Molecular Diversity (2022) 26:1501–1517
https://doi.org/10.1007/s11030-021-10282-8

ORIGINAL ARTICLE

Machine learning models for predicting the activity of AChE and BACE1 dual inhibitors for the treatment of Alzheimer's disease

G. Dhamodharan[1] · C. Gopi Mohan[1]

Received: 1 April 2021 / Accepted: 19 July 2021 / Published online: 29 July 2021

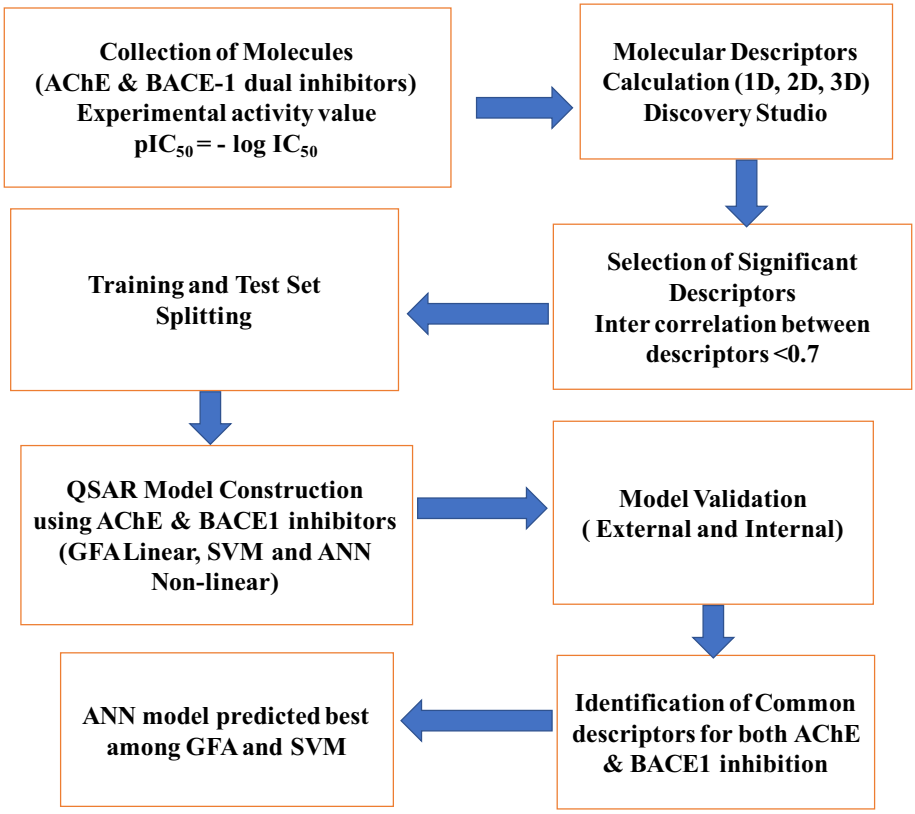
Abstract

Multi-target directed ligand-based 2D-QSAR models were developed using different N-benzyl piperidine derivatives showing inhibitory activity toward acetylcholinesterase (AChE) and β-Site amyloid precursor protein cleaving enzyme (BACE1). Five different classes of molecular descriptors belonging to spatial, structural, thermodynamics, electro-topological and *E*-state indices were used for machine learning by linear method, genetic function approximation (GFA) and nonlinear method, support vector machine (SVM) and artificial neural network (ANN). Dataset used for QSAR model development includes 57 AChE and 53 BACE1 inhibitors. Statistically significant models were developed for AChE ($R^2 = 0.8688$, $q^2 = 0.8600$) and BACE1 ($R^2 = 0.8177$, $q^2 = 0.7888$) enzyme inhibitors. Each model was generated with an optimum five significant molecular descriptors such as electro-topological (ES_Count_aaCH and ES_Count_dssC), structural (QED_HBD, Num_Terminal-Rotomers), spatial (JURS_FNSA_1) for AChE and structural (Cl_Count, Num_Terminal Rotomers), electro-topological (ES_Count_dO), electronic (Dipole_Z) and spatial (Shadow_nu) for BACE1 enzyme, determining the key role in its enzyme inhibitory activity. The predictive ability of the generated machine learning models was validated using the leave-one-out, Fischer (*F*) statistics and predictions based on the test set of 11 AChE ($r^2 = 0.8469$, $r^2_{\text{pred}} = 0.8138$) and BACE1 ($r^2 = 0.7805$, $r^2_{\text{pred}} = 0.7128$) inhibitors. Further, nonlinear machine learning methods such as ANN and SVM predicted better than the linear method GFA. These molecular descriptors are very important in describing the inhibitory activity of AChE and BACE1 enzymes and should be used further for the rational design of multi-targeted anti-Alzheimer's lead molecules.

✉ C. Gopi Mohan
cgmohan@aims.amrita.edu; cgopimohan@yahoo.com

[1] Computational Biology and Bioinformatics Lab, Center for Nanosciences and Molecular Medicine, Amrita Vishwa Vidyapeetham, Kochi, Kerala 682041, India

Graphic abstract



Keywords Acetylcholinesterase · BACE1 · Alzheimer's disease · Descriptor · QSAR · SVM · ANN

Introduction

Alzheimer's disease (AD) is a multifactorial disorder of the brain and is dependent on the person genetic, age and environment parameters. It is involved in disruption of the neuronal cell functions in cerebral cortex and hippocampus. AD initial signs include the continuous loss of intellectual capacity accompanied by short-term memory. This in turn leads to depression, aggression and other abnormal mental behavior [1]. Different hypotheses for AD pathogenesis were reported which include amyloid hypothesis, cholinergic hypothesis and tau hypothesis. These hypotheses play a significant role in understanding the AD severity. Due to its complex nature, researchers focused on developing multi-target directed drugs for curing AD. Progression of this disease is mainly associated with the cholinergic and amyloid hypothesis [2–4], which in turn causes enhanced production of Aβ peptide. Aggregation and accumulation of the peptide in the brain increase neuronal toxicity leading to nerve damage and degeneration, an intermediate stage of dementia. Production of Aβ peptide is by proteolytic processing of amyloid precursor protein (APP). This involves two signaling pathways: (1) cleavage of α- and γ-secretases generating soluble APP fragments and (2) sequential cleavage of β- and γ-secretases generating neurotoxic Aβ peptides. Further, loss of cholinergic neurotransmission by the decreased level of neurotransmitter causes Aβ aggregation in the brain leading to its damage and memory loss [5–7]. Thus, AChE and BACE1 were the most promising druggable targets for the treatment of AD [8–11]. FDA approved drugs such as tacrine, rivastigmine, galantamine and donepezil target AChE enzyme [5, 6]. AD treatments using these drugs are only symptomatic, and none of them have a complete cure or control on its progress. However, the complete AD cure or its precise treatment strategy is still unknown. Thus, in the present scenario there is an urgent need to develop new drugs or disease treatment strategies. Recently, different N-benzyl piperidine derivatives were reported to have potent inhibitory activity against both AChE and BACE1 enzymes [12, 13]. The AD drug discovery is very challenging, and the clinical trial towards BACE1 small molecule inhibitors raised many obstacles for drugs adverse effect. The multifaceted nature of AD should be taken into account while designing new inhibitors, which have the ability to

suppress multiple signaling pathways. So the novelty of the present work was developing machine learning models for the discovery of dual targeting inhibitors of BACE1 and AChE enzymes, and future clinical trial should focus on multi-targeting aspects for the treatment of AD more effectively with less adverse events.

Computer-aided molecular design techniques are successfully used in pharmaceutical research towards drug discovery program. Different metric systems from hit to lead identification including ADMET predictions were optimized in this process. Computer-aided design of new chemical entities, in the preclinical drug discovery stages, lead to reduction in the discovery time and cost. Quantitative structure–activity relationship (QSAR) is a key Chemometrics technique used nowadays in molecular modelling and drug discovery program. In this technique, molecular descriptors and its influence on biochemical and pharmacological effects are closely dependent. These in turn are computed from its chemical structure using different predictive machine learning methods. Thus, using the concept of 2D-QSAR one can rely on its prediction toward molecular activities and descriptors role in the pharmacological process. QSAR studies have established further the mechanisms of drug actions in different diseased targets [14–27]. From these observations, we performed QSAR studies using GFA, SVM and ANN machine learning methods to understand the AChE and BACE1 dual enzyme inhibitory activity.

Genetic function approximation (GFA) is a very popular method in QSAR studies. Rogers and Hopfinger developed GFA method, by employing the statistical analysis to compute molecular descriptors for generating different QSAR models. QSAR models' sensitivity analysis is the next step, and the best significant model developed was employed in predicting the test set molecules [10–14]. Several QSAR models were developed in the past to predict the potential inhibitors for AChE and BACE1 enzyme inhibition [10–20]. The predictive models in this reported work depend on the molecular mechanisms of action toward single enzyme, i.e., either AChE or BACE1. There are different computational strategies and methods reported, including machine learning, artificial intelligence and deep learning for designing potent anti-AD drug-like molecules [28–31]. It is important to mention at this juncture that very few research works were reported for dual targeting of AChE and BACE1 enzymes using small molecule drugs. So, the present work would be valuable in addressing the predictive models for dual targeting of AChE and BACE1 enzymes for the discovery of small molecule drugs against AD. The main objective of the present study was to determine the potential molecular descriptors involved in inhibitory activity in the structurally diverse class of AChE and BACE1 dual target inhibitors by developing highly robust and statistically significant 2D-QSAR models using different machine learning methods.

Materials and methods

Dataset for QSAR analysis

The biological dataset used for the present study includes N-benzyl piperidine analogs showing inhibitory activity against both AChE and BACE1 enzymes. We selected 57 molecules of N-benzyl piperidine analogs from studies reporting experimental inhibitory activity values (IC_{50}) expressed in micro-molar (μM) range [12, 13]. In order to optimize the biological activity, the IC_{50} value in micro-molar (μM) range was converted into molar (M) range, and its logarithm ($-\ pIC_{50}$) as the dependent variable, i.e., ($pIC_{50} = -\log_{10}IC_{50}$), was employed for developing the 2D-QSAR models. The chemical structure of the 57 molecules under study was built using ChemBio3D Ultra software version 14.0 [32], and their conformational energy was minimized using Merck Molecular Force Field (MMFF94), as shown in Supplementary material Fig. 1S. For developing 2D-QSAR model, the main dataset is partitioned into training and test sets based on the random selection method in DS software. Different sets of biological data in AChE and BACE1 were chosen on knowledge basis by taking into account the most potent and least potent compounds in AChE and BACE1 training and test datasets. This in turn addressed the maximum inhibitory activity divergence in different datasets. 2D-QSAR model for AChE was developed by partitioning the dataset into 46 training set and 11 test set molecules, while BACE1 was developed using 42 training set and 11 test set molecules. The chemical structure of these dataset molecules is depicted in Supplementary material Fig. 1S.

Molecular descriptor calculation and selection

We obtained 594 molecular descriptors (variables) from the AChE and BACE1 datasets containing 57 molecules (Supplementary material Fig. 1S) using module "Calculate Molecular Properties" in Discovery Studio (DS) software [33]. These include 1D, 2D and 3D molecular descriptors, for predicting the AChE and BACE1 enzyme inhibitory activity. At first, it was essential to decide the best possible number of variables that produce statistically significant QSAR models. Next step is to link the molecular structure of these compounds with their AChE and BACE1 inhibitory activity. In general, the 2D-QSAR model was developed using the training set compounds, and further, the developed models were validated using the test set and external set of compounds.

Correlation of the variables was computed which refers to the scaled form of covariance and could determine

both the strength and direction of the linear relationship between the two variables, i.e., pIC_{50} and molecular descriptors from AChE and BACE1 datasets. In the present study, the number of molecular descriptors used in developing 2D-QSAR equations with respect to the AChE and BACE1 inhibitory activities was initially optimized and the corresponding correlation matrix was computed. Correlation matrix shows the distribution of the magnitude and direction of the multivariate data and the highly correlated variables > 0.7 were not selected in order to prevent its over-fitting. Also, molecular descriptors with constant or zero values and the least correlated descriptors with the AChE and BACE1 enzyme inhibitory activity (pIC_{50}) were removed. Finally, 39 significant molecular descriptors among 594 were selected to develop the robust and predictive 2D-QSAR model for both AChE and BACE1 dual enzyme inhibitions. Further, using QSAR model equations the variance-adjusted coefficient and the predicted variance were computed for the training and test datasets of AChE and BACE1 enzymes described in results section.

2D-QSAR model development and validation

In the present study, our main aim was to develop dual-target directed ligand-based 2D-QSAR models for AChE and BACE1 enzymes using linear method (GFA) and two nonlinear methods (SVM and ANN). GFA is a combination of Holland's genetic algorithm (GA) and Friedman's multivariate adaptive regressive splines (MARS), which helps in selecting the variables randomly through linear method [34]. It works by the process of evolution, where the multiple models (equations) were generated with random number of variables (molecular descriptors). In the present study, under the "Create Genetic Function Approximation Model" protocol, 100 parent equations were generated randomly. Then, the pairs of these parent equations were used to generate the offspring equations by crossover operations. GFA study was performed using the DS software [33].

A support vector machine (SVM) is a supervised nonlinear machine learning methods developed by Vapnik and Cortes for the classification and regression problems implemented in WEKA software ver. 3.8.4 [35, 36]. It works on the principle of structural risk minimization to prevent the over-fitting problem by balancing the dataset complexity. SVM method showed robust prediction in the QSAR model development by introducing a parameter, epsilon-insensitive loss function. It was solved by sequential minimal optimization (SMO), an iterative algorithm proposed by Smola and Scholkopf, which computes the nonlinear regression problems. SVM regression models were finally developed by optimizing epsilon, gamma (RBF kernel) and complexity parameters [24, 36].

Artificial neural network (ANN) is a popular nonlinear machine learning method, based on the biological neural network implemented in WEKA software ver. 3.8.4. It has interconnected three neural layers, which include input, hidden and output layers. ANN takes information from the molecular descriptors through the input layer and moves the information to the hidden layer for processing, and finalized results were sent to the output layers. Based on the dataset volume, a number of hidden layers were increased or decreased [25, 36]. The present network architecture had five neurons as input layer, corresponding to five different molecular descriptors in the model selection. Output layer was fixed with one neuron and consists of AChE and BACE1 enzyme inhibitory activity values (pIC_{50}). Initially, the ANN parameters that include the number of descriptors, the number of hidden layers, the learning rate (η) and the momentum (μ) were optimized by the trial-and-error method by checking the RMSE value of the predicted model. Also μ and rate of weight adjustment will avoid abrupt RMSE changes, and using the learning parameter (η) the corresponding corrections were made for the ANN model. Finally**,** the best ANN model was judged by the high squared correlation coefficient (R^2) with a very low RMSE between the actual and predicted molecule inhibitory activity values (pIC_{50}) for AChE and BACE1 enzymes.

The quality-of-fit model determines the 2D-QSAR model robustness and its significance, which depend mainly on three statistical parameters: coefficient of determination (R^2), Pearson's correlation (R), explained variance (R^2_a) and degree of statistical confidence (F-test). Expressions for R^2, R^2_a and the F-test are shown below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_{\text{exp}} - \hat{Y}_{\text{pred}}\right)^2}{\sum_{i=1}^{n} \left(Y_{\text{exp}} - \hat{Y}_{\text{train}}\right)^2} \tag{1a}$$

$$R = \frac{\sum (Y_{\text{exp}} - \overline{Y}_{\text{exp}}).\left(\hat{Y}_{\text{pred}} - \overline{\hat{Y}}_{\text{pred}}\right)}{\sqrt{\sum (Y_{\text{exp}} - \overline{Y}_{\text{exp}})^2}.\sqrt{\sum (\hat{Y}_{\text{pred}} - \overline{\hat{Y}}_{\text{pred}})^2}} \tag{1b}$$

$$R^2_a = \frac{(N-1)R^2 - d - 1}{N - d - 1} \tag{2}$$

$$F = \frac{(R^2/d)}{[\left(1 - R^2\right)/(N - d - 1)]} \tag{3}$$

where Y_{exp}, $\hat{Y}_{\text{pred}}$ and $\overline{Y}_{\text{train}}$ are the experimental activity, predicted activity and average value of the experimental activity in the training set. N and d are the number of training set molecules and number of chemical descriptors of 2D-QSAR equation.

The external validation involves the use of external test set of 11 molecules on developed training set models for both AChE and BACE1 inhibitors. The predictive ability of the generated model was estimated by using different statistical parameters such as predictive correlation coefficient (R^2_{pred}), squared correlation coefficient between actual and predicted values with (r^2) and without intercept ($r_0{}^2$) of test set molecules,$r_0^{'2}$ is the squared correlation coefficient of predicted activity against experimental activity at zero intercept and their corresponding slopes. k is the slope of experimental activity against predicted activity values at zero intercept, k' is the slope of predicted activity value against experimental activity value at zero intercept. The r_m^2 metric is an important validation strategy in analyzing the models ability precisely that also have regression through origin methods, which includes-$r_m^{'2}, \overline{r_m^2}, \left|r_m^2\right|, (r^2-)/r^2 and (r^2-)/r^2$.

$$R^2_{pred} = 1 - \frac{\sum(Y_{\text{exp (test)}} - \hat{Y}_{\text{pred(test)}})^2}{\sum(Y_{\text{exp (test)}} - \overline{Y}_{\text{training}})^2} \tag{4}$$

$$r^2 = \frac{\left[\sum\left(y_{\text{exp}} - \bar{y}_{\text{exp}}\right).\sum\left(y_{\text{pred}} - \bar{y}_{\text{pred}}\right)\right]^2}{\sum(y_{\text{pred}} - \bar{y}_{\text{pred}})^2.\sum(y_{\text{exp}} - \bar{y}_{\text{exp}})^2} \tag{5}$$

$$r_0^2 = 1 - \frac{\sum(y_{\text{exp}} - k * \hat{y}_{\text{pred}})^2}{\sum(y_{\text{exp}} - \bar{y}_{\text{exp}})^2} \tag{6}$$

$$r_0^{'2} = 1 - \frac{\sum(y_{\text{pred}} - k' * y_{\text{exp}})^2}{\sum(y_{\text{pred}} - \bar{y}_{\text{pred}})^2} \tag{7}$$

$$k = \frac{\sum\left(y_{\text{exp}} * y_{\text{pred}}\right)}{\sum(y_{\text{pred}})^2} \tag{8}$$

$$k' = \frac{\sum\left(y_{\text{exp}} * y_{\text{pred}}\right)}{\sum(y_{\text{exp}})^2} \tag{9}$$

$$r_m^2 = r^2 * \left[1 - \sqrt{(r^2 - r_0^2)}\right] \tag{10}$$

$$r_m^{'2} = r^2 * \left[1 - \sqrt{(r^2 - r_0^{'2})}\right] \tag{11}$$

$$\overline{r_m^2} = \frac{\left(r_m^2 + r_m^{'2}\right)}{2} \tag{12}$$

$$\left|r_m^2\right| = \left|r_m^2 - r_m^{'2}\right| \tag{13}$$

where $Y_{exp(\text{test})}$, $\hat{Y}_{\text{pred(test)}}$ and $\overline{Y}_{\text{training}}$ are the experimental activity value of the test set, predicted activity value of the test set and mean value of the experimental activity value of the training set molecules, respectively [37, 38].

In 2D-QSAR technique, the cross-validation is a key method to understand the model robustness and its predictive ability. The evaluation of QSAR models using leave-one-out (LOO) or leave-many-out (LMO) methods is a standard procedure. In both these cross-validation methods, a single molecule or many molecules from the training dataset were removed at a time in a sequential manner. The QSAR model is further confirmed on the different data set, and the inhibitory activity value is predicted by the modified QSAR model. This computational protocol is repeated for the training data set molecules. Cross-validation results are further validated through R^2_{LOO}, PRESS and RMSE to reveal further the QSAR models robustness. Cross-validated correlation coefficient R^2_{LOO}, S_{PRESS} and RMSE were calculated as.,

$$R^2_{\text{LOO}} = 1 - \frac{\text{PRESS}}{\text{SSY}} \tag{14}$$

$$S_{\text{PRESS}} = \sqrt{\frac{\text{PRESS}}{(N - d - 1)}} \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{\text{PRESS}}{N}} \tag{16}$$

In Equations (14–16), the term PRESS denotes the predictive residual sum of squares, $\sum(Y_{\text{obs}} - Y_{\text{pred}})^2$ for the developed model; SSY is the sum of squared differences between the experimental activity and the average experimental activity of training set, $\sum(Y_{\text{obs}} - Y_{\text{aver.obs}})^2$. R^2_{LOO} or Q^2 value higher than (> 0.5) correlation shows that the developed 2D-QSAR model was robust with high predictive ability.

Results and discussion

2D-QSAR and 3D-QSAR techniques are extensively used in Chemoinformatics and computer-aided molecular design. Different case studies were successfully reported using these techniques for designing new chemical entities in the drug discovery process. 2D-QSAR technique can address both the linear and nonlinear cases, which are conformational independent as well as structurally different molecular scaffolds. It also does not depend on the molecular dataset alignment for structure–activity relationship studies. However, 3D-QSAR approaches mainly depend on the bioactive conformation toward its biological target, which can

be obtained by molecular docking technique and experimentally by X-ray crystallography, NMR or cryoelectron microscopy. It further depends on 3D alignment of the dataset molecules during the model development, such as comparative molecular field analysis CoMFA/comparative molecular similarity indices analysis-CoMSIA. 3D-QSAR method can further predict better the molecular mechanisms of ligand–receptor interactions. Also, the 2D-QSAR model involves a set of statistically significant molecular descriptors, which will reveal the mode of molecular interactions with respect to its target. However, this method does not take into account the 3D conformation or geometric features of the molecules, which would further limit understanding of the protein–ligand interaction phenomenon. Both these computational strategies are well established in the drug discovery program [39–44].

In the present study, we adopted different machine learning models to design potent inhibitors against AChE and BACE1 targets using linear and nonlinear methods. The novelty of the present work involves developing linear and nonlinear machine learning models using both AChE and BACE1 inhibitors containing dataset, which would in turn address in predicting the new dual targeted inhibitors by addressing the corresponding signaling pathways for AD therapy.

Genetic function approximation model

Multi-target directed ligand-based 2D-QSAR model was developed using genetic function approximation (GFA) method for AChE and BACE1 inhibition. For developing GFA model, a dataset of 57 AChE inhibitors was divided into 46 training set and 11 test set molecules and dataset of 53 BACE1 inhibitors was divided into 42 training set and 11 test set molecules, as shown in Supplementary material Fig. 1S. GFA model was developed using training set molecules, and its validation was done by the corresponding test set molecules. DS software was used to compute 594 molecular descriptors from these datasets. Among these descriptors, we removed 555 molecular descriptors showing high intercorrelation and zero standard deviation to check the models over-fitness and significance. In other words, 39 molecular descriptors showing significant influence on AChE and BACE1 enzyme inhibitory activity were finally selected for 2D-QSAR model development.

Using GFA method, the best 2D-QSAR model was finally selected on the basis of the statistical parameters- which include squared correlation coefficient (R^2), explained variance (R^2_{adj}), cross-validated correlation coefficient (q^2) and F-test. The total number of molecular descriptors in the 2D-QSAR model development was further optimized by sensitivity analysis.

Sensitivity study of 2D-QSAR model

Sensitivity of 2D-QSAR model was tested by varying the number of molecular descriptors and its effect on q^2 (cross-validated correlation coefficient) parameter [22, 39]. To optimize the total number of descriptors in the 2D-QSAR model, a set of five models were generated by varying the number of descriptors from three to seven. A graph was plotted between the number of molecular descriptors and parameter q^2. q^2 value showed an increase by increasing from three to five descriptors and later on, increasing further the descriptors, the q^2 value decreased showing the model's sensitivity in Fig. 1. Finally, optimum five significant molecular descriptors for developing robust 2D-QSAR models for AChE and BACE1 were identified (Fig. 1).

Using GFA method, the 2D-QSAR equation developed has five significant molecular descriptors for AChE enzyme inhibitory activity.,

$$\begin{aligned} \mathrm{pIC}_{50} = {} & 4.4751 - 0.13146 * \mathrm{ES\_Count\_aaCH} \\ & - 0.59381 * \mathrm{ES\_Count\_dssC} \\ & + 1.1031 * \mathrm{QED\_HBD} + 0.29138 \\ & * \mathrm{Num\_Terminal\ Rotomers} + 5.2785 \\ & * \mathrm{Jurs\_FNSA\_1} \end{aligned} \tag{17}$$

$$N_{(\text{Training set})} = 46, R^2 = 0.8190,\ \text{Pearson} - R = 0.9050,$$
$$R^2_{\text{adj}} = 0.7964, q^2 = 0.7617, \text{RMSE} = 0.2612,$$
$$F - \text{test} = 36.2$$

$$n_{(\text{test})} = 11,\ r^2 = 0.8067, R^2_{\text{pred}} = 0.8052, r^2_0 = 0.8057, r^2_m = 0.7221$$

Using the GFA method, the 2D-QSAR equation developed has five significant molecular descriptors for inhibiting the activity of BACE1 enzyme.,

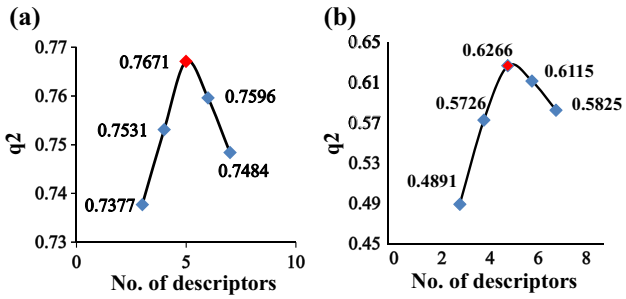


Fig. 1 Sensitivity test of 2D-QSAR models by varying the number of descriptors and its corresponding q^2 using AChE (**a**) and BACE1 (**b**) enzyme inhibitors. Red dot indicates an optimum number of molecular descriptors with high q^2 value for developing the best 2D-QSAR model

$$\begin{aligned} \mathrm{pIC}_{50} &= 4.1576 + 0.25297 * \mathrm{Cl\_Count} - 0.33064 * \mathrm{ES\_Count\_dO} \\ &\quad + 0.94614 * \mathrm{Num\_Terminal\ Rotomers} \\ &\quad - 0.11168 * \mathrm{Dipole\_Z} + 0.49099 * \mathrm{Shadow\_nu} \end{aligned} \tag{18}$$

$$N_{(\text{Training set})} = 42, R^2 = 0.7436, \text{Pearson} - R = 0.8623,$$
$$R^2_{\text{adj}} = 0.6802, q^2 = 0.6266,$$
$$\text{RMSE} = 0.2981, F - \text{test} = 20.89$$

$$n_{(\text{test})} = 11,\ r^2 = 0.7363, R^2_{\text{pred}} = 0.6730, r_0^2 = 0.7235, r_m^2 = 0.6530$$

R^2 is the training set squared correlation coefficient, R^2_{adj} is the square of adjusted correlation coefficient, q^2 is the cross-validated correlation coefficient of the training set, and F-test is the related variance statistics. The correlation matrix showing significance of five molecular descriptors is presented in Table 1 for developing the BACE1 2D-QSAR model.

Equations (17) and (18) explain 79.6% and 68% (variance-adjusted coefficient), and predicted variance of the AChE and BACE1 dataset was 76.1% and 62.6%. The inter-correlation descriptors (variables) in these datasets were checked for over-fitting. The best five selected significant molecular descriptors in AChE and BACE1 2D-QSAR equations are less correlated as shown in Tables 1 and 2. Scatter plot of the actual vs predicted pIC50 activity values for the AChE and BACE1 inhibitors training and test sets are depicted in Fig. 2.

Predictive correlation (R^2_{pred}) value for the AChE test dataset was 0.756 and for BACE1 test dataset was 0.6730. Test set molecules actual and predicted values with (r^2) and without (r_0^2) intercepts are 0.8067 and 0.8057, for AChE inhibitors and 0.7363 and 0.7235 for BACE1 inhibitors. These indicate that there are some differences in the r_m^2 values 0.7811 and 0.6530 for AChE and BACE1 datasets. The above statistical analysis clearly depicts that the AChE and BACE1 inhibitory activity prediction values from GFA method (Eqs. 17 and 18) for the test dataset molecules are in better agreement with the actual (experimental) values, shown in Tables 3 and 4.

Chemical structure of some of the representative most active and least active molecules from the AChE and BACE1 inhibitors datasets reported by Sharma et al. [12, 13] to understand its structure–activity relationships is shown in Fig. 3.

Most active molecules

Support vector machine model

Support vector machine (SVM) models were developed by using the five significant molecular descriptors selected by the GFA method with the same AChE and BACE1 training and test dataset. To develop an accurate SVM-QSAR model, the capacity C (a regularization parameter), epsilon (noise variance) and gamma (RBF kernel parameter) were optimized by checking the corresponding lowest RMSE values. An optimum value of C, epsilon (ε) and gamma (γ) was obtained using the L5O cross-validation method for SVM model. Initially, these three parameters were fixed to default values and the C value was varied from 0.1 to 1000 for its optimization with the lowest RMSE value. The γ value was changed from 0.001 to 0.035, and γ versus RMSE plot shown in Fig. 4a. An optimum value for γ 0.027 for the lowest RMSE value was obtained by choosing an optimum value $C = 10$. To optimize the ε parameter, it was varied from 0.01 to 0.1 and the corresponding RMSE values are plotted in Fig. 4b. Finally an optimum value for ε fixed was 0.076, by choosing $C = 10$. Finally, for the best SVM model prediction the optimized values of γ, ε and C are 0.027, 0.076 and 10.

AChE training data set inhibitory activity (pIC_{50}) value by SVM model showed statistically significant $R^2 = 0.8204$, Pearson-$R = 0.9058$, $q^2 = 0.7628$ and RMSE = 0.2471 predictions, while for its test set molecules the R^2_{pred} was 0.8137. Further, AChE test set molecules r^2 between its experimental and predicted values with r^2 and without r_0^2 intercepts obtained were 0.8175 and 0.814. Test set molecules r_m^2 were 0.7697, indicating some changes in the numerical value between its actual and predicted values. The experimental inhibitory activity vs predicted activity (pIC_{50}) of the AChE training and test data set molecules is based on SVM model shown in Table 3.

Table 1 The correlation matrix of the molecular descriptors computed by GFA method for developing 2D-QSAR model using AChE inhibitors

Property	ES_Count_aaCH	ES_Count_dssC	QED_HBD	Num_TerminalRotomers	JURS_FNSA_1
ES_Count_aaCH	1				
ES_Count_dssC	− 0.0330	1			
QED_HBD	0.1280	− 0.4954	1		
Num_Terminal Rotomers	0.2837	0.1424	0.0110	1	
JURS_FNSA_1	0.1227	0.1814	0.0683	0.5701	1

Table 2 Experimental AChE inhibitory activity (pIC_{50}) and the predicted activity value of molecules along with the residuals calculated using GFA, SVM and ANN methods

Molecule no.	Exp. IC_{50} (μM)[a]	Actual pIC_{50}	Predicted pIC_{50} (GFA)	Residual (GFA)	Predicted pIC_{50} (SVM)	Residual (SVM)	Predicted pIC_{50} (ANN)	Residual (ANN)
Training set (46)								
19	15	4.823	5.134	− 0.311	5.196	− 0.373	4.934	− 0.111
20	2.8	5.552	5.618	− 0.066	5.658	− 0.106	5.572	− 0.02
22	2.6	5.585	5.688	− 0.103	5.700	− 0.115	5.654	− 0.069
23	15	4.823	5.14	− 0.317	5.217	− 0.394	5.174	−0.351
24	4.4	5.356	5.105	0.251	5.185	0.171	5.108	0.248
25	0.9	6.045	6.045	0.000	6.067	− 0.022	5.887	0.158
27	1.0	6.000	6.317	− 0.317	6.315	− 0.315	6.269	− 0.269
28	1.5	5.823	5.621	0.202	5.648	0.175	5.624	0.199
29	1.3	5.886	6.098	− 0.212	6.073	− 0.187	5.79	0.096
30	2.9	5.537	5.469	0.068	5.511	0.026	5.557	− 0.02
31	1.3	5.886	5.685	0.201	5.695	0.191	5.652	0.234
32	2.2	5.657	5.412	0.245	5.47	0.187	5.341	0.316
33	1.7	5.769	5.519	0.25	5.582	0.187	5.525	0.244
34	2.2	5.657	5.553	0.104	5.59	0.067	5.698	− 0.041
35	2.4	5.619	5.478	0.141	5.521	0.098	5.517	0.102
36	4.0	5.397	5.567	− 0.17	5.601	− 0.204	5.574	− 0.177
37	2.3	5.638	5.352	0.286	5.402	0.236	5.483	0.155
39	3.7	5.431	5.536	− 0.105	5.586	− 0.155	5.514	− 0.083
40	0.72	6.142	6.502	− 0.36	6.463	− 0.321	6.394	− 0.252
41	0.11	6.958	6.659	0.299	6.605	0.353	6.713	0.245
42	0.59	6.229	6.73	− 0.501	6.669	− 0.44	6.851	− 0.622
44	0.62	6.207	6.553	− 0.346	6.49	− 0.283	6.692	− 0.485
45	1.2	5.92	5.951	− 0.031	5.955	− 0.035	5.705	0.215
46	0.71	6.148	6.157	− 0.009	6.138	0.01	6.111	0.037
5c	1.56	5.806	5.771	0.035	5.807	− 0.001	5.679	0.127
5d	2.03	5.692	5.433	0.259	5.505	0.187	5.407	0.285
5e	4.2	5.376	5.387	− 0.011	5.471	− 0.095	5.299	0.077
5f	2.1	5.677	5.785	− 0.108	5.817	− 0.14	5.702	− 0.025
5 g	1.78	5.749	5.915	− 0.166	5.934	− 0.185	5.892	− 0.143
5 h	1.05	5.978	5.985	− 0.007	5.996	− 0.018	6.024	− 0.046
6a	1.32	5.879	5.791	0.088	5.821	0.058	5.631	0.248
6b	0.783	6.106	6.247	− 0.141	6.221	− 0.115	5.823	0.283
6c	0.149	6.826	6.731	0.095	6.67	0.156	6.853	− 0.027
6d	0.264	6.578	6.351	0.227	6.308	0.27	6.381	0.197
6e	0.828	6.081	6.295	− 0.214	6.264	− 0.183	5.85	0.231
6f	0.075	7.124	6.747	0.377	6.665	0.459	6.957	0.167
6 g	0.055	7.259	6.819	0.44	6.749	0.51	7.001	0.258
10d	0.356	6.448	6.373	0.075	6.328	0.12	6.415	0.033
10e	1.64	5.785	6.358	− 0.573	6.321	− 0.536	5.891	− 0.106
10f	0.086	7.065	6.734	0.331	6.654	0.411	6.942	0.123
10 g	0.144	6.841	6.88	− 0.039	6.803	0.038	7.083	− 0.242
10 h	0.119	6.924	6.836	0.088	6.764	0.16	7.025	− 0.101
13b	1.37	5.863	5.77	0.093	5.784	0.079	5.66	0.203
13d	3.49	5.457	5.922	− 0.465	5.912	− 0.455	5.729	− 0.272
13 g	0.207	6.684	6.519	0.165	6.499	0.185	6.68	0.004
13 h	0.162	6.79	6.515	0.275	6.495	0.295	6.672	0.118

Table 2 (continued)

Molecule no.	Exp. IC_{50} (μM)[a]	Actual pIC_{50}	Predicted pIC_{50} (GFA)	Residual (GFA)	Predicted pIC_{50} (SVM)	Residual (SVM)	Predicted pIC_{50} (ANN)	Residual (ANN)
Test Set (11)								
17	15	4.823	5.023	− 0.200	5.103	− 0.28	4.823	0.000
18	2.1	5.677	5.059	0.618	5.161	0.516	5.135	0.542
21	4.8	5.318	5.756	− 0.438	5.78	− 0.462	5.638	− 0.32
26	0.44	6.356	6.164	0.192	6.175	0.181	6.028	0.328
38	4.1	5.387	5.555	− 0.168	5.602	− 0.215	5.529	− 0.142
43	0.85	6.07	6.109	− 0.039	6.097	− 0.027	5.760	0.31
6 h	0.096	7.017	6.938	0.079	6.855	0.162	7.149	− 0.132
10a	1.70	5.769	5.863	− 0.094	5.885	− 0.116	5.645	0.124
10b	1.05	5.978	6.244	− 0.266	6.219	− 0.241	5.822	0.156
10c	0.222	6.653	6.748	− 0.095	6.685	− 0.032	6.884	− 0.231
13c	0.235	6.628	6.354	0.274	6.349	0.279	6.339	0.289

[a]Experimental AChE inhibitory activity of molecules from Ref. [12, 13]

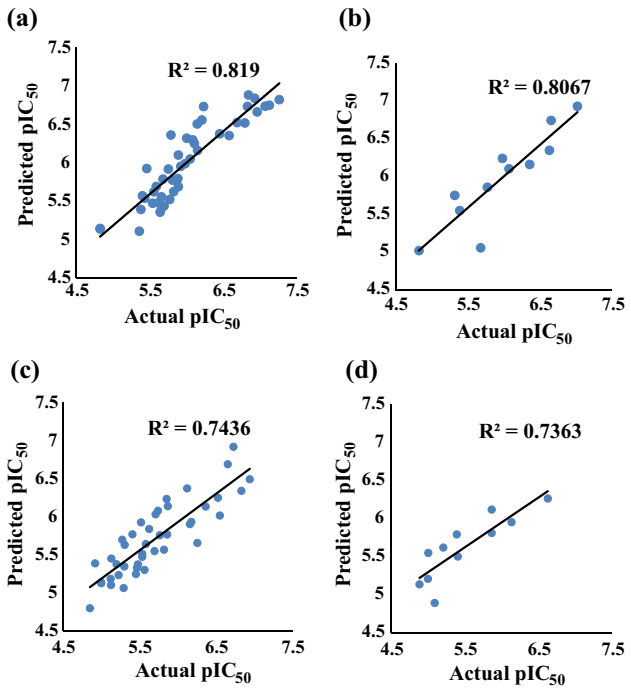


Fig. 2 Scatter plot of the actual vs predicted pIC_{50} activity values for the AChE inhibitors training set (**a**) and test set (**b**) and BACE1 inhibitors training set (**c**) and test set (**d**)

Similarly, SVM-based QSAR model was also developed for BACE1 inhibitors consisting of 42 training set and 11 test set molecules. The model optimization was based on L5O cross-validation method by varying the capacity parameter C, gamma (γ) and epsilon (ε). The C value was varied from 0.1 to 1000 to get an optimized value with the lowest RMSE value. By fixing the optimized value $C = 100$, the γ value was varied from 0 to 0.02 and the corresponding RMSE value was computed. A graph was plotted between γ and RMSE showing the optimized γ value of 0.016 for developing an SVM model with the least RMSE value shown in Fig. 4c.

The epsilon (ε) value was varied from 0 to 0.2, and 0.12 was chosen as an optimized value having the least RMSE value shown in Fig. 4d. Further, SVM model is developed using training dataset of 42 BACE1 inhibitors with statistically significant parameters of $R^2 = 0.7719$, Pearson-$R = 0.8785$, $q^2 = 0.7664$ and RMSE = 0.2635. The optimized SVM model was further checked to predict the activity of the BACE1 test set molecules and R^2_{pred} obtained was 0.7272. Actual and predicted values of the squared correlation coefficient (r^2) of the test set molecules with (r^2) and without (r_0^2) intercepts are 0.7623 and 0.7553, respectively. The test set molecules r_m^2 (0.7697) show the numerical relationship between the actual and predicted values. Tables 3 and

Table 3 The correlation matrix of the molecular descriptors computed by the GFA method for developing 2D-QSAR model using BACE1 inhibitors

Property	Cl_Count	ES_Count_dO	Num_TerminalRotomers	Dipole_Z	Shadow_nu
Cl_Count	1				
ES_Count_dO	− 0.1001	1			
Num_TerminalRotomers	− 0.3001	0.5719	1		
Dipole_Z	− 0.0097	0.3351	0.2550	1	
Shadow_nu	0.2112	0.1685	0.1185	− 0.1737	1

Table 4 Experimental BACE1 inhibitory activity (pIC_{50}) and the predicted activity value of molecules along with the residuals calculated using GFA, SVM and ANN methods

Molecule No.	Exp. IC_{50} (μM)[a]	Actual pIC_{50}	Predicted pIC_{50} (GFA)	Residual (GFA)	Predicted pIC_{50} (SVM)	Residual (SVM)	Predicted pIC_{50} (ANN)	Residual (ANN)
Training Set (42)								
17	6.3	5.200	5.372	− 0.172	5.339	− 0.139	5.498	− 0.298
18	3.4	5.468	5.324	0.144	5.286	0.182	5.438	0.030
19	14	4.853	4.795	0.058	4.706	0.147	4.722	0.131
20	2.7	5.568	5.298	0.270	5.266	0.302	5.416	0.152
21	12	4.920	5.384	− 0.464	5.348	− 0.428	5.504	− 0.584
23	3.3	5.481	5.370	0.111	5.339	0.142	5.501	−0.020
26	0.28	6.552	6.011	0.541	5.981	0.571	6.098	0.454
27	1.8	5.744	6.076	− 0.332	6.005	− 0.261	6.015	− 0.271
28	1.7	5.769	5.753	0.016	5.734	0.035	5.925	− 0.156
29	3.9	5.408	5.765	− 0.357	5.661	− 0.253	5.515	− 0.107
30	1.5	5.823	5.564	0.259	5.554	0.269	5.661	0.162
31	3.5	5.455	5.242	0.213	5.203	0.252	5.322	0.133
32	5.3	5.275	5.695	− 0.420	5.710	− 0.435	5.829	− 0.554
33	2.9	5.537	5.471	0.066	5.420	0.117	5.543	− 0.006
35	2.0	5.698	5.545	0.153	5.506	0.192	5.59	0.108
37	5.9	5.229	5.231	−0.002	5.177	0.052	5.281	− 0.052
38	2.6	5.585	5.637	− 0.052	5.596	− 0.011	5.615	− 0.03
40	0.43	6.366	6.13	0.236	6.114	0.252	6.233	0.133
41	0.22	6.657	6.685	− 0.028	6.618	0.039	6.924	− 0.267
43	0.66	6.180	5.928	0.252	5.929	0.251	6.243	− 0.063
44	3.0	5.522	5.920	− 0.398	5.864	− 0.342	5.628	− 0.106
45	0.55	6.259	5.652	0.607	5.622	0.637	5.637	0.622
46	2.9	5.537	5.507	0.030	5.457	0.080	5.561	− 0.024
5c	5.07	5.294	5.061	0.233	5.043	0.251	5.287	0.007
5e	10.0	5.000	5.123	− 0.123	5.087	− 0.087	5.132	− 0.132
5f	4.97	5.303	5.344	− 0.041	5.290	0.013	5.421	− 0.118
5 h	1.37	5.863	5.763	0.100	5.700	0.163	5.775	0.088
6a	7.33	5.134	5.447	− 0.313	5.387	− 0.253	5.519	− 0.385
6c	2.35	5.628	5.836	− 0.208	5.798	− 0.170	5.745	− 0.117
6 g	0.146	6.835	6.336	0.499	6.313	0.522	6.482	0.353
6 h	1.34	5.872	6.134	− 0.262	6.123	− 0.251	6.232	− 0.360
10a	4.92	5.308	5.626	− 0.318	5.557	− 0.249	5.581	− 0.273
10b	1.94	5.712	6.027	− 0.315	5.964	− 0.252	5.850	− 0.138
10c	1.39	5.856	6.228	− 0.372	6.116	− 0.260	6.166	− 0.310
10e	0.296	6.528	6.246	0.282	6.275	0.253	6.718	− 0.190
10f	0.114	6.943	6.488	0.455	6.560	0.383	7.301	− 0.358
10 g	0.186	6.730	6.915	− 0.185	6.774	− 0.044	6.89	− 0.160
10 h	0.751	6.124	6.369	− 0.245	6.335	− 0.211	6.496	− 0.372
13b	7.51	5.124	5.18	− 0.056	5.118	0.006	5.184	− 0.060
13c	7.44	5.128	5.099	0.029	5.121	0.007	5.157	− 0.029
13d	10.0	5	5.132	− 0.132	5.079	− 0.079	5.113	− 0.113
13e	0.69	6.161	5.899	0.262	5.896	0.265	6.194	− 0.033
Test Set (11)								
22	10.0	5	5.208	− 0.208	5.175	− 0.175	5.279	− 0.279
25	0.73	6.136	5.952	0.184	5.994	0.142	6.072	0.064
34	13.0	4.886	5.135	− 0.249	5.041	− 0.155	5.076	− 0.190

Table 4 (continued)

Molecule No.	Exp. IC_{50} (μM)[a]	Actual pIC_{50}	Predicted pIC_{50} (GFA)	Residual (GFA)	Predicted pIC_{50} (SVM)	Residual (SVM)	Predicted pIC_{50} (ANN)	Residual (ANN)
36	9.90	5.004	5.547	− 0.543	5.500	− 0.496	5.580	−0.576
5d	8.10	5.091	4.891	0.200	4.856	0.235	4.939	0.152
5 g	1.36	5.866	5.809	0.057	5.732	0.134	5.766	0.100
6b	6.17	5.209	5.619	− 0.410	5.587	− 0.378	5.626	− 0.417
6d	3.93	5.405	5.501	− 0.096	5.464	− 0.059	5.577	− 0.172
6e	4.06	5.391	5.789	− 0.398	5.764	− 0.373	5.862	− 0.471
6f	0.236	6.627	6.262	0.365	6.291	0.336	6.466	0.161
10d	1.36	5.866	6.117	− 0.251	6.090	− 0.224	6.163	− 0.297

[a]Experimental BACE1 inhibitory activity of molecules from Ref. [12, 13]

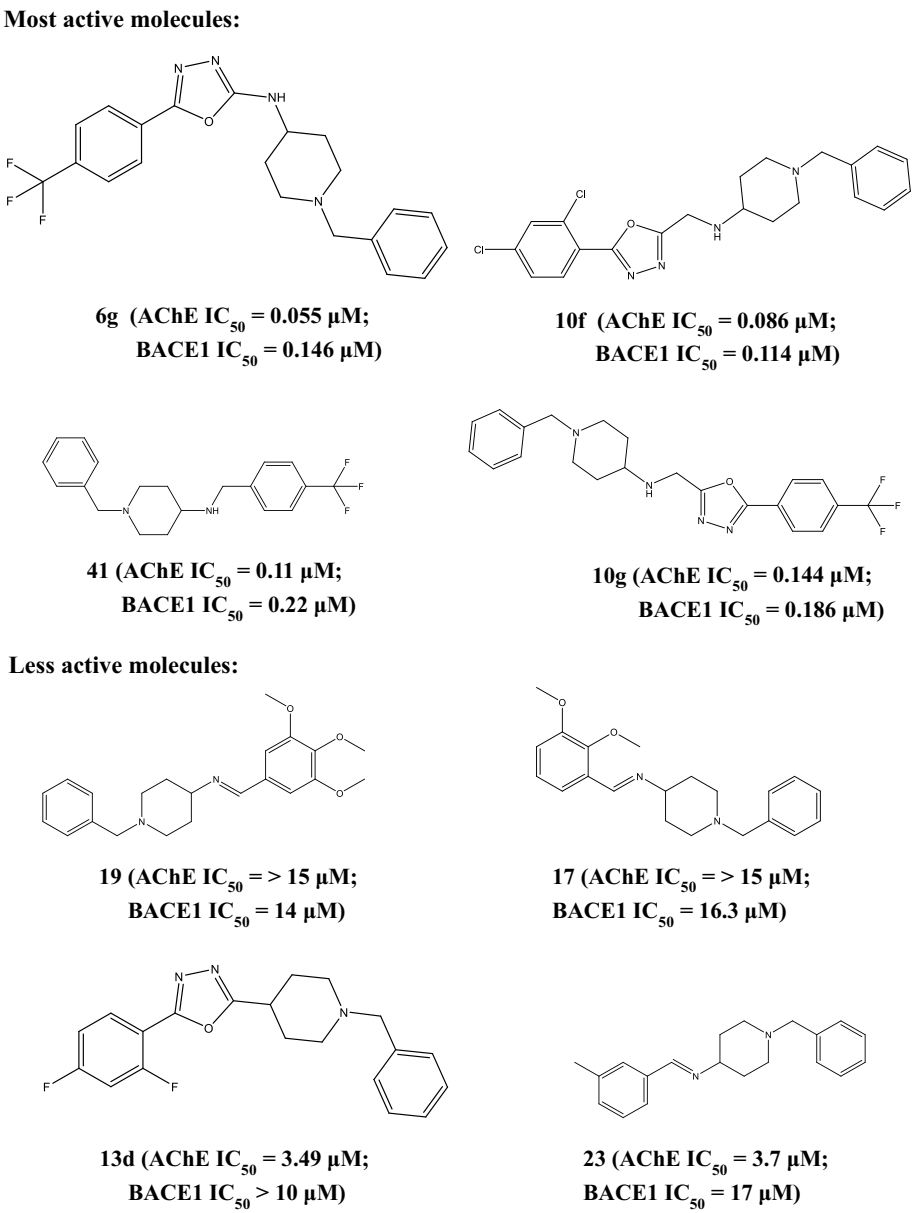


Fig. 3 Some of the representative most active molecules (**6 g, 10f, 41, 10 g**) and less active molecules (**19, 17, 13d, 23**) chemical structure from AChE and BACE1 inhibitors datasets [12, 13]

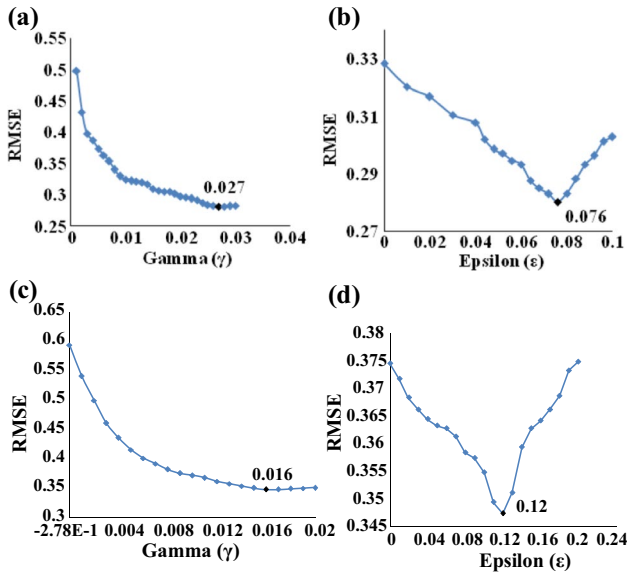


Fig. 4 Using SVM method, the variation of γ versus RMSE (C = 10, γ = 0.027) (**a**) and epsilon (ε) versus RMSE ($C = 10$, $\varepsilon = 0.076$) (**b**) of AChE inhibitors training dataset and variation of γ versus RMSE ($C = 100$, $\gamma = 0.016$) (**c**) and epsilon (ε) versus RMSE ($C = 100$, $\varepsilon = 0.12$) (**d**) of BACE1 inhibitors training dataset

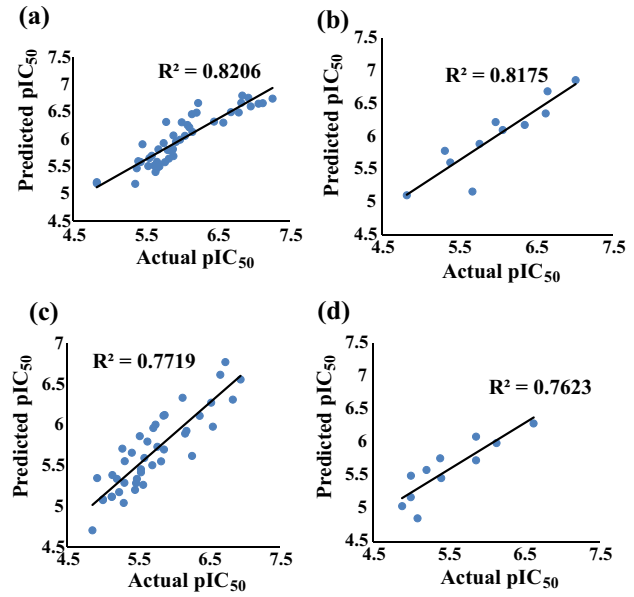


Fig. 5 SVM method, scatter plot of actual vs predicted pIC_{50} activity values for AChE training (**a**) and test set (**b**) molecules and BACE1 training (**c**) and test set (**d**) molecules

4 show the experimental and SVM model-based predicted activity of the AChE and BACE1 training and test dataset molecules, and the corresponding scatter plot is shown in Fig. 5a–d.

Artificial neural network model

Artificial neural network (ANN), a nonlinear machine learning method for developing QSAR model, is the most popular technique nowadays. In the present study, structurally diverse AChE and BACE1 dual inhibitors were used, for developing the robust 2D-QSAR model using ANN technique. Initially, ANN method involves three neuron layers to build the QSAR model using training set molecules. ANN parameters such as number of hidden layers, learning rate (η), and momentum (μ) were optimized before performing the enzyme inhibitory activity prediction. Using training data set, an optimum value of η and μ parameter was determined by L5O cross-validation showing the lowest RMSE value. Further, parameter with lowest RMSE was identified for an optimal value of $\eta = 0.3$ and $\mu = 0.2$ for the ANN model development. Number of neurons in the ANN model hidden layer was computed by changing from 2 to 10. An optimal value of 3 neurons hidden layer gave the lowest RMSE value. Using an optimized ANN model, the training data set showed a statistically significant $R^2 = 0.8688$, Pearson-$R = 0.9321$, RMSE = 0.2143 and $q^2 = 0.8600$ values, while the R^2_{pred} value for the test dataset was 0.8138. Correlation coefficient (r^2) for the test set molecules obtained was 0.8469, and at intercept zero the corresponding correlation coefficient was $r_0^2 = 0.8274$. It was interesting to note that r^2 and r_0^2 were close to each other, and a satisfactory value of r_m^2 (0.7284) was also obtained.

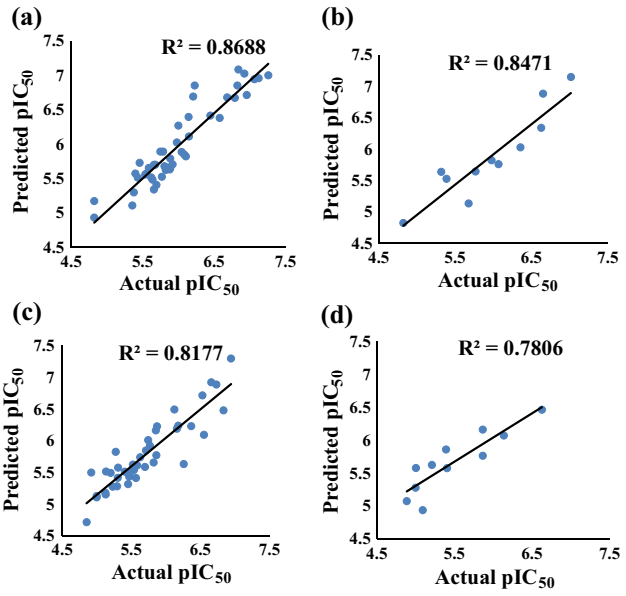


Fig. 6 ANN method, scatter plot of actual vs predicted pIC_{50} activity values for AChE training (**a**) and test set (**b**) molecules and BACE1 training (**c**) and test set (**d**) molecules

By using the same optimization parameters, $\eta = 0.3$, $\mu = 0.2$ and hidden neuron layer = 3, ANN models were developed for BACE1 inhibitors. Using an optimized ANN model, the BACE1 training dataset showed $R^2 = 0.8177$, Pearson-$R = 0.9042$, RMSE = 0.2505 and $q^2 = 0.7888$ values, and R^2_{pred} value was 0.7128 for the test dataset. Also, r^2 value for the test set molecules was 0.7805 and r_0^2 on intercept setting to zero was 0.7748. Significant value of r_m^2 (0.7216) was also obtained, as the value of r^2 and r_0^2 was not much different. Tables 3 and 4 show the experimental and ANN model-based predicted activity of AChE and BACE1 training and test dataset molecules, and the corresponding scatter plot is shown in Fig. 6a–d.

2D-QSAR model validation for AChE and BACE1 inhibitors

The significance of the developed linear (GFA) and nonlinear (SVM, ANN) QSAR models were evaluated using different cross-validation techniques which include cross-validated (R^2_{LOO} or q^2) method, predictive residual error sum of squares (PRESS), S_{PRESS} and RMSE. These machine learning models with predictive ability and its robustness for AChE and BACE1 inhibitors dataset are shown in Table 5.

In addition to the cross-validation study, the machine learning models were subjected to other key validation parameters using the test dataset. The calculation of r^2 between the experimental and predicted biological activity of the test dataset measures the strength of the developed machine learning models, and it should be close to 1. Similarly, different statistical parameters were calculated: regression coefficient through the origin, whether to check r'^2 or $r'_0{}^2$ close to r^2. The regression slope at the origin (k or k') was close to 1. The calculated parameters for the different machine learning models were satisfied for the following conditions, i.e., the correlation coefficient (r^2) > 0.6, cross-validated correlation (R^2_{pred} or q^2) > 0.5, r'^2 or $r'_0{}^2$ close to r^2, such that $\overline{r_m^2} > 0.5$, $|r'_m{}^2| < 0.2$, $(r^2 - r'_o{}^2)/r^2$ or $(r^2 - r_o'^2)/r^2$ should be less than 0.1. Also, k or k' is in the range of 0.85 to 1.15 for developing the good predictive models [37, 38] and is presented in Table 6 for AChE and BACE1 inhibitors.

Importance of molecular descriptors

Nine key molecular descriptors belonging to four different classes were generated by GFA method using AChE and BACE1 inhibitors, which include electro-topological (ES_Count_aaCH and ES_Count_dssC, ES_Count_dO), structural (QED_HBD, Cl_Count, Num_Terminal Rotomers), spatial (JURS_FNSA_1, Shadow_nu), and electronic (Dipole_Z). Among these, QED_HBD, Cl_Count, Num_Terminal Rotomers, JURS_FNSA_1, and Shadow_nu are correlated positively, while ES_Count_aaCH, ES_Count_dssC, ES_Count_dO and Dipole_Z molecular descriptors correlated negatively with the AChE and BACE1 enzymes inhibitory activity. Thus, QSAR equations generated have significant molecular descriptors belonging to 1D, 2D and

Table 5 LOO/L5O cross-validation parameter from GFA, SVM and ANN techniques for AChE and BACE1 inhibitors

2D-QSAR-Models	R^2 (Training set)	R^2_{LOO} or R^2_{L5O}	PRESS	S_{PRESS}	RMSE
AChE dataset					
GFA	0.8190	0.7617	2.729	0.068	0.2612
SVM	0.8204	0.7628	2.806	0.070	0.2471
ANN	0.8688	0.8600	2.111	0.052	0.2143
BACE1 dataset					
GFA	0.7436	0.6266	3.200	0.088	0.2981
SVM	0.7719	0.7664	2.916	0.081	0.2635
ANN	0.8177	0.7888	2.636	0.073	0.2505

Table 6 2D-QSAR model validation parameters for AChE and BACE1 inhibitors test dataset computed using GFA, SVM, and ANN methods

Statistical parameters	GFA AChE	BACE1	SVM AChE	BACE1	ANN AChE	BACE1
r^2	0.8067	0.7363	0.8175	0.7623	0.8469	0.7805
R^2_{pred}	0.8052	0.6730	0.8137	0.7272	0.8138	0.7128
r_o^2	0.8057	0.7235	0.8140	0.7553	0.8274	0.7748
r'^2_0	0.7750	0.6033	0.8133	0.6092	0.8466	0.7556
k	0.9975	0.9788	0.9969	0.9841	1.0126	0.9697
k'	1.0003	1.0189	1.0009	1.0138	0.9856	1.0291
r_m^2	0.7811	0.6530	0.7697	0.6985	0.7284	0.7216
r'^2_m	0.6631	0.4678	0.7645	0.4640	0.8300	0.6574
$\overline{r_m^2}$	0.7221	0.5603	0.7671	0.5813	0.7792	0.6895
$\lvert r_m^2 \rvert$	0.118	0.1852	0.0052	0.2345	0.101	0.0642
$(r^2 - r_o^2)/r^2$	0.0012	0.0173	0.0042	0.0092	0.0231	0.072
$(r^2 - r_o'^2)/r^2$	0.0393	0.180	0.0051	0.2008	0.0004	0.0318

3D category. Further, the protonation state of the molecules having piperidine moiety in the structure and conformation ensembles is broadly handled by 3D and higher dimensionality types of descriptors in the studied dataset. Molecular descriptors, which include Jurs_FNSA_1, QED_HBD, ES_Count_atom types and Shadow_nu, belong to the higher-dimensional category and could broadly handle these chemical features in the dataset. In addition, the electrotopological and structural class of molecular descriptors showed importance by appearing in both the GFA-based 2D-QSAR model of AChE and BACE1 dual-targeted inhibitors.

The electrotopological descriptor defines the combination of electronic features and topological environment for any given atoms and influences the involvement of its molecular fragments toward various endpoints [16, 39]. The electrotopological state indices descriptor; ES_Count_aaCH represents the total number of carbon atoms connected with hydrogen along with the 2 aromatic bonds; ES_Count_dssC the total number of carbon atoms connected with one double and two single bonds and ES_Count_dO the electro-topological state indices for the number of oxygen atoms connected with one double bond. Electronic descriptor Dipole_Z represents induced dipole moment along the Z_axis and indicates that higher polarizability of the solvent molecule increases the solubility. Spatial descriptor Jurs_FNSA_1 denotes the fractional charged partial surface areas and characterizes the molecules by its shape and electronic features. Shadow_nu descriptor, which is also spatial, describes the extent of molecule shadows by calculating the ratio of largest to smallest dimensions in the molecule along the XYZ axis. Cl_Count and Num_Terminal Rotomers belong to structural descriptor, which is positively correlated in our model and defines the 'Cl Element counts' and number of rotatable bonds in the molecules. Interestingly, Num_Terminal Rotomers is a common descriptor obtained in both AChE and BACE1 2D-QSAR models and describes the presence of non-terminal atoms which is connected by CF_3, CCl_3 or NOO in the molecules, showing higher activity toward these enzymes. QED_HBD, another structural descriptor, represents the quantitative estimation of the hydrogen bond contribution toward the biological activity. This in turn supports the molecular fragments at its terminal end as suggested by the positive correlation of the common descriptor Num_Terminal Rotomers present in both AChE and BACE1 2D-QSAR models.

Also, molecular descriptor importance can be understood from their QSAR model coefficients, and different types of interactions which include steric, electrostatic, lipophilic and hydrophobic forces do play a critical role in describing the binding affinity of AChE and BACE1 inhibitors. Electrotopological state atom indices and ES_Count_atom types encode electronic and topological information, revealing the importance of the degree of branching, connectivity of atoms, unsaturation in the molecule, which could be significantly correlated with the electronic properties (electrostatic and hydrophobic) of AChE and BACE1 inhibitors. Also, negative correlation of the electro-topological descriptors in the developed 2D-QSAR model can be influenced by the weak inhibitory activity of **17**, **19–24, 32** and **34–39** molecules, which have the presence of electron-donating group (–OH, CH_3 and OCH_3), respectively. However, in the BACE1 enzyme the molecules **20**, **23**, **35** and **38** showed better inhibitory activity when electron-donating groups are substituted at the ortho- and para-position and was further predicted well with our developed AChE and BACE1 QSAR models (Tables 3 and 4). Exception of these was seen in molecules **19** and **34,** in which substitution at the fourth position significantly reduced enzymes inhibition. It is interesting to note that the electron-withdrawing group (–Cl and –F) containing molecules (**25, 26**, **29, 31, 40, 41, 42, 44** and **46**) having di or tri substitutions showed marked improvements in AChE and BACE1 inhibition, which in turn can be influenced by the positive correlation of the Num_Terminal Rotomers and Halogen_Element counts molecular descriptor with the enzymes activity. These molecules are predicted well with our developed 2D-QSAR models (Tables 3 and 4). Finally, molecules **26** and **41** were selected for biochemical studies by Sharma et al. [12, 13], and this molecule showed excellent predicted activity residual values in our training and test AChE and BACE1 datasets (Tables 3 and 4). FinallyIn addition to these interesting observations, the present machine learning results are in consonance with the earlier report, indicating the electron-withdrawing group(s) of the atoms in different structural environment was essential for better AChE/BACE1 dual inhibitory activity, which in turn contribute in enhancing the blood–brain barrier uptake of drugs [12, 13, 15]. Further, these molecules are reported to have promising molecular interactions with the peripheral anionic site and catalytic active site of AChE enzyme and catalytic dyad residues (Asp32 and Asp228) in BACE1 enzyme obtained by molecular docking and dynamic simulations studies [12, 13].

In addition to these interesting observations, another series of AChE and BACE1 dual inhibitors were also studied by modification in the X-spacer and substitution of electron-withdrawing group(s) at the phenyl terminal end [12, 13]. These dual inhibitors were also included in our QSAR dataset to develop different machine learning models. Linker –NH series modification (**6a–6 h**), $NHCH_2$ series (**10a–10 h**) and direct linking (**13a–13 h**) molecules showed interesting inhibitory activity trend in the order of (**6a–6 h**) > (**10a–10 h**) > (**13a–13 h**) except **13e** molecule. Our machine learning models also predicted the corresponding activity of the molecules in a similar manner except the molecule **13e** as an outlier. Also, computed molecular spatial descriptor Jurs_FNSA_1 and structural descriptor Num_Terminal Rotomers

which is positively correlated in machine learning models do support the structure–activity relationships of these molecules. However, few of these molecules in the series (**6b–6e**, **10c**, **10d**, **13c**, **13 h** and **13e**) were shown week inhibitory activity [12, 13]. Further, the un-substituted molecules (**5a, 6a, 10a** and **13a**) showed moderate-to-poor activity, which in turn was supported by the positively correlated "halogen_count" descriptor in the models developed. Substitutions from 4-CN to 4-NO_2 electron-withdrawing group at the terminal phenyl group enhanced the inhibitory activity of the molecules **6f** and **10f** in which chlorine di-substitutions were made at positions 2 and 4; molecules with lipophilic and electron-withdrawing groups 4-CF_3, 4-OCF_3, 4NO_2, 4-CF_3 and 4-OCF_3 (**6 g, 10 g, 6 h** and **10 h; 13c, 13 g** and **13 h**) also showed higher inhibitory activity, which in turn was predicted well from our developed machine learning models (Tables 3 and 4). These results revealed that the computed descriptor Halogen_count seems to be highly significant in predicting the activity of the dataset molecules, which was in good agreement with its experimental activity values as described above. Further, molecules **6 g** and **10f** were found to be promising showing dual binding affinity against both AChE and BACE1 enzymes active sites by molecular docking and dynamics simulation studies [12, 13]. These results are in concordance with our machine learning models. Thus, the molecular descriptors obtained in our AChE and BACE1 targeted inhibitors could further be deciphered in designing better molecules, and the corresponding dual inhibitory activities can be predicted using the robust QSAR models developed in the present study.

Recently, Ponzoni et al. have reported diverse type of BACE1 inhibitors with different scaffolds covering a wide chemical space for robust QSAR classification model development using ANN, random forest (RF), and the random committee (RC) methods implemented in Weka software. They initially tested the selected compounds on the basis of Lipinski's rule of five using Qikprop module and applied the combination strategy of hybridization with forward/backward approach for improving the quality of the developed QSAR model. Also, the information redundancy in the model was tested using the most significant descriptors visualization along with the final model showing low cardinality with the best compound classification [17]. Wong et al. developed different QSAR models to analyze the AChE inhibitory activity of variety of datasets (A to D) containing Tacrine and its derivatives/hybrids. The experimental dataset of these new tacrine-related inhibitors used was measured under different experimental conditions from variety of sources. Interestingly, the QSAR model developed using different dragon descriptors was robust and has the predictive capacity using the linear approach of replacement method in all these developed models. These models were further validated using R^2_{loo} and S_{loo} (square of the correlation coefficient and standard deviation of leave-one-out) parameters and also the Y-Randomization method to measure the model stability and quality [20]. Both these groups adopted either BACE1 or AChE diverse class of inhibitors for developing the robust QSAR model, and the future users should implement applicability domain methods. In the present work, dual targeting strategy was adopted to identify inhibitors against both BACE1 and AChE enzymes. We also used three different machine learning strategies GFA, ANN and SVM from DS and WEKA software for the AChE and BACE1 dataset containing dual inhibitors, and the best model was obtained using the ANN method. Our developed models were validated using cross-validation and external test set molecules. These models also depend on an applicability domain method similar to Ponzoni et al. work. As a future work for predicting dual inhibitory activity, we plan to evaluate this machine learning methodology in the development of better models by incorporating the classification model concepts as well as reported earlier [17, 20]. This would enhance the development of 2D-QSAR models in much diverse class of AChE and BACE1 dual inhibitors. These integrated in silico methodologies would further enhance the virtual screening of large chemical space to discover more diverse scaffold containing dual binding site inhibitors.

Conclusions

Nowadays, machine learning modeling is an important established technique in the modern drug discovery program. It determines that the molecular structure changes can affect the physicochemical properties of a molecule leading to its efficacy. Dual-targeted inhibitor-based 2D-QSAR models were developed using GFA, SVM and ANN techniques against both AChE and BACE1 enzymes inhibition. The developed machine learning models were validated by cross-validation and external test set molecules. Among the four different classes of molecular descriptors, QED_HBD, Cl_Count, Num_Terminal Rotomers, JURS_FNSA_1, and Shadow_nu showed positive correlation, while ES_Count_aaCH, ES_Count_dssC, ES_Count_dO and Dipole_Z molecular descriptor showed negative correlation in developing the robust 2D-QSAR models using inhibitors of AChE and BACE1 enzymes. Further, the descriptor (Num_Terminal Rotomers), which was obtained commonly in both AChE and BACE1 enzymes based 2D-QSAR models emphasized the need for a higher number of rotatable bonds with electron-withdrawing groups in the drug-like molecules for its dual inhibition. Among the three different machine learning models developed, ANN method showed better predictive ability and statistical significance in comparison with GFA and SVM methodologies.

These theoretical models would be useful for the future rational design of multi-targeted inhibitors against both AChE and BACE1 enzymes. Models would also provide guidance for chemical synthesis of the best predicted molecules to undertake successfully its *in vitro/ in vivo* experimental validations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-021-10282-8.

Acknowledgements The authors acknowledge Computational biology and Bioinformatics Lab, Centre for Nanosciences and Molecular Medicine, Amrita Vishwa Vidyapeetham, Kochi, for the computing infrastructure support. CGM gratefully acknowledges the Department of Biotechnology (DBT), India, for providing financial support in procuring the Discovery studio license used in the present study (Grant No. BT/PR21018/BID/7/776/2016).

Declarations

Conflict of interest Authors declare no competing interests.

References

1. Hou Y, Dan X, Babbar M, Wei Y, Hasselbalch SG, Croteau DL, Bohr VA (2019) Ageing as a risk factor for neurodegenerative disease. Nat Reviews Neurol 15(10):565–581. https://doi.org/10.1038/s41582-019-0244-7
2. Wenk GL (2003) Neuropathologic changes in Alzheimer's disease. J Clini Psychia 64(suppl 9):7–10 (**PMID: 12934968**)
3. Mattson MP (2004) Pathways towards and away from Alzheimer's disease. Nature 430(7000):631–639. https://doi.org/10.1038/nature02621
4. Schelterns P, Feldman H (2003) Treatment of Alzheimer's disease; current status and new perspectives. Lanc Neurol 2(9):539–547. https://doi.org/10.1016/s1474-4422(03)00502-7
5. Briggs R, Kennelly SP, O'Neill D (2016) Drug treatments in Alzheimer's disease. Clini Med 16(3):247. https://doi.org/10.7861/clinmedicine.16-3-247
6. Conrado DJ, Duvvuri S, Geerts H, Burton J, Biesdorf C, Ahamadi M, Macha S, Hather G, Francisco Morales J, Podichetty J, Nicholas T, Stephenson D, Trame M, Romero K, Corrigan B (2020) Drug development tools in the alzheimer disease continuum (ddt-ad) working group. challenges in Alzheimer's Disease drug discovery and development: the role of modeling, simulation, and open data. Clin Pharmacol Ther 107(4):796–805. doi: https://doi.org/10.1002/cpt.1782.
7. Cacabelos R (2018) Have there been improvements in Alzheimer's disease drug discovery over the past 5 years? Expert Opin Drug Discov 13(6):523–538. https://doi.org/10.1080/17460441.2018.1457645
8. Cheung J, Rudolph MJ, Burshteyn F, Cassidy MS, Gary EN, Love J, Franklin MC, Height JJ (2012) Structures of human acetylcholinesterase in complex with pharmacologically important ligands. J Med Chem 55(22):10282–10286. https://doi.org/10.1021/jm300871x
9. Tran TS, Tran TD, Tran TH, Mai TT, Nguyen NL, Thai KM, Le MT (2020) Synthesis, In silico and in vitro evaluation of some flavone derivatives for acetylcholinesterase and BACE-1 inhibitory activity. Molecules 25(18):4064. https://doi.org/10.3390/molecules25184064
10. Huang W, Tang L, Shi Y, Huang S, Xu L, Sheng R, Wu P, Li J, Zhou N, Hu Y (2011) Searching for the multi-target-directed ligands against Alzheimer's disease: discovery of quinoxaline-based hybrid compounds with AChE, H3R and BACE 1 inhibitory activities. Bioorg Med Chem 19(23):7158–7167. https://doi.org/10.1016/j.bmc.2011.09.061
11. Dvir H, Silman I, Harel M, Rosenberry TL, Sussman JL (2010) Acetylcholinesterase: from 3D structure to function. Chemico-Biol Interact 187(1–3):10–22. https://doi.org/10.1016/j.cbi.2010.01.042
12. Sharma P, Tripathi A, Tripathi PN, Prajapati SK, Seth A, Tripathi MK, Srivastava P, Tiwari V, Krishnamurthy S, Shrivastava SK (2019) Design and development of multitarget-directed N-Benzylpiperidine analogs as potential candidates for the treatment of Alzheimer's disease. Europ J Med Chem 167:510–524. https://doi.org/10.1016/j.ejmech.2019.02.030
13. Sharma P, Tripathi A, Tripathi PN, Singh SS, Singh SP, Shrivastava SK (2019) Novel molecular hybrids of n-benzylpiperidine and 1, 3, 4-oxadiazole as multitargeted therapeutics to treat alzheimer's disease. ACS Chem Neurosci 10(10):4361–4384. https://doi.org/10.1021/acschemneuro.9b00430
14. Andersson CD, Hillgren JM, Lindgren C et al (2015) Benefits of statistical molecular design, covariance analysis, and reference models in QSAR: a case study on acetylcholinesterase. J Comput Aided Mol Des 29:199–215. https://doi.org/10.1007/s10822-014-9808-1
15. Bahuguna A, Bharatam PV, Rawat DS (2021) 3D QSAR studies on amphiphilic indoles for antimycobacterial activity. J Biochem Mol Toxicol 35(3):e22675. https://doi.org/10.1002/jbt.22675
16. Roy K, Mitra I (2012) Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. Curr Comput Aided Drug Des 8(2):135–158. https://doi.org/10.2174/157340912800492366
17. Ponzoni I, Sebastián-Pérez V, Martínez MJ et al (2019) QSAR Classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease. Sci Rep 9:9102. https://doi.org/10.1038/s41598-019-45522-3
18. Kumar V, Saha A, Roy K (2020) In silico modeling for dual inhibition of acetylcholinesterase (AChE) and butyrylcholinesterase (BuChE) enzymes in Alzheimer's disease. Comput Biol Chem 88:107355. https://doi.org/10.1016/j.compbiolchem.2020.107355
19. Roy KK, Dixit A, Saxena AK (2008) An investigation of structurally diverse carbamates for acetylcholinesterase (AChE) inhibition using 3D-QSAR analysis. J Mol Graph Model 27(2):197–208. https://doi.org/10.1016/j.jmgm.2008.04.006
20. Wong KY, Mercader AG, Saavedra LM et al (2014) QSAR analysis on tacrine-related acetylcholinesterase inhibitors. J Biomed Sci 21:84. https://doi.org/10.1186/s12929-014-0084-0
21. Anju CP, Subhramanian S, Sizochenko N, Melge AR, Leszczynski J, Mohan CG (2019) Multiple e-Pharmacophore modeling to identify a single molecule that could target both streptomycin and paromomycin binding sites for 30S ribosomal subunit inhibition. J Biomol Struct Dyn 37(6):1582–1596. https://doi.org/10.1080/07391102.2018.1462731
22. Mungalpara J, Pandey A, Jain V, Mohan CG (2010) Molecular modelling and QSAR analysis of some structurally diverse N-type calcium channel blockers. J Mol Model 16(4):629–644. https://doi.org/10.1007/s00894-009-0591-1
23. Pandey A, Mungalpara J, Mohan CG (2010) Comparative molecular field analysis and comparative molecular similarity indices analysis of hydroxyethylamine derivatives as selective human BACE-1 inhibitor. Mol Divers 14(1):39–49. https://doi.org/10.1007/s11030-009-9139-7

24. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R (2014) QSAR modeling: Where have you been? Where are you going to? J Med Chem 57(12):4977–5010. https://doi.org/10.1021/jm4004285
25. Shahlaei M (2013) Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chem Rev 113(10):8093–8103. https://doi.org/10.1021/jm4004285
26. Gupta S, Fallarero A, Järvinen P, Karlsson D, Johnson MS, Vuorela PM, Mohan CG (2011) Discovery of dual binding site acetylcholinesterase inhibitors identified by pharmacophore modeling and sequential virtual screening techniques. Bioorg Med Chem Lett 21(4):1105–1112. https://doi.org/10.1016/j.bmcl.2010.12.131
27. Shailesh KP, Indira G (2019) In silico structure based prediction of Receptor-ligand binding ainity. Current progress and challenges, In: Mohan CG (ed) Structural Bioinformatics: applications in preclinical drug discovery process, 1st edn, Springer Nature (USA), pp 109–175
28. Carpenter KA, Huang X (2018) Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a review. Curr Pharm Des 24(28):3347–3358. https://doi.org/10.2174/1381612824666180607124038
29. Mishra R, Li B (2020) The application of artificial intelligence in the genetic study of Alzheimer's disease. Aging Dis 11(6):1567–1584. https://doi.org/10.14336/AD.2020.0312
30. Yang X, Wang Y, Byrne R, Schneider G, Yang S (2019) Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev 119(18):10520–10594. https://doi.org/10.1021/acs.chemrev.8b00728
31. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20(3):318–331. https://doi.org/10.1016/j.drudis.2014.10.012
32. ChemBio3D Ultra software ver. 14.0 (2014) PerkinElmer, Inc. ChemBioOffice
33. BIOVIA DS, (2018) BIOVIA Discovery studio. Dassault systems, San Diego
34. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
35. WEKA software, Ver.3.8.4 (2020)
36. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
37. Roy K, Kar S (2014) The rm2 metrics and regression through origin approach: Reliable and useful validation tools for predictive QSAR models (Commentary on "Is regression through origin useful in external validation of QSAR models?"). Europ J Pharm Sci 62:111–114
38. Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharma Design 13(34):3494–3504
39. Gupta S, Fallarero A, Vainio MJ, Saravanan P, Santeri Puranen J, Järvinen P, Johnson MS, Vuorela PM, Mohan CG (2011) Molecular docking guided comparative GFA, G/PLS, SVM and ANN models of structurally diverse dual binding site acetylcholinesterase inhibitors. Mol Inform 30(8):689–706. https://doi.org/10.1002/minf.201100029
40. Recanatini M, Cavalli A, Hansch C (1997) A comparative QSAR analysis of acetylcholinesterase inhibitors currently studied for the treatment of Alzheimer's disease. Chem Biol Interact 105(3):199–228. https://doi.org/10.1016/s0009-2797(97)00047-1
41. Mungalpara J, Pandey A, Jain V, Mohan CG (2010) Molecular modeling and QSAR analysis on some structurally diverse N-type calcium channel blockers. J Mol Model 16(4):629–644. https://doi.org/10.1007/s00894-009-0591-1
42. Pandey A, Jignesh M, Mohan CG (2010) Comparative molecular field analysis and comparative molecular similarity indices analysis of hydroxyethylamine derivatives as selective human BACE-1 inhibitor. Mol Divers 14(1):39–49. https://doi.org/10.1007/s11030-009-9139-7
43. Awale M, Mohan CG (2008) 3D QSAR CoMFA analysis of C5 substituted Pyrrolotriazines as HER2 (ErbB2) inhibitors. J Mol Graphics Model 26(7):1169–1178. https://doi.org/10.1016/j.jmgm.2007.10.008
44. Niu B, Zhao M, Su Q, Zhang M, Lv W, Chen Q, Chen F, Chu D, Du D, Zhang Y (2017) 2D-SAR and 3D-QSAR analyses for acetylcholinesterase inhibitors. Mol Divers 21(2):413–426. https://doi.org/10.1007/s11030-017-9732-0