutation,” *N. Engl. J. Med.*, vol. 364, no. 26, pp. 2507–2516, Jun. 2011.
[28] “Downloads,” *Anaconda*. [Online]. Available: https://www.anaconda.com/download/. [Accessed: 04-Apr-2018].
[29] “Installation — Matplotlib 2.2.2 documentation.” [Online]. Available: https://matplotlib.org/. [Accessed: 27-Mar-2018].
[30] “1. Supervised learning — scikit-learn 0.19.1 documentation.” [Online]. Available: http://scikit-learn.org/stable/supervised_learning.html. [Accessed: 28-Mar-2018].