Fall 2022 CS/BIOL 123A Bioinformatics Project Report

Project Title: Small Molecule Drug Development for the BRAF V600 Mutation

Names:

Michelle Quach <Molecular Biology>

Deven Shah <Software Engineering>

Harshmeet Singh <Molecular Biology>

Ethan Vrooman <Computer Science>

ABSTRACT

This report presents the findings behind the use of computational or in-silico methods to find therapeutic targets allows for the effective integration of the massive amounts of data currently available and the accurate prediction of the effectiveness of a given target molecule that could potentially inhibit the expression of the most common B-Raf Proto-Oncogene, Serine/Threonine Kinase (BRAF) mutation. In order to find small chemical molecules that may prevent the expression of the most prevalent BRAF oncogenic mutation, machine-learning algorithms, such as the SVM (Support Vector Machine). An SVM model utilizes support vectors to adjust the threshold of the hyperplane to categorize data points and is widely used for classification models. Complemented with a Random Forest Classifier, the linear SVM model was able to use a dataset with 243 different compounds to achieve an average of 0.976 precision, 0.975 recall, 0.966 accuracies, and a 0.962 area under the receiving operating characteristic curve across 50 independent iterations. 10 common features were present in all 50 iterations, which provides computational evidence that these features directly affect the identification of the model. The model is not limited to strictly identifying compounds, as it affords the ability to determine if certain features truly affect the identification. This model may be used to conclude whether a QuaSAR descriptor truly correlates with the potential of a compound to inhibit the expression of the BRAF mutation. The model consistently achieved optimal performance with each iteration.

Future work will implement an improved feature selection process to achieve perfect performance, a deeper analysis of feature importances, and use alternative classification models.

Key Words: BRAF-V600E, machine learning, SVM, Random Forest Classifier, QuaSAR

Table of Contents

List of Tables

List of Figures

INTRODUCTION

For our project, the problem that we would like to address has to do with the identification of small molecules that have the ability to inhibit the BRAF mutation V600E. The BRAF V600E mutation is a genetic alteration that is most common in papillary thyroid carcinomas (PTC), which are generally responsible for over 80 - 90% of all thyroid cancers (One example of cancer affected by the mutation). We chose this problem due to the benefits that can be obtained if we are able to make accurate identification of the molecules that play a role in the inhibition of the BRAF V600E mutation. Our study of this problem has, intrinsically, multiple purposes. Firstly, the real-world relevance of the problem. Using machine learning algorithms, in bioinformatics, to quickly and accurately extract and analyze data from databases is on the frontier of developing knowledge. For example, a study done by our professor deals with the viability and legitimacy of using machine learning algorithms in the scope of bioinformatics. It compares the accuracy of two specific algorithms (SVM and 3D-QSAR to be exact) and developed a new statistical metric (EPP or Expected Predictive Performance) made specifically to measure the effectiveness of machine learning algorithms against each other regarding the datasets being used as input (*Wesley et al., 2016*). At the end of the study, one of the disclaimers or retrospective thoughts was that there needs to continue to study or use machine learning(ML) algorithms in the field of bioinformatics. This is because the efficiency and effectiveness of ML algorithms is still not a topic well documented, it is most definitely more cost and time efficient than the traditional bioassay tests, but it has not undergone enough documented research to be widely applicable. Our team chose this problem, to add to the small but gradually growing pool of research done on ML algorithms in bioinformatics.

A reason why this problem is important to be addressed is the fact that it can give us insight into how to prevent cancer before it reaches the terminal state. In the article “BRAF(V600E) mutation and the biology of papillary thyroid cancer”, (*Frasca (2008)*), the authors studied the correlation of BRAF-V600E and the features of PTCs. The paper found about 16 studies with a positive correlation to the overall amount of PTCs and 12 studies with no correlation to the overall amount of PTCs, with a grand total of 2276 PTCs for the positive correlation and 1165 for the negative correlation. During the

study, they studied the association of BRAF-V600E to PTC in terms of the tumor, host, and environmental variables. Their sample of the patients involved all had gone through surgery for the thyroid during the years 2002-2005. After the confirmation for each of the patients involved in the study, tall cell variant papillary carcinomas (TCV-PTCs) were characterized by the papillae lined to be a single layer of tall cells. Finally, they measured the PTCs at 10mm or less than the maximum diameters that are considered micro-PTCs. The other method in this study was Laser Capture Microdissection (LCM). They washed the micro-PTCs with ethanol and rehydrated them in deionized water for the prep. Confirmation of the presence of the tumor was shown in the first and the last section of each section series. Another method in this study was Immunohistochemistry or IHC. This method involved the process of staining thyroid tissues with the avidin-biotin-peroxidase complex(ABC) method. The sections were then treated with normal serum for 20 minutes followed by the application of monoclonal antibodies against MMP-2 or MMP-9. The reaction products were developed in a 3-diaminobenzidine tetrahydrochloride solution with 0.03% H_2O_2. With all of this data, the study made use of statistical analysis to go over all of the data found from the studies and obtained a result that there was a correlation between BRAF V600E mutation and the clinicopathological features of PTCs.

If our machine learning algorithm is properly able to identify the molecules responsible for the inhibition of the mutation, we can use this algorithm to be able to make treatment methods that are able to inhibit the mutation with pinpoint accuracy. Solving this problem is extremely important due to the application that can be used if our algorithm can be used accurately. If we are able to identify the molecules that can be used to inhibit the mutations of BRAF V600E protein, we can potentially manipulate the algorithm to the point that it can identify molecules used for the inhibition of all other proteins, as long as we are able to fill in the different parameters needed. The significance of being able to identify proteins that cause or enable mutations (like BRAF V600E) is that it could lead to methods to prevent or treat said mutations better. According to a study (*Elisei et al., 2008*) the BRAF mutation in PTC patients has a statistically relevant correlation to being significant to its lethality, in the prone to fatal cases of PTC. Perhaps being able to identify the small chemical molecules that can

inhibit the expression of such a troublesome protein may cause a lower mortality rate in the aforementioned population of PTC patients. The BRAF protein mutation is not only observable in PTC either, it is conducive to the development of a number of different cancers, including melanoma and colorectal cancer (*Muling et al., 2013*). So research on this singular mutation can lead to progress on multiple fronts of research within the field of cancer. This impact on the biological fronts of research, coupled with its potential impact on the bioinformatic front of research, are factors of significance to solving this problem.

These potential benefits were only mentioned within the scope of the BRAF gene and its role in mutations that lead to cancer. Machine-learning algorithms that are able to identify small chemical molecules to inhibit gene expression would also be able to do the opposite. Assuming a mutation or disease would require the exhibition of a gene for treatment, this same research and method could be applied to that situation.

BACKGROUND

The BRAF gene is located on chromosome 7 and is used as a good target for drug development. When mutated, this can cause the overgrowth of cells, leading to cancer developing as a result. Usually, the BRAF mutation would be a point mutation, specifically, having the “T” nucleotide change to an “A” nucleotide in a single-base pair shift. Numerous malignancies, including melanoma and colorectal cancer, have been shown to express mutations for this gene, notably those that happen at the 600th amino acid position.

The BRAF protein, which is expressed by the BRAF gene, is involved in the MAPk pathway that is used to help mediate cellular growth in response to growth signaling by the organism. This protein will interact with a GTPase molecule, RAS, to help activate other proteins and kinases inside the cell as the RAS molecule relays the signal to do so. Once activated, BRAF will activate or make other proteins/kinases and then go into the nucleus of the cell to perform different transcription factors. If there is a mutation, the transcription factors would recognize the perfect promoter for the gene to

produce proteins which would cause the cells to divide and grow. This would cause the BRAF to be hyperactive resulting in the activation of all the kinases in the cells and the cells turn into tumors.

Some of the known FDA-approved drugs that are known to treat cancers caused by the BRAFV600E have similar genetic variations to each other that would help including Bortezomib, Carfilzomib, Thalidomide, Lenalidomide, Dexamethasone, Pomalidomide, Vemurafenib, Dabrafenib, Sorafenib, and Encrafenib. Each of these drugs plays a certain role in the inhibition process of the BRAFV600E mutation. For example, Bortezomib and Carfilzomib are small-molecule antineoplastic proteasome inhibitors that can inhibit mutations such as BRAF V600E. Vemurafenib and Dabrafenib classify as Small Molecule Antineoplastic BRAF kinase inhibitors, which means that they were made for the sole purpose of being able to inhibit the BRAF V600E mutation which is extremely beneficial for the treatment of cancer that is caused by the mutation.

A comparable study to ours is the study done in 2016 by Wesley et al. The BRAF-V600 gene is analyzed by machine learning algorithms to find inhibitors of it, but this is not the sole focus of the study. The study's goal is to take the predictive data done on the inhibitors of the BRAF-V600 gene and HIV integrase and then compare the data to find a statistical measure to help assess the goodness of machine learning algorithms. There were differences and similarities between our analysis of the BRAF-V600 gene, on a more material-based scope our software and data sets used are different. The scikit sklearn SVM classifier version used in the prior study is 0.17.1, our current version is 1.4.1, and the python version previously used was 2.7.3 compared to our current 3.10.7. In terms of methodology, the comparative study had multiple methods of analyses using the two ML algorithms it used (SVM and 3D-QSAR), it has a “Best Possible Model” approach where each algorithm would simply have its descriptors or hyperparameters tuned to provide for maximum accuracy(*Wesley et al., 2016*). It also contained a “Constrained to MOE Descriptors” that restricted the range of the hyperparameters of the SVM model to match the smaller set of hyperparameters available in the 3D-QSAR algorithm. Our study will not need a restriction on the descriptors of our singular SVM model, our study will focus more heavily on the fine-

tuning of the SVM hyperparameters rather than their comparison to another ML algorithm.

A support vector machine (SVM) is a classification model that utilizes hyperplanes and support vectors to classify data points. The hyperplane serves as a strict boundary between the two classifications, positive and negative. The support vectors are data points that lie relatively close to the hyperplane and serve as an aid to the position and orientation of the hyperplane. The kernel hyperparameter optimizes the hyperplane to best separate the data points. Possible kernels include linear, polynomial, and radial. The initial implementation of the SVM model utilized the entire provided dataset and had no hyperparameter optimization. This model was built on Python as it offers multiple modules to assist in feature selection, constructing the model, and evaluating the identifications. Various system performance metrics are used depending on the machine learning model type. The basic performance metrics used most commonly are precision, recall, f1-score, accuracy, and the receiving operating characteristic (ROC). Precision assesses the number of compounds that are truly positive, with respect to the number of compounds that were identified by the model as positive. Recall evaluates the number of compounds that are truly positive, with respect to all classifications from the model. Because there exists a tradeoff between precision and recall, a consolidated metric called the f1-score was introduced. The f1-score is a single metric that balances the precision and recall metrics. Accuracy calculates the number of classifications that were identified correctly, with respect to all instances. The ROC is a graphical representation that displays the true positive rate with respect to the false positive rate. The metric used to evaluate the ROC curve is the area under the curve (AUC-ROC).

The data set used in the previous study contained 303 compounds of which 243 were used as a training set and 60 were used as a test set. We have a dataset of 243 compounds in which we plan to use an 80/20 training/testing split. The reasoning for the narrowed dataset of the prior study is to achieve a higher degree of accuracy as well as a better relation to the competing ML algorithm because the competing ML algorithm needed a certain type of data.

DATA COLLECTED / ACCESSED

For the SVM model, a significant amount of data was required to properly train the model, with a wide range of features. Most data points were collected from PubChem, a commonly used database that contains seamlessly accessible information about chemical compounds. This includes data from the atom count, and density, to various scoring metrics. The dataset used was provided by Dr. Leonard Wesley as a comma-separated values (CSV) file. This dataset contains 243 unique chemical compounds that have the potential to inhibit the BRAF V600E mutation. Each compound has 356 different QuaSAR descriptors. The “Class” column is a binary feature that denotes “0” as a compound that does not have the potential to inhibit the mutation and vice versa for “1”. The amount of data provided is sufficient to create a model using an 80/20 training-testing sample split. Using Python and the Pandas module, the CSV file containing the dataset is initialized in a Pandas data frame.

Non-programmers were tasked to identify genetic variations of different FDA-approved small molecule drugs that hold an important role in the inhibition of BRAF V600E mutation. For this portion, the main database that we made use of was PDR (Prescriber’s Digital Reference). This database provided us with great deals of data on particular drugs. For our research purpose and due to the massive amount of data that was shown through this database, we picked specific bits of data from the database to better be able to organize and comprehend our research. First, we got the SMILES Annotation of the drug molecule that we were researching. This provided us with a relative visual representation of what molecules were in our molecule. We could use this data to compare different drug molecules to see if there was a pattern to observe for drugs involved in the inhibition of BRAF V600E mutation. Next, we recorded the Class of the molecule. This was an overall representation of the drug's purpose. Some of the drugs in our list targeted certain molecules, while other drugs targeted specific mutations. We then recorded the Mechanism of Action of each and every drug. This provided us with a detailed explanation of how the drug was able to identify and attack its target. It gave us insight into its overall procedure and mechanism. Finally, we

recorded the pharmacokinetics of each of the drugs researched. This, like the Mechanism of Action, provided us with a smaller and more targeted approach of the overall mechanism of the drug and its movement in the body once consumed.

APPROACH AND METHOD

The objective of this project is to engineer a machine-learning model that accurately identifies small molecules that have the potential to inhibit the expression of the most common BRAF oncogenic mutation and to identify which small molecule drugs would have the most effect on the BRAF V600E mutation. With the information obtained from various databases on drugs with the potential to inhibit BRAF V600E mutation, the machine learning model would use the list of drugs researched to obtain which sets of drugs would have the most impact on the mutation. Initially, the four members involved in this project were split into pairs based on expertise. The members who specialized in Biology were in one pair, and the members who specialized in Computer Science were in another pair. The objective of the Biology-focused pair was to research and identify genetic variations for which the FDA-approved drugs can address cancers that are caused by BRAF-V600E mutations. Using the information obtained, the Computer Science-focused pair implemented an SVM model to identify which compounds have the potential to inhibit the expression of the BRAF mutation.

The objective of the support vector machine was to identify which compounds have the ability to inhibit the BRAF V600E mutation and to programmatically determine which attributes affect the identification. The provided dataset consisting of various compounds and attributes retrieved from PubChem was used to evaluate the model. Python 3.10 was the language used to implement the model. Along with modules available from the Python Standard Library, Scikit-Learn and MatPlotLib were used. Scikit-Learn provides scripts to construct the model and MatPlotLib provides an avenue to display results.

The initial phase was to import and optimize the provided dataset. All compounds with missing features were removed from the dataset. Compounds with missing attributes yield inaccurate results when determining the features that impacted the classification. After all of the necessary compounds were removed, all features with an unknown data type were encoded categorically. As the minimally optimized dataset was established, it was split into features and targets. The 'Class' column in the dataset is the only target in this model. Recall that the "Class" column is a binary feature that denotes "0" as a compound that does not have the potential to inhibit the mutation and vice versa for "1". The rest of the columns were features that potentially impacted the target. To further optimize the dataset, all features with constant or quasi-constant values were removed.

The next phase was to determine the features that impacted the target the most during the execution. A Random Forest Classifier was used to determine the most important features. The most important features returned from the Random Forest Classifier were the features used in the feature dataset. All other features were removed. Note that the most important features may differ with each iteration. The feature and target datasets were then divided into an 80/20 train-test split. In other words, 80% of randomly selected features and targets were used for training the model, and the remaining 20% were used to test/validate the model.

As the dataset preparation was finalized, the SVM model was engineered. A support vector classifier (SVC) was used to identify each compound. The default kernel for this classifier was a radial basis function (RBF). Using the SVC, along with the training features and targets, the classifier was fitted. The testing features were used with the fitted classifier to identify the remaining compounds. The initial SVM model had the following optimizations of the four hyperparameters:

- Number of iterations: 5
- Number of features: 200
- Variance threshold: 0.005
- Kernel type: RBF

The number of iterations determines the number of times the model is to be executed. With the sacrifice of time, the higher the number of iterations, the more accurate the performance results will be. The number of features details the number of important features established from the Random Forest Classifier to be used. In this case, the 200 most important features were used. The variance threshold sets a restriction to the minimum variance for each feature. The kernel type is simply the kernel function used by the SVM to classify the data points.

The optimal hyperparameter tuning was determined through a sequential testing process. Records of performance at this stage are found in Appendix A. First, the best kernel type was found. To find the most optimal kernel type, 5 iterations of the model (using 200 features and a 0.005 variance threshold) were executed for each of the four valid kernel types: RBF, linear, polynomial, and Sigmoid. The linear kernel type yielded the highest average f1-score and accuracy (refer to Table A1). Next, the variance threshold was determined. Using a linear kernel and 200 features, the model was executed 5 times across 6 different variance thresholds: 0.0, 0.001, 0.01, 0.1, 1, and 10. The variance threshold that yielded the highest average f1-score and accuracy was 0.001 (refer to Table A2). Finally, the optimal number of important features was evaluated. The model with a linear kernel and a 0.001 variance threshold was executed across 5 iterations for 5 different numbers of features to use: 200, 150, 100, 50, and 10. 10 features yielded the highest average f1-score and accuracy (refer to Table A3), but this was due to overfitting. As a result, 50 features were decided to be the number of features to use with each iteration. The model is to be executed multiple times; thus, a relatively high number of iterations was required to yield the most accurate performance metrics. The number of iterations chosen was 50. The hyperparameters subsequent to optimization have the following values:

- Number of iterations: 50
- Number of features: 50
- Variance threshold: 0.001
- Kernel type: linear

The performance of each iteration of the model was recorded using precision, recall, accuracy, and AUC-ROC. Precision assesses the ratio of compounds that the model correctly identified as inhibiting the BRAF V600E mutation, with respect to all compounds that were identified by the model to inhibit the BRAF V600E mutation. Recall evaluates the ratio of compounds that the model correctly identified as inhibiting BRAF V600E mutation, with respect to all compounds identified by the model. Accuracy determines the ratio of compounds that were identified by the model to inhibit the BRAF V600E mutation, with respect to all identified compounds. The ROC curve graphically represents the false positive rate (FPR), with respect to the true positive rate (TPR). The FPR represents the ratio of compounds that were incorrectly identified to inhibit the BRAF V600E mutation, with respect to all compounds that truly do not inhibit the BRAF V600E mutation. The TPR represents the ratio of compounds that were correctly identified to inhibit the BRAF V600E mutation, with respect to all compounds that truly do not inhibit the BRAF V600E mutation. The AUC-ROC evaluates the ability of the model to differentiate between the two classifications. Along with the performance, the features used in each iteration were recorded. After all of the iterations were executed successfully, the performance of the model across all iterations was evaluated by taking the average precision, recall, accuracy, and AUC-ROC. The higher the performance metrics were, the better the model was performing. Additionally, a set of features that were used in all iterations was recorded. These features were the most common ones to affect the classification.

EVALUATION OF RESULTS

The model was executed 50 times, with each iteration using 50 features, a 0.001 variance threshold, and a linear kernel type. Out of the 50 iterations, 6 iterations achieved a perfect performance (Iterations 2, 3, 13, 29, 37, and 44). A perfect performance is described by having precision, recall, accuracy, and AUC-ROC of exactly 1.0. Figure 1 displays the ROC curve of Iteration 2, one of the six perfectly performing iterations. Iteration 1 had the lowest precision, accuracy, and AUC-ROC of

approximately 0.886, 0.917, and 0.882, respectively. Iteration 11 had the lowest recall of approximately 0.914. Iteration 1 was established as the iteration with the lowest performance. Figure 2 displays the ROC curve of Iteration 1, the lowest-performing iteration.

Figure 1: ROC of Iteration 2.

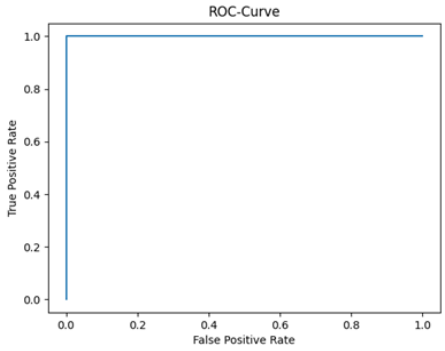
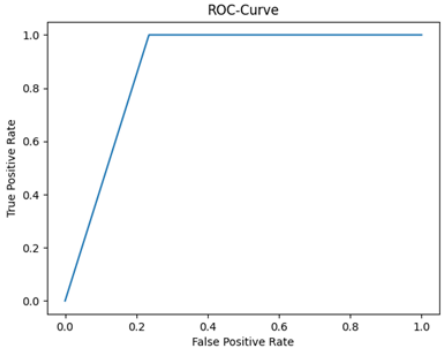


Figure 2: ROC of Iteration 1.



Nevertheless, the performance of the model across all 50 iterations yielded an optimal performance. The performance of all consolidated iterations had an approximate average precision of 0.976, recall of 0.975, the accuracy of 0.966, and AUC-ROC of 0.962. Table 1 contains the approximate performance of the highest-performing iteration, lowest-performing iteration, and the average performance across all 50 iterations. Appendix B provides a visualization of the precision, recall, accuracy, and AUC-ROC of all iterations.

Table 1: Highest, lowest, and average performance of the model.

	Iteration	**Precision**	**Recall**	**Accuracy**	**AUC-ROC**
Highest-Performing Iteration	2	1.000	1.000	1.000	1.000
Lowest-Performing Iteration	1	0.886	1.000	0.917	0.882

Average Performance	*Uses all 50 iterations*	0.976	0.975	0.966	0.962

The average precision concludes that 97.6% of all compounds that the model identified as inhibiting the BRAF V600E mutation are correct. The average recall establishes that 97.5% of all compounds were correctly identified as inhibiting the BRAF V600E mutation. The average accuracy proves that the model correctly identifies 96.6% of all compounds. The average AUC-ROC means that there is approximately a 96.2% chance that the model will be able to distinguish between the two identifications successfully.

The features present in all 50 iterations were: AM1_LUMO, E_ang, GCUT_PEOE_1, GCUT_PEOE_3, GCUT_SMR_3, MNDO_LUMO, PEOE_VSA-4, PM3_LUMO, SlogP_VSA5, and logS. Table 2 details the QuaSAR descriptors of the said features.

Table 2: QuaSAR descriptors of features.

Feature	Description
AM1_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital is calculated using the AM1 Hamiltonian [MOPAC].
E_ang	Angle bends potential energy.
GCUT_PEOE_1 & 3	The GCUT descriptors are calculated from the eigenvalues of a modified graph distance adjacency matrix. Each ij entry of the adjacency matrix takes the value 1/sqr(dij) where dij is the (modified) graph distance between atoms i and j. The diagonal takes the value of the PEOE partial charges. The resulting eigenvalues are sorted and the smallest, 1/3-ile, 2/3-ile, and largest eigenvalues are reported.
GCUT_SMR_3	The GCUT descriptors use the atomic contribution to molar refractivity (using the Wildman and Crippen SMR method)

	instead of partial charge.
MNDO_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital is calculated using the MNDO Hamiltonian [MOPAC].
PEOE_VSA-4	Sum of v*i* where q*i* is in the range [-0.25,-0.20).*
PM3_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital is calculated using the PM3 Hamiltonian [MOPAC].
SlogP_VSA5	Sum of v*i* such that L*i* is in (0.15,0.20].**
logS	Log of the aqueous solubility (mol/L). This property is calculated from an atom contribution linear atom type model [Hou 2004] with r2 = 0.90, ~1,200 molecules.

**PEOE. The Partial Equalization of Orbital Electronegativities (PEOE) method of calculating atomic partial charges [Gasteiger 1980] is a method in which charge is transferred between bonded atoms until equilibrium. To guarantee convergence, the amount of charge transferred at each iteration is damped with an exponentially decreasing scale factor. Let qi denote the partial charge of atom i as defined above. Let vi be the van der Waals surface area (Å2) of atom i (as calculated by a connection table approximation)(CCGI, 2008)*

***SlogP. Log of the octanol/water partition coefficient (including implicit hydrogens). This property is an atomic contribution model [Crippen 1999] that calculates logP from the given structure; i.e., the correct protonation state (washed structures). Results may vary from the logP(o/w) descriptor. The training set for SlogP was ~7000 structures.The Subdivided Surface Areas are descriptors based on an approximate accessible van der Waals surface area (in Å2) calculation for each atom, vi along with some other atomic property, pi. The vi are calculated using a connection table approximation. Each descriptor in a series is defined to be the sum of the vi over all atoms i such that pi is in a specified range (a,b).Li denotes the contribution to logP(o/w) for atom i as calculated in the SlogP descriptor(CCGI, 2008)*

Table 3: BRAF Inhibitor Drugs

DRUG	SMILES	CLASS
Bortezomib	B([C@@H](CC(C)C)NC(=O)[C@@H](CC1=CC=CC=C1)NC(=O)C2=NC=CN=C2)(O)O	Small Molecule Antineoplastic Proteasome Inhibitors used for the treatment of multiple myeloma and mantle cell lymphoma
Carfilzomib	CC(C)C[C@@H](C(=O)[C@]1(CO1)C)NC(=O)[C@H](CC2=CC=CC=C2)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCC3=CC=CC=C3)NC(=O)CN4CCOCC4	Small Molecule Antineoplastic Proteasome Inhibitors used for the treatment of relapsed or refractory multiple myeloma as a single-agent and in combination with other anti-myeloma therapies
Thalidomide	C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O	Agents of Leprosy, Immunomodulators, Angiogenesis Inhibitors
Lenalidomide	C1CC(=O)NC(=O)C1N2CC3=C(C2=O)C=CC=C3N	Immunomodulators, Angiogenesis Inhibitors
Dexamethasone	C[C@@H]1C[C@H]2[C@@H]3CCC4=CC(=O)C=C[C@@]4([C@]3([C@H](C[C@@]2([C@]1(C(=O)CO)O)C)O)F)C	Ophthalmological Corticosteroids Respiratory Corticosteroids Systemic Corticosteroids Systemic Corticosteroids, Plain
Pomalidomide	C1CC(=O)NC(=O)C1N2CC3=C(C2=O)C=CC=C3N	Immunomodulators, Angiogenesis Inhibitors
Vemurafenib	CCCS(=O)(=O)NC1=C(C(=C(C=C1)F)C(=O)C2=CNC3=NC=C(C=C23)C4=CC=C(C=C4)Cl)F	Small Molecule Antineoplastic BRAF kinase inhibitors
Dabrafenib	N/A	Small Molecules Antineoplastic BRAF kinase Inhibitors
Sorafenib	CNC(=O)C1=NC=CC(=C1)OC2=CC=C(C=C2)NC(=O)NC3=CC(=C(C=C3)Cl)C(F)(F)F	Small Molecule Antineoplastic Multikinase Inhibitors
Encorafenib	C[C@@H](CNC1=NC=CC(=N1)C2=CN(N=C2C3=CC(=CC(=C3F)NS(=O)(=O)C)Cl)C(C)C)NC(=O)OC	Small Molecule Antineoplastic BRAF kinase Inhibitors

CONCLUSION AND DISCUSSION

Taking into consideration the given data set only contained small chemical molecules that had potential to inhibit the BRAF V600E gene, the SVM model was able to quickly and accurately narrow down the dataset to those molecules that will inhibit the targeted gene along with the most significant traits of the molecules. The SVM model, once tuned, was able to achieve an average precision of 97.6%, recall of 97.5%, and accuracy of 96.6%. These percentages lend statistical significance to the results of the

model, as they exemplify the small possibilities of false positives, negatives, and erroneous data. These results can also be achieved within minutes of running the model, this is taking into account the program running the input dataset through 50 iterations. The dataset contained 243 small chemical molecules that had the potential to inhibit the targeted gene, this helped as all data was relevant and it is recommended that only relevant data should be run through a machine learning algorithm for efficiency's sake. These 243 molecules contained 356 descriptors that were used as data points for the SVM model.

The SVM model takes four parameters as input to determine how to analyze the given dataset. These four parameters can be sequentially tested to discern which gives the best results. This simple sequential testing and tuning of hyperparameters allow for a sizable dataset (243 rows of molecules with 356 columns of descriptors) to be analyzed in a succinct and concise manner. Along with the precision, recall, and accuracy statistics, the SVM model was also able to discern 10 out of the 356 descriptors that were the most common among the small molecule compounds that were found to properly inhibit the BRAF V600E gene. Finding that over 50 iterations of a decent dataset 10 descriptors were found to be the most prevalent can be significant to know what properties of these small molecules are causing the inhibition of BRAF V600E. Correlation may not always be causation but it cannot be ignored in a scientific setting where statistics can prove coincidences are not mere coincidences. The 10 descriptors found to be most common could shrink to an even narrower number given if more iterations were to be performed. Conducting further testing, with more iterations and perhaps even more data could narrow 10 to 5 descriptors and these descriptors could be used to accurately draw conclusions on why small molecule compounds are able to inhibit one of the most commonly found gene expressions that accompany a number of cancers.

To connect with the SVM model, we also were able to research some particular drugs from the PDR database to identify any patterns between small molecule drugs that can inhibit the BRAF V600E mutation. When going through the database, we found 10 small molecule drugs that fit the purpose of BRAF mutation inhibition. These were

Bortezomib, Carfilzomib, Thalidomide, Lenalidomide, Dexamethasone, Pomalidomide, Vemurafenib, Dabrafenib, Sorafenib, and Encrafenib. When researching these drugs, we recorded the Smiles, Classes, and Pharmacokinetics. Each of these sections in the database provided us with relative information on patterns that can be observed between drugs that inhibit a particular mutation. With this, we hope to we can look up more drugs with the ability to inhibit the BRAF V600E mutation and make a database for it so that our SVM model would be able to provide more insight into whether the drugs in the database differ in terms of effectiveness for the inhibition of the mutation.

FUTURE WORK

Possible future implementations to the presented model include improvements to feature selection, the use of multiple machine learning classification algorithms, and a more comprehensive analysis of the features selected. The current implementation of the SVM model derives feature importances through a Random Forest Classifier. However, this results in an average of 96% accuracy. As there is a causation relationship between the features and the target, it is possible to achieve perfect performance with each iteration. One potential future addition to the project is to use an optimized classifier to find features that consistently affect the identification. Moreover, along with improving the feature selection process, a detailed analysis may be accompanied by detailing an ordered list of the most and least impactful features. Another possible future implementation is to utilize alternate classification algorithms with the use of ensemble learning. K-Nearest Neighbors (KNN) and Decision Trees are among the most accurate machine-learning classification models. Ensemble learning allows the model to select the best-performing iteration. An SVM, KNN, and Decision Tree may be used in unison to find which features consistently appear to influence the identification

REFERENCES

Note: Most references and readings were split among team members prior to completing the project to gain background knowledge of the proposed project.

1. CANCER THERAPY: PRECLINICAL| JANUARY 31 2013

2. Chemical Computing Group Inc.(CCGI) (2008). *QuaSAR-Descriptor*. Cadaster.eu. https://cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm

3. Elisei, R., Ugolini, C., Viola, D., Lupi, C., Biagini, A., Giannini, R., Romei, C., Miccoli, P., Pinchera, A., & Basolo, F. (2008). BRAFV600E mutation and outcome of patients with papillary thyroid carcinoma: A 15-year median follow-up study. *The Journal of Clinical Endocrinology & Metabolism*, *93*(10), 3943–3949. https://doi.org/10.1210/jc.2008-0607

4. Frasca, F., Nucera, C., Pellegriti, G., Gangemi, P., Attard, M., Stella, M., Loda, M., Vella, V., Giordano, C., Trimarchi, F., Mazzon, E., Belfiore, A., & Vigneri, R. (2008). BRAF(V600E) mutation and the biology of papillary thyroid cancer. *Endocrine Related Cancer*, *15*(1), 191–205. https://doi.org/10.1677/erc-07-0212

5. Fu, G., Batchelor, C., Dumontier, M., Hastings, J., Willighagen, E., & Bolton, E. (2015). PubChemRDF: Towards the semantic annotation of PubChem compound and substance databases. *Journal of Cheminformatics*, *7*(1). https://doi.org/10.1186/s13321-015-0084-4

6. Li, C., Lee, K. C., Schneider, E. B., & Zeiger, M. A. (2012). *braf*V600E mutation and its association with Clinicopathological features of papillary thyroid cancer: A meta-analysis. *The Journal of Clinical Endocrinology & Metabolism*, *97*(12), 4559–4570. https://doi.org/10.1210/jc.2012-2104

7. Muling Mao; Feng Tian; John M. Mariadason; Chun C. Tsao; Robert Lemos, Jr; Farshid Dayyani; Y.N. Vashisht Gopal; Zhi-Qin Jiang; Ignacio I. Wistuba; Xi M.

Tang; William G. Bornman; Gideon Bollag; Gordon B. Mills; Garth Powis; Jayesh Desai; Gary E. Gallick; Michael A. Davies; Scott Kopetz(2013).Resistance to BRAF Inhibition in BRAF-Mutant Colon Cancer Can Be Overcome with PI3K Inhibition or Demethylating Agents. https://doi.org/10.1158/1078-0432.CCR-11-1446

8. Wesley, L., Veerapaneni, S., Desai, R., McGee, F., Joglekar, N., Rao, S., & Kamal, Z. (2016). 3D-QSAR and SVM prediction of BRAF-V600E and HIV integrase inhibitors: A comparative study and characterization of performance with a new expected prediction performance metric. *American Journal of Biochemistry and Biotechnology*, *12*(4), 253–262. https://doi.org/10.3844/ajbbsp.2016.253.262

APPENDICES

Appendix A: Initial SVM Optimization Performance Records

The following tables contain the f1-score and accuracy of various different hyperparameter values, excluding the number of iterations. The testing vector format is established as <*number of features*, *variance threshold*, *kernel type*>.

Table A1: Performance of various kernel types. Test vector: <200, 0.005, kernel type>. Linear kernel type was most optimal.

Kernel	F1-Score	Accuracy
RBF	0.8513758136020299	0.7416666666666666
linear	0.9302733573156108	0.9041666666666666
poly	0.8314568299317422	0.7125
sigmoid	0.7379079111578196	0.5875

Table A2: Performance of various variance thresholds: Test vector: <200, variance threshold, linear>. 0.001 variance threshold was most optimal.

Variance Threshold	F1-Score	Accuracy
0.0	0.9465678657815204	0.925
0.001	0.953960484383661	0.9375
0.01	0.8965079365079365	0.8666666666666666
0.1	0.9281677347146925	0.9
1	0.8945250607752107	0.8541666666666666

10	0.9228353299440887	0.8958333333333334

Table A3: Performance of various numbers of features used. Test vector: <number of features, 0.001, linear>. 50 features to use was most optimal to prevent overfitting.

Number of Features Used	F1-Score	Accuracy
200	0.9531902224155745	0.9291666666666668
150	0.9457908593525032	0.9166666666666667
100	0.950539287545538	0.9333333333333332
50	0.9713119327136199	0.9583333333333333
10	0.9722348336594913	0.9583333333333333

Appendix B: Performance of All 50 Model Iterations

The figures below visualize the precision, recall, accuracy, and AUC-ROC of each of the 50 iterations in the form of a graph.

Figure B1: Precision of all 50 iterations.

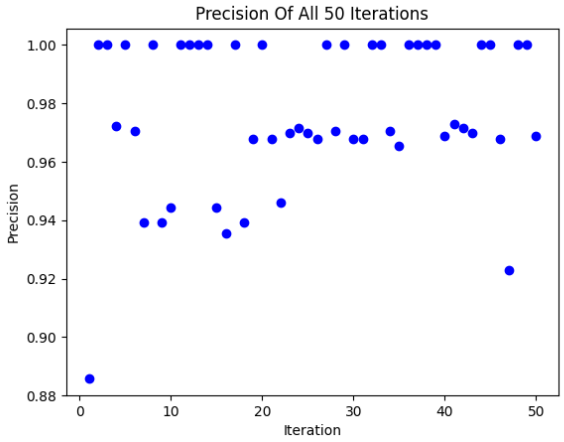


Figure B2: Recall of all 50 iterations.

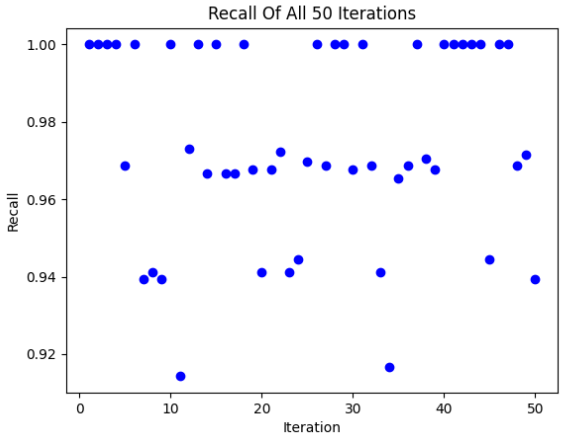


Figure B3: Accuracy of all 50 iterations.

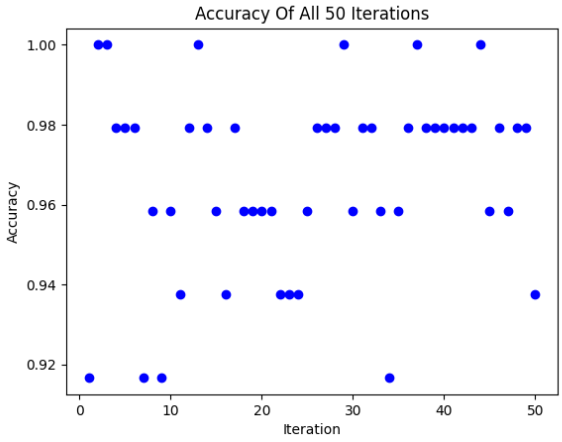


Figure B4: AUC-ROC of all 50 iterations.