Chapter 3
Classification of BACE-1 inhibitors by using machine learning methods

Summary:

The beta-site amyloid precursor protein cleaving enzyme 1 (BACE-1) is a transmembrane aspartyl-protease, that cleaves amyloid precursor protein (APP) at the β-site. The sequential proteolytic cleavage of APP, first by β-secretase and then by γ-secretase complex, leads to the production and release of amyloid-β peptide, a pathological hallmark of AD. BACE-1 inhibitors are reported to possess considerable potential in decreasing the level of amyloid-β in the brain and preventing the progression of AD. A classification study has been conducted on 3536 diverse BACE-1 inhibitors, obtained from the Binding DB database, by extracting two types of descriptors, that is, molecular property (Mordred) and fingerprints (Pubchem, MACCS and KRFP). Furthermore, based on the descriptors, various machine learning algorithms such as Naïve Bayesian (NB), nearest known neighbors (kNN), support vector machine (SVM), random forest (RF) and gradient-boosted algorithms (XGB) were applied to develop classification models. The performance of the models was evaluated using accuracy, precision, recall, and the F1 score of the test set. The best NB, kNN, SVM, RF and XGB classifiers had F1 scores of 0.74, 0.85, 0.86, 0.87 and 0.87, respectively. The diverse 3536 BACE-1 inhibitors were clustered into 11 subsets, and the structural features of each subset were evaluated. The important fragments present in active and inactive compounds were also identified. The model developed in the study would serve as a valuable tool for the design of BACE-1 inhibitors and the virtual screening of molecules to identify these.

3 Classification of BCAE-1 inhibitors using machine learning methods

3.1 Introduction

The β-site amyloid precursor protein cleaving enzyme 1 (BACE-1), also known as β-secretase, is mainly expressed in the neurons of the brain. It is a transmembrane aspartyl-protease responsible for producing toxic Aβ from amyloid precursor protein (APP), which is a pathological hallmark in AD [37]. The cascade proteolytic cleavage of APP is led by BACE-1 and followed by γ-secretase, which leads to the production and release of amyloid-β peptide [38]. The initial cleavage of APP by BACE-1 yields a membrane-bound C-terminal fragment, C-99, which is also a rate-limiting step in the formation of Aβ peptide. This membrane-bound fragment is further cleaved by γ-secretase to produce 39- to 43-amino acid Aβ peptide fragments [39].

BACE-1, a 501 amino acid long protease, was first cloned in 1999 by five research groups simultaneously [40]. Its gene is located on chromosome 1. BACE-1 predominantly cleaves the β site of APP at M596-D597, and also at Y606-E607, but to a lesser extent. It operates mainly in an acidic environment. The BACE-1 possesses hallmark features of eukaryotic aspartic proteases of the pepsin family with two active site motifs, i.e., asp-thr-gly-ser (residues 93-96) and asp-ser-gly-thr (residues 289-292) [41]. The fact that the protease accounted for most of the β-secretase activity in the brain, was supported by subsequent demonstrations in BACE-1 knockout mice that did not produce Aβ peptides in the brain [42].

The X-ray co-crystal structure of BACE-1 with peptidomimetic inhibitors showed important interactions and was a key step in the further development of BACE-1 inhibitors [43]. The R289 and hydrophobic pockets formed by the active site played an important role in substrate binding [44]. Efforts are being made to develop compounds that specifically inhibit BACE-1, in view of its strong *in-vivo* and *in-vitro* validation as a

major β-secretase enzyme in the brain. Many drug candidates reached the clinical trials but failed mainly due to safety concerns **(Table 3.1)**. The other factor of failure may be the very late initiation of the treatment to make any impact on Aβ production [45]. Further, the size of BACE-1 active site (twenty-eight amino acids) is also relatively large and having small molecules to occupy such an active site is a challenge [46].

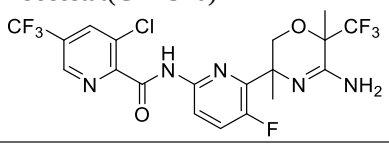
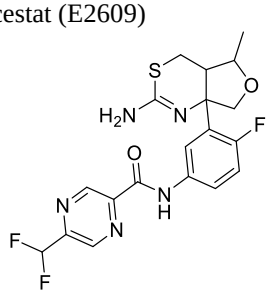
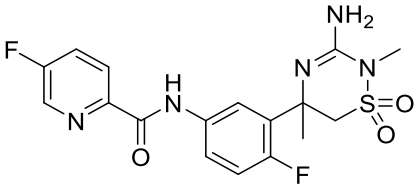
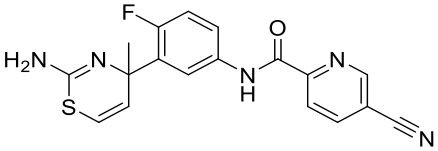
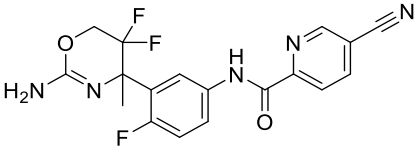
The blood-brain barrier permeability is another major challenge. Many of the BACE-1 inhibitors developed were prone to efflux by P-gp protein. Therefore, the process of drug entry into the brain is complex, even the drug has successfully crossed BBB [47]. Despite the challenges, various research groups have been able to design, develop, and synthesize selective, potent, and bioavailable inhibitors.

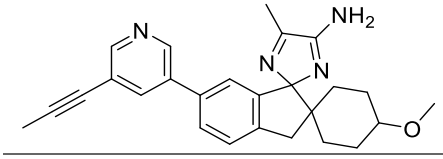
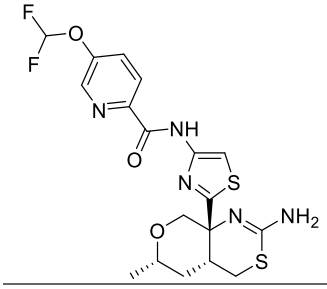
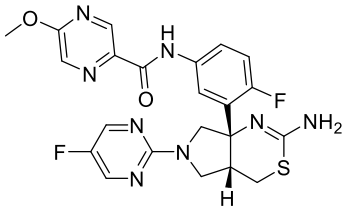
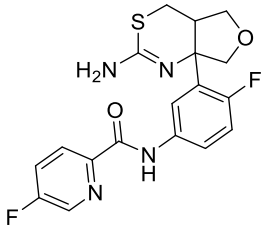
Several efforts were made to develop predictive models to classify BACE-1 inhibitors. Ponzoni *et al*. used a chemically diverse dataset with 215 molecules and calculated chemical descriptors to classify BACE-1 inhibitors. The best result was obtained with random forest algorithm (85% accuracy) [48]. Huang *et al.*, performed 3D-QSAR of BACE-1 inhibitors based on topomer CoMFA on a dataset of 125 compounds. The correlation coefficient of the fitting modeling was 0.966 [49]. Subramaniam *et al.* published a QSAR model with molecular descriptors that encompass molecular fingerprint, 1-dimensional, 2-dimensional, constitutional, physicochemical descriptors, topological descriptors and 3-dimensional molecular fields. The group performed both qualitative classification and quantitative regression involving linear, nonlinear and deep neural network methods. The best accuracy (81% accuracy) was observed for the model created on deep neural network using Constitutional, physicochemical and topological descriptors [50]. In another study by Kumar *et al*., a PLS-regression-based 2D-QSAR model, from 98 diverse compounds having defined BACE-1 enzyme inhibitory activity, was used to explore the essential structural requirements or molecular properties for

enzyme inhibition. The correlation coefficient of best model was found to be 0.826. In the same study, the authors performed a qualitative study using 3D QSAR pharmacophore modeling. The test set accuracy was found to be 73.43% [51].

In the present study, we used machine learning classification algorithms to build models on a very large dataset to identify active BACE-1 inhibitors. These models would serve as practical screening tools for designing and virtual screening of molecules to identify BACE-1 inhibitors.

Table 3.1 Clinical drug candidates as BACE-1 inhibitors

Name	**Sponsor**	**Phase**	**Clinical Trial accession number**	**Status**
Umibecestat(CNP520) 	Amgen, Novartis	Phase 2/3	NCT03131453	Discontinued in July, 2019
Elenbecestat (E2609) 	Biogen, Eisai	Phase 3	NCT03036280	Discontinued in September, 2019
Verubecestat (MK-8931) 	Merck	Phase 3	NCT01953601	Discontinued in April, 2018
Atabecestat (JNJ-54861911) 	Janssen, Shionogi Pharma	Phase 3	NCT02569398	Discontinued in May, 2018
RG7129 (RO5508887) 	Roche	Phase 1	NCT01664143	Discontinued in October 2013

Lanabecestat (AZD3293) 	AstraZeneca Eli Lilly & Co.	Phase 3	NCT02783573	Discontinued in June, 2018
PF-06751979 	Pfizer	Phase 1	NCT03126721	Completed in April, 2017
LY3202626 	Eli Lilly & Co.	Phase 2	NCT02791191	Discontinued in October 2018
LY2886721 	Eli Lilly & Co.	Phase 2	NCT01561430	Discontinued in June, 2013
BI 1181181 (VTP 37948)	Boehringer Ingelheim, Vitae Pharmaceuticals	Phase 1	NCT02044406	Completed in September, 2014. Discontinued in 2015.

3.2 Computational Method

3.2.1 Deriving experimental data for BACE-1 inhibitors

Molecular structures of BACE-1 inhibitors and their IC_{50} values were obtained from Binding DB by using the following criteria: (i) data of human BACE-1 inhibition assay (ii) data of IC_{50} assay, and (iii) filtration of compounds with multiple entries or without assay data [52]. The process resulted in the identification of a total of 4960 compounds, and their IC_{50} values ranged from 0.011 to 50,000 nM. The compounds having activity within the range of 50-500 nM, based on the differences in *in-vitro* studies, were further

filtered to avoid any bias (N=1424). The compounds having IC_{50} values less than 50 nM were marked as active (designated as 1), and those with IC_{50} values more than 500 nM were marked as inactive (designated as 0). This resulted in 1527 active and 2009 inactive compounds.

3.2.2 Ligand pre-treatment and preparation

The obtained dataset was further checked for any duplicity. The data from Binding DB was imported into the RDkit Python module. The molecules were standardized using the SaitizeMol module of RDkit. The standardization process involved removals of salts, mixtures, metal ions and correction of the geometry of the chemical structure. The correct protonation states were assigned at pH 7.4 using Open Babel.

3.2.3 Molecular descriptor calculation and feature selection

Mordred 1.2 was used to calculate molecular descriptors for all the ligands. It calculated a total 1826 descriptors, out of which 213 were 3-dimensional. In the present study, we used only 2D descriptors for the model development to avoid any complexity from 3D geometry considerations. The calculated descriptors were normalized using Sci-kit Learn Standard Scaler. Pearson correlation coefficients between the descriptors and log-transformed IC_{50} values were calculated to identify descriptors related to the activity. All the properties with a correlation coefficient of more than 0.25 and less than -0.25 were selected, resulting in 215 descriptors.

A correlation matrix among the descriptors was obtained to remove highly co-related features. One of the two highly co-related features i.e., Pearson correlation coefficient greater than 0.8 and lesser than -0.8 was also removed. This resulted in the identification of 43 features. The feature selection was based on the backward feature elimination using a linear regression model and the calculation of the P-value. Further, the threshold value was set to 0.05 [53]. Finally, there were 22 features left that were used for building

models, apart from molecular descriptors and fingerprint descriptors, viz. MACCS, PubChem fingerprint, and Klekota–Roth fingerprint (KRFP) were also calculated using the CDK descriptor calculator. All the features with zero variance were removed **(Table S2 of appendix**).

3.2.4 BACE-1 inhibitor training and test data sets

BACE-1 inhibitor dataset was split into training (70%), validation (15%) and test (15%) sets, by stratified *train_test_split* of python module *scikit learn* using *random state* as 1 into 2476 compounds in the training set and 530 compounds each in validation and test sets. Principal component analysis (PCA) was performed on 2D descriptors data to examine the chemical space and ensure uniform data distribution across training, validation and test sets. The number of components was kept to three (n_component=3).

3.2.5 Machine learning classification algorithms

Five machine learning algorithms viz. Naïve Bayesian (NB), Random Forest (RF), support vector machine (SVM), gradient boosting machine (XGB) and k- nearest neighbors (k-NN) were used to obtain classification models. All the machine learning classification algorithms were employed using Python library *scikit learn*-0.22.1 on Python 3.7. Grid search was performed to identify the optimal combination of parameters using *GridsearchCV* of Scikit Learn. Five-fold cross-validation (CV) accuracy was compared to determine the optimal combination of parameters.

3.2.5.1 Naïve Bayesian classifier

Gaussian type NB and Bernoulli type NB classifiers were used for Mordred 1.2 and fingerprint-based descriptors, respectively. The former is preferably used for continuous type of data, while the latter is used for Boolean or binary features. A grid search was performed to obtain the hyperparameter of the model. Gaussian type NB Models were

built for different values of *var_smoothing* ranging from 10^{-3} to 10^{-10} with a stepwise increment of 10^{-1}.

3.2.5.2 k-Nearest neighbours

Grid search was performed by using different combinations of parameters to optimize the performance of k-NN models. The parameter included:

- Number of neighbours- All the odd numbers between 3 to 20
- Distance metric- Euclidean, Manhattan and Minkowski distance

Thus, 27 models were built for each set of descriptors, and the best model was selected based on 5-fold CV accuracy.

3.2.5.3 Support vector classifier

Grid search was performed by using different combinations of parameters to optimize the performance of k-NN models. The parameters included:

- Regularization parameters (C)- 0.1, 1.0, 10, 100 and 1000
- Kernel coefficient (gamma)- 0.0001, 0.001, 0.01, 0.1 and 1.0
- Kernel- Poly, rbf and sigmoid

The combinations of the above parameters resulted in 75 models for each set of descriptors. The best model was selected on the basis of 5-fold CV accuracy.

3.2.5.4 Random forest classifier

Random forest (RF) is determined by three parameters, i.e., n estimator (number of trees in forest), max depth (maximum depth of the tree) and min sample split (minimum number of samples required to split an internal node). A grid search was performed to obtain parameters for maximum 5-fold CV accuracy by using the following parameters.

- n estimators 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000
- Max depth ranges from 5 to 50 with a stepwise increment of 5
- Min sample split ranges from 2 to 10

3.2.5.5 Gradient boosting algorithm

The parameters of Gradient boosting algorithm (XGB) were similar to RF along with loss function. A grid search was performed to obtain hypertuned parameters and best model was selected on the basis of accuracy score. Since the XGB is computationally expensive, a small grid search was performed using the following parameters.

- n estimators 500, 600, 700, 800, 900 and 1000
- Max depth ranges from 5 to 10
- Min sample split ranges from 2 to 10

3.2.6 Validation of the performance of the models

The performance of the models was evaluated by using the following metrics: Accuracy (A), Cross validated accuracies (5-CV, 10-CV, and leave-one-out), true positive (TP), false positive (FP), true negative (TN), false positive (FP), false negative (FN), precision (PE), recall (RC), and F1 score.

$$A = \frac{TP + TN}{TP + FN + TN + FP} X\ 100$$

$$PE = \frac{TP}{TP + FP}$$

$$RC = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * PE * RC}{PE + RC}$$

3.2.7 Comparison of the performance of models created on 2D descriptors and fingerprints with 3D descriptors

3.2.7.1 Compound alignment and calculation of 3D descriptors

All the molecules were aligned to the co-crystallized ligand of BACE-1 in their bioactive conformation inside the binding site retrieved from the protein data bank (https://www.rcsb.org/). The bound conformation of compound 5HA (PDB ID-3tpp) had

a relatively better resolution (1.60 Å) and was used as a template for the alignment of molecular structures. The alignment was performed using the GetCrippenO3A module of RDkit. The number of conformers was set to 100, and the best score index for each molecule was selected from the multiple conformers. The 3D descriptors of the selected conformers were calculated using RDkit.

3.2.7.2 Evaluation of the performance of models build on 3D descriptors.

The models were trained and evaluated by using the methods discussed above in sections **3.2.5** and **3.2.6,** respectively.

3.3 Results and Discussion

3.3.1 Chemical space exploration and dataset distribution of BACE-1 inhibitors

The complete BACE-1 inhibitor dataset is represented in the figure with a pair plot of molecular property descriptors viz. Molecular weight, SlogP, TPSA, and number of rotatable bonds**.** Both active and inactive compounds were found to be well distributed in figure. There was no bias in the selection of active and inactive compounds for preparing the data sets, as was evident from the overlapping **(Figure 3.1)**.

3.3.2 Feature selection of calculated descriptors

The feature selection process contributed more to the prediction output. It was an important step to increase and improve the accuracy of the classification model. Mordred 1.2 calculated total 1826 molecular descriptors and building a classification model on such a large number of features would decrease the performance of a model. Pearson co-relation coefficient is a measure of relationship between the two variables. Since, the highly co-related features are not meaningful therefore, any one of features was selected. The co-relation of the activities and selected molecular properties are listed in table **(Figure 3.2).** The corresponding p-values indicated that all the co-relations were statistically significant and, could be used for building models in machine learning [54].

The final number of features used for building model has been summarised in table **(Table S1 of appendix)**.

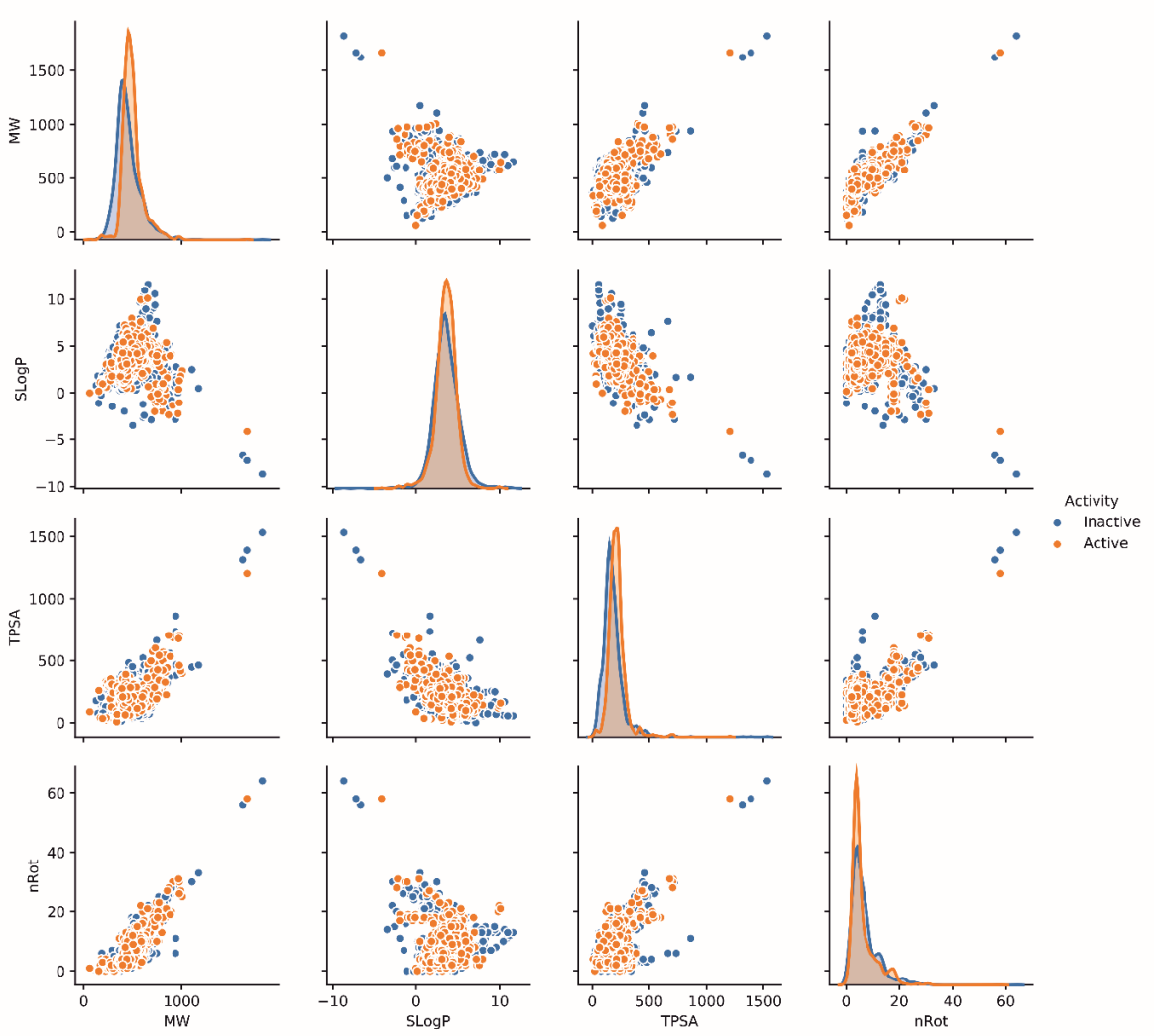


Figure 3.1 Pair plot of BACE-1 inhibitor dataset. The diagonal represents the frequency of distribution. (MW=Molecular weight, TPSA=Total Polar Surface Area, nRot=Number of rotatable bond)

3.3.3 BACE-1 inhibitor training, validation and test sets

The test-train split of Python toolkit, *scikit learn*, divided actives and in actives into test, validation and training sets in equal proportions **(Figure 3.3A)**. In an ideal splitting, the chemical space of training set should overlap to that of test set and validation set. The 3D scatter plot of PC1, PC2 and PC3 represented on X, Y and Z-axis respectively is depicted in figure **(Figure 3.3B)**. Three principal components obtained after performing PCA were

able to explain more than 50 % variance of the dataset. The scatter plot shows homogenous distribution (overlap) for the validation and test sets around the training set compounds **(Figure 3.3B)**.

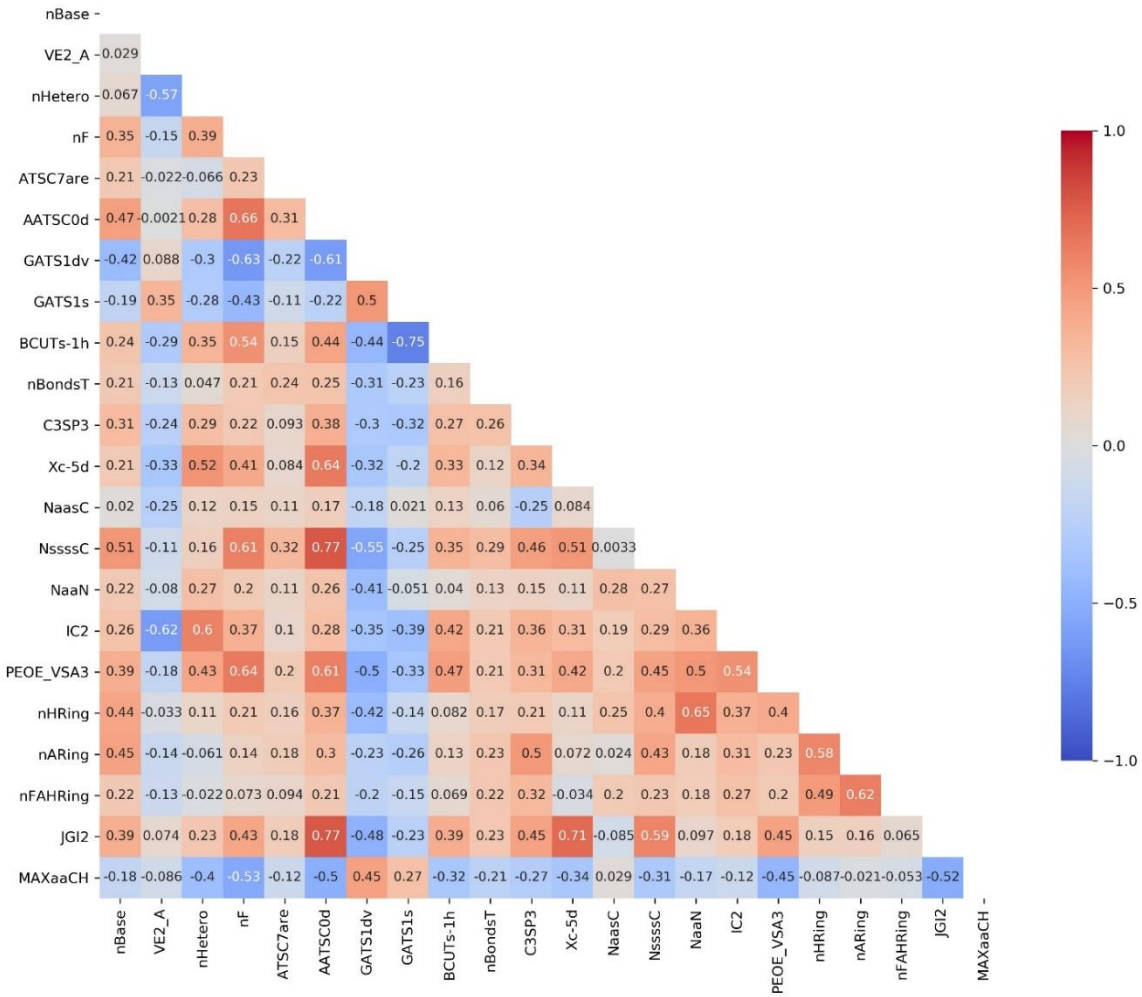


Figure 3.2 Pearson correlation coefficient of calculated descriptors after performing feature selection

Table 3.2 Pearson correlation coefficient between selected molecular properties and BACE-1 inhibition.

Descriptor code	Description	Pearson correlation coefficient*
nBase	Basic group count	-0.33
VE2_A	VE2 of adjacency matrix	0.30
nHetero	Number of hetero atoms	-0.27
nF	Number of F atoms	-0.33
ATSC7are	Centered moreau-broto autocorrelation of lag 7 weighted by allred-rocow EN	-0.25
AATSC0d	Averaged and centered moreau-broto autocorrelation of lag 0 weighted by sigma electrons	-0.37
GATS1dv	Geary coefficient of lag 1 weighted by valence electrons	0.30
GATS1s	Geary coefficient of lag 1 weighted by intrinsic state	0.26
BCUTs-1h	First highest eigenvalue of Burden matrix weighted by intrinsic state	-0.31
nBondsT	Number of triple bonds in non-kekulized structure	-0.29
C3SP3	SP3 carbon bound to 3 other carbons	-0.25
Xc-5d	5-ordered Chi cluster weighted by sigma electrons	-0.33
NaasC	Number of aasC	-0.25
NssssC	Number of ssssC	-0.40
NaaN	umber of aaN	-0.26
IC2	2-ordered neighborhood information content	-0.43
PEOE_VSA3	MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20)	-0.40
nHRing	Hetero ring count	-0.35
nARing	aromatic ring count	-0.27
nFAHRing	aromatic fused hetero ring count	-0.25
JGI2	2-ordered mean topological charge	0.33
MAXaaCH	max of aaCH	0.25

*P<0.05

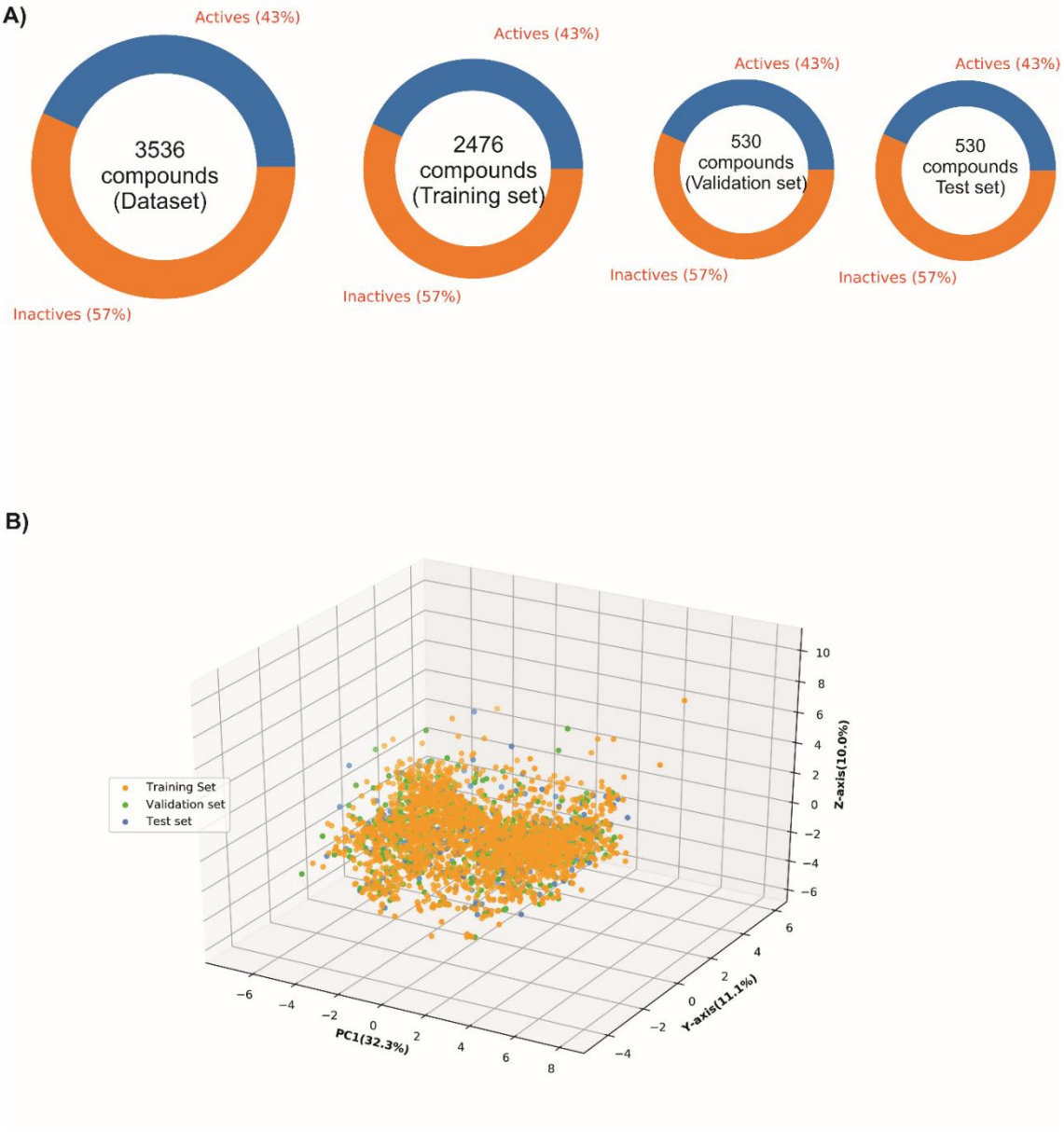


Figure 3.3 Data distribution for (A) complete BACE-1 dataset and (B) Chemical space of training, validation and test set as represented by PCA plot

3.3.4 Machine learning classification algorithms

3.3.4.1 Naïve Bayesian classifier

Naïve Bayesian classifier is based on Bayesian theorem, where it searches through each feature having ability to separate in an unbiased way. A molecule is represented in the form of 1-D array of molecular descriptors and structural fragments. The probability of a

molecule belonging to active or inactive class can be predicted by combining the-calculated probabilities of all descriptors using the following Bayesian theorem:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Where, A represents the activity class of a compound and B represents a molecular property; P(A) is prior probability of compound activity; P(B) is probability that a property is owned by a compound in sample data; P(B|A) is conditional probability of likelihood of the event B, when A has already occurred. On performing grid search, it was observed that the parameter *var_scaling* had no impact on the performance of the model. The four NB models, built on dataset from Pubchem fingerprints i.e., NB-4, showed better performance than the rest. The accuracies of training and validation sets were 74.82% and 74.51%. The result has been summarised in table **(Table 3.3).**

Table 3.3 Parameters and performance of classification model built by NB classifier

Model number	**Type**	**Var smoothing**	**Descriptors**	**Training set**			**Validation set**			
				Q	5-CV	10-CV	Q	PR	RC	F1
NB-1	Gaussian	10^{-9}	Mordred	74.26	75.47	75.25	75.14	0.70	0.72	0.71
NB-2	Bernoulli	10^{-9}	KRFP	73.81	72.72	74.33	74.38	0.68	0.77	0.72
NB-3	Bernoulli	10^{-9}	MACCS	70.78	70.62	70.54	70.62	0.64	0.73	0.68
NB-4	Bernoulli	10^{-9}	Pubchem	74.82	74.62	76.79	74.51	0.70	0.75	0.72

Q-Accuracy, CV-cross validation, PR- precision, RC- Recall

3.3.4.2 k-Nearest neighbours

The k-NN algorithm is one of the simplest machine learning algorithms. The data is represented in a high-dimensional feature space and labels from the closest nodes are assigned to query. Value k specifies the number of closest neighbours and is kept variable; too high or too low value may result in overfitting [55]. The optimum values of parameters have been summarized in **Table 3.4.** The dataset of KRFP descriptors i.e.,

KNN-2, showed best performance in kNN models, with training and validation set accuracies of 91.03% and 88.51%, respectively. The results of the models are summarized in **Table 3.4.**

Table 3.4 Parameters and performance of classification model built by kNN classifier

Model number	**N-neighbors**	**Distance metrics**	**Descriptors**	**Training set**			**Validation set**			
				Q	5-CV	10-CV	Q	PR	RC	F1
KNN-1	5	Manhattan	Mordred	90.10	84.24	84.44	85.49	0.82	0.83	0.83
KNN-2	5	Manhattan	KRFP	91.03	86.78	87.31	88.51	0.87	0.86	0.86
KNN-3	7	Manhattan	MACCS	89.05	83.79	84.24	85.87	0.81	0.82	0.82
KNN-4	5	Manhattan	Pubchem	89.93	84.84	84.52	83.61	0.80	0.81	0.81

3.3.4.3 Support vector machines model

SVM model solves the classification problem by using a linear or non-linear kernel function to map data into high-dimensional space by finding an optimally separating hyperplane [56]. It can be determined by three parameters i.e., a penalty parameter (C), a kernel function and a kernel coefficient.

The performance of SVM model, built on Mordred 1.2 descriptors, was not as good as the corresponding models on fingerprint descriptors. The model raised on KRFP fingerprints SVM-2, showed best validation set accuracy of 91.33% having precision, recall and F1 scores of 0.91, 0.88 and 0.89, respectively. The other models built on Modred descriptors, MACCS and Pubchem fingerprints showed comparable performances. The parameters and performance of model having maximum accuracy for each dataset is included in **Table 3.5.**

Table 3.5 Parameters and performance of classification model built by SVM classifier

Model number	Parameters			Descriptors	Training set			Validation set			
	C	gamma	kernel		Q	5-CV	10-CV	Q	PR	RC	F1
SVM-1	10	0.1	Rbf	Mordred	96.68	86.50	87.11	88.88	0.87	0.86	0.86
SVM-2	100	0.1	Rbf	KRFP	95.95	89.13	89.33	91.33	0.91	0.88	0.89
SVM-3	1	0.1	Rbf	MACCS	93.73	86.90	87.02	90.01	0.91	0.85	0.88
SVM-4	10	0.01	Rbf	Pubchem	95.11	87.03	87.03	88.13	0.86	0.86	0.86

3.3.4.4 Random-forest classifier

Random forest is an ensemble learning method in which multiple decision trees are built on the training dataset. It is not a linear modelling algorithm. On each split node of tree, different subset of descriptor is evaluated for best split. Thus, a forest of such decision tree is created and the final prediction for a new sample is the average prediction of all the trees in a forest [57]. The parameters and performance of the model having maximum accuracy for each dataset is summarised in **Table 3.6.**

Table 3.6 Parameters and performance of classification model built by RF classifier

Model number	Parameters			Descriptors	Training set			Validation set			
	N estimator	Max depth	Min samples split		Q	5-CV	10-CV	Q	PR	RC	F1
RF-1	500	10	4	Mordred	94.98	85.29	85.73	86.62	0.84	0.84	0.84
RF-2	700	20	5	KRFP	95.67	88.00	88.44	89.83	0.91	0.84	0.87
RF-3	500	15	5	MACCS	94.86	86.94	87.39	89.26	0.89	0.84	0.87
RF-4	800	30	5	Pubchem	96.04	86.86	87.19	87.94	0.89	0.81	0.85

The performance of the random forest models, built on the Mordred dataset, was not as good as the model built on fingerprint descriptors. For the four random forest models,

RF-2 created on KRFP dataset showed maximum test set accuracy, precision, recall and F1 scores of 90.96%, 0.88, 0.90 and 0.89, respectively.

3.3.4.5 Gradient boosting algorithm

The ensemble method combined prediction from multiple decision trees to generate final predictions. In gradient boosting (XGB) machines, the learning procedure fits new models after each iteration, to provide a more accurate estimate of the response variable. The key idea behind the algorithm is to develop new base-learners, to be utterly correlated with the negative gradient of the loss function and to be associated with the whole ensemble [58]. The performance of model GB-4, built on the Pubchem dataset showed maximum validation set accuracy of 89.45%. The parameters and performance of the best model, corresponding to every dataset has been represented in **Table 3.7**.

Table 3.7 Parameters and performance of classification model built by XGB classifier

Model	**Parameters**			**Descriptors**	**Training set**			**Validation set**			
number	N estimator	Max depth	Min samples split		Q	5-CV	10-CV	Q	PR	RC	F1
GB-1	500	3	7	Mordred	98.98	85.74	85.73	86.25	0.85	0.81	0.85
GB-2	500	4	10	KRFP	97.85	85.73	85.73	89.07	0.89	0.84	0.86
GB-3	700	4	10	MACCS	93.17	86.22	86.26	87.75	0.88	0.82	0.85
GB-4	500	3	8	Pubchem	96.72	87.72	87.28	89.45	0.88	0.86	0.87

3.3.5 Prediction of the test set data

A test set was used to validate the performance of all the 20 models. The performance of models on test set is summarised in Table 3.8**.** All the 20 classification models showed overall accuracy between 71.69 - 89.62% on the test set. The model SVM2, built on KRFP fingerprint by SVM classifier, showed maximum accuracy, precision and recall score (0.87). The models showed precision score in the range of 0.65 – 0.90. Further, model RF2, built on KRFP fingerprint by RF classifier also showed precision score of 0.90.

3.3.6 Comparison of the performance of models build on 2D descriptors and fingerprints with 3D descriptors

The AUC of ROC plot for the best model of all the classification algorithms has been summarized in table (**Table S3 of appendix** and **Figure S1 of appendix)**. The best AUC of ROC was shown by the model build on RF classifier, RF-5 having score of 0.93, which was less than the models built on 2D descriptors and fingerprint datasets.

3.3.7 Structural diversity of BACE-1 inhibitors

The dataset of BACE-1 inhibitors consisted of diversity of structures such as benzopyran, imidazole, thiazolidines, pyrimidine etc. To evaluate the structural diversity between every two ligand pairs, we calculated Tanimoto coefficient-based distance matrix from their MACCS fingerprints. In general, two inhibitors were considered similar, if they have a Tanimoto coefficient of more than 0.70. It was observed that only 6.25% of the ligands had Tanimoto coefficient greater than 0.70, indicating that the dataset is structurally diverse. The frequency distribution of Tanimoto coefficient of each ligand pair is depicted in **Figure 3.4.**

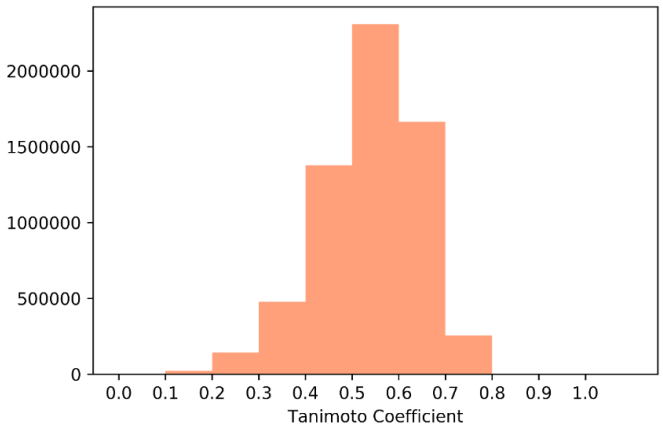


Figure 3.4 Frequency distribution histogram of Tanimoto similarity coefficient for pair of BACE-1 inhibitor based on MACCS fingerprints.

Table 3.8 Performance of 20 classification models on test set.

Model	TN	FP	FN	TP	Q	PR	RC	F1	AUC of ROC
NB-1	227	74	59	170	74.90	0.69	0.74	0.71	0.81
NB-2	219	82	47	182	75.66	0.68	0.79	0.73	0.85
NB-3	207	94	56	173	71.69	0.65	0.75	0.69	0.81
NB-4	231	70	52	177	76.98	0.71	0.77	0.74	0.82
KNN-1	263	38	31	198	86.98	0.83	0.86	0.85	0.92
KNN-2	269	32	36	193	87.16	0.85	0.84	0.85	0.93
KNN-3	256	45	31	198	85.66	0.81	0.86	0.83	0.93
KNN-4	264	37	42	187	85.09	0.83	0.81	0.82	0.92
SVM-1	271	30	37	192	87.35	0.86	0.83	0.85	0.91
SVM-2	280	21	34	195	89.62	0.90	0.85	0.87	0.95
SVM-3	276	25	37	192	88.30	0.88	0.83	0.86	0.93
SVM-4	274	27	31	198	89.05	0.88	0.86	0.87	0.94
RF-1	266	35	33	196	87.16	0.84	0.85	0.85	0.94
RF-2	281	20	37	192	89.24	0.90	0.83	0.87	0.94
RF-3	276	25	39	190	87.92	0.88	0.82	0.85	0.95
RF-4	279	22	34	195	89.43	0.89	0.85	0.87	0.95
GB-1	266	35	41	188	85.66	0.84	0.82	0.83	0.93
GB-2	275	26	35	194	88.49	0.88	0.84	0.86	0.94
GB-3	276	25	38	191	88.11	0.88	0.83	0.85	0.94
GB-4	275	26	30	199	89.43	0.88	0.86	0.87	0.94

3.3.8 k-Means clustering of BACE-1 inhibitors

In order to explore the structural features of 3536 inhibitors in the dataset, clustering was performed on the MACCS fingerprints data. The k-Means clustering is a method to classify data into k-groups, by minimizing distances within-group to the centroid. Further, most commonly used Euclidean distance metric was employed in the evaluation. The optimum value of k was determined by using the elbow method and silhouette score. For MACCS data with 11 clusters, the silhouette score was found to be only 0.14. Therefore, dimensionality reduction was carried out by using t-distributed stochastic neighbours embedding (t-SNE). The data was reduced to two-dimensions with the help of t-SNE, and was designated as t-SNE1 and t-SNE2. The silhouette score was found to be 0.52, after clustering on this dataset. The clusters and their centroids are included in **Figure 3.5.** Pairwise distances were calculated by using Euclidean distance metric, to visualize the structures closest to the centroid of cluster. The structure of compounds closest to the

centroids and activity class of all the 11 clusters have been represented in **Figure 3.6.** Further, subset 0 consisted of 187 compounds, with 63.64 % actives. The cluster mainly consisted of compounds with napthyridin and thiazine moieties. The subset 1 consisted of 322 compounds, with 31.06% of actives haveing oxopyrrolidin, pyrrolidine and sulphone moieties. The subset 2 had 453 compounds with similar numbers of active and inactive compounds. This cluster had a few azaspiro and pyridyl containing compounds too. Subset 3 consisted of compounds containing napthaquinone ring. This subset of 253 compounds, had only 8.3 % of actives. The compounds in cluster 4 had groups such as azabicyclo and fluorophenyl. This cluster had 424 compounds in which 59.67 % of compounds were active. In cluster 5, there were 209 compounds with 92.82 % actives. This cluster had the maximum percentage of active compounds among the clusters. Most of the compounds in this cluster possessed nitrile group, phenyl pyrrolidine ring and aza-

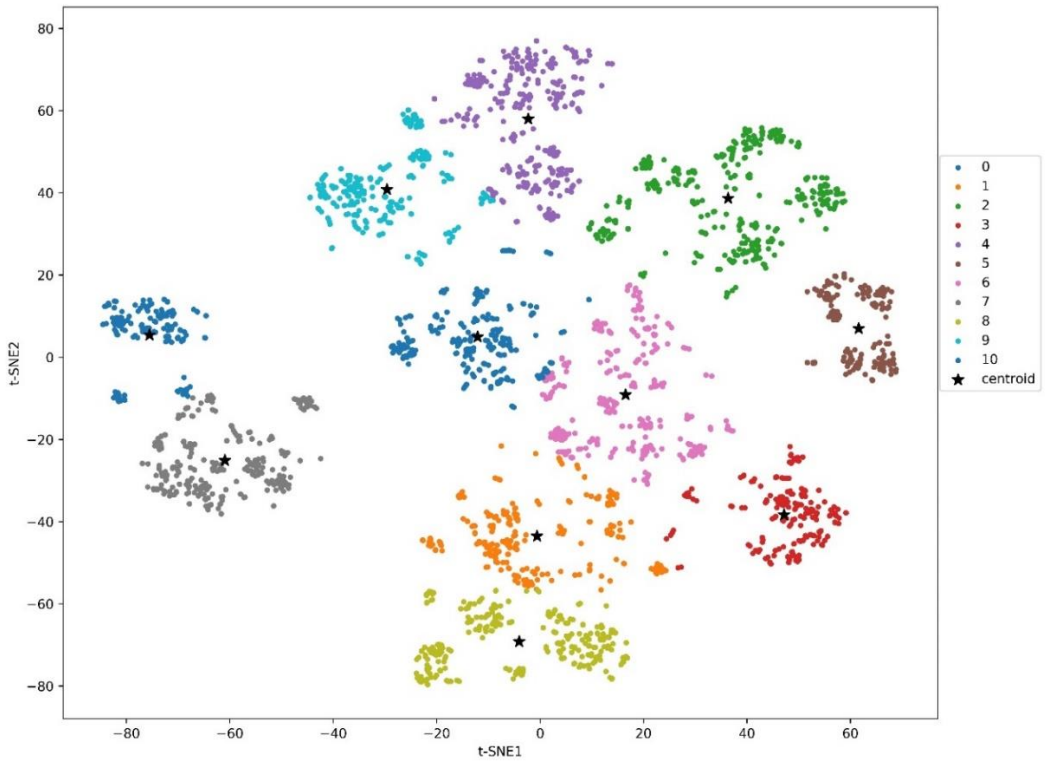


Figure 3.5 Clustering of compounds in 11-subsets along with their centroids (represented by ★). t-SNE1 and t-SNE2 are the two dimensions reduced from 166 dimensions of MACCS fingerprints.

bicyclo ring. The cluster 6 had 390 compounds with only 15.38 % of active compounds. The compounds consisted of groups such as aminopyridine, pyrrole ring and N-cycloalkylbenzamide. Cluster 7 was composed of compounds containing piperazinyl, isoxazolyl and sulfonyl groups. The cluster had 45.27 % of active compounds out of total 349 compounds. In cluster 8, there were 353 compounds in which 41.93 % of compounds were active. The compounds had fluorophenyl, hydroxy and acetamide groups. Cluster 9 had 287 compounds and 51.22 % of them were active. The percentage of active and inactive compounds in this subset was almost similar. The compounds in this subset consisted of oxazine ring and hydroxypicolinamide group. Finally, cluster 10 had 310 compounds in which 33.87% of compounds were active. The compounds in this subset had thiophene, imidazolyl and fluoropyridyl groups.

3.3.9 Comparison of performance of classification algorithms and descriptors

3.3.9.1 Molecular descriptors

Performance of the models built on molecular descriptor dataset was evaluated by using area under curve (AUC) of the receiver operating characteristic (ROC) plot. The values were found to be 0.81, 0.92, 0.91, 0.94 and 0.93 for NB, kNN, SVC, RF and GB algorithms, respectively **(Figure 3.7A)**. The performance of random forest classifier was observed to be best considering the AUC and F1 score. The class specific feature importance and overall feature importance of top 10 properties have been summarized in **Table 3.9**.

3.3.9.2 KRFP fingerprints

The AUC of ROC plots for machine learning algorithms built on KRFP fingerprint dataset revealed that the performance of NB classifier (AUC=0.85) was not as good as other classification algorithms i.e., kNN, SVC, RF and GB. The AUC of ROC for kNN. Further, SVC, RF and GB was found to be 0.93, 0.95, 0.95 and 0.94, respectively **(Table**

3.7 B). The F1-score of RF classifier was also finest among the other classification algorithms. The top ten good and bad fingerprints obtained from random forest classifiers are included in **Figure 3.8.**

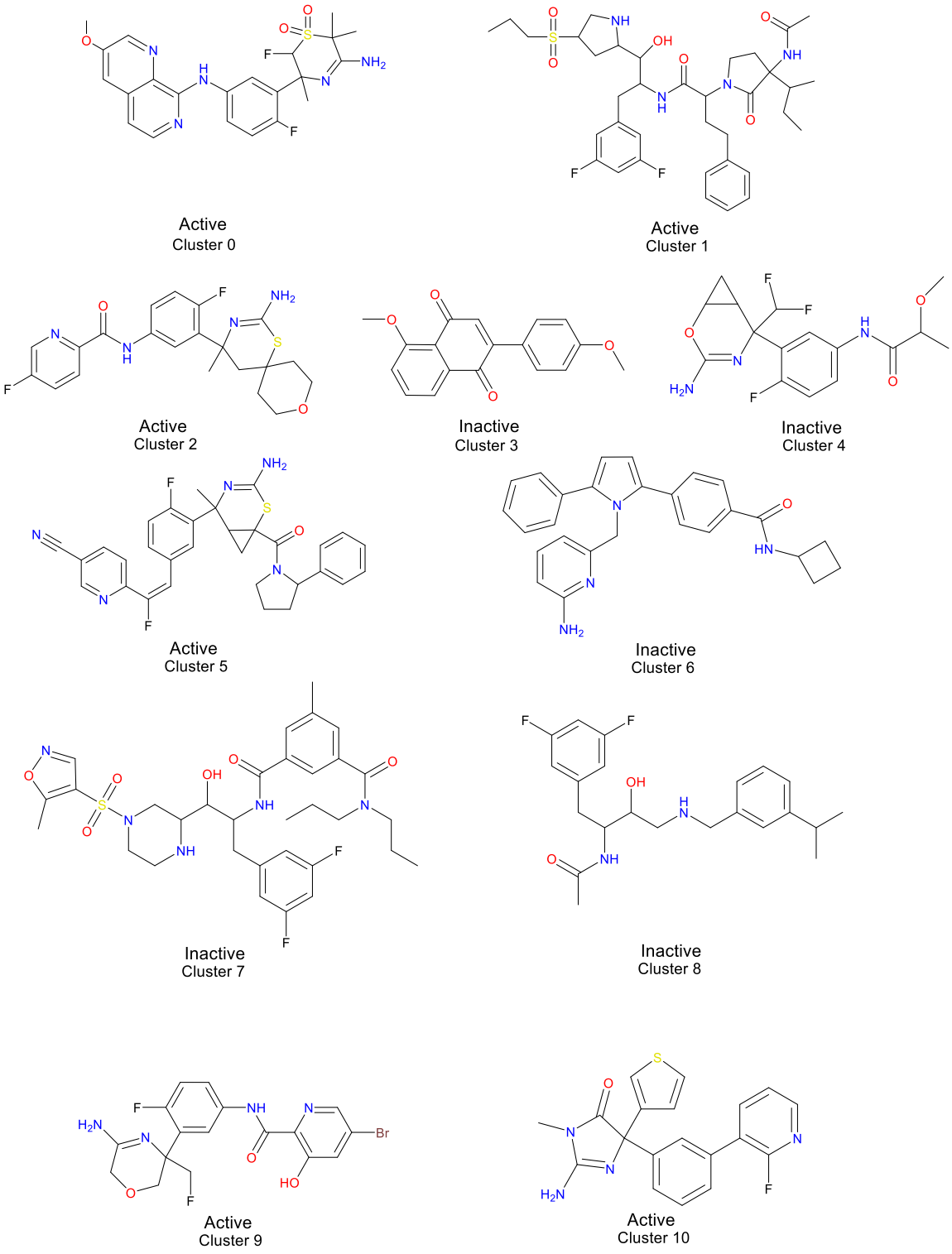


Figure 3.6 Central compounds and their corresponding activities in the eleven subsets

3.3.9.3 Pubchem fingerprints

In case of Pubchem fingerprints, the performance of kNN, SVC, RF and GB classifiers were almost similar, but the performance of NB classifier was less than the remaining

classifiers. The AUC of ROC was found to be 0.82, 0.92, 0.94,0.95 and 0.94 for NB, kNN, SVC, RF and GB, respectively **(Table 3.7)**. On comparing the other accuracy metrics, RF classifier was found to be the best. The top ten good and bad fragments are represented in **Figure 3.9.**

3.3.9.4 MACCS fingerprint

In case of MACCS fingerprint, NB classifier showed poor result (AUC of ROC=0.81), but rest of the classifiers showed better result. AUC of ROC was found to be 0.93 for kNN, 0.94 for SVC, 0.95 for RF and 0.94 for XGB algorithm **(Figure 3.7 D)**. The top ten important features, along with their class specific feature importance were calculated. The fragments contributing to active and inactive BACE-1 inhibitors were identified. The active compounds were mainly found to contain a fluorine and an aromatic nitrogen atom **Table 3.10**.

Table 3.9 Top ten features from Mordred dataset with their feature importance values from RF classifier

S. No.	Descriptor code	Feature importance (Active class)	Feature importance (Inactive class	Feature importance (Overall)
1	BCUTs-1h	1.826	1.750	0.128
2	nHetero	0.295	0.382	0.104
3	nBase	0.525	0.581	0.089
4	GATS1dv	0.077	0.073	0.081
5	nF	0.080	0.007	0.081
6	ATSC7are	0.052	0.115	0.080
7	GATS1s	0.052	0.047	0.072
8	NaasC	0.138	0.059	0.067
9	Xc-5d	0.029	0.021	0.066
10	C3SP3	0.036	0.028	0.062

Table 3.10 Top ten features from MACCS dataset with their feature importance values from RF classifier

S.No.	MACCS Key	Description	Feature Importance (Active Class)	Feature importance (Inactive class)	Overall feature importance
1	MACCSFP42	F	0.03292	0.015881	0.041926
2	MACCSFP65	C%N	0.031471	0.018215	0.040214
3	MACCSFP112	AA(A)(A)A	0.030114	0.019467	0.032893
4	MACCSFP66	CC(C)(C)A	0.02617	0.015269	0.032021
5	MACCSFP144	Anot%A%Anot%A	0.017212	0.013341	0.022599
6	MACCSFP92	OC(N)C	0.019617	0.014014	0.022523
7	MACCSFP154	C=O	0.016109	0.013349	0.018649
8	MACCSFP19	7M RING	0.004582	0.001327	0.017893
9	MACCSFP134	X (HALOGEN)	0.014445	0.00942	0.017506
10	MACCSFP95	NAAO	0.012847	0.009494	0.015532

(F-fluorine, %- aromatic query bond, A- Any valid periodic table element symbol, =- double bond)

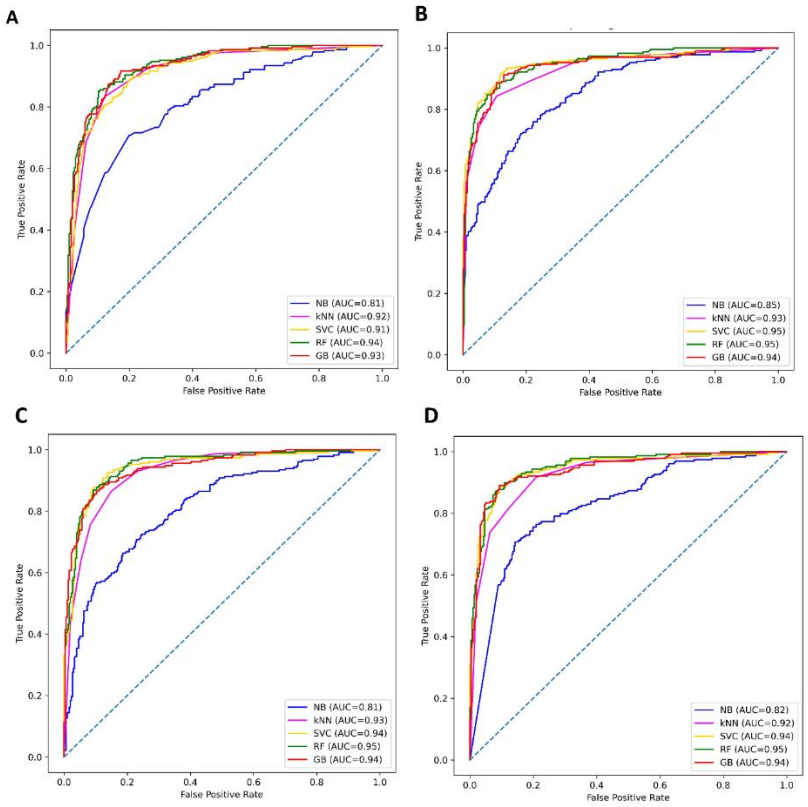


Figure 3.7 Receiver operating characteristics (ROC) curves of models build on (A) Mordred, (B) KRFP fingerprints, (C) MACCS fingerprints and (D) Pubchem fingerprints datasets

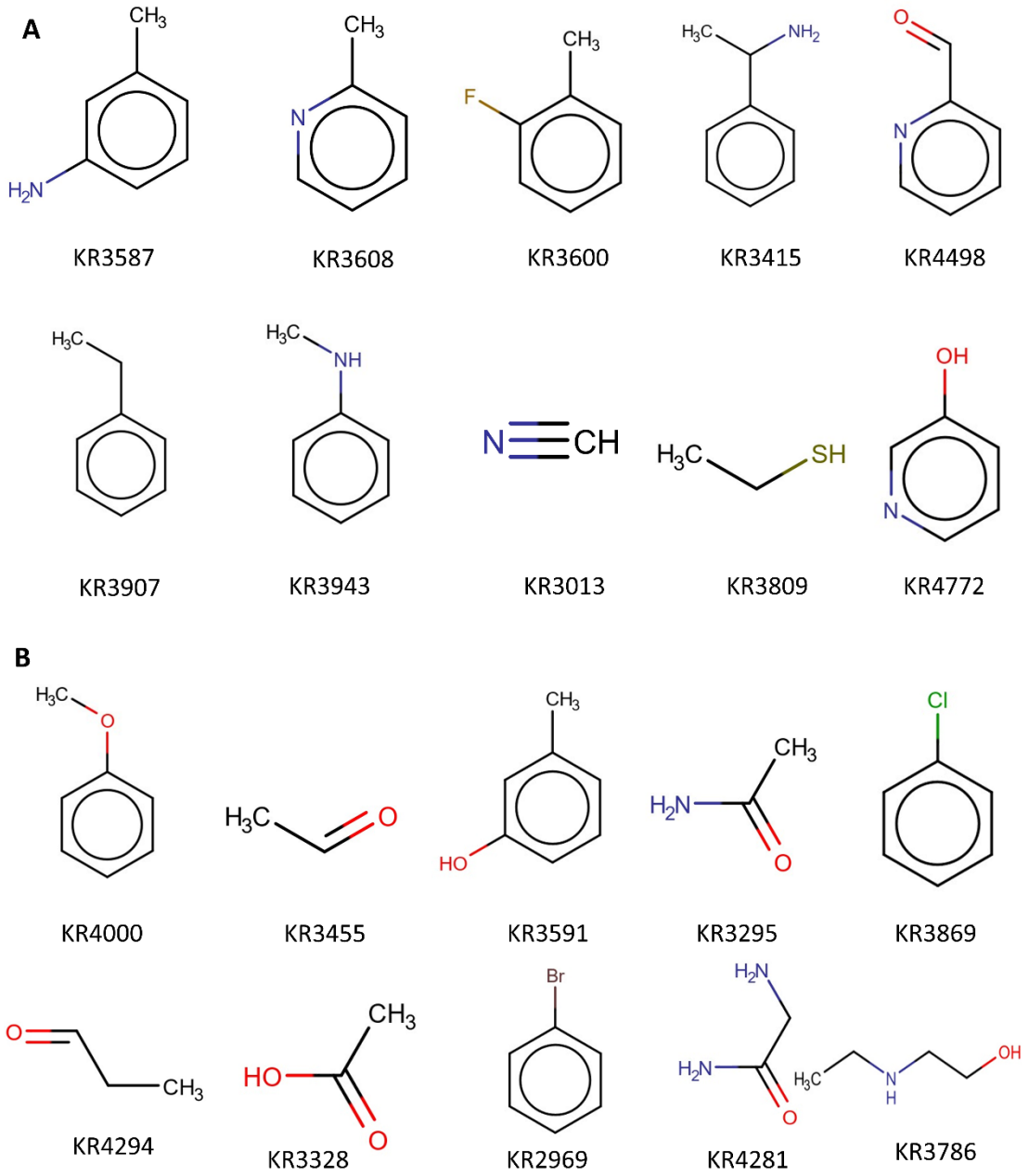


Figure 3.8 KRFP fingerprints (A) beneficial and (B) adversely affecting the BACE-1 inhibitory activity obtained from RF classifier

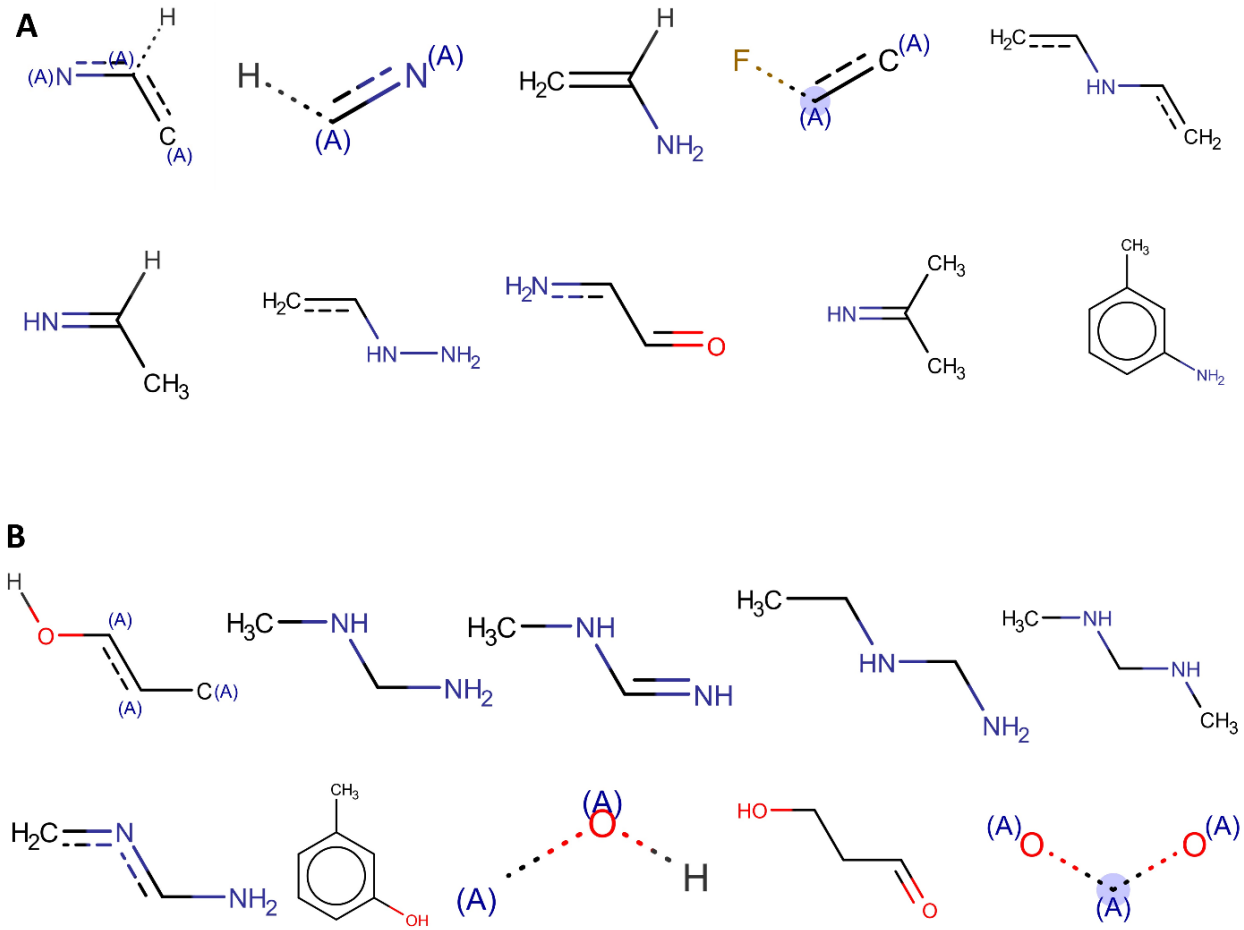


Figure 3.9 Pubchem fingerprints (A) beneficial and (B) adversely affecting the BACE-1 inhibitory activity obtained from RF classifier

3.3.10 Defining applicability domain of models

The applicability domain of a model specifies the limit within which the prediction is considered reliable. This restricts the applicability of a model to predict the compounds, which are similar to the training set used in model building. Similarity measurement was used to define the applicability domain. The applicability domain of the model was defined using the following equation:

$$AD = \gamma + \sigma Z$$

At first, the average pairwise Euclidean distance between all the training dataset based on Mordred descriptors was calculated. Then, the pairs with distances lower than the average were formulated and γ and σ were calculated as the average and standard deviation,

respectively. *Z*, an empirical cut-off value, was set to 0.5 [59]. The applicability domain of our model was found to be 9.74.

3.4 Conclusion

In the study, we developed twenty classification models, based on 3536 BACE-1 inhibitors. The molecular structures were quite diverse according to the Tanimoto coefficients. The dataset was divided into training, validation and test sets in the ratio of 70:15:15. Two types of descriptors i.e., molecular properties (Mordred 1.2) and chemical fingerprints (KRFP, Pubchem and MACCS) were calculated. Five machine learning algorithms were used to develop classification models viz. NB, SVM, kNN, RF and XGB. The performance of NB classifier was not up to the mark in comparison to the other classifiers, as was evident from the F1 score and AUC of ROC of all the classification algorithms.

The classification-based machine learning models developed in the study, showed better accuracy of 89.62% than the models developed by Ponzoni *et. al.* The accuracy of the best model in the earlier study, built by using RF algorithm, was 85% [59].The contribution of various fingerprints was analysed through the feature importance of *scikit learn*. The fragments such as aromatic amine, pyridine ring, fluoropyridine and nitrile groups were important for the active BACE-1 inhibitors. The model and the dataset used for its development can be accessed at https://github.com/ravisingh15/BACE-1_inhibitor. The broad chemical diversity of the compounds used in the dataset and the structural details would be certainly helpful in designing active BACE-1 inhibitors. The classification model developed could also be used as a virtual screening tool for the identification of BACE-1 inhibitors.