Chapter 4
Machine Learning based screening of *in-house* database to identify BACE-1 inhibitors

Summary:

The β-site APP cleaving enzyme-1 (BACE-1) is one of the key targets for novel drugs to treat Alzheimer's disease (AD). The BACE-1 plays a key role in the amyloidogenic process, leading to the production of amyloid-β (Aβ) plaques in the brain. In the present work, an ML model based on the sulfonamides dataset was developed. The best ML model was built using the XGBoost algorithm on PubChem fingerprints. The model had an accuracy, precision, recall and F1 score of 0.89, 0.88, 0.99 and 0.93, respectively, on the validation set. The same model was used to screen the database of previously synthesized and reported in-house compounds. The screening resulted in the identification of two hits, i.e., compound 28 and compound 37. Both the compounds were screened for their BACE-1 inhibitor activity. The IC_{50} value of compound 28 was found to be 0.431 ± 0.006 μM, and compound 37 showed an IC_{50} value of 0.272 ± 0.019 μM. The docking study revealed that compound 37 also showed interactions with the catalytic dyad of BACE-1, i.e., Asp32 and Asp228.

4 Machine Learning based screening of in-house database to identify BACE-1 inhibitors

4.1 Introduction

BACE-1 is an aspartyl protease of the pepsin family, discovered in 1999. BACE-1initiates the production of Aβ, which represents the rate-limiting enzyme in the amyloidogenic pathway. BACE-1 cleaves the Aβ precursor protein (APP) to its membrane-bound C-terminus fragment C99 (CTF) and soluble APPβ fragment. The BACE-1 is essential for the generation of all monomeric units of Aβ, including $A\beta_{42}$, which plays a crucial role in the pathogenesis of AD. The concentrations and activity rates of BACE-1 are actively increased in AD brains and body fluids. Therefore, BACE-1 emerged as a primary drug target for decreasing the production of Aβ in the AD brain [60]. BACE-1 is a type-1 transmembrane protein that is different from other peptidases of the same family. The catalytic domains of BACE have two significant motifs of the sequence DTGS and DSGT that together forms the active site of the enzyme [61]. BACE-1 consists of metal binding sites; it has a copper-binding site in its cytosolic domain [62]. The crystal structure of BACE-1 reveals that its proteolytic pocket is relatively large and is less hydrophobic; therefore, it becomes challenging for developing small-molecule inhibitors using high-throughput virtual screening [63, 64].

4.1.1 Sulphonamides as BACE-1 Inhibitors in Human Clinical Trials

Non-peptide BACE-1 inhibitors such as sulphonamides had some success in pre-clinical studies as some of the drugs were seen in various phases of clinical trials as well.

BACE-1 inhibitor MK-8931 (Verubecestat) entered the Phase-3 of clinical trial conducted in mild to moderate Alzheimer's patients and was terminated as it failed to show efficacy over the placebo. MK-8931 reduced the levels of $A\beta_{40}$ in healthy

participants, whereas it showed a decrease in cognitive performance compared to the placebo [65].

Phase I clinical trial study of SUVN-502 (Masupirdine) revealed that it is well-tolerated by healthy young and old adult participants. Phase II clinical trial (NCT02580305) for SUVN-502 in mild to moderate AD patients in combination with donepezil and memantine was completed but failed to show significant benefits.

Phase I clinical trial of SAM-760 was completed and well tolerated in healthy subjects and AD patients. Further, Phase II was terminated as it failed to show significant benefits [66].

Bertini *et al*. developed a series of substituted aryl sulphonamides **(I, Figure 4.1)** as BACE-1 inhibitors where the highest potency of a compound was found to be 1.6 µM [67]. Kang *et al* synthesized a series of sulphonamide chalcones (**II, Figure 4.1**) as dual inhibitor of BACE-1 and acetylcholinesterase. The compounds showed activity in the micromolar range; the best activity was 0.62 µM. Li *et al*. identified some sulphonamide derivatives via virtual screening as BACE-1 and PPARγ inhibitors (**III, Figure 4.1**). The IC_{50} value of one of the identified hits was found to be 1.24 µM. Zou *et al*. developed a series of pyrazole and sulphonamide-based BACE-1 inhibitors with potent activity. The best compound showed an IC_{50} value of 0.036 µM **(IV**, **Figure 4.1).**

Over the last decade, several research has been done on the therapeutic potential of BACE-1 inhibition. However, despite the fact that inhibitors effectively reduce Aβ levels, clinical trials still fail to show benefits in cognitive function when given to patients with mild-to-moderate AD. This raises concerns about the true value of these putative anti-AD medications as well as the design of the clinical trials. Recent research indicates that starting BACE-1 inhibitor therapy as soon as possible is the best course of action. A critical problem that may help to explain some of the prior failures is the best time to

begin using BACE-1 inhibitors [68]. Furthermore, recent studies report multitarget approaches focused on BACE-1, whose ligands are synthesized as small molecules that can be used to alter both BACE-1 and other AD-related targets through synergistic pathways due to the complex nature of AD.

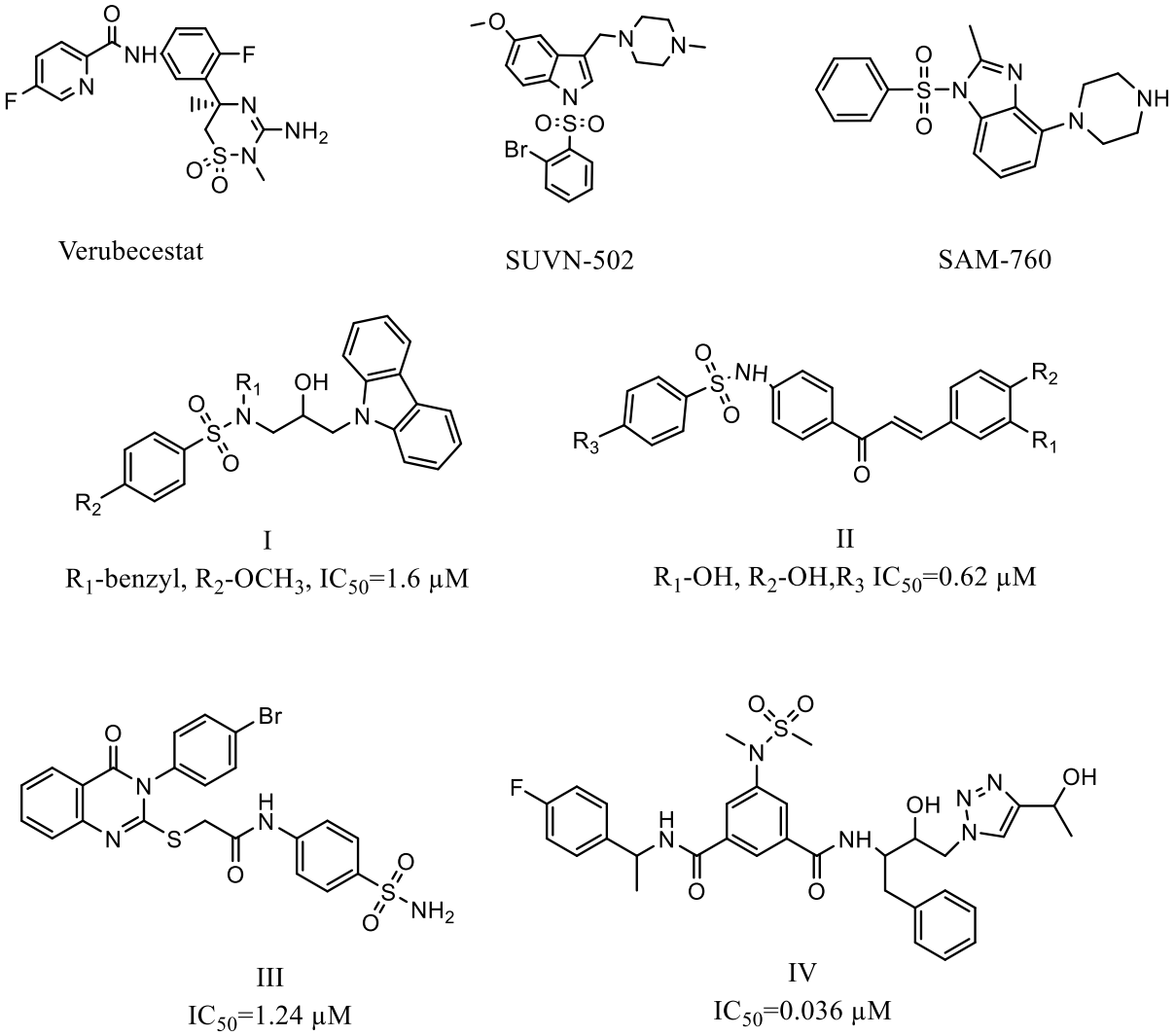


Figure 4.1 Sulphonamides as BACE-1 inhibitors in clinical and pre-clinical studies.

4.1.2 Machine Learning in drug discovery

ML techniques have been increasing and widely adopted in the early stages of drug discovery processes. ML is the branch of AI that focuses on developing and applying computer algorithms that use raw and unprocessed data to perform a specific task [69].

In the field of drug discovery, the applications of ML are growing enormously among a large number of pharmaceutical companies. The goal is to minimize the need for animal testing and primarily use high-throughput screening techniques to reduce the work and assist medication disclosure [70]. ML is classified into four groups based on the methodologies as: Supervised, Semi-supervised, Unsupervised and Reinforcement learning. These techniques increase decision-making, QSAR analyses, hit discoveries and *de novo* drug designs more accurately. In the ML methodology of drug discovery, there are the following steps in the experimental setup: 1) data collection; 2) generation of descriptors; 3) searching best subset of variables; 4) model training; and 5) model validation [69].

4.1.3 Machine learning algorithms

4.1.3.1 Random Forest

RF is a supervised learning method which is composed by the combination of tree predictors such that each tree depends on the values of a random vector independently and with the same layout for each tree in the forest [71]. Each tree in Random Forest is transverse in a particular way:

1. Giving a training dataset *N*, *n* random samples with repetition taken as training set (Bootstrap).
2. For each node of the tree, *M* input variables are determined, where $m<<M$ and the value of *m* remains constant. The node used are the randomly chosen variables.
3. Every tree is generated to its maximum expansion.

4.1.3.2 XGBoost classifier

XGBoost stands for Extreme Gradient Boosting, is an efficient and scalable machine learning Classifier model based on the Gradient Boosting Machine (GBM), providing parallel tree boosting and enhancing performance by using subsampling ratio, learning

rate, and maximum tree depth to avoid overfitting. XGBoost defines additional features such as handling missing data with nodes, default directions and specifying efficiently splitting thresholds during split node [72]. XGBoost produces comparable and better predictive accuracy and supports the inherent ability to handle highly diverse and complex descriptors [73].

4.1.3.3 LightGBM

LightGBM is another scalable and flexible GBM approach that shows comparable performance with the other existing boosting tools by learning efficiency and accuracy with lower consumption of memory [74]. LightGBM is a fast, high-performance tree-based learning algorithm, used for both classification and regression tasks. It can reduce the cost of the gain for each split-up in training. In LightGBM, the tree grows vertically and leaf-wise, while most decision-tree learning algorithms grow horizontally and level-wise [75].

In the present work, we have collected a dataset of sulphonamides as BACE-1 inhibitors and then developed and validated an ML model to classify the BACE-1 inhibitors and used this model to screen our *in-house* library of sulphonamides. The identified hits were then screened for BACE-1 activity using an *in-vitro* assay.

4.2 Materials and Methods

4.2.1 Dataset Collection

The dataset for BACE-1 was obtained from BindingDB (https://www.bindingdb.org/), a public web-accessible database [76]. Only the compounds containing the sulphonamide group were selected further. The KNIME analytical tool was used to filter the compounds with multiple entries and IC_{50} values. The compounds having IC_{50} values less than 500 nM were marked as active (recognized as 1), while compounds with IC_{50} more than 500

nM were marked as inactive (recognized as 0). Hence, total of 327 actives and 194 inactive compounds were obtained [77].

4.2.2 Fingerprint Descriptors

KNIME analytical tools were used to generate the fingerprint descriptors for the BACE-1 dataset using Fingerprints and Fingerprints expander nodes. The five fingerprint descriptors viz. MACCS, Estate, PubChem, ECFP4, and ECFP6 were obtained.

4.2.3 Data Splitting

The dataset of BACE-1 inhibitors were split into training (80%), validation (10%), and test (10%) sets by train_test_split by using the scikit learn python module having a random state of 2529. Training dataset was used for model development and other two subsets (i.e., test and validation) were used to evaluate training model performance against new data.

4.2.4 Machine learning classification algorithms

Random Forest (RF), gradient boosting machine (XGBoost), and LightGBM machine learning algorithms were used for classification models using Python library *Scikit learn*. Grid search using *GridsearchCV* was performed to identify the optimal combination of values for the hyperparameters.

4.2.4.1 Random-Forest Classifier

Three different parameter combinations were used to determine the RF, that is the number of trees in the random forest (n_estimators), maximum depth of the tree (max_depth), and minimum number of samples required to split an internal node (min_samples_leaf) (*scikit-learn 1.2.2*). A grid search was performed to obtain the maximum accuracy using following parameters:

- n estimators- 50, 100, 200, 300, 400, and 500

- Maximum depth ranges from 5 to 50 with an increment of 5.
- Minimum sample split ranges from 2 to 10.

4.2.4.2 XG Boost Classifier

XGBoost or Extreme Gradient boosting classifier can work well in smaller datasets (*XGBoost 1.7.5*). A grid search was performed to tune hyperparameters, and based on accuracy score the best model was selected. XGBoost provides large range of hyperparameters such as:

- Maximum depth of a tree (max_depth)- 5,7,9,11,13, and 15.
- Learning rate ranges from 0.01 to 0.1.
- Gamma- 0, 0.25, and 1.
- Lambda (reg_lambda) ranges from 0 to 15.
- scale_pos_weight used for imbalanced classes having values 3,5,7,9, and 11.
- Subsample is the ratio of training instances having a value of 0.8.
- colsample_bytree is the subsample ratio of the column having value of 0.5.
- Tree construction algorithm (tree_method) used 'gpu_hist'.

4.2.4.3 LightGBM Classifier

LightGBM works on a histogram-based algorithm that results in faster and more accurate results compared to XGBoost (LightGBM 3.2.2). The most critical hyperparameters used by the LightGBM are:

- 'num_leaves': 10-50,
- 'reg_alpha': [0.1, 0.5],
- 'lambda_l1': [0-5],
- 'lambda_l2': [0, 1],

- 'min_data_in_leaf': 30-100,
- 'learning_rate': 0.9-0.001

4.2.4.4 Performance Evaluation of models

The methods of performance evaluation of models have been mentioned in the section **3.2.6.**

4.2.5 Screening Database preparation

The *in-house* database of previously synthesized and reported sulphonamides in our lab were prepared using DataWarrior V5.5.0 [78-80]. The database consists of 129 reported sulphonamide derivatives. The database was screened with the best model to identify the hits [81].

4.2.6 BACE-1 inhibition assay

The identified hits were evaluated for their BACE-1 inhibition potential using fluorescence resonance energy transfer (FRET) based BACE-1 fluorescence assay kit (Catalog No. CS0010, Sigma-Aldrich). The kit consists of fluorescent assay buffer, stop solution, substrate (7-Methoxycumarin-4-acetyl [Asn670, Lue671]-Amyloid βA4 Precursor Protein 770 Fragment 667-676-(2,4 dinitrophenyl) Lys-Arg-Arg amide trifluoroacetate salt) and BACE-1 enzyme. Different concentrations of test compounds were prepared. The fluorescence intensity was measured immediately after the addition of BACE-1 enzyme with the wavelength of excitation and emission was set at 320 nm and 405 nm, respectively. All the measurements were performed in triplicate. The percentage inhibition was calculated using the following formulae: $[(Io–Ii)/Io] \times 100$, where Io and Ii are the fluorescence intensities obtained in the absence and presence of an inhibitor, respectively and the IC_{50} values were calculated using linear regression graph (GraphPad Prism 5.1, GraphPad Software Inc.).

4.2.7 Docking study

The docking study was performed to study the binding pose and interaction of the identified hits with the BACE-1 protein.

4.2.7.1 Grid generation and validation

The amino acid residues involved in the protein-ligand interactions of the selected protein (PDB ID-6EQM) were identified by using BIOVIA Discovery studio visualizer. The identified residues were used to construct a grid box around the active site as the reference points. The Autogrid 4.0 was used to calculate grid maps of interaction energies having various atom types present in the ligand (A, C, HD, N, NA, S, OA, Br, Cl and I). The grid size was set to xyz points at 60×60×54, having a grid spacing of 0.336 Å, and the grid centers were placed at the coordinates X: 28.936, Y: 79.442, Z: 18.584, respectively. Further, the obtained grid was validated by redocking ligand (BUH) and the Root Mean Square Deviation (RMSD) value was calculated between experimentally obtained co-crystalized ligand and docked pose using Maestro. The RMSD was found to be 0.389 Å. Precision docking was performed using AutoDock 4.2 by engaging Lamarckian Genetic Algorithm (LGA) with the genetic algorithm runs kept at 100.

4.3 Result and Discussion

4.3.1 Machine learning models

The training data set had total of 521 compounds out of which 416 were taken for training set and remaining compounds were equally divided into test set and validation set using stratified splitting**.**

4.3.1.1 Random Forest classifier

Random forest is an ensemble of decision trees. The **Table 4.1** and **Table 4.2** summarizes the performance of Random Forest classifiers build using different fingerprints on the training and test set. The summary of hyperparameters of all the models have been

summarized in **Table S4 of appendix).** The result indicates that the model build using PubChem fingerprints had the best F1 score of 0.91. The model had an accuracy, precision and recall scores of 0.86, 0.84 and 0.98 respectively. The model build using Estate fingerprint showed recall score of 1.0 on training and test set but the precision score was low.

4.3.1.2 XGBoost Classifier

It is an ensemble of several weak classifier that uses a gradient boosting framework. The hyperparameters of the best model for every descriptor has been summarized in **Table S5** of **appendix.** The **Table 4.1** summarizes the performance of XGBoost classifier on training and test set. The accuracy and F1 score of models build using XGBoost classifier was better than that of RF classifier when evaluated on test set. The models built using PubChem fingerprint showed the best F1 score on the test set.

4.3.1.3 LightGBM classifier

It is also a boosting algorithm based on decision tree. It is considered to be fast and less computational memory intensive. The hyperparameters corresponding to each fingerprint for LightGBM model has been summarized in **Table S6 of appendix.** The model build using PubChem showed the best accuracy and F1 score on the test set i.e., 0.87 and 0.92 respectively. The model performed better than the other two algorithms. The summary of performance of models has been given is the **Table 4.1**.

4.3.2 Performance of ML models on validation set

In order to check the robustness of machine learning models it is necessary to evaluate the models on independent validation set. The performance of every algorithm on different type of fingerprint has been summarized in **Table 4.2.** The result indicates that the model build using XGB classifier with PubChem fingerprints showed the best performance on the external validation set with accuracy, precision, recall and F1 score

of 0.89, 0.88, 0.99 and 0.93, respectively. The best model was selected and feature importance were calculated. The top 20 features and their importance has been summarized in a figure **(Figure S2 of appendix)** and the top ten fragments have been shown in **Figure 4.2**.

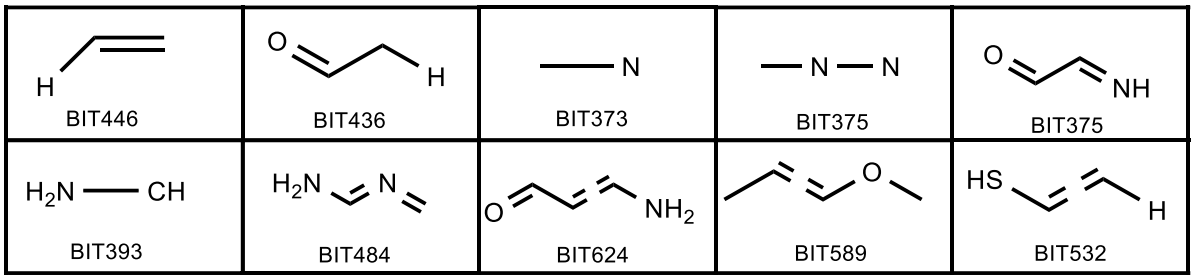


Figure 4.2 Top 10 important PubChem fingerprints

4.3.3 Screening of *in-house* library

The *in-house* library of compounds was screened virtually using XGB classifier build using PubChem fingerprints. The cut-off was kept to 0.5 which is the default value for several binary classification algorithms. The compounds having score more than the cut-off were marked

Table 4.1 Performance of classification models on the training set

Fingerprints	XGBoost classifier				LightGBM classifier				RF classifier			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
MACCS	0.91	0.85	0.98	0.91	0.96	0.96	0.98	0.97	0.88	0.81	0.98	0.89
ECFP-4	0.91	0.90	0.98	0.91	0.99	0.99	1.00	0.99	0.88	0.87	0.97	0.91
ECFP-6	0.91	0.90	0.98	0.91	0.99	0.99	1.00	1.00	0.76	0.78	0.96	0.86
PubChem	0.95	0.94	1.00	0.97	0.98	0.98	0.99	0.99	0.86	0.85	0.99	0.91
Estate	0.83	0.81	1.00	0.89	0.86	0.97	0.96	0.91	0.78	0.78	1.00	0.87

Table 4.2 Performance of classification models on the test set

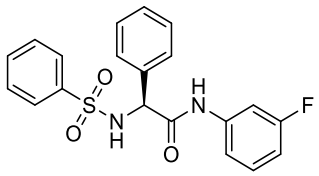
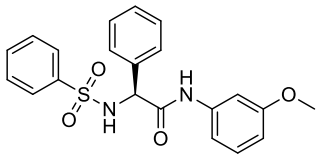
Fingerprints	RF classifier				XGBoost classifier				LightGBM classifier			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
MACCS	0.84	0.81	0.98	0.90	0.88	0.84	0.99	0.91	0.85	0.87	0.93	0.90
ECFP-4	0.84	0.86	0.93	0.89	0.88	0.87	0.93	0.90	0.90	0.92	0.92	0.93
ECFP-6	0.79	0.87	0.93	0.88	0.88	0.87	0.93	0.90	0.88	0.92	0.92	0.92
PubChem	0.86	0.84	0.98	0.91	0.87	0.87	0.97	0.92	0.88	0.89	0.94	0.92
Estate	0.80	0.79	1.00	0.88	0.82	0.80	1.00	0.89	0.79	0.82	0.92	0.87

Table 4.3 Performance of ML models on validation set

Classifier	Descriptors	Accuracy	Precision	Recall	F1 score
RF	MACCS	0.85	0.83	0.98	0.90
	ECFP-4	0.86	0.86	0.93	0.89
	ECFP-6	0.75	0.78	0.93	0.85
	PubChem	0.83	0.82	0.98	0.89
	Estate	0.80	0.78	1.00	0.88
XGB	MACCS	0.87	0.83	0.97	0.89
	ECFP-4	0.87	0.88	0.94	0.91
	ECFP-6	0.87	0.88	0.94	0.91
	PubChem	**0.89**	**0.88**	**0.99**	**0.93**
	Estate	0.83	0.82	0.97	0.89
LightGBM	MACCS	0.82	0.86	0.90	0.88
	ECFP-4	0.88	0.85	0.89	0.87
	ECFP-6	0.89	0.85	0.89	0.87
	PubChem	0.88	0.89	0.96	0.92
	Estate	0.80	0.83	0.91	0.87

active and were selected as hit. The screening resulted in the identification of two virtual hits i.e., compound-28 (*(S)-(+)-N-(3-fluorophenyl)-2-phenyl-2-(phenylsulfonamido) acetamide*) and compound-37 *((S)-(+)-N-(3-methoxyphenyl)-2-phenyl-2-(phenylsulfonamido) acetamide*), which were previously reported for acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) activity [82]. The summary of the reported properties of both the hits has been given in **Table 4.4**.

Table 4.4 Summary of reported properties for identified hits

Compound code	Structure	% Inhibition at a concentration of 50 µM	
		BChE	**AChE**
28		01.23±0.84	05.82 ± 0.78
37		41.77±0.62	14.03 ± 0.85

4.3.4 *In-vitro* BACE-1 inhibitory activity

The identified hits were evaluated for their BACE-1 inhibition using FRET-based assay kit. The molecules were initially screened to determine the percentage inhibition at 1 µM and then they were screened at different concentrations to determine their IC_{50} values. The compound 28, containing 3-fluorophenyl group, showed IC_{50} of 0.431±0.006 µM and the compound 37, containing 3-methoxyphenyl group showed IC_{50} value of 0.272±0.019 µM (**Table 4.5)**.

Table 4.5 Summary of in-vitro and docking result of ligands with BACE-1 (PDB ID-6EQM)

Compound code	**hBACE—1 IC_{50}(µM)±S.D.[a]**	**Binding energy (Kcal/mol)**	**Ligand efficiency (Kcal/mol)**	**Interactions (PDB ID-6EQM)**
Compound 28	0.431±0.006	-7.66	-0.284	Arg307 (H-bond), Thr232 (H-bond), Gly230 (H-bond), Tyr71 (Pi-Pi T-shaped), Ala396 (Pi-Sigma)
Compound 37	0.272±0.019	-7.58	-0.271	Asp32 (H-bond), Leu30 (Pi-alkyl), Phe108 (Pi-Sulfur), Asp228 (Pi-anion), Tyr198 (Pi-Pi T-shaped)

[a]Data expressed in Mean±S.D.(n=3)

4.3.5 Docking study

The grid validation was performed by redocking the co-crystallized ligand and calculating the RMSD between the docked pose and co-crystallized ligand. The RMSD value was found to be 0.389 Å. The docked pose and co-crystallized ligand have been represented in **Figure S3 of appendix** The docking study revealed that the compound 28 and compound 37 had binding energy of -7.66 and -7.58 Kcal $mol^{-1,}$ respectively. Their interaction diagram revealed that the compound 28 showed H-bond interaction with Arg307 and Thr232. The compound 37 showed interaction with the catalytic dyad i.e., Asp32 and Asp228 via. H-bond and Pi-anion interactions, respectively. The summary of docking result containing binding energy, ligand efficiency and interactions has been represented in **Table 4.5**.

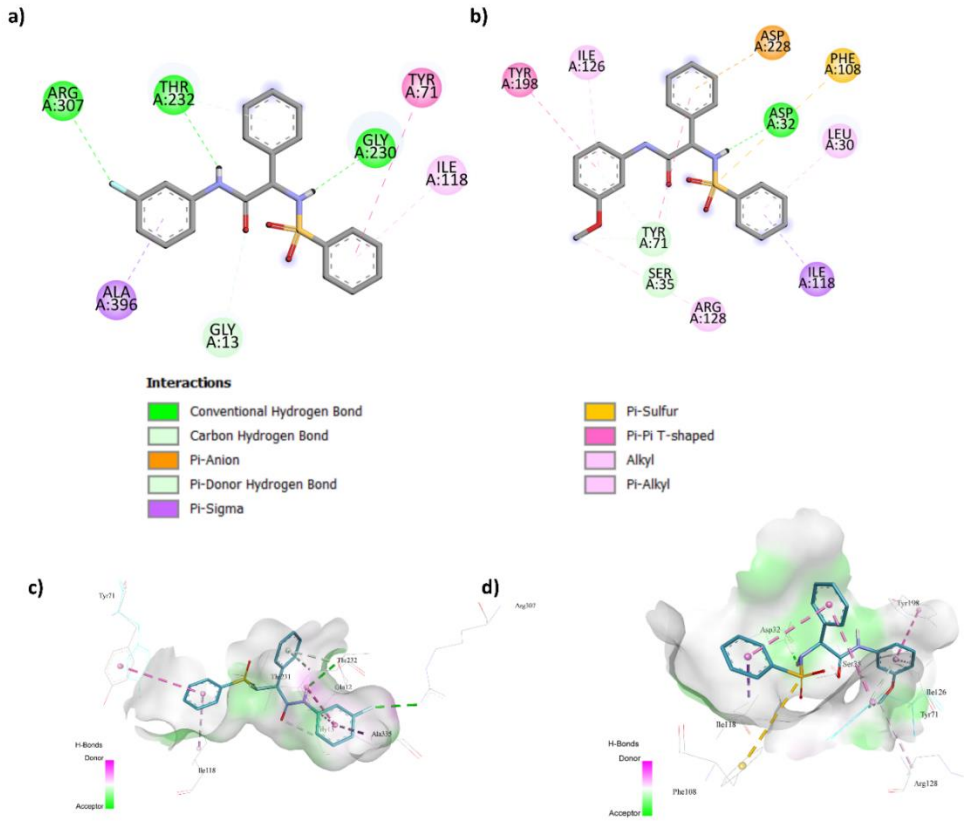


Figure 4.3 2D interaction diagram of (a) compound 28 and (b) compound 37 and 3D interaction diagram of (c) compound 28 and (d) compound 37.

4.4 Conclusion

BACE-1 is a promising target for the treatment of AD. Several sulphonamide-based BACE-1 inhibitors have shown potential for decelerating the long-term progression of AD. Drug discovery pipelines are extremely long and complicated process. In this study, a ML model was developed using to classify the BACE-1 inhibitors. The classification was based on the range of IC_{50} value. The compounds having IC_{50} value less than 500 nM were marked as active and the compound having IC_{50} value more than 500 nM were marked as inactive. The best ML model had accuracy, precision, recall and F1 score of 0.89, 0.88, 0.99 and 0.93 on the validation set. The model was built using the XGboost algorithm on PubChem fingerprints. The model was used to screen the *in-house* library of potential sulphonamides as BACE-1 inhibitors. Upon screening, we obtained two hits, i.e., Compound 28 and compound 37, which were previously reported as weak AChE and BuChE inhibitors. Both compounds were evaluated for their *in-vitro* BACE-1 activity. The compound 28 showed an IC_{50} value of 0.431±0.006 µM and compound 37 showed an IC_{50} value of 0.272±0.019 µM. The docking study revealed compound 37 interacted with the catalytic dyad of BACE-1, i.e., Asp32 and Asp228. Thus, the developed model has shown reliable prediction and further studies can be done on the identified hits to make them potential lead molecules.