# **Internship Report:** Predicting Drug Activity Based on Gene Expression

By: Aaron Mohammed

University of South Florida
Department of Pharmacology

## I. Summary

During the summer of 2022, I worked with Dr. Feng Cheng from the Department of Pharmaceutical Science at the University of South Florida. The goal of our research project was to determine if gene expression could be used to predict drug activity on cancer cell lines from the NCI-60 panel. The R programming language was used to carry out this research. First, the drug activity dataset and gene expression datasets were obtained from the CellMiner database. The drugs were filtered and only the compounds with a max activity greater than 8 and standard deviation greater than 1 were kept. Correlation coefficients were calculated between drug activity and each of the gene expression datasets. Based on those results, the gene expression dataset that resulted in the highest correlations was chosen to be used in machine learning models. The correlation results were used to gain insight into the underlying biology at play by utilizing DAVID. The 'WGCNA' R package was used to cluster the genes into module eigengenes since the gene expression dataset that was chosen contained over 10,000 genes. The 'e1071' R package was then used to create SVM machine learning models where the response vectors was the drug activity data and the data matrix was eigengene expression. After training and testing, it was found that gene expression can predict drug activity with high accuracy for certain drugs.

## II. Background

During the late 1980's, the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI) began the development of an anticancer drug screening called the NCI-60 cell line panel. The NCI-60 is made up of 60 cancer cell lines from various types of cancer like skin, colon, and breast cancer. So far, over 80,000 compounds have been screened *in vitro*. The NCI-60 panel plays a very important role in drug discovery, in fact, most FDA approved drugs were screened in the NCI-60 panel at least twice[1]. Gene expression data for each of these cell lines were downloaded from CellMiner, a database created by the DTP. Examples of the kind of gene expression data that can be downloaded include but are not limited to RNAseq, microRNA, and exon data.

The NCI-60 drug activity data that was used for this project was also downloaded from CellMiner. The raw "Compound activity: DTP NCI-60" dataset was used. It contained over 50,000 different compounds, of which 158 were FDA approved drugs and 79 were in clinical trials. In this dataset, drug activity is expressed as the negative log of GI50[2]. GI50 is the concentration of the drug at which there is a 50% reduction in growth of the cancer cells[3]. This means that a high value indicates only a small concentration of the drug was needed to inhibit the proliferation of the cells by 50%. The compound activity value of 13 is the max value in this dataset, while -4 is the lowest. 3 gene expression datasets were downloaded from CellMiner and the dataset that had the best correlations with drug activity was used for machine learning. The RNAseq, Affy HuEx 1.0, and Agilent mRNA datasets were downloaded.

## III. Tasks

### Correlation Coefficients

After downloading the drug activity and expression data, R was used to calculate correlation coefficients. This was done in order to determine which gene expression dataset would serve as the better predictor. The drug activity dataset contained over 80,000 rows since it contained multiple entries/experiments for most of the compounds. In order to reduce runtime, the size of the drug activity dataset was reduced by removing the drug activity data from past experiments and keeping the data from only the most recent experiments for each drug. This was accomplished by utilizing the rev() and duplicate() functions, and the not, !, operator to subset the data frame based on the column that contained the NSC IDs for each drug. This reduced the number of drugs from 83,680 to 56,461. To reduce the size further, the data frame was filtered and only the compounds that were FDA approved, had a max activity value >= 8, and a standard deviation >= 1 were extracted. This reduced the number of drugs to 21.

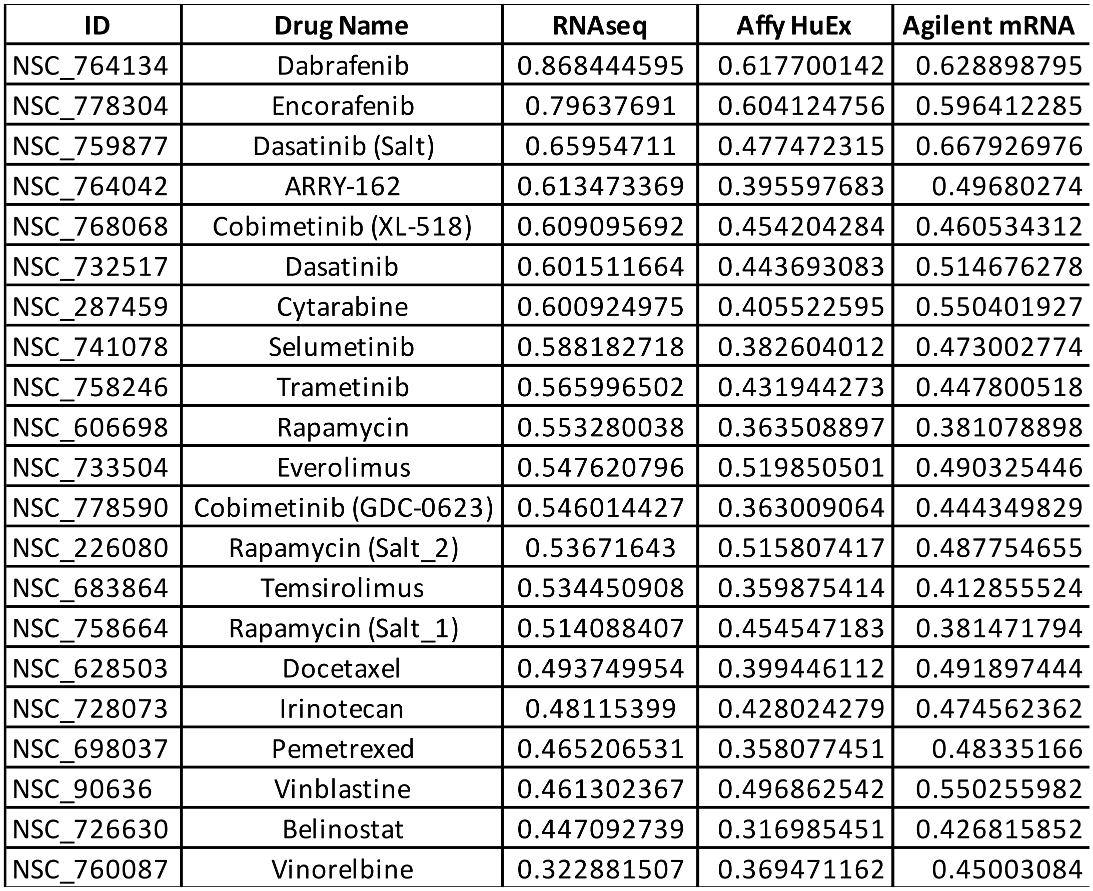
The gene expression datasets were very large. In order to reduce runtime, they were filtered so that only highly expressed genes were kept. This was determined by calculating average expression for each gene across the cell lines and keeping the ones with average expression greater than or equal to a certain value. For the RNAseq dataset, genes that had average expression of at least 1 were kept, reducing the number of genes from 23,808 to 10,131. For Affy HuEx, the average expression cutoff was 6 and for Agilent mRNA, the cutoff was 5. The number of genes for Affy HuEx were reduced from 1,048,575 to 268 and 41,090 to 845 for Agilent mRNA.

Once the datasets were filtered, it was then time to get the correlation coefficients between drug activity and gene expression. At first, a nested for loop containing the cor.test() function was used, but Dr. Cheng recommended that I get in the habit of using the “apply” functions since those functions are more efficient than for loops for large amounts of data. I accomplished this by creating a functional or "nested function", which is a function of functions. First, the datasets were all transposed so that the drug IDs and gene IDs became columns. The names of drug and gene IDs were stored in string vectors, called drug_name and G respectively, to be used as column indexes. Two lapply() functions and a cor.test() function were used, each nested within the other. That functional is shown below,

```
Corr <- lapply(drug_name, function(drug_name, drugs_df){
        lapply(gene_name, function(gene_name, genes_df){
               cor.test(as.numeric(drugs_df[[drug_name]]),
                        as.numeric(genes_df[[gene_name]])) %>%
               tidy()}, genes_df) %>%
        bind_rows() %>%
        mutate(Gene = gene_name) %>%
        select(Gene, estimate, p.value) %>%
        as.data.frame()}, drugs_df)
```

where Corr is the data frame the output is stored in, drugs_df is the drug activity dataset, and genes_df is the gene expression dataset. In order to save the output from this functional, the 'broom' and 'dplyr' packages were used. The pipe operator, %>%, and the tidy() function from the 'broom' package were used to store the statistical output of cor.test() into a tibble, which is essentially a simple version of a data frame that minimizes memory. The bind_rows(), mutate(), and select() functions from the 'dplyr' package were used to manipulate the tibbles and prepare them for conversion to data frames. The output of this nested lapply() function is a list of data frames, so the ldply() function from the 'plyr' package was used to take values from each data frame stored in the list and place them into a single data frame.

The correlation coefficients and p values from the list were extracted and stored in separate data frames using for loops. Another data frame containing the max correlation values, standard deviation of activity, max activity, and names for each drug was created to be exported to excel. This process was repeated for each gene expression dataset, using both the Pearson and Spearman methods. In total, there were 6 resulting datasets. The RNAseq dataset was chosen since it was more correlated with drug activity than the other gene expression datasets. The max correlation coefficients using the Pearson method is shown in Table 1. The drug with the highest Pearson correlation coefficient was Dabrafenib, 0.868 with a p-value of 1.093e-14.



| ID | Drug Name | RNAseq | Affy HuEx | Agilent mRNA |
|---|---|---|---|---|
| NSC_764134 | Dabrafenib | 0.868444595 | 0.617700142 | 0.628898795 |
| NSC_778304 | Encorafenib | 0.79637691 | 0.604124756 | 0.596412285 |
| NSC_759877 | Dasatinib (Salt) | 0.65954711 | 0.477472315 | 0.667926976 |
| NSC_764042 | ARRY-162 | 0.613473369 | 0.395597683 | 0.49680274 |
| NSC_768068 | Cobimetinib (XL-518) | 0.609095692 | 0.454204284 | 0.460534312 |
| NSC_732517 | Dasatinib | 0.601511664 | 0.443693083 | 0.514676278 |
| NSC_287459 | Cytarabine | 0.600924975 | 0.405522595 | 0.550401927 |
| NSC_741078 | Selumetinib | 0.588182718 | 0.382604012 | 0.473002774 |
| NSC_758246 | Trametinib | 0.565996502 | 0.431944273 | 0.447800518 |
| NSC_606698 | Rapamycin | 0.553280038 | 0.363508897 | 0.381078898 |
| NSC_733504 | Everolimus | 0.547620796 | 0.519850501 | 0.490325446 |
| NSC_778590 | Cobimetinib (GDC-0623) | 0.546014427 | 0.363009064 | 0.444349829 |
| NSC_226080 | Rapamycin (Salt_2) | 0.53671643 | 0.515807417 | 0.487754655 |
| NSC_683864 | Temsirolimus | 0.534450908 | 0.359875414 | 0.412855524 |
| NSC_758664 | Rapamycin (Salt_1) | 0.514088407 | 0.454547183 | 0.381471794 |
| NSC_628503 | Docetaxel | 0.493749954 | 0.399446112 | 0.491897444 |
| NSC_728073 | Irinotecan | 0.48115399 | 0.428024279 | 0.474562362 |
| NSC_698037 | Pemetrexed | 0.465206531 | 0.358077451 | 0.48335166 |
| NSC_90636 | Vinblastine | 0.461302367 | 0.496862542 | 0.550255982 |
| NSC_726630 | Belinostat | 0.447092739 | 0.316985451 | 0.426815852 |
| NSC_760087 | Vinorelbine | 0.322881507 | 0.369471162 | 0.45003084 |

**Table 1 –** Max correlation coefficient values between drug activity and gene expression.

## Biology Underlying the Data

In order to gain an understanding of the biology at play, hierarchal clustering of drug activity was used to determine which drugs behave similarly. A cluster dendrogram of drug activity was created in R and is shown in Figure 1. The first step in making this plot was calculating the distances between all possible pairs of drug activity values between all drugs and placing them in a distance matrix. This was done by using the dist() function. The distance matrix was then inputted into the hclust() function which uses complete linkage clustering to form the dendrogram. Two clusters were produced, one with 7 drugs and the other with 14. To find out if these clusters were formed because the drugs in each cluster had similar mechanisms, a webserver called DAVID was used to investigate the pathways associated with genes that were positively and negatively correlated with drug activity. DAVID is an abbreviation for Database for Annotation, Visualization, and Integrated Discovery, it allows a user to access tools and databases for

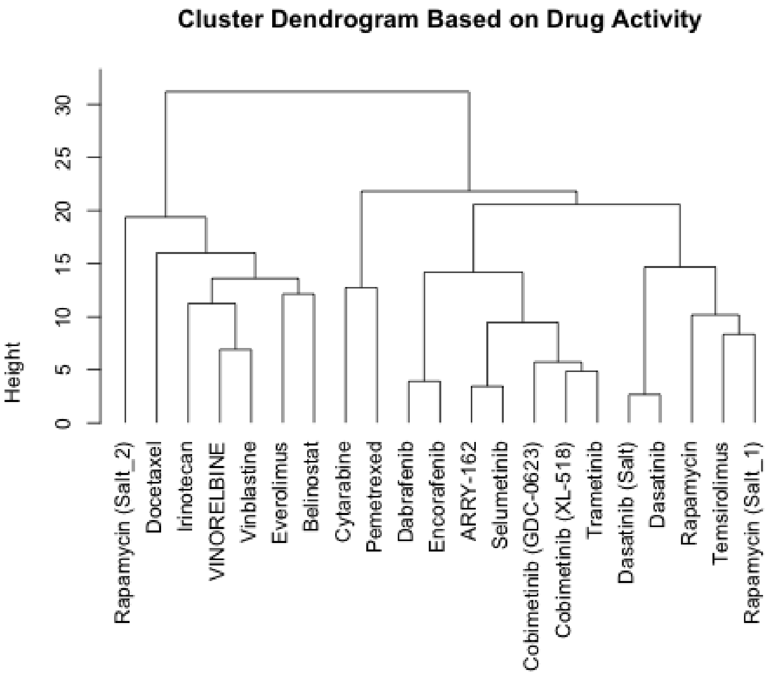
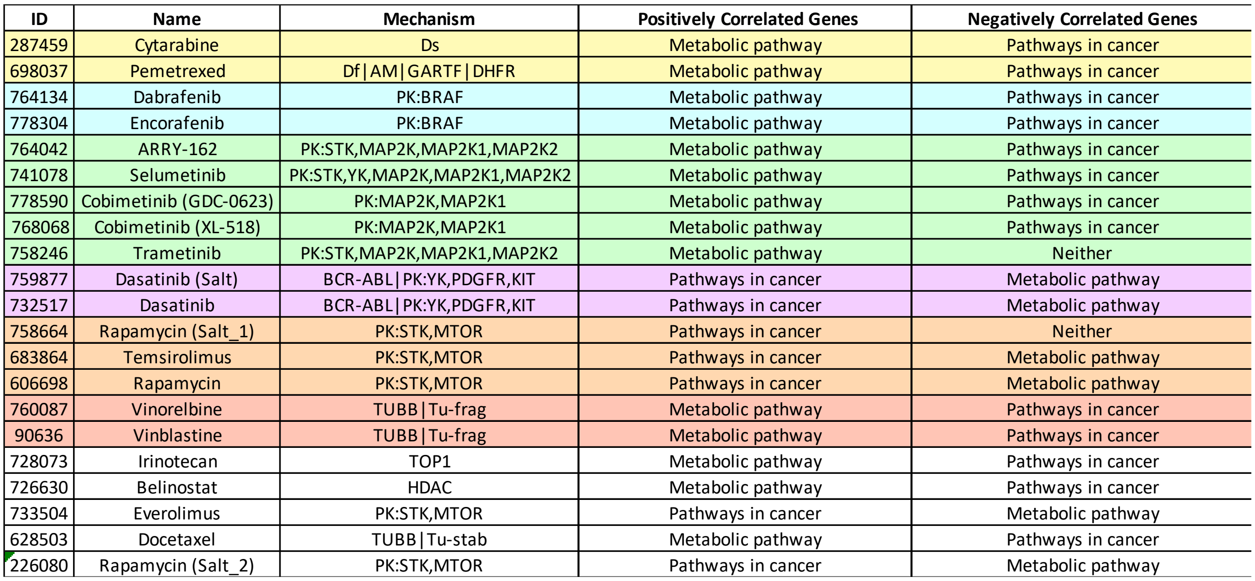


**Figure 1 –** This map displays how similar each drug's activity is with each other.

functional annotation analysis[4]. Genes that were highly correlated with drug activity could influence drug sensitivity and genes that had low correlation could be responsible for drug resistance when highly expressed. To investigate the pathways associated with drug sensitivity, gene IDs for the top 500 positive correlated genes for each drug were inputted into DAVID. For pathways associated with drug resistance, gene IDs for the lowest 500 genes with negative correlation were inputted into DAVID. The KEGG pathway database within DAVID was selected for this analysis. KEGG stands for Kyoto Encyclopedia of Genes and Genomes, and it contains annotations of gene functions[5]. A table was created to keep track of whether the genes were associated with either metabolic pathways or pathways in cancer. It was found that 14 drugs had positively correlated genes that were primarily associated with metabolic pathways and 7 had positively correlated genes that were primarily associated with pathways in cancer. 13 drugs had negatively correlated genes that were primarily associated with pathways in cancer and 6 that had negatively correlated genes that were primarily associated with metabolic pathways. There were two drugs who's lowest 500 correlated genes were not associated with metabolic pathways or pathways in cancer. The results are shown in Table 2 and include the genes involved with the mechanism of action for each drug that were included in the NCI-60 drug activity dataset. The colored sets in Table 2 represent clusters with a height less than 10. Clearly, drugs that are within these clusters at that height either share the same mechanisms or have the same type of correlation with genes of certain pathways. The drugs that are not in clusters that have a height less than 10 are the exceptions.



| ID | Name | Mechanism | Positively Correlated Genes | Negatively Correlated Genes |
|---|---|---|---|---|
| 287459 | Cytarabine | Ds | Metabolic pathway | Pathways in cancer |
| 698037 | Pemetrexed | Df\|AM\|GARTF\|DHFR | Metabolic pathway | Pathways in cancer |
| 764134 | Dabrafenib | PK:BRAF | Metabolic pathway | Pathways in cancer |
| 778304 | Encorafenib | PK:BRAF | Metabolic pathway | Pathways in cancer |
| 764042 | ARRY-162 | PK:STK,MAP2K,MAP2K1,MAP2K2 | Metabolic pathway | Pathways in cancer |
| 741078 | Selumetinib | PK:STK,YK,MAP2K,MAP2K1,MAP2K2 | Metabolic pathway | Pathways in cancer |
| 778590 | Cobimetinib (GDC-0623) | PK:MAP2K,MAP2K1 | Metabolic pathway | Pathways in cancer |
| 768068 | Cobimetinib (XL-518) | PK:MAP2K,MAP2K1 | Metabolic pathway | Pathways in cancer |
| 758246 | Trametinib | PK:STK,MAP2K,MAP2K1,MAP2K2 | Metabolic pathway | Neither |
| 759877 | Dasatinib (Salt) | BCR-ABL\|PK:YK,PDGFR,KIT | Pathways in cancer | Metabolic pathway |
| 732517 | Dasatinib | BCR-ABL\|PK:YK,PDGFR,KIT | Pathways in cancer | Metabolic pathway |
| 758664 | Rapamycin (Salt_1) | PK:STK,MTOR | Pathways in cancer | Neither |
| 683864 | Temsirolimus | PK:STK,MTOR | Pathways in cancer | Metabolic pathway |
| 606698 | Rapamycin | PK:STK,MTOR | Pathways in cancer | Metabolic pathway |
| 760087 | Vinorelbine | TUBB\|Tu-frag | Metabolic pathway | Pathways in cancer |
| 90636 | Vinblastine | TUBB\|Tu-frag | Metabolic pathway | Pathways in cancer |
| 728073 | Irinotecan | TOP1 | Metabolic pathway | Pathways in cancer |
| 726630 | Belinostat | HDAC | Metabolic pathway | Pathways in cancer |
| 733504 | Everolimus | PK:STK,MTOR | Pathways in cancer | Metabolic pathway |
| 628503 | Docetaxel | TUBB\|Tu-stab | Metabolic pathway | Pathways in cancer |
| 226080 | Rapamycin (Salt_2) | PK:STK,MTOR | Pathways in cancer | Metabolic pathway |

**Table 2 –** This table includes biological information related to positive and negative correlations. Pathways associated with positive correlation involve genes that allow the drug to be affective when they are highly expressed. Colored sets are drugs within clusters that have a height less than 10.

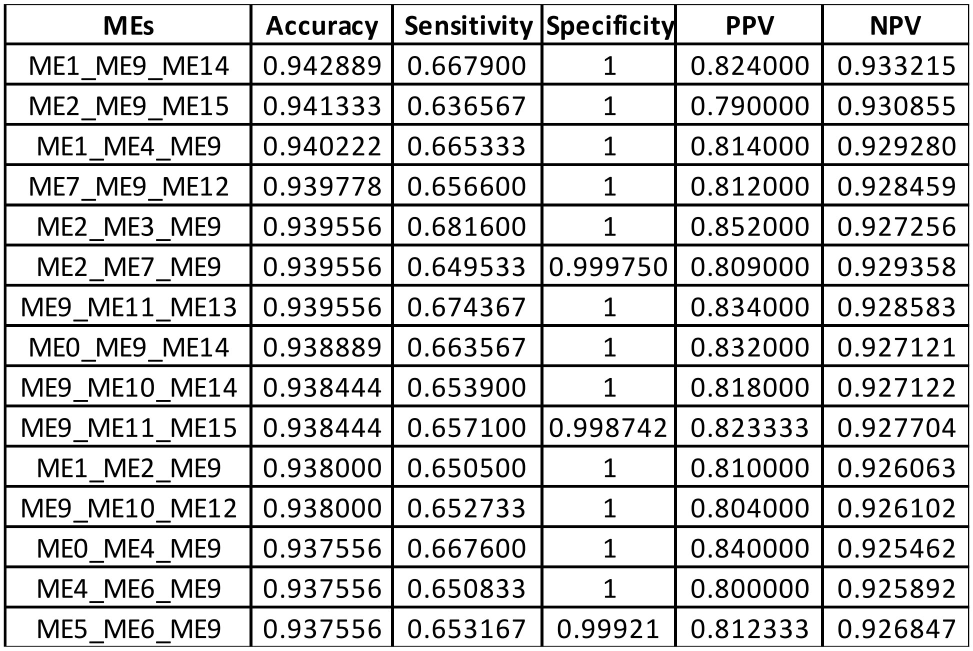
Dabrafenib and Encorafenib were the 2 drugs with the highest amount of correlation with gene expression and they are both BRAF inhibitors.

## Weighted Correlation Network Analysis

Since the RNAseq dataset had over 10,000 genes, the ‘WGCNA’ R package was used to cluster the genes and form module eigengenes. This was necessary because SVMs are not suitable for large datasets. Module Eigengenes are a collection of genes that have similar expression patterns across multiple samples. WGCNA stands for weighted correlation network analysis and its based on the correlations between multiple genes based on how their gene expression changes in different samples.

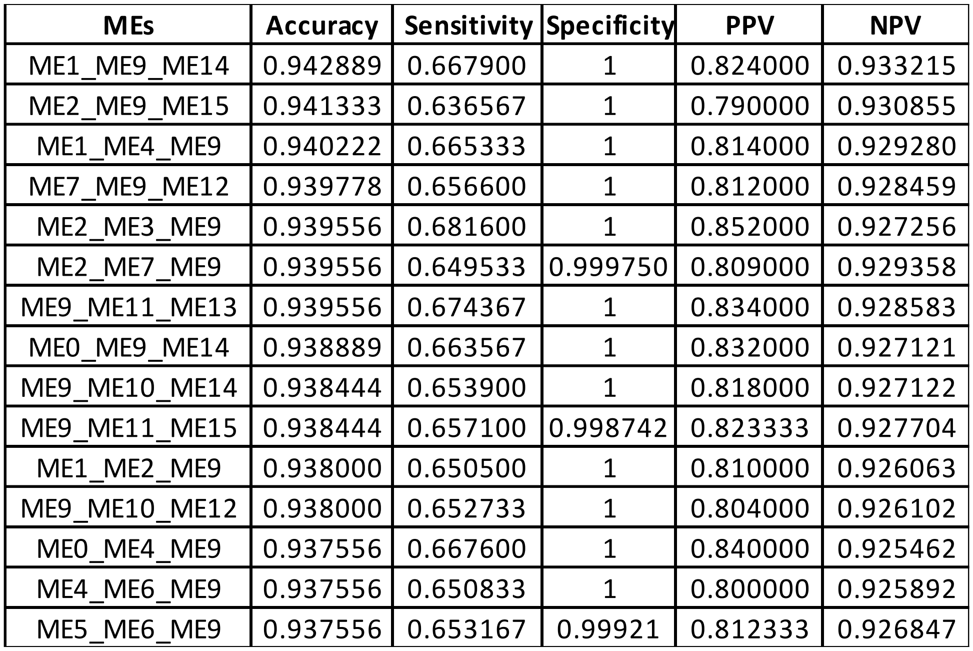
## Machine Learning

After generating the module eigen genes, the ‘e1071’ package was used to train and test SVMs. SVM stands for support-vector machine and it’s a type of machine learning model that can be used for classification.



| MEs | Accuracy | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| ME1_ME9_ME14 | 0.942889 | 0.667900 | 1 | 0.824000 | 0.933215 |
| ME2_ME9_ME15 | 0.941333 | 0.636567 | 1 | 0.790000 | 0.930855 |
| ME1_ME4_ME9 | 0.940222 | 0.665333 | 1 | 0.814000 | 0.929280 |
| ME7_ME9_ME12 | 0.939778 | 0.656600 | 1 | 0.812000 | 0.928459 |
| ME2_ME3_ME9 | 0.939556 | 0.681600 | 1 | 0.852000 | 0.927256 |
| ME2_ME7_ME9 | 0.939556 | 0.649533 | 0.999750 | 0.809000 | 0.929358 |
| ME9_ME11_ME13 | 0.939556 | 0.674367 | 1 | 0.834000 | 0.928583 |
| ME0_ME9_ME14 | 0.938889 | 0.663567 | 1 | 0.832000 | 0.927121 |
| ME9_ME10_ME14 | 0.938444 | 0.653900 | 1 | 0.818000 | 0.927122 |
| ME9_ME11_ME15 | 0.938444 | 0.657100 | 0.998742 | 0.823333 | 0.927704 |
| ME1_ME2_ME9 | 0.938000 | 0.650500 | 1 | 0.810000 | 0.926063 |
| ME9_ME10_ME12 | 0.938000 | 0.652733 | 1 | 0.804000 | 0.926102 |
| ME0_ME4_ME9 | 0.937556 | 0.667600 | 1 | 0.840000 | 0.925462 |
| ME4_ME6_ME9 | 0.937556 | 0.650833 | 1 | 0.800000 | 0.925892 |
| ME5_ME6_ME9 | 0.937556 | 0.653167 | 0.99921 | 0.812333 | 0.926847 |

Figure 2 Dabrafenib



| MEs | Accuracy | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| ME1_ME9_ME14 | 0.942889 | 0.667900 | 1 | 0.824000 | 0.933215 |
| ME2_ME9_ME15 | 0.941333 | 0.636567 | 1 | 0.790000 | 0.930855 |
| ME1_ME4_ME9 | 0.940222 | 0.665333 | 1 | 0.814000 | 0.929280 |
| ME7_ME9_ME12 | 0.939778 | 0.656600 | 1 | 0.812000 | 0.928459 |
| ME2_ME3_ME9 | 0.939556 | 0.681600 | 1 | 0.852000 | 0.927256 |
| ME2_ME7_ME9 | 0.939556 | 0.649533 | 0.999750 | 0.809000 | 0.929358 |
| ME9_ME11_ME13 | 0.939556 | 0.674367 | 1 | 0.834000 | 0.928583 |
| ME0_ME9_ME14 | 0.938889 | 0.663567 | 1 | 0.832000 | 0.927121 |
| ME9_ME10_ME14 | 0.938444 | 0.653900 | 1 | 0.818000 | 0.927122 |
| ME9_ME11_ME15 | 0.938444 | 0.657100 | 0.998742 | 0.823333 | 0.927704 |
| ME1_ME2_ME9 | 0.938000 | 0.650500 | 1 | 0.810000 | 0.926063 |
| ME9_ME10_ME12 | 0.938000 | 0.652733 | 1 | 0.804000 | 0.926102 |
| ME0_ME4_ME9 | 0.937556 | 0.667600 | 1 | 0.840000 | 0.925462 |
| ME4_ME6_ME9 | 0.937556 | 0.650833 | 1 | 0.800000 | 0.925892 |
| ME5_ME6_ME9 | 0.937556 | 0.653167 | 0.99921 | 0.812333 | 0.926847 |

Figure 3 Encorafenib

## Discussion

## Citations

1. Holbeck, S. L., Collins, J. M. & Doroshow, J. H. Analysis of Food and Drug Administration–Approved Anticancer Agents in the NCI60 Panel of Human Tumor Cell Lines. *Mol. Cancer Ther.* **9**, 1451–1460 (2010).
2. CellMiner - Datasets. https://discover.nci.nih.gov/cellminer/datasets.do.
3. Reinhold, W. C. *et al.* CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res.* **72**, 3499–3511 (2012).
4. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* gkac194 (2022) doi:10.1093/nar/gkac194.
5. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).