Vol. 13(1), pp. 1-17, January-June 2024
DOI: 10.5897/JCBBR2023.0137
Article Number: 1429E1371934
ISSN 2141-2227

http://www.academicjournals.org/JCBBR

ACADEMIC JOURNALS
expand your knowledge

Journal of Computational Biology and Bioinformatics Research

Full Length Research Paper

Computational prediction of small molecules targeting Lassa fever drug target using quantitative structure activity relationship (QSAR) and random forest algorithm

Angela Makolo[1] and Pelumi Stephen Gboyega[2]*

[1]Computer Science Department, University of Ibadan, Ibadan, Oyo State, Nigeria.
[2]ICT and Bioinformatics Department, National Biotechnology Development Agency, Ilorin, Kwara State, Nigeria.

Received 8 December, 2023 Accepted 28 February, 2024

Lassa fever, an endemic viral hemorrhagic fever in West Africa, is attributed to the Lassa virus as its causative agent, and this disease has led to the untimely death of many people in the affected areas. At present, the available treatment options for Lassa fever are limited and there is need for new drugs. This study aims to use computational tools to predict the efficacy of small molecules that can target the Lassa fever virus glycoprotein which is essential for viral entry into host cells. This study uses quantitative structure activity relationship (QSAR) to reduce the cost and time of preclinical evaluation of potential drugs. This study retrieves 7620 molecules that can inhibit Lassa virus glycoprotein from ChEMBL database and builds a regression model with random forest algorithm. Its performance was compared with other regression models by using lazy predict, and random forest performed better than most of the regression models. The coefficient of determination r^2 are 0.93 and 0.56 for the training and test set and root mean square error (RMSE) of 0.32 and 0.77 for the training set and test set, respectively. In conclusion, the model satisfies the acceptable QSAR model.

Key words: Quantitative structure-activity relationship, bioactivity, drug-likeness, drug target.

INTRODUCTION

Lassa fever is listed among the diseases that pose significant public health threats (Klitting et al., 2021). Lassa fever is an acute viral illness in West Africa that is contracted by humans through contact with animals (Minari et al., 2021). Lassa fever was first identified in Nigeria in 1969 following the tragic death of two missionary nurses (Id et al., 2022). Subsequently, it extended its transmission to other West African nations, namely Nigeria, Benin, Togo, Mali, Guinea, Sierra Leone, and Liberia through its animal carrier, the “multimammate rat” (*Mastomys natalensis*), leading to endemicity in these regions (Kikiowo, 2021). Approximately 5,000 deaths are documented per year in West Africa due to incidents of Lassa virus infection, with estimates ranging from

*Corresponding author: E-mail: stephenpelummy@gmail.com.

100,000 to 300,000 cases (Arefin et al., 2021; James, 2020).

Eradicating Lassa fever within the West Africa sub-region has become a complex endeavor due to the unpredictable nature of recovery from the illness. This uncertainty arises from the virus's capability to persist in human bodily fluids, such as semen, even after a person has recuperated (Arefin et al., 2021; Oloniniyi et al., 2016). Therefore, there is a need to research drug molecules to eliminate the Lassa virus in human body fluids.

In drug discovery, virtual screening (VS) and quantitative structure-activity relationship (QSAR) represent pivotal approaches that effectively complement laboratory processes, aiming to mitigate challenges encountered in experiments (Chakravarti and Alla, 2019). Therefore, virtual screening is now adopted in pharmaceutical companies and various academic and industrial projects to predict the biological activities of new molecules. The major goal of the QSAR is to identify optimal chemical structures, along with their most suitable conformer for effective interaction with a drug target (Abdel-ilah et al., 2017).

A drug target is a molecule within the body that the drug targets to produce a therapeutic effect.

QSAR utilize both three-dimensional (3D) and two-dimensional (2D) molecular representations as input models for forecasting biological activities (Chakravarti and Alla, 2019). Biological activity refers to the advantageous or harmful impacts a medication can impose on living organisms, and its significance is pivotal in the realm of medical uses (Alberga et al., 2019).

To perform QSAR, small molecules will potentially bind to a variety of drug targets which can be protein or cell targets (Alberga et al., 2019).

With the help of online repositories of well-annotated biological activity, such as ChEMBL and BindingDB, it has become easy to perform QSAR (Alberga et al., 2019).

Evaluating the potential bioactivity of a molecule involves comparing its chemical structure and properties with those of molecules that already exhibit known activity (Kwon et al., 2019).

Serendipity and systematic screening were important in drug discovery in the early years. Nevertheless, in recent times, it has expanded its scope to evolve from nature-inspired drug design to a more systematically guided approach known as rational drug design. In the pursuit of expediting drug discovery, numerous methodologies have been employed by researchers. Notably, (Makolo and Ajiboye, 2023) utilized nucleotide sequencing to identify and delineate Corona Virus potential drug targets. Within this progression, both drug design methodologies, namely nature-inspired and rational, now revolve around a fundamental concept called QSAR (Abdel-ilah et al., 2017). The QSAR framework has swiftly advanced, enabling the swift *in-silico* prediction of molecular characteristics and biological behaviors of new compounds, often without the need for extensive experimental testing. This approach effectively conserves resources, time, labor, and expenses.

The period following genomics has led to the emergence and accumulation of a wide array of QSAR models. The physical and chemical characteristics of each compound are typically determined through quantum chemical calculations of each compound because they can effectively capture both the molecule's global and specific properties (Nantasenamat et al., 2010). Supplementary sets of descriptors can be extracted from commercially accessible or freely available molecular property software packages designed for academic purposes. The fusion of multiple sources of molecular descriptors often yields a vast number of descriptors for subsequent analysis.

Tackling this extensive dataset necessitates computer algorithms capable of modeling its inherently intricate and multidimensional nature. A broad spectrum of learning algorithms exists to unveil concealed patterns within this substantial data, and adeptly selecting and optimizing learning parameters becomes pivotal for the triumph of modeling and prediction efforts (Hasan et al., 2022).

The drug discovery process follows a similar pattern, involving screening an extensive collection of compounds through high-throughput methods, resulting in substantial data generation. The QSAR approach emerges as a hopeful technology capable of establishing connections between a compound's structural attributes and its corresponding biological effects for streamlining the drug discovery workflow (Abdel-ilah et al., 2017). The evolution of a chemical structure entails the spatial positioning of atoms within a molecule, along with the chemical bonds that link these atoms. These characteristics can be used computationally to anticipate potential biological interactions.

To extract, analyze, discover, and predict the effect of drug molecules on a drug target, machine learning (ML) is being increasingly used. Machine learning has become a prominent computational technique widely utilized in the field of drug discovery (Bosc et al., 2021).

HEK293 cells, widely used in biomedical research, have been implicated in the replication and propagation of various viruses, including lassa virus (Tan et al., 2021).

A research article by Creative Biolab (2024) described the production of recombinant lineage IV Lassa in mammalian HEK293 cells assembled with z, GPC and N antigens and its usefulness in the development of LASV IV diagnostics and vaccine development.

This research utilizes machine learning to analyze and predict the behavior of drug molecules for Lassa fever target protein. This study also delves into drug-like qualities, encompassing factors such as water and fat solubility, effectiveness at the biological target, ligand efficiency, lipophilic efficiency, and molecular weight.

The explicit objective of the study is to employ Virtual

Screening (VS), QSAR, and Machine Learning (ML) to analyze and predict the behavior of drug molecules targeting the Lassa fever virus protein while exploring crucial drug-like qualities.

Scope and limitation

This model performed only the computational aspect of the drug discovery by using a QSAR and machine learning algorithms to discover drug candidates.

MATERIALS AND METHODS

In this work, QSAR model using random forest regressor was used to predict bioactivity of molecules that can inhibit lassa virus drug target.

The techniques and methods used in developing our model can be broadly divided into four phases, as shown in Figure 1.

Data collection and processing

Figure 2 shows the method applied in the data collection and processing of the bioactivity data that are used for the building of QSAR model.

A data set of inhibitors against human HEK293, the expression system of recombinant Lassa fever virus GP2 glycoprotein, was downloaded from ChEMBL database. ChEMBL is a 'chemogenomic' database that integrates chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. The ChEMBL ID of the dataset is ChEMBLID614818 shown in Table 1. The dataset obtained from ChEMBL comprises various measurements for bioactivity including IC50, Ki, % activity, % inhibition, EC50 and IC50. The bioactivity data with IC50 as their measurement units are selected for further investigation. They contain 7620 compounds.

Labelling molecules as active and inactive

Molecules with IC50 standard value greater than or equal to 10000 nM were labelled inactive, while those with IC50 value less than or equal to 1000 nM were labelled as active. The IC50 standard values are converted to PIC50 to ensure no uneven data points. PIC50 is the negative logarithm of IC50. After the conversion of IC50, any value greater than or equal to 6 was labelled as active, and any value less than 6 was labelled as inactive.

Calculation of Lipinski's descriptors

All small organic compounds and salts were removed from the compounds, and Lipinski descriptors (Molecule weight, log p, no of Hydrogen bond donors and no of Hydrogen bond acceptor) were computed in order to calculate the likelihood of being a drug-like molecule as shown in Figure 4.

Lipinski's descriptor's rule:

(1) Molecular weight < 500 Dalton
(2) Octanol-water partition coefficient (LogP) < 5
(3) Hydrogen acceptors < 10
(4) Hydrogen donors < 5

Descriptor calculation

The canonical SMILE of the molecules was encoded by a vector of fingerprint descriptors accounting for its molecular constituents and standardized using the built-in function of the PaDEL-Descriptor as a toolkit. This descriptor helped to capture the feature space of chemical compounds and represented the molecule substructures in binary digits. Figure 3 gives more details on the description calculation.

Collinearity is a condition where descriptor pairs are closely related to one another, and this will not only add complexity to the model but potentially give rise to bias. Collinearity was handled using the panda drop () function to drop one of the two features from a highly correlated pair.

Model building

The model building is in three stages: (i) data splitting, (ii) model building by Random Forest, (iii) and evaluation of the model as shown in Figure 4.

Data splitting

The data derived from the descriptors was divided into two sets to develop machine learning algorithms. 80% of the data was allocated for the training dataset, while the remaining 20% was designated as the test set.

Model building using Random Forest

Random Forest (RF) is an ensemble algorithm composed of several decision trees (Simeon et al., 2016). The fundamental concept behind Random Forest is to avoid constructing an overly complex decision tree with an excessive number of nodes, which could lead to overfitting and excessive adaptation to the data. Instead, it creates multiple trees to reduce variance (Simeon et al., 2016).

This model was built using Random Forest regressors, the n estimator was set to 100, and the remaining parameters were set to default, as shown in Figure 5.

Performance evaluation

The performance of this model was evaluated using the coefficient of determination (R^2), and root mean square error (RMSE) as the performance metric. In the QSAR model, R^2 is used to determine the model's goodness of fit.

A QSAR model is acceptable when it has a coefficient of determination R^2 value > 0.6 for the training set and R^2> 0.5 for the test set (Valeria Catalani et al., 2021).

RMSE is a measure of the prediction error exhibited by the trained model.

RESULTS AND DISCUSSION

The data set used for this model building is the HEK293 cell line. It contains 81788 bioactivity data of standard types potency, IC50, EC50, CC50, and Ki and only the standard type of IC50 was retained for further investigation. The data set that has IC50 as its standard type contains 7620 compounds. The coefficient of determination (r^2) and RMSE are the performance metrics used to measure the performance of the model

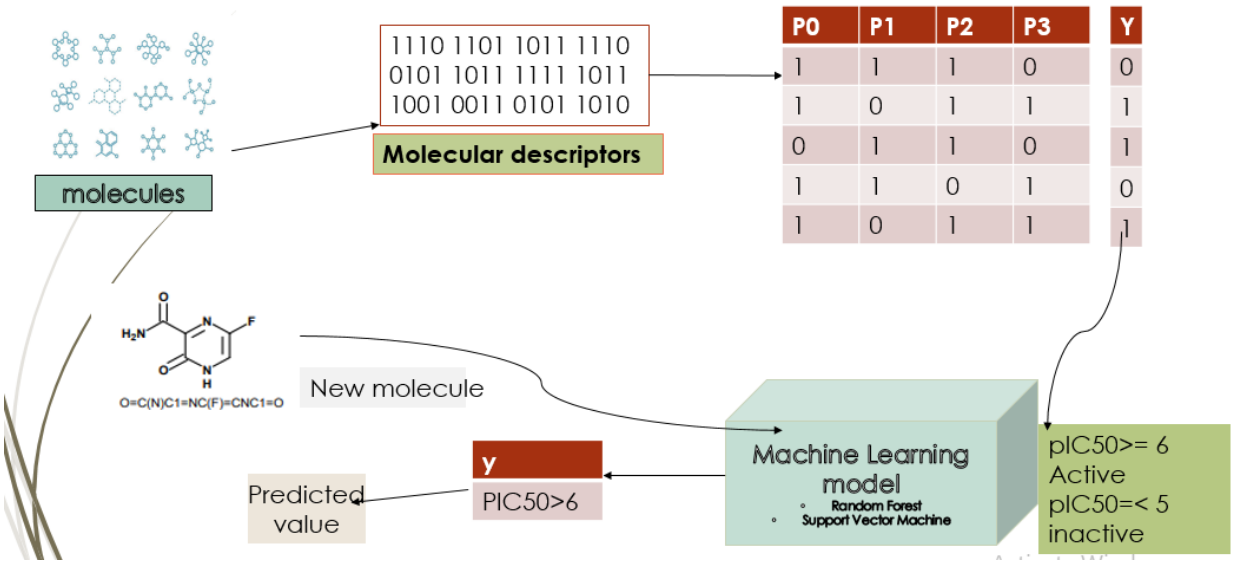

Figure 1. Generic model of techniques and method QSAR and machine learning used.

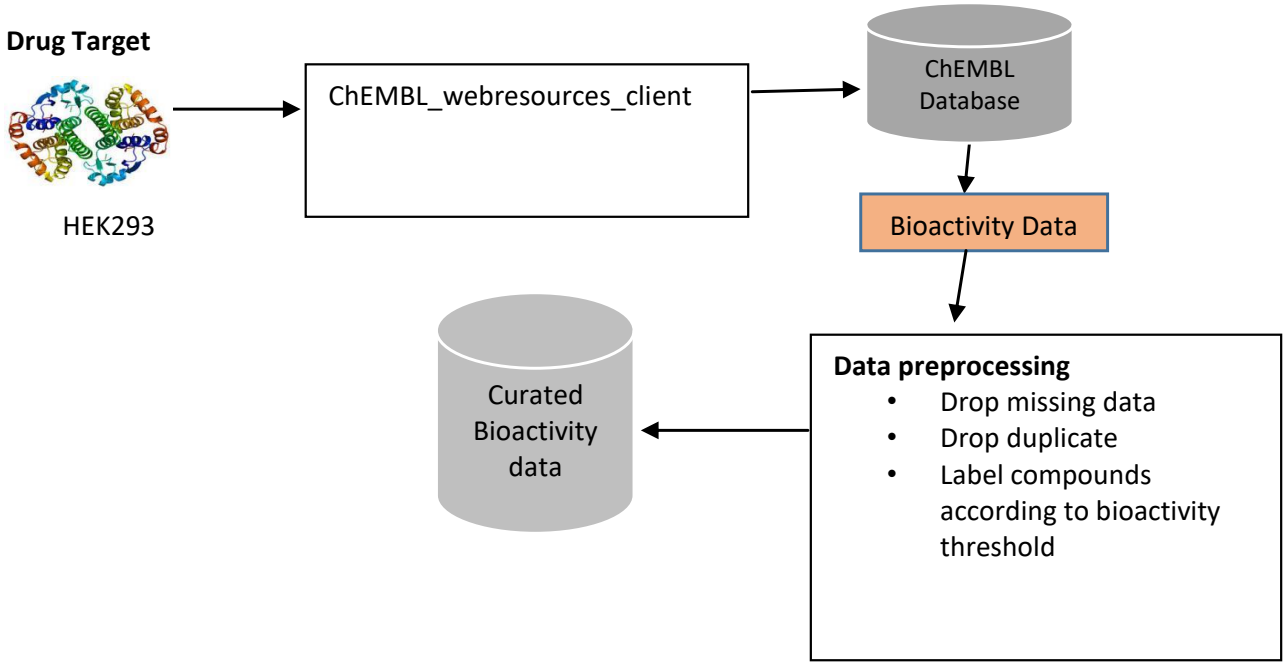

Figure 2. Data collection process.

built.

The coefficient of the determinant (r^2) defines the goodness of fit of the QSAR model. A QSAR model is acceptable when it has the value $r^2 > 0.6$ for the training set and $r^2 > 0.5$ for the test set. This model has the coefficient of determination r^2 of 0.93 and 0.56 (approximately) for the training and test set and RMSE of 0.32 and 0.77 for the training set and test set, respectively.

Calculation of descriptors

The values generated from Lipinski's descriptor were combined with the labeled bioactivity data and frequency plot to show the distribution of active and inactive

Table 1. Dataset.

Datasets source link	Name	No. of molecules
ChEMBL614818 [Target Report Card (ebi.ac.uk)	HEK293	7620 IC50 value

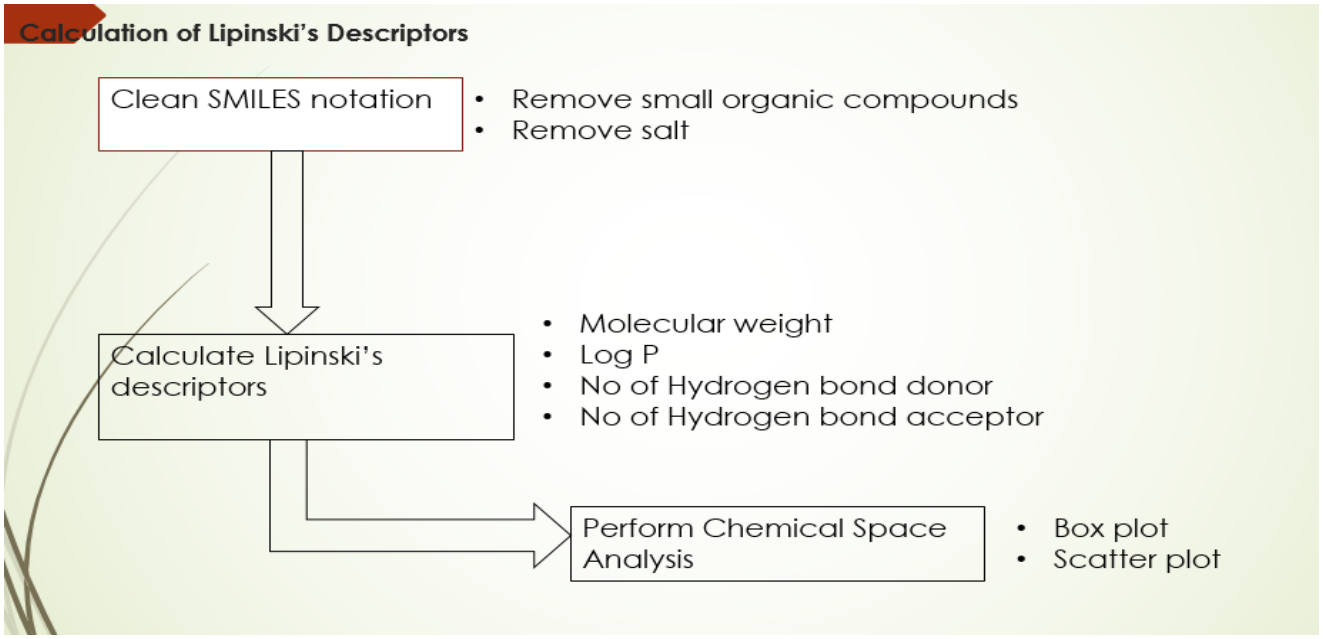

Figure 3. Calculation of Lipinski's descriptors.

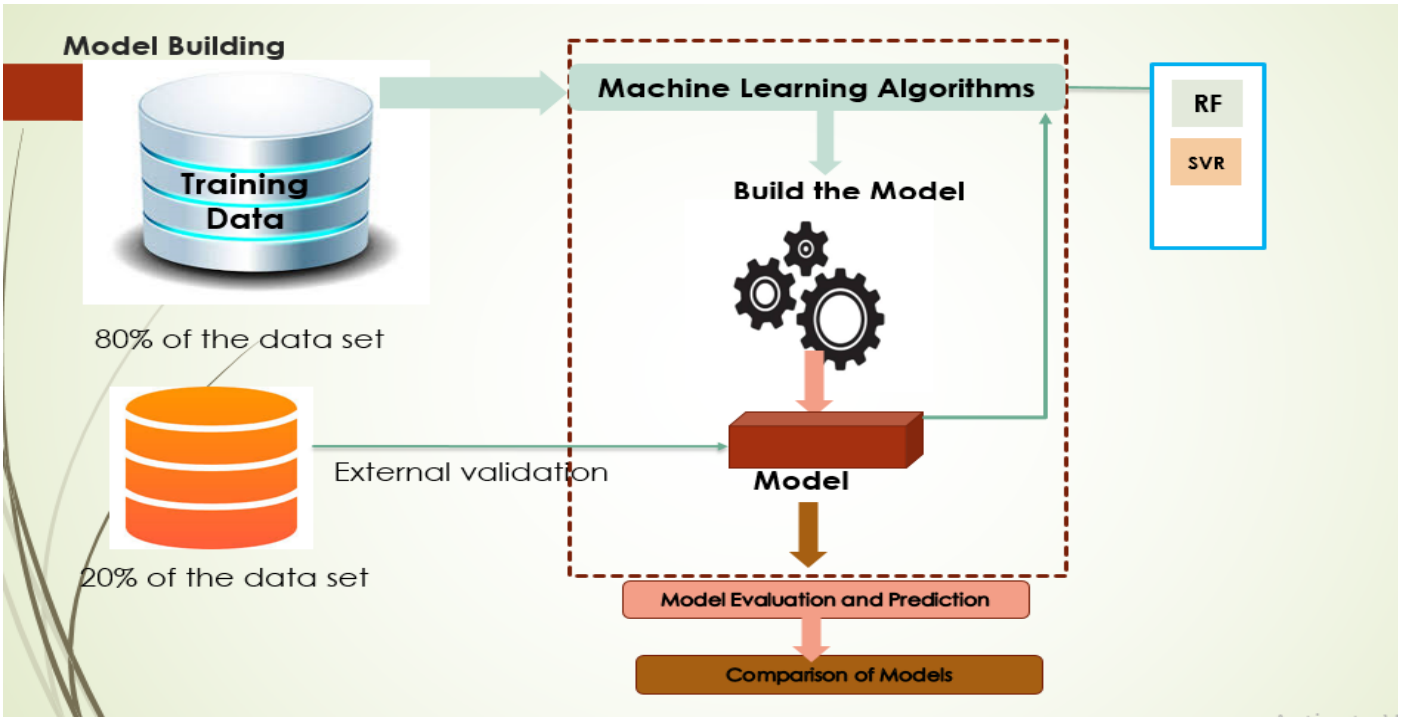

Figure 4. Model building using machine learning algorithms.

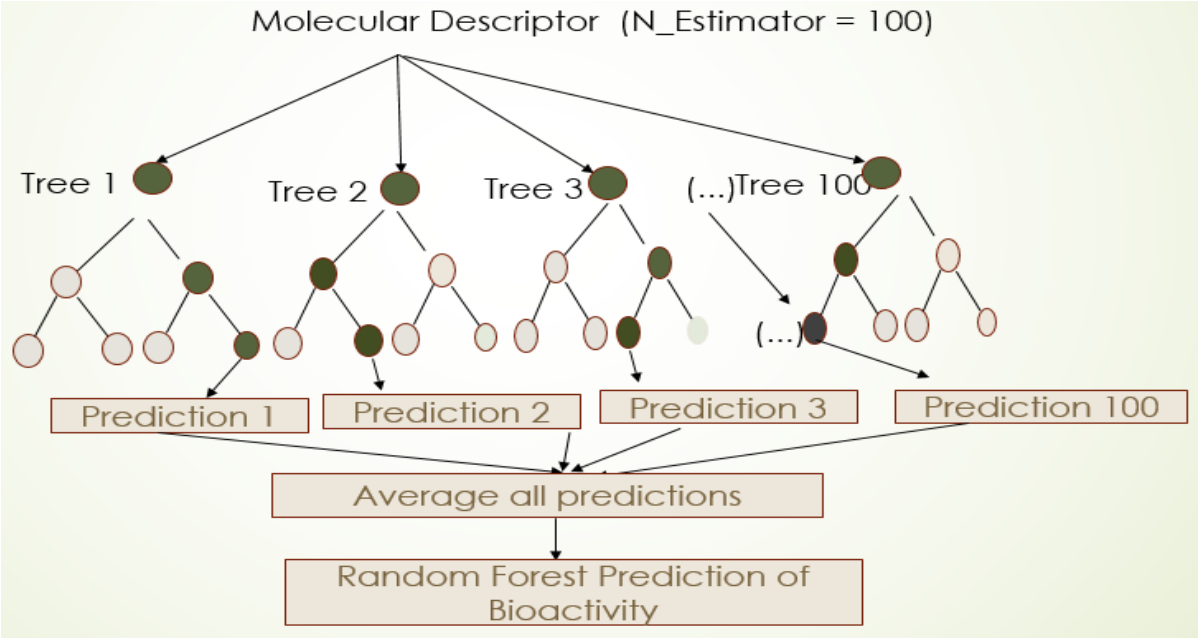

Figure 5. Random forest algorithms.
Source: Attanasi and Coburn (2023)

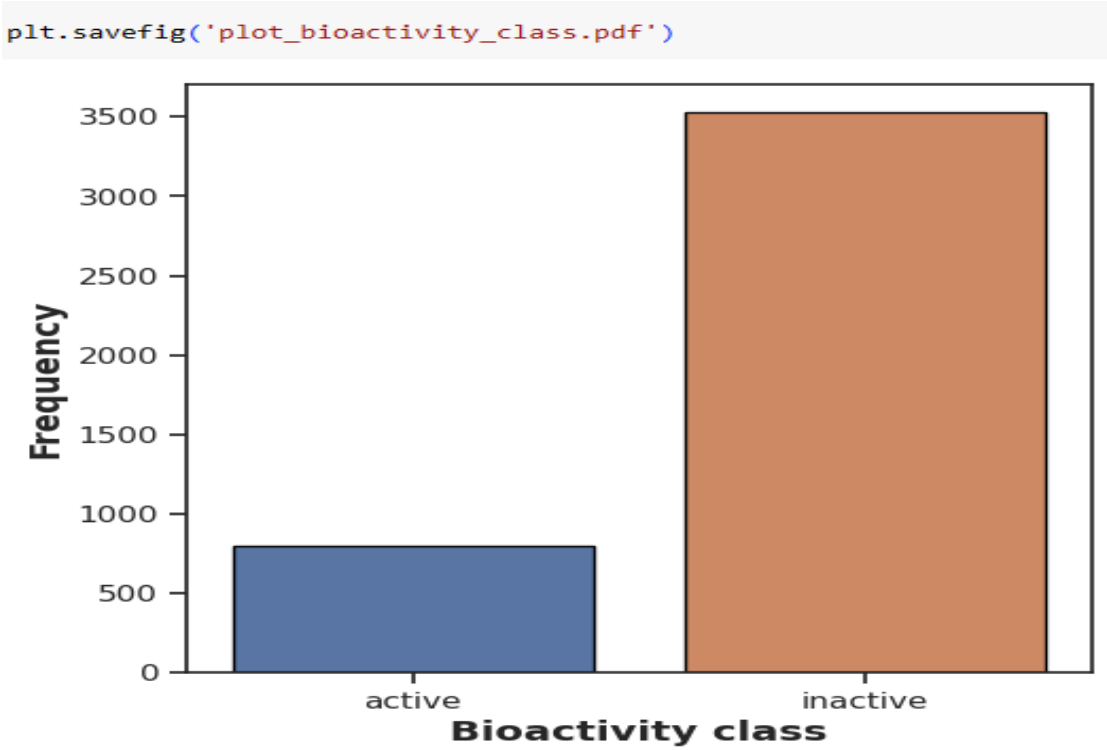

Figure 6. Frequency plot to show active and inactive molecules.

classes, as shown in Figure 6. Based on the set threshold, the frequency plot shows we have more inactive than active molecules. The standard value in IC50 was later converted to PIC50 to ensure that there

```
plt.ylabel( LogP , fontsize=14, fontweight= bold )
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0)
plt.savefig('plot_MW_vs_LogP.pdf')
```

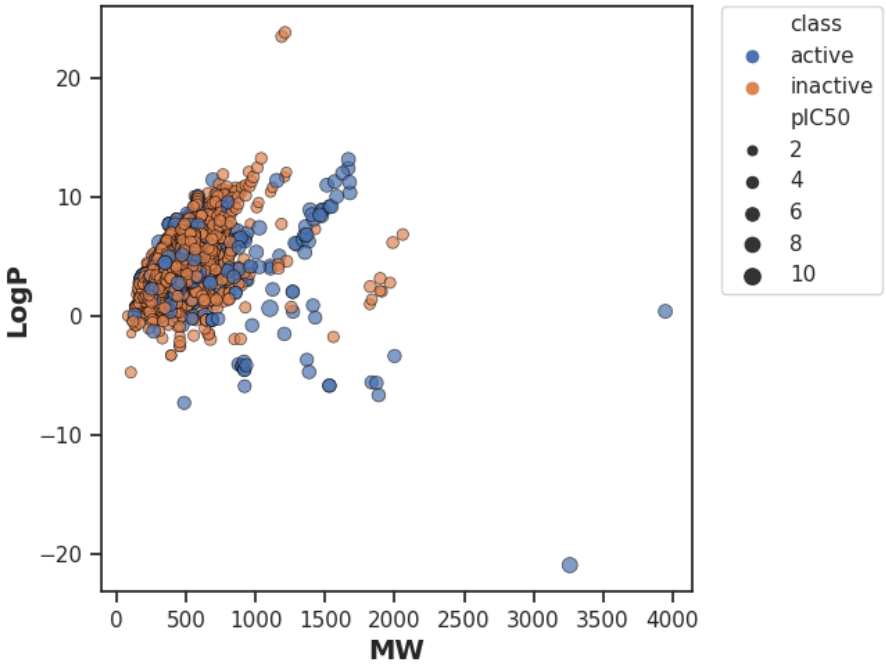

Figure 7. Scatter plot of LogP vs molecular weight.

was no uneven distribution of data points by taking the negative logarithm of IC50, and the active molecules now have a threshold of PIC50 greater than 6 and inactive molecules the threshold is PIC50 less than 5. The statistical description of the bioactivity dataset of 4984 after the conversion to PIC50: mean 4.9, standard deviation 1.172096, minimum value 1.207468, 25% 4.157531, 50% 4.698970, 75% 5.208485, max 10.920819 as shown in Figure 7 shows that the data was evenly distributed.

Exploratory data analysis of inhibitors via Lipinski's descriptor

Exploratory data analysis of inhibitors was performed to gain insights into the structure-activity relationship by analysing Lipinski's rule-of-five descriptors (Ursu et al., 2011). This provides important knowledge on the general character of compounds governing inhibitory properties of compounds. Exploratory data analysis was performed using Lipinski's rule-of-five descriptors comprising MW, LogP, numHDonors and numHAcceptors. MW represents the molecular weight of a compound and is commonly used because it can be used to determine the dosage of the drug and formulation of the appropriate compound size that is important for its passage via the lipid membrane. LogP detects if a compound can cross the cell membrane and reach its target. numHDonors and numHAcceptors describe the number of hydrogen bond donors and hydrogen bond acceptors, respectively, which are used to measure hydrogen bonding capacity. Visualisation of the chemical space of LogP as a function of MW is as shown in Figure 8. LogP is a measure of the lipophilicity of a compound, which affects its ability to cross biological membranes. MW is a measure of the size and complexity of a compound, which affects its solubility and transport.

Active compounds are those that have a desired biological effect, such as inhibiting a target enzyme or binding to a receptor. Inactive compounds are those that do not have the desired effect or have undesirable side effects.

The scatter plot suggests that active compounds tend to have higher LogP and lower MW than inactive compounds. This means that active compounds are more lipophilic and smaller than inactive compounds, which may make them more likely to reach and interact with their targets.

The scatter plot also suggests that there is a trade-off between LogP and MW, as increasing one tends to decrease the other. This means that there is a balance between lipophilicity and size that affects the activity of a compound.

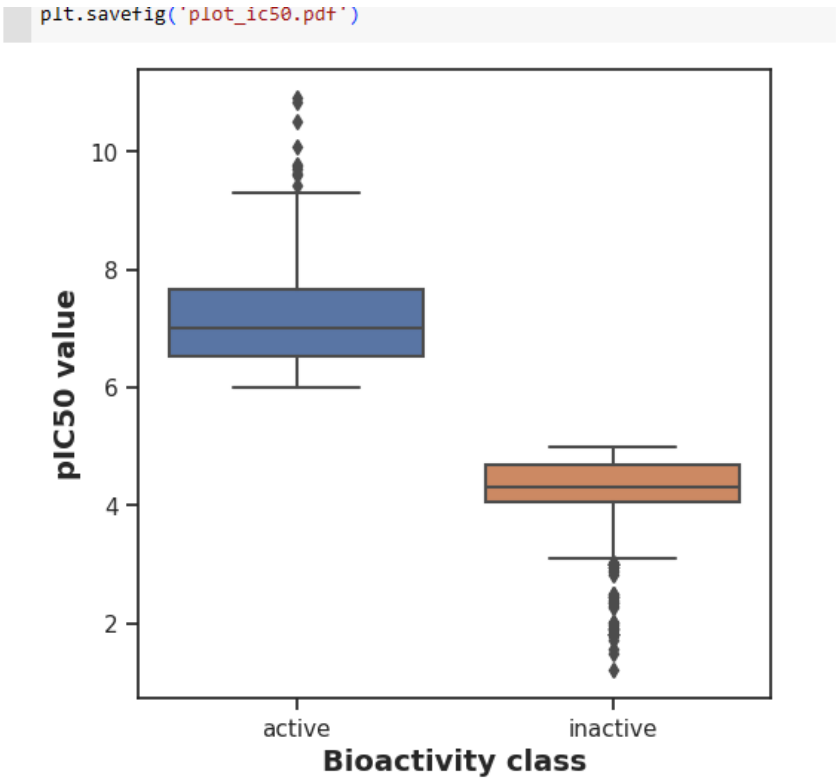

Figure 8. Box plot of PIC50 vs bioactivity class.

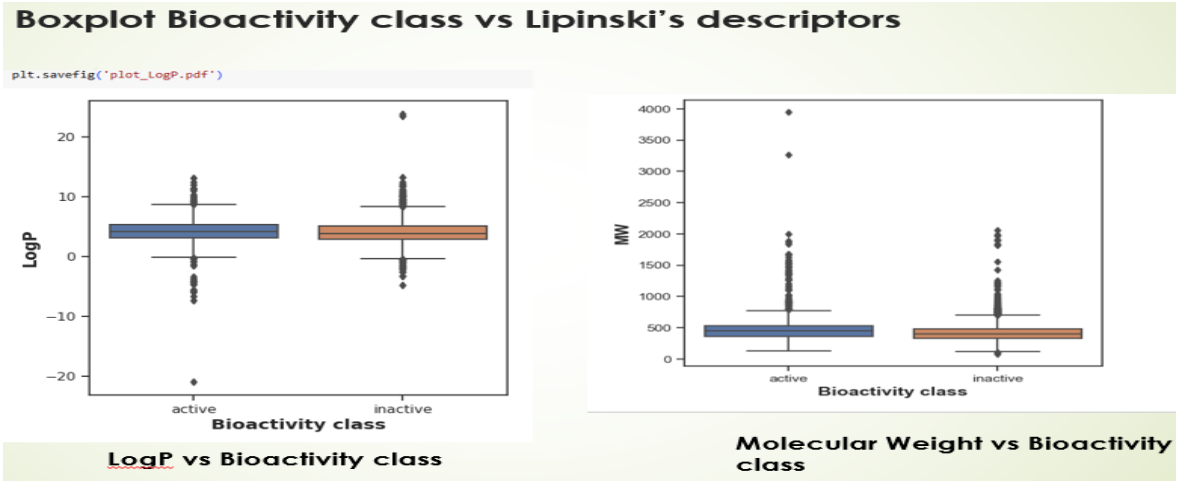

Figure 9. Box plot of bioactivity class and Lipinski's descriptors.

The PIC50 value is a measure of the potency of a compound, which is the inverse of the concentration required to achieve 50% of the maximum effect. Higher PIC50 values indicate higher potency, meaning that less compound is needed to achieve the same effect.

The scatter plot shows that the PIC50 values vary within each class of compounds, as indicated by the different shades of blue and orange. Darker shades indicate higher PIC50 values, meaning higher potency. The scatter plot shows that some active compounds have higher potency than others, and some inactive compounds have lower potency than others.

Also, Figure 9 shows the box plot graph of the distribution of PIC50 values for two different bioactivity

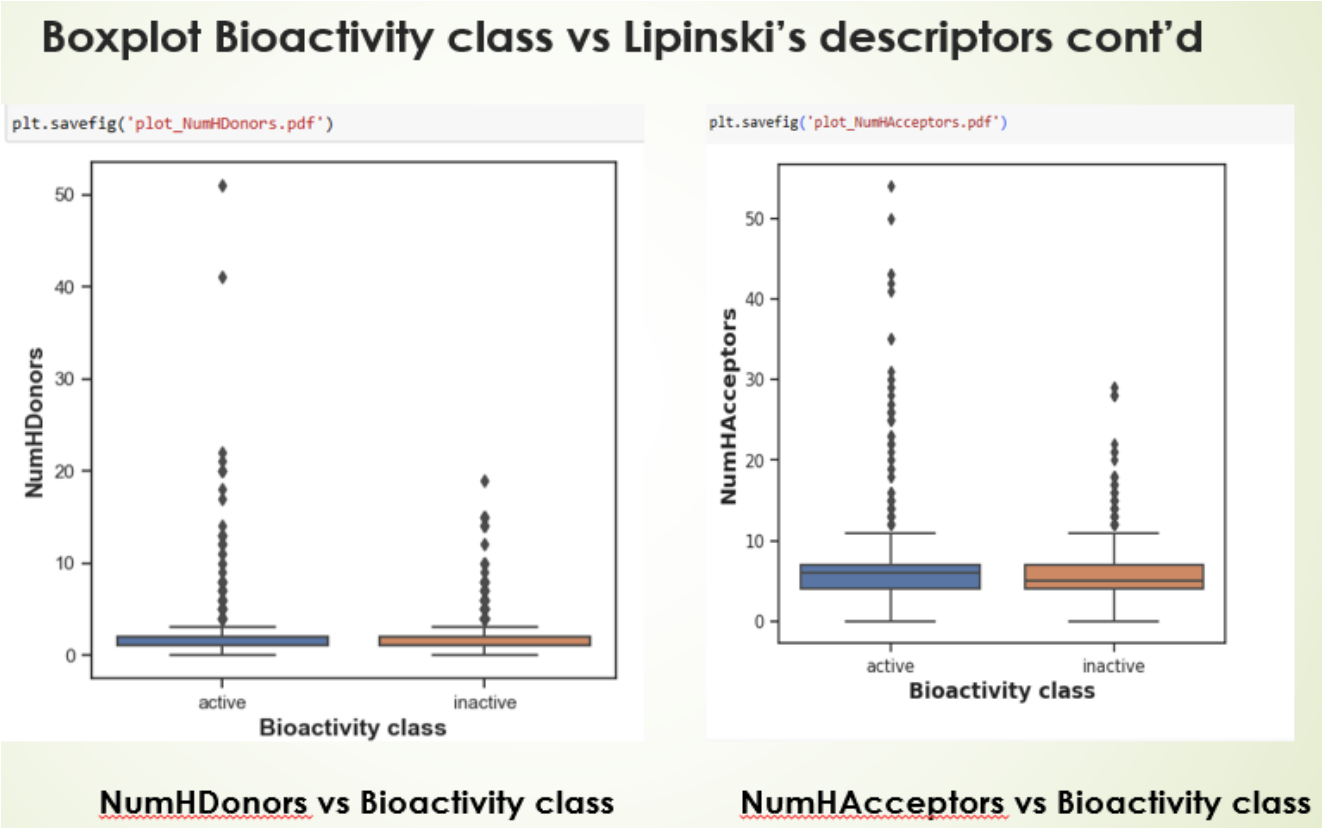

Figure 10. Box plot of bioactivity class and Hydrogen bond donor and acceptors.

classes: active and inactive. The PIC50 value is a measure of the potency or efficacy of a biochemical substance, such as a drug or a ligand, in interacting with a biological target, such as protein or a receptor. A higher PIC50 value means that the substance has a higher affinity or a lower concentration required to achieve a certain effect.

The graph has two categories on the x-axis: active and inactive, representing the substances that have a bioactivity above or below a certain threshold, respectively. The y-axis shows the PIC50 values ranging from 0 to 10. The graph has two boxes: one blue for the active substances and one orange for the inactive substances. Each box represents the interquartile range (IQR) of the PIC50 values for each category, which is the difference between 25 and 75th percentiles. The line inside each box is the median PIC50 value for each category, which is the middle value when the data is sorted in ascending order. The whiskers are the outliers extend from the boxes to the minimum and maximum values within 1.5 times the IQR bioactive class active class versus PIC50 to visualize active and inactive inhibitors; the active class has PIC50 values greater than six and the inactive class less than or equal to 5. The dot is the outliers found in both the active and inactive classes.

In addition, the box plot of Lipinski's descriptors shown in Figure 10 the graph of logP versus bioactivities class shows that the active compounds tend to have a wider range of lipophilicity, which may suggest that they have a different mode of action or target different receptors. The inactive compounds are more clustered around a neutral LogP value, which may imply that they are less likely to cross biological membranes or bind to hydrophobic sites.

Also in the same Figure 10, the graph of molecular weight (MW) versus bioactivity class shows the distribution of MW for active and inactive classes. MW is a measure of the size and complexity of a compound, which can affect its pharmacokinetic and pharmacodynamic properties. In this plot, the median of MW for active compounds is around 500, while for inactive compounds it is slightly higher but still below 1000. This implies that the active compounds are smaller and simpler than the inactive ones, which may make them more likely to penetrate biological barriers and reach their targets.

The “inactive” class has a more compact distribution with fewer outliers, implying that the inactive compounds have similar MW and are less diverse. These compounds may be too large or too complex to interact with the desired targets, or they may have unfavorable properties that limit their bioavailability or efficacy.

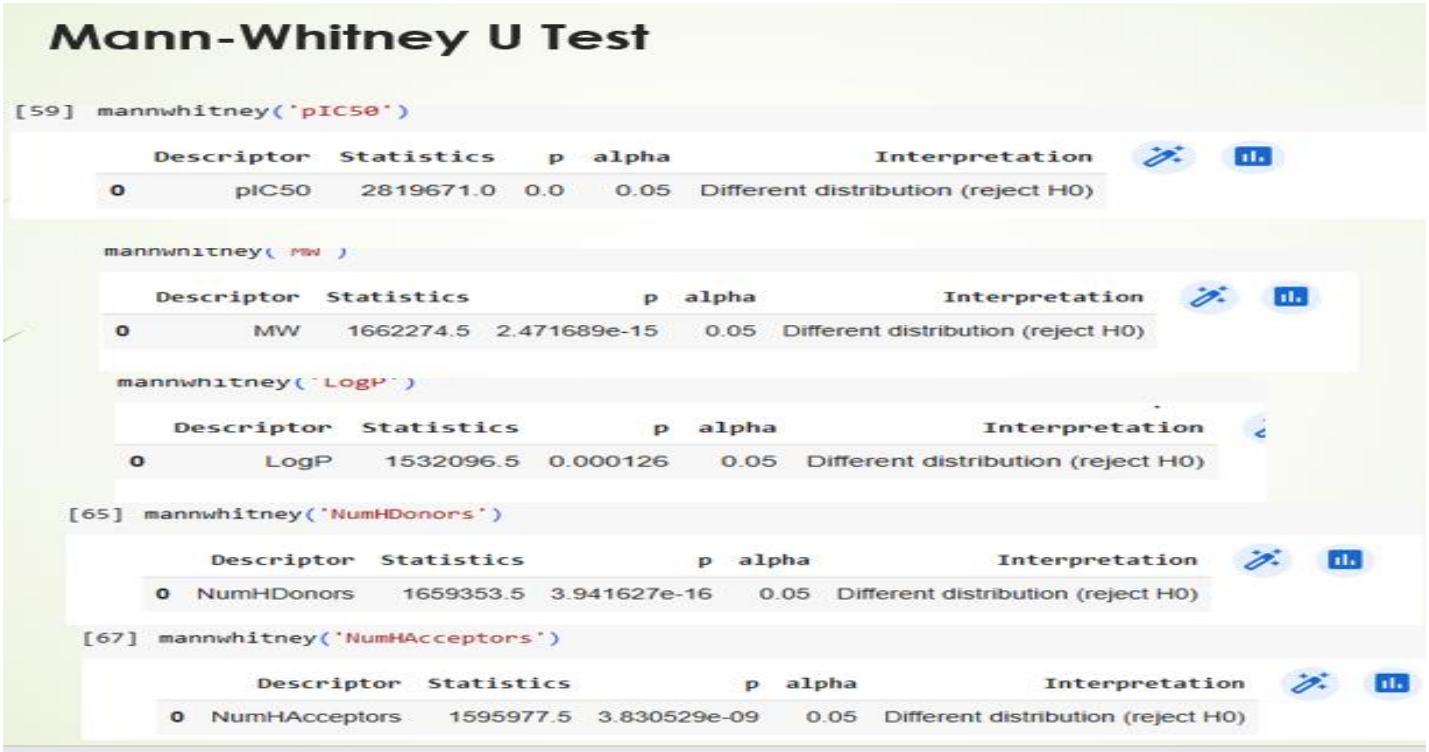

Figure 11. Mann-Whitney U test for each Lipinski's rule of drug likeness.

In Figure 11, the graph of NumHDonors versus Bioactivity classes shows the distribution of a chemical property called NumHDonors in active and inactive classes. NumHDonors are the number of atoms that can donate hydrogen bonds to other molecules. These bonds are important for the interaction between drugs and their targets. The graph suggests that the active class has lower and less variable NumHDonors than the inactive class. This means that molecules with fewer hydrogen bond donors are more likely to be biologically active. However, there are also many outliers in both classes, meaning that there are exceptions to this trend.

In Figure 11, the graph of NumHAcceptors versus Bioactivity class represents the distribution of NumHAcceptors in two different bioactivity classes: active and inactive. The NumHAcceptors is a measure of the number of hydrogen atoms that can form hydrogen bonds with other molecules, which can affect the solubility and permeability of a compound. In this plot, the "active" class has a higher average NumHAcceptors compared to the "inactive" class, indicating that the active compounds are more polar and can form more hydrogen bonds. The "active" class also has a wider interquartile range (IQR) and more outliers, suggesting that there is more diversity and variability in the NumHAcceptors of the active compounds. The "inactive" class has a more compact distribution with fewer outliers, implying that the inactive compounds have similar NumHAcceptors and are less polar. The active compounds have a higher tendency to interact with polar targets or environments, while the inactive compounds are more likely to be excluded or rejected by them.

A Mann-Whitney test is a non-parametric statistical test that compares the distributions of two independent groups of data. It tests whether there is a significant difference in the median values of the groups. The null hypothesis (H0) is that the two groups have the same distribution, and the alternative hypothesis (H1) is that they have different distributions.

Mann-Whitney tests were conducted on four descriptors: 'pIC50', 'MW', 'NumHDonors', and 'NumHAcceptors', comparing two independent groups of data. The results revealed significant differences in the distributions of these descriptors between the groups, with p-values much smaller than 0.05 as shown in Figure 12. This suggests that the median values of the descriptors likely differ between the two groups.

Molecular descriptor calculation from PADEL-Descriptors

PADEL-descriptor does the calculation of the local properties of the canonical smile in binary digits. The output of this Descriptor is used as an X-input to build the machine learning. The output of the result is as shown in

```
df3_X = df3_X.drop(columns=['Name'])
df3_X
```

	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9	...	PubchemFP871	PubchemFP872	PubchemFP873	PubchemFP874	Pubchem
0	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
1	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
2	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
3	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
4	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4979	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
4980	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
4981	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
4982	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	
4983	1	1	1	0	0	0	0	0	0	1	...	0	0	0	0	

4984 rows x 881 columns

Figure 12. Conversion of canonical SMILES to Pubchem binary digits' fingerprint.

Figure 13.

The molecular descriptors have 4984 rows and 881 columns.

Model building

Splitting of data

The dimension of the data set was checked before splitting the data. The dimension of X is (4984, 881), and the dimension of Y is (4984).

The data is split into a training set and a test set; 80% of the data is used for the training set and 20% for the test set. The dimension of the data is viewed after the splitting of the data: X_train (3987, 881), Y_train (3987), X_test (997, 881) and Y_test (997).

Building a regression model using Random Forest

The regression model is built using Random Forest Regressor, n_estimators is 100, and the train set gives the coefficient of determination (R^2) equals 0.93 approximately and the prediction test is 0.56 approximately, which is an acceptable QSAR according to Valeria Catalani et al. (2021). The result of the model is as shown in Figure 14. The model's fitness is visualised by plotting the scatter plot of the Predicted PIC50 value versus the Experimental PIC50 value, as shown in Figure 15.

In Figure 15, the graph shows how well the predicted PIC50 values match the experimental PIC50 values. The PIC50 value is a measure of how potent a substance can inhibit a biological function. The higher the PIC50 value, the more potent the substance is. The graph compares the experimental and predicted values for different substances using a scatter plot. The x-axis shows the experimental PIC50 values and the y-axis shows the predicted PIC50 values. The blue line is the line of best fit, which shows the average trend of the data. The closer the dots are to the line, the more accurate the predictions are. The graph suggests that there is a good correlation between the experimental and predicted values, as most of the dots are close to or on the line. This means that the prediction method is reliable and can be used to estimate the potency of new substances.

Comparison with other models

This model is compared with other models using the library lazy predict, some lazy predict regressors with good performance in QSAR modelling are selected, as shown in Figure 16, and their performance was compared with Random Forest Regressors. The

```
#Examine the dimension of x
X.shape
```

```
(4984, 881)
```

```
[15] Y.shape
```

```
(4984,)
```

Data split (80/20 ratio)

```
[16] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

data dimension examination

```
[17] X_train.shape, Y_train.shape
```

```
((3987, 881), (3987,))
```

```
[18] X_test.shape, Y_test.shape
```

```
((997, 881), (997,))
```

Building a Regression Model using Random Forest

```
[19]
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train, Y_train)
r2 = model.score(X_train, Y_train)
r2
```

```
0.9251573167005389
```

```
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train, Y_train)
r2 = model.score(X_test, Y_test)
r2
```

```
0.5597194869230097
```

Figure 13. Regression model using random forest.

```
ax.figure.set_size_inches(5, 5)
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

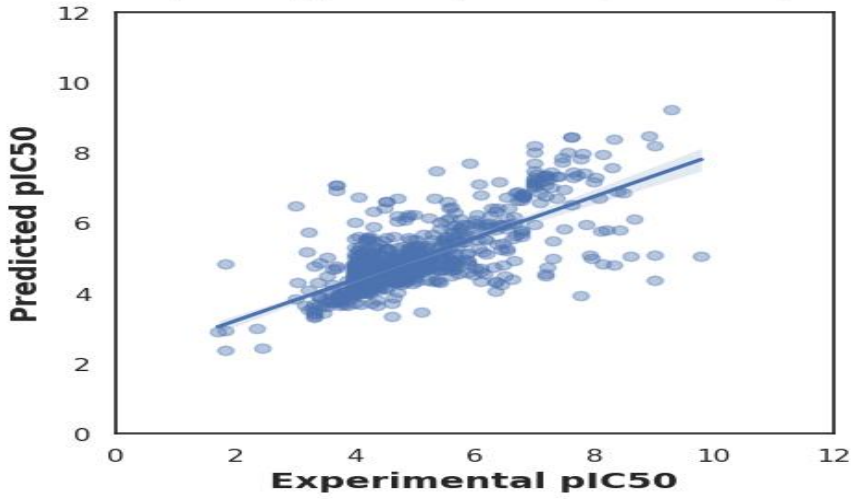

Figure 14. Scatter plot of experimental versus predicted PIC50 values.

```
[ ]  # Defines and builds the lazyclassifier
     from sklearn.utils import all_estimators
     from sklearn.base import RegressorMixin
     chosen_regressors = [
         'SVR',
         'BaggingRegressor',
         'RandomForestRegressor',
         'DecisionTreeRegressor',
         'MLPRegressor',
         'LGBMRegressor',
         'KNeighborsRegressor',
         'GradientBoostingRegressor',
         'LinearRegression',
         'BayesianRidge'
     ]

     REGRESSORS = [
         est
         for est in all_estimators()
         if (issubclass(est[1], RegressorMixin) and (est[0] in chosen_regressors))
     ]

     clf = LazyRegressor(verbose=1, ignore_warnings=False, custom_metric=None, regressors=REGRESSORS)
     models_train,predictions_train = clf.fit(X_train, X_train, Y_train, Y_train)
     models_test,predictions_test = clf.fit(X_train, X_test, Y_train, Y_test)
```

Figure 15. Comparison of machine learning models using lazy predict.

```
# Performance table of the training set (80% subset)
predictions_train
```

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
DecisionTreeRegressor	0.96	0.97	0.20	0.75
RandomForestRegressor	0.90	0.92	0.32	27.44
BaggingRegressor	0.88	0.90	0.36	2.96
MLPRegressor	0.87	0.90	0.37	17.39
SVR	0.61	0.70	0.65	22.42
KNeighborsRegressor	0.58	0.68	0.67	1.11
LinearRegression	0.42	0.55	0.79	0.90
GradientBoostingRegressor	0.39	0.53	0.80	6.38
BayesianRidge	0.35	0.49	0.83	1.68

Figure 16. Performance table of the training set.

```
[17] # Performance table of the test set (20% subset)
     predictions_test
```

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
RandomForestRegressor	-2.70	0.57	0.77	26.69
SVR	-2.95	0.54	0.80	11.63
BaggingRegressor	-3.14	0.52	0.82	3.70
KNeighborsRegressor	-3.24	0.51	0.83	0.41
GradientBoostingRegressor	-3.80	0.45	0.88	5.46
BayesianRidge	-4.25	0.39	0.92	1.60
MLPRegressor	-5.15	0.29	1.00	18.24
DecisionTreeRegressor	-5.88	0.21	1.05	0.57
LinearRegression	-525985204345501328080896.00	-607312233933058762340.00	291115294133.39	0.83

Figure 17. Performance table of the test set.

DecisionTreeRegressor performs best on the training set with an adjusted R-squared value of 0.96, R-squared value of 0.97, RMSE value of 0.20 and time taken of 0.75, and RandomForestRegressors is the second model that performs best on the training set with adjusted R-squared value of 0.90, R-squared value of 0.92, RMSE value of 0.32 and time taken of 27.44, the detail is as shown in Figure 17, while on the test set RandomForestRegressor performs best on the test set with adjusted R-squared value of -2.70, R-squared value of 0.57, RMSE value of 0.77 and time taken of 26.69. At the same time, SVR is the second model that performs best on the test set with an adjusted R-squared value of -2.95, R-squared value of 0.54, RMSE value of 0.80 and time taken of 11.63; the detail is as shown in Figure 18.

The visualisation of the model's performance metrics given in Figure 18 shows the R-Squared values of different regression models. Random Forest Regressor has a higher R-Squared value than most of the other models, except for Decision Tree Regressor. This means that Random Forest is able to capture the relationship between the input and output variables better than most of the other models. Figure 19 shows that RMSE value of DecisionTreeRegressor with value of 0.20, RandomForestRegressor with value of 0.32, BaggingRegressor with value of 0.36 and MLP with value of 0.37 satisfied the acceptable QSAR value of 0.5 for RMSE and Figure 20 shows that RandomForestRegressor, MLPRegressor and SVR took more time to build the model than other machine learning algorithms.

Conclusion

In this study, the experiment was performed with 7620 datasets, and after the data cleaning, it was reduced to 4984, and these 4894 datasets were used to perform QSAR. The canonical smile of molecule is used to build this model after its conversion by PADEL-Description to binary digits, and the model built by random regression satisfied the threshold of an acceptable QSAR model.

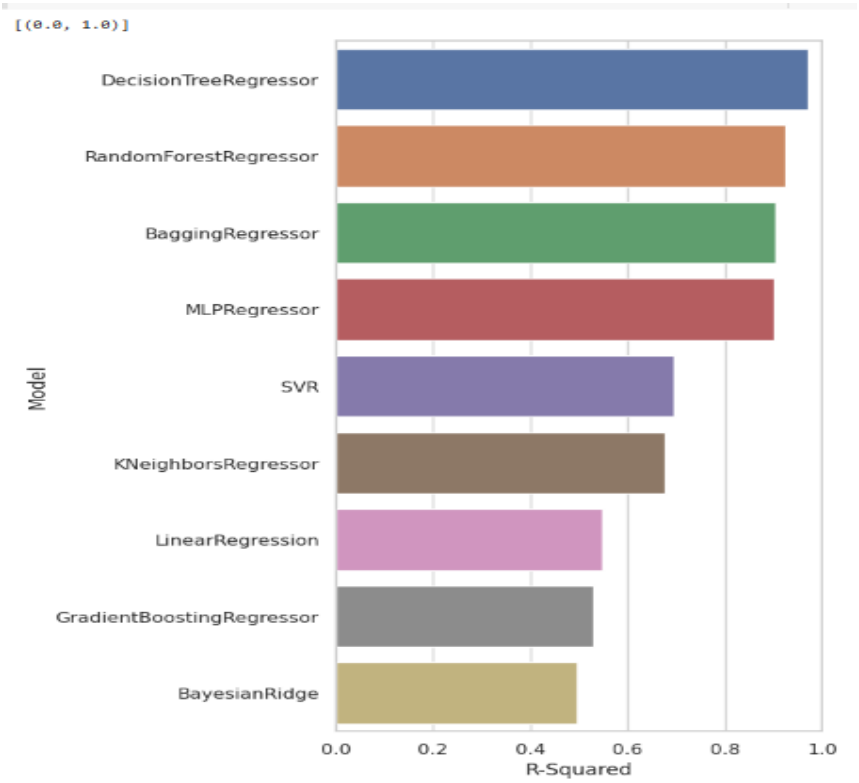

Figure 18. Visualisation of R-square.

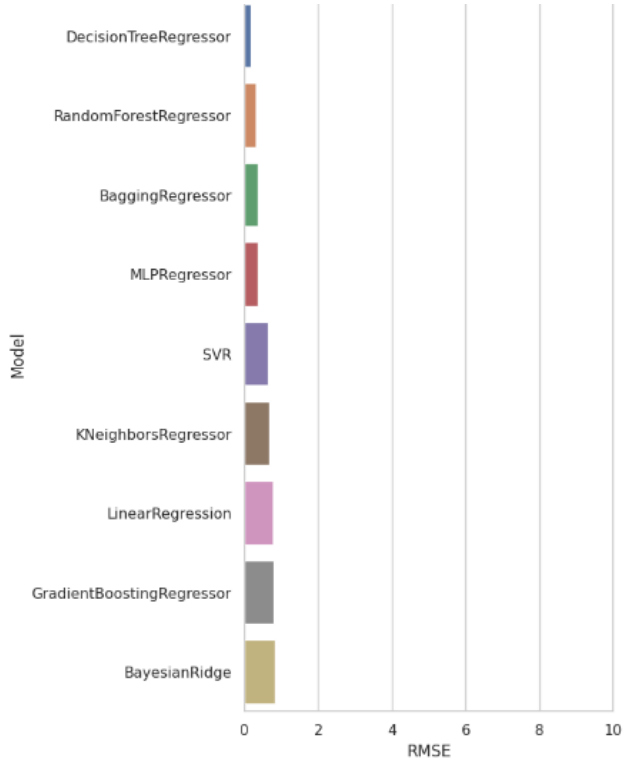

Figure 19. Visualisation of RMSE result.

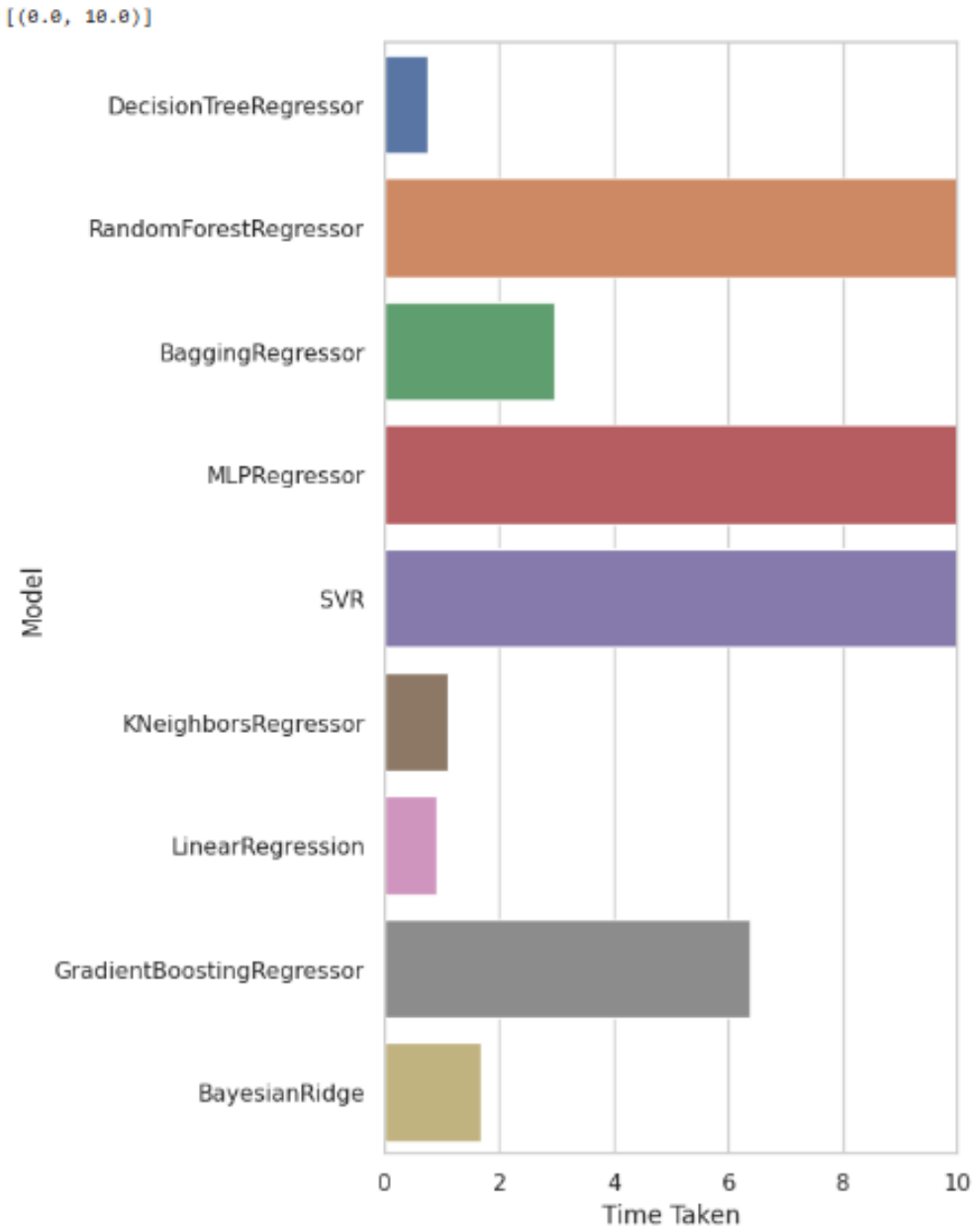

Figure 20. Time taken by the models.

Hence, this model can predict newly discovered drug candidates that can inhibit Lassa Fever Virus GP2 glycoprotein in HEK293.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES

Abdel-ilah L, Veljović E, Gurbeta L, Badnjević A (2017). Applications of QSAR Study in Drug Design. International Journal of Engineering Research and Technology 6(06):582-587.

Alberga D, Trisciuzzi D, Montaruli M, Leonetti F, Mangiatordi GF, Nicolotti O (2019). A New Approach for Drug Target and Bioactivity Prediction: The Multifingerprint Similarity Search Algorithm (MuSSeL). Journal of Chemical Information and Modeling 59(1):586-596.

Arefin A, Hossen S, Islam T, Islam A, Biswas P, Nu A (2021). Target specificity of selective bioactive compounds in blocking α-dystroglycan receptor to suppress Lassa virus infection: an in silico approach. Free full text Target specificity of selective bioactive compounds in blocking α- dystroglycan receptor to suppress Lassa virus infection: an in silico approach. Journal of biomedical research, 35(6):459.

Attanasi ED, Coburn TC (2023). Random Forest. In: Daya Sagar, B.S., Cheng, Q., McKinley, J., Agterberg, F. (eds) Encyclopedia of

Mathematical Geosciences. Encyclopedia of Earth Sciences Series. Springer, Cham.
Bosc N, Felix E, Arcila R, Mendez D, Saunders MR, Green DVS, Ochoada J, Shelat AA, Martin EJ, Iyer P, Engkvist O, Verras A, Duffy J, Burrows J, Gardner JMF, Leach AR (2021). MAIP: a web service for predicting blood - stage malaria inhibitors. Journal of Cheminformatics pp. 1-14.
Chakravarti SK, Alla SRM (2019). Descriptor Free QSAR Modeling Using Deep Learning with Long Short-Term Memory Neural Networks. Frontiers in Artificial Intelligence 2:1-18.
Creative Biolabs (2024). Recombinant Lassa IV Virus-like Particles (LASV IV VLPs). Available at: https://www.creative-biolabs.com/vaccine/recombinant-lassa-lineage-iv-virus-like-particles-lasv-iv-vlps.htm
Hasan R, Alsaiari AA, Fakhurji BZ, Habibur M, Molla R, Asseri AH, Sumon AA, Park MN, Ahammad F, Kim B (2022). Application of Mathematical Modeling and Computational tools in the modern drug design and development process. Molecules 27(13):4169.
Id APS, Duvignaud A, Jaspard M, Malvy D, Carroll M, Tarning J, Olliaro PL, Horby PW (2022). Ribavirin for treating Lassa fever: A systematic review of pre-clinical studies and implications for human dosing. PLoS Neglected Tropical Diseases pp. 1-18. https://doi.org/10.1371/journal.pntd.0010289
James O (2020). Modelling and optimal control analysis of Lassa fever disease. Informatics in Medicine Unlocked 20:100419.
Kikiowo B (2021). Molecular Interaction and Inhibitory Activity of Dandelion's Compounds on Nucleoprotein: A Therapeutic Intervention in Lassa Fever. Biointerface Research in Applied Chemistry 11(5):12573-12583.
Klitting R, Kafetzopoulou LE, Thiery W, Dudas G, Gryseels S, Kotamarthi A, Vrancken B, Gangavarapu K, Momoh M, Sandi JD, Goba A, Alhasan F, Grant DS, Garry RF, Smither AR, Zeller M, Pauthner MG, Mcgraw M (2021). Predicting the evolution of Lassa Virus endemic area and population at risk over the next decades. Nature Communications, pp. 1-17.
Kwon S, Bae H, Jo J, Yoon S (2019). Comprehensive ensemble in QSAR prediction for drug discovery. BMC Bioinformatics pp. 1-12.
Minari J, Agho E, Adebiyi F, Rotimi O, Sholaja B, Adejumo J (2021). Molecular Docking and Identification of Candidate Blockers for Endonuclease Domain of Lassa Virus Polymerase as Potential Drugs. Journal of Applied Sciences and Environmental Management 25(11):1899-1907.
Makolo AU, Ajiboye F (2023). Prediction of Genomic Signature of Ngs Sequences and Comparative Drug-Likeness. American Scientific Research Journal for Engineering, Technology and Sciences 90(1):573-589.
Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010). Advances in computational methods to predict the biological activity of compounds. Expert Opinion on Drug Discovery 5(7):633-654.
Oloniniyi OK, Unigwe US, Okada S, Kimura M, Koyano S, Miyazaki Y, Yasuda J. (2016). Genetic characterization of Lassa virus strains isolated from 2012 to 2016 in southeastern Nigeria Author summary. https://doi.org/10.1371/journal.pntd.0006971
Simeon S, Anuwongcharoen N, Shoombuatong W (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. https://doi.org/10.7717/peerj.2322
Tan E, Sze C, Chin H, Feng Z, Lim S, Ng SK (2021). HEK293 Cell Line as a Platform to Produce Recombinant Proteins and Viral Vectors. Frontiers in Bioengineering and Biotechnology 9:1-9.
Ursu O, Rayan A, Goldblum A, Oprea T (2011). Understanding drug-likeness. Wiley Interdisciplinary Reviews: Computational Molecular Science https://doi.org/10.1002/wcms.52