



# Machine learning-based screening of in-house database to identify BACE-1 inhibitors

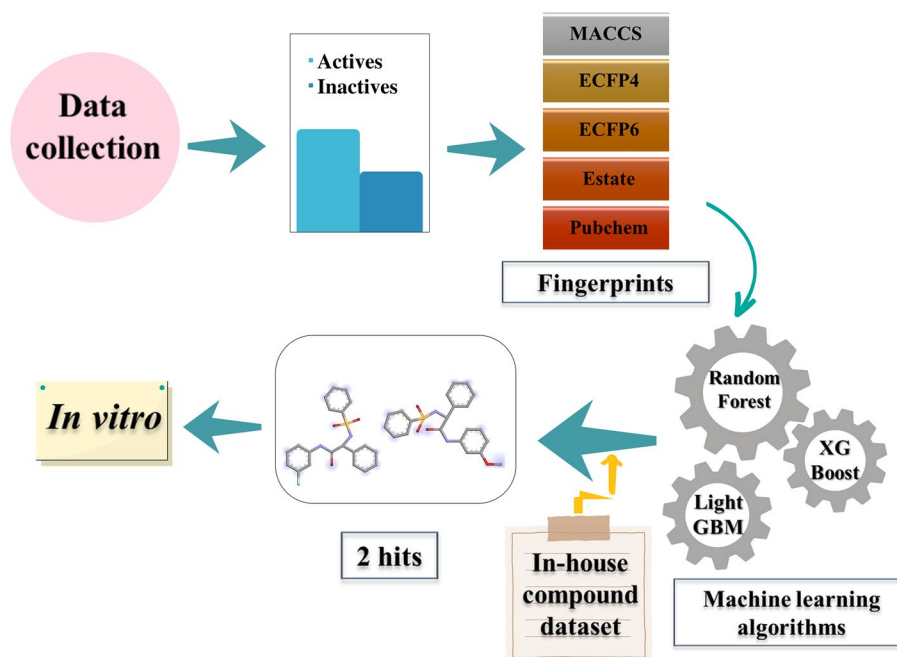
Ravi Singh<sup>1</sup> · Asha Anand<sup>1</sup> · Ankit Ganeshpurkar<sup>2</sup> · Powsali Ghosh<sup>1</sup> · Tushar Chaurasia<sup>1</sup> · Ravi Bhushan Singh<sup>3</sup> · Dileep Kumar<sup>2</sup> · Sushil Kumar Singh<sup>1</sup> · Ashok Kumar<sup>1</sup>

Received: 28 February 2023 / Accepted: 14 July 2023 / Published online: 26 July 2023  
© Institute of Chemistry, Slovak Academy of Sciences 2023

## Abstract

The  $\beta$ -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- $\beta$  ( $A\beta$ ) plaques in the brain. In the present work, we have developed an ML model based on the sulfonamides dataset. 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  $0.431 \pm 0.006 \mu\text{M}$ , and compound 37 showed an  $IC_{50}$  value of  $0.272 \pm 0.019 \mu\text{M}$ . The docking study revealed that compound 37 also showed interactions with the catalytic dyad of BACE-1, i.e., Asp32 and Asp228.

## Graphical abstract



**Keywords** BACE-1 · Machine learning · Drug discovery · Autodock · Alzheimer's disease

## Introduction

B-site APP cleaving enzyme 1 (BACE1) is an aspartyl protease of the pepsin family, discovered in 1999. BACE1 initiates the production of A $\beta$ , which represents the rate-limiting enzyme in the amyloidogenic pathway. BACE1 cleaves the A $\beta$  precursor protein (APP) to its membrane-bound C terminus fragment C99 (CTF) and soluble APP $\beta$  fragment. The BACE1 is essential for the generation of all monomeric units of A $\beta$ , including A $\beta$ <sub>42</sub>, which plays a crucial role in the pathogenesis of Alzheimer's disease (AD). The concentrations and activity rates of BACE1 are actively increased in AD brains and body fluids. Therefore, BACE1 emerged as a primary drug target for decreasing the production of A $\beta$  in the AD brain (Hampel et al. 2021). BACE1 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 (Vassar 2014). BACE1 consists of metal binding sites; it has a copper-binding site in its cytosolic domain (Hung et al. 2010). The crystal structure of BACE1 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 (Turner et al. 2001; Ghosh and Osswald 2014).

### Sulfonamides as BACE1 inhibitors in human clinical trials

Non-peptide BACE1 inhibitors such as sulfonamides had some success in preclinical studies as some of the drugs were seen in various phases of clinical trials as well.

BACE1 inhibitor MK-8931 (Verubecestat) entered the Phase III 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 $\beta$ <sub>40</sub> in healthy participants, whereas it showed a decrease in cognitive performance compared to the placebo (Kennedy, et al. 2016).

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 (Sastre, et al. 2017).

Bertini et al. developed a series of substituted aryl sulfonamides (I, Fig. 1) as BACE1 inhibitors where the highest potency of a compound was found to be 1.6  $\mu$ M (Bertini et al. 2017). Kang et al. synthesized a series of sulfonamide chalcones (II, Fig. 1) as dual inhibitor of BACE1 and acetylcholinesterase. The compounds showed activity in the micromolar range; the best activity was 0.62  $\mu$ M. Li et al. identified some sulfonamide derivatives via virtual screening as BACE1 and PPAR $\gamma$  inhibitors (III, Fig. 1). The IC<sub>50</sub> value of one of the identified hits was found to be 1.24  $\mu$ M. Zou et al. developed a series of pyrazole and sulfonamide-based BACE-1 inhibitors with potent activity. The best compound showed an IC<sub>50</sub> value of 0.036  $\mu$ M (IV, Fig. 1).

Over the last decade, several research has been done on the therapeutic potential of BACE1 inhibition. However, despite the fact that inhibitors effectively reduce A $\beta$  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 BACE1 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 BACE1 inhibitors (Voytyuk et al. 2018). Furthermore, recent studies report multitarget approaches focused on BACE1, whose ligands are synthesized as small molecules that can be used to alter both BACE1 and other AD-related targets through synergistic pathways due to the complex nature of AD.

### Machine learning in drug discovery

Machine learning (ML) techniques have been increasing and widely adopted in the early stages of drug discovery processes. ML is the branch of artificial intelligence (AI) that focuses on developing and applying computer algorithms that use raw and unprocessed data to perform a specific task (Carracedo-Reboredo et al. 2021). 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 (Gupta et al. 2021). 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 (Carracedo-Reboredo et al. 2021).

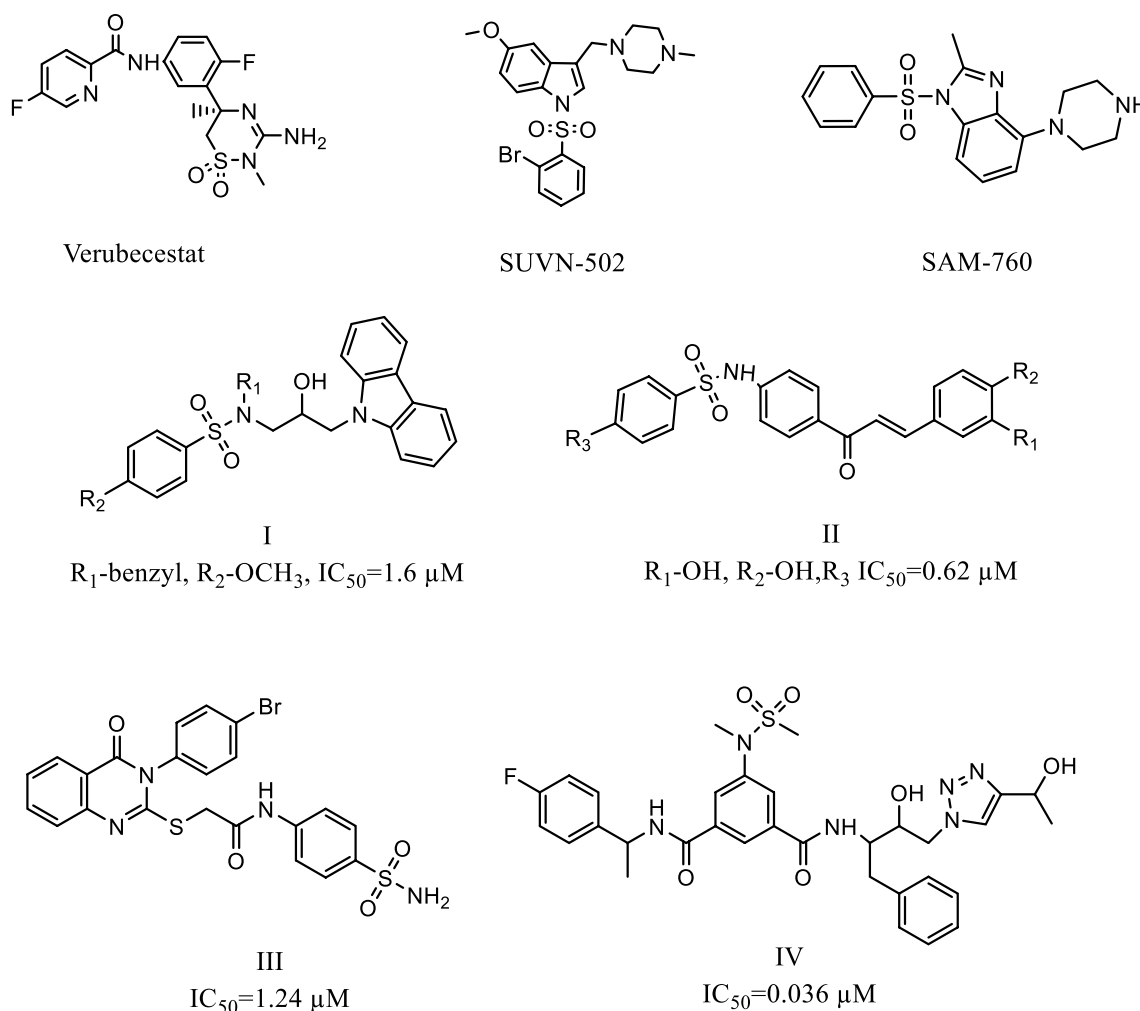


Fig. 1 Sulfonamides as BACE1 inhibitors in clinical and preclinical studies

## Machine learning algorithms

### Random forest

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 (Breiman 2001). 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  $m \ll M$  and the value of  $m$  remains constant. The node used is the randomly chosen variables.

- (3) Every tree is generated to its maximum expansion.

### XGBoost classifier

XGBoost stands for extreme gradient boosting and 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 (Sagi and Rokach 2021). XGBoost produces comparable and better predictive accuracy and supports the inherent ability to handle highly diverse and complex descriptors (Babajide Mustapha and Saeed 2016).

## 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 (Du et al. 2022). 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 (Zhao et al. 2019).

In the present work, we have collected a dataset of sulfonamides 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 sulfonamides. The identified hits were then screened for BACE-1 activity using an in vitro assay.

## Materials and methods

### Dataset collection

The dataset for BACE1 was obtained from BindingDB (<https://www.bindingdb.org/>), a public web-accessible database (Gilson et al. 2016). Only the compounds containing the sulfonamide 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 (Berthold 2009).

### Fingerprint descriptors

KNIME analytical tools were used to generate the fingerprints descriptors for the BACE1 dataset using Fingerprints and Fingerprints expander nodes. The five fingerprint descriptors, viz. MACCS, Estate, PubChem, ECFP4 and ECFP6, were obtained.

### Data splitting

The dataset of BACE1 inhibitors was 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.

## 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.

### 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.

### XGBoost 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.'

### 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 hyper-parameters 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

## Performance evaluation

**Accuracy:** Accuracy is the percentage of the total correctly classified outcomes from the total outcomes.

$$\text{Accuracy (\%)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100$$

**Precision:** Precision refers to the number of true positives divided by the total number of the positive predictions (i.e., sum of true positives and false positives). Precision indicates the quality of the positive predictions made by a model.

$$\text{Precision (\%)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100$$

**Recall:** Recall is the ratio between the true positives to the sum of true positives and false negatives.

$$\text{Recall (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

**F1 score:** F1 score measures the model’s accuracy for a dataset. It combines the scores of precisions and recall of a model and made a correct prediction for the entire dataset.

$$\text{F1 (\%)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$$

where TP is the true positive, FP is the false positive, TN is the true negative and FN is the false negative.

## Screening database preparation

The in-house database of previously synthesized and reported sulfonamides in our laboratory was prepared using DataWarrior V5.5.0 (Swetha et al. 2019; Ganeshpurkar et al. 2018; Kumar et al. 2018). The database consists of 129 reported sulfonamide derivatives. The database was screened with the best model to identify the hits (Sander et al. 2015).

## BACE-1 inhibition assay

The identified hits were evaluated for their BACE1 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  $\beta$ A4 precursor protein 770 fragment 667-676-(2,4 dinitrophenyl) Lys-Arg-Arg amide trifluoroacetate salt) and BACE1 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:  $[(I_0 - I_i)/I_0] \times 100$ , where  $I_0$  and  $I_i$  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.).

## Docking study

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

## 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) (Hampel et al. 2021). The grid size was set to xyz points at  $60 \times 60 \times 54$ , 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-crystallized 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.

## Result and discussion

### Machine learning models

The training dataset 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 (Table S1 of S.I).

### Random forest classifier

Random forest is an ensemble of decision trees. Table 1 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 is summarized in Table S2 of Supplementary Information (S.I.). 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 score 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.

### 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 is summarized in Table S3 of S.I. Table 2 summarizes the performance of XGBoost classifier on training and test set. The accuracy and F1 score of models build using XGBoost classifier were 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.

### 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 is summarized in Table S4 of S.I. 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 is given in Table 1.

**Table 1** 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 2** 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
XGB	Estate	0.80	0.78	1.00	0.88
	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
LightGBM	<b>PubChem</b>	<b>0.89</b>	<b>0.88</b>	<b>0.99</b>	<b>0.93</b>
	Estate	0.83	0.82	0.97	0.89
	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

The Bold signifies the performance of the best model

### 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 is summarized in Table 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 was calculated. The top 20 features and their importance are summarized in a figure (Figure S1 of S.I.), and the top ten fragments are shown in Fig. 2.

### Screening of in-house library

The in-house library of compounds was screened virtually using XGB classifier build using PubChem fingerprints. The cutoff was kept to 0.5 which is the default value for several binary classification algorithms. The

BIT446	BIT436	BIT373	BIT375	BIT375
BIT393	BIT484	BIT624	BIT589	BIT532

Fig. 2 Top 10 important PubChem fingerprints

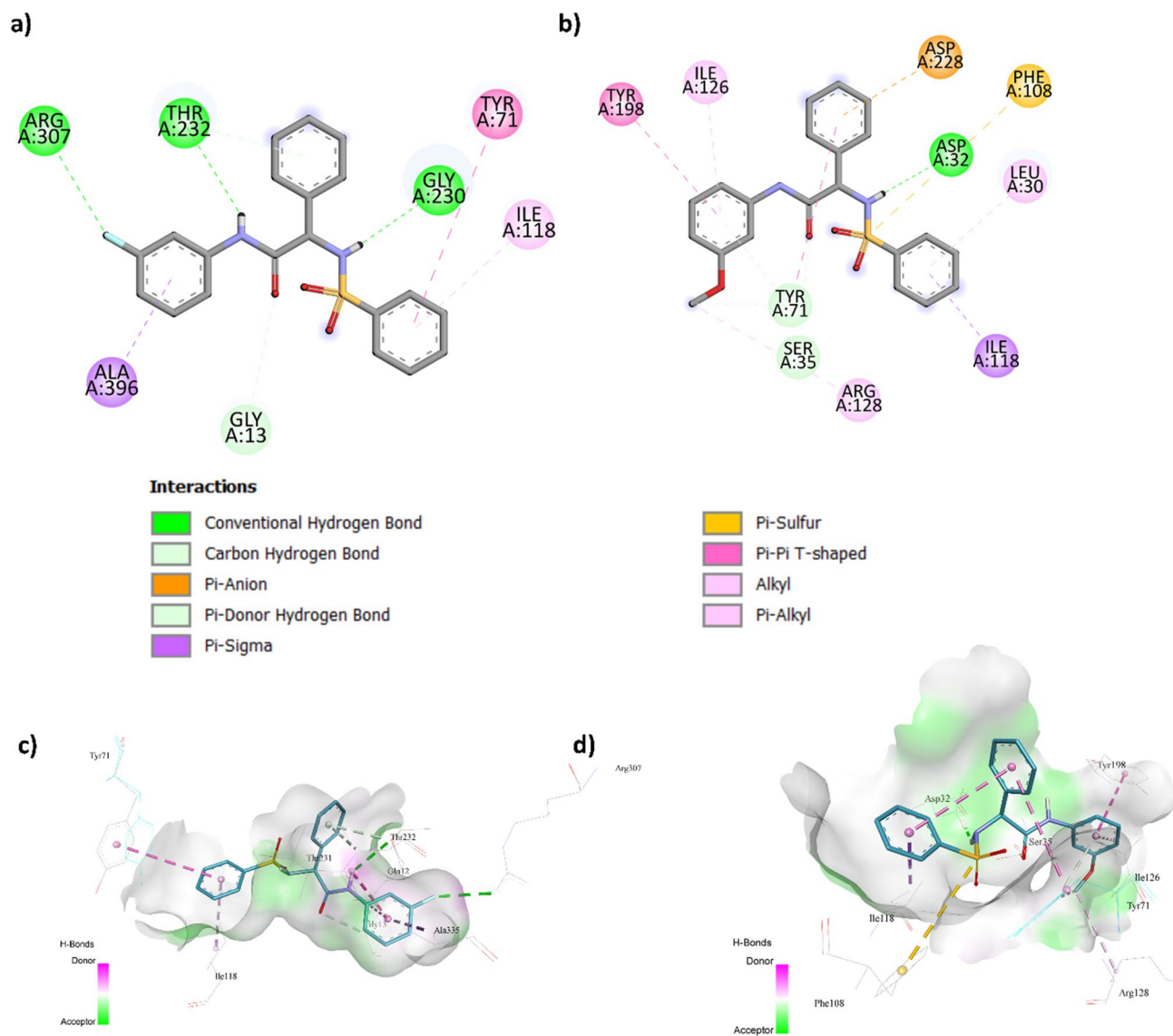
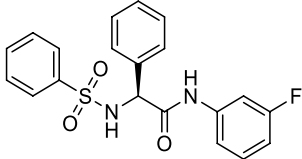
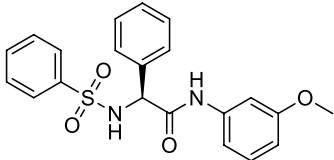


Fig. 3 2D interaction diagram of (a) compound 28 and (b) compound 37 and 3D interaction diagram of (c) compound 28 and (d) compound 37

compounds having score more than the cutoff were marked as 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)

**Table 3** Summary of reported properties for identified hits

Compound code	Structure	% Inhibition at a concentration of 50 $\mu\text{M}$	
		BChE	AChE
28		01.23 $\pm$ 0.84	05.82 $\pm$ 0.78
37		41.77 $\pm$ 0.62	14.03 $\pm$ 0.85

*acetamide*), which were previously reported for acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) activity (Ganeshpurkar et al. 2022). The summary of the reported properties of both the hits is given in Table 3.

### 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  $\mu\text{M}$ , and then, they were screened at different concentrations to determine their  $\text{IC}_{50}$  values. The compound 28, containing 3-fluorophenyl group, showed  $\text{IC}_{50}$  of  $0.431 \pm 0.006 \mu\text{M}$  and the compound 37, containing 3-methoxyphenyl group, showed  $\text{IC}_{50}$  value of  $0.272 \pm 0.019 \mu\text{M}$  (Table 4).

### 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 are represented in Fig. S2 of S.I. The docking study revealed that the compound 28 and compound 37 had binding energy of  $-7.66$  and  $-7.58 \text{ kcal mol}^{-1}$ , 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 (Fig. 3). The summary of docking result containing binding energy, ligand efficiency and interactions is represented in Table 4.

### Conclusion

BACE-1 is a promising target for the treatment of AD. Several sulfonamide-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  $\text{IC}_{50}$  value. The compounds

**Table 4** Summary of in vitro and docking result of ligands with BACE-1 (PDB ID-6EQM)

Compound code	hBACE—1 $\text{IC}_{50}(\mu\text{M}) \pm \text{S.D.}^{\text{a}}$	Binding energy (Kcal/mol)	Ligand efficiency (Kcal/mol)	Interactions (PDB ID- 6EQM)
Compound 28	$0.431 \pm 0.006$	$-7.66$	$-0.284$	Arg307 (H-bond), Thr232 (H-bond), Gly230 (H-bond), Tyr71 (Pi-Pi T-shaped), Ala396 (Pi-Sigma)
Compound 37	$0.272 \pm 0.019 \mu\text{M}$	$-7.58$	$-0.271$	Asp32 (H-bond), Leu30 (Pi-alkyl), Phe108 (Pi-Sulfur), Asp228 (Pi-anion), Tyr198 (Pi-Pi T-shaped)

<sup>a</sup>Data expressed in mean  $\pm$  S.D. ( $n = 3$ )



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 sulfonamides 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 the compounds were evaluated for their in-vitro BACE-1 activity. The compound 28 showed an  $IC_{50}$  value of  $0.431 \pm 0.006 \mu\text{M}$ , and compound 37 showed an  $IC_{50}$  value of  $0.272 \pm 0.019 \mu\text{M}$ . Docking study revealed that the compound 37 showed interaction with the catalytic dyad of BACE-1, i.e., Asp32 and Asp228. Thus, the developed model has shown reliable prediction and further studies and optimizations can be done on the identified hits to make them potential lead molecules.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11696-023-02982-2>.

**Acknowledgements** The authors would like to acknowledge the financial support from the Ministry of Education (MoE), New Delhi, India, in the form of a teaching assistantship to RS, PG and Asha. Ravi Bhushan Singh would like to thank DST, SERB, New Delhi, for his TARE grant award.

## Declarations

**Conflict of interest** No potential conflict of interest was reported by the author(s).

## References

- Babajide Mustapha I, Saeed F (2016) Bioactive molecule prediction using extreme gradient boosting. *Molecules* 21(8):983
- Berthold MR et al (2009) KNIME-the Konstanz information miner: version 20 and beyond. *SIGKDD Explor Newsl* 11(1):26–31
- Bertini S et al (2017) Sulfonamido-derivatives of unsubstituted carbazoles as BACE1 inhibitors. *Bioorg Med Chem Lett* 27(21):4812–4816
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Carracedo-Reboredo P et al (2021) A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J* 19:4538–4558
- Du Z et al (2022) Inference of gene regulatory networks based on the light gradient boosting machine. *Comput Biol Chem* 101:107769
- Ganeshpurkar A, Kumar D, Singh SK (2018) Design, synthesis and collagenase inhibitory activity of some novel phenylglycine derivatives as metalloproteinase inhibitors. *Int J Biol Macromol* 107:1491–1500
- Ganeshpurkar A et al (2022) Identification of sulfonamide based butyrylcholinesterase inhibitors through scaffold hopping approach. *Int J Biol Macromol* 203:195–211
- Ghosh AK, Osswald HL (2014) BACE1 ( $\beta$ -secretase) inhibitors for the treatment of Alzheimer's disease. *Chem Soc Rev* 43(19):6765–6813
- Gilson MK et al (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44(D1):D1045–D1053
- Gupta R et al (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Diversity* 25(3):1315–1360
- Hampel H et al (2021) The  $\beta$ -secretase BACE1 in Alzheimer's disease. *Biol Psychiat* 89(8):745–756
- Hung YH, Bush AI, Cherny RA (2010) Copper in the brain and Alzheimer's disease. *J Biol Inorg Chem* 15(1):61–76
- Kennedy ME et al (2016) The BACE1 inhibitor verubecestat (MK-8931) reduces CNS  $\beta$ -amyloid in animal models and in Alzheimer's disease patients. *Sci Trans Med* 8(363):363ra150
- Kumar D et al (2018) Development of Piperazinediones as dual inhibitor for treatment of Alzheimer's disease. *Eur J Med Chem* 150:87–101
- Sagi O, Rokach L (2021) Approximating XGBoost with an interpretable decision tree. *Inf Sci* 572:522–542
- Sander T et al (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model* 55(2):460–473
- Sastre AA et al (2017) Effect of the treatment of type 2 diabetes mellitus on the development of cognitive impairment and dementia. *Cochrane Database Syst Rev*
- Swetha R et al (2019) Multifunctional hybrid sulfonamides as novel therapeutic agents for Alzheimer's disease. *Future Med Chem* 11(24):3161–3178
- Turner RT et al (2001) Subsite specificity of memapsin 2 ( $\beta$ -secretase): implications for inhibitor design. *Biochemistry* 40(34):10001–10006
- Vassar R (2014) BACE1 inhibitor drugs in clinical trials for Alzheimer's disease. *Alzheimer's Res Therapy* 6(9):1–14
- Voytyuk I, De Strooper B, Chávez-Gutiérrez L (2018) Modulation of  $\gamma$ - and  $\beta$ -Secretases as early prevention against Alzheimer's disease. *Biol Psychiat* 83(4):320–327
- Zhao Q et al (2019) Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharmaceutica Sinica B* 9(6):1241–1252

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Ravi Singh<sup>1</sup> · Asha Anand<sup>1</sup> · Ankit Ganeshpurkar<sup>2</sup> · Powsali Ghosh<sup>1</sup> · Tushar Chaurasia<sup>1</sup> · Ravi Bhushan Singh<sup>3</sup> · Dileep Kumar<sup>2</sup> · Sushil Kumar Singh<sup>1</sup> · Ashok Kumar<sup>1</sup> 

✉ Ashok Kumar  
akmaurya.rs.phe@iitbhu.ac.in

<sup>1</sup> Pharmaceutical Chemistry Research Laboratory 1,  
Department of Pharmaceutical Engineering & Technology,  
Indian Institute of Technology (Banaras Hindu University),  
Varanasi 221005, India

<sup>2</sup> Department of Pharmaceutical Chemistry, Poona College  
of Pharmacy, Bharti Vidyapeeth Erandwane, Pune, India

<sup>3</sup> Institute of Pharmacy, Harish Chandra PG College, Varanasi,  
India