Rationalizing fragment based drug discovery for BACE1: insights from FB-QSAR, FB-QSSR, multi objective (MO-QSPR) and MIF studies

Prabu Manoharan · R. S. K. Vijayan ·
Nanda Ghoshal

Abstract The ability to identify fragments that interact with a biological target is a key step in FBDD. To date, the concept of fragment based drug design (FBDD) is increasingly driven by bio-physical methods. To expand the boundaries of QSAR paradigm, and to rationalize FBDD using *In silico* approach, we propose a fragment based QSAR methodology referred here in as FB-QSAR. The FB-QSAR methodology was validated on a dataset consisting of 52 Hydroxy ethylamine (HEA) inhibitors, disclosed by *GlaxoSmithKline* Pharmaceuticals as potential anti-Alzheimer agents. To address the issue of target selectivity, a major confounding factor in the development of selective BACE1 inhibitors, FB-QSSR models were developed using the reported off target activity values. A heat map constructed, based on the activity and selectivity profile of the individual R-group fragments, and was in turn used to identify superior R-group fragments. Further, simultaneous optimization of multiple properties, an issue encountered in real-world drug discovery scenario, and often overlooked in QSAR approaches, was addressed using a Multi Objective (MO-QSPR) method that balances properties, based on the defined objectives. MO-QSPR was implemented using Derringer and Suich desirability algorithm to identify the optimal level of independent variables (*X*) that could confer a trade-off between selectivity and activity. The results obtained from FB-QSAR were further substantiated using MIF (Molecular Interaction Fields) studies. To exemplify the potentials of FB-QSAR and MO-QSPR in a pragmatic fashion, the insights gleaned from the MO-QSPR study was reverse engineered using Inverse-QSAR in a combinatorial fashion to enumerate some prospective novel, potent and selective BACE1 inhibitors.

Keywords FB-QSAR · FB-QSSR · MO-QSPR · BACE1 · BACE2 · Cat-D

Introduction

It has been about 45 years since the seminal contributions of Hansch that laid the foundation for quantitative structure–activity relationships (QSAR). Hansch's original LFER [1] approach, that emerged as a method of quantitatively correlating physicochemical properties of molecules with their biological activity, has metamorphosed into a widely used tool, substantially contributing to the drug discovery process. Though many valuable extensions to the classical QSAR paradigm has been explored, nevertheless, the application of fragment based and multi objective principles in QSAR approaches is sparse. Lead discovery using Fragnomics [2, 3] is an emerging paradigm in drug discovery that utilizes smaller molecules (fragments) to identify fragments that bind specifically but with low affinity to the target receptor. This emerging field is mainly driven by biophysical approaches (NMR, X-Ray, SPR and MS) and by few structure based computational techniques like MCSS [4–6], GRID [7]. Fragment screening, conducted using biochemical assay or a biophysical technique, is labour intensive and could also end up with weak fragment hits. Another important pre-requisite to carry out fragment screening is a robust assay system capable of

P. Manoharan · R. S. K. Vijayan · N. Ghoshal (✉)
Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology, (A unit of CSIR), 4 Raja S.C. Mullick Road, Jadavpur, Kolkata 700032, India
e-mail: nghoshal@iicb.res.in

quantifying weak binding. Jencks concept of additivity phenomena [8], that laid the foundation for FBDD and the additivity phenomena assumed in classical Free Wilson [9] QSAR modelling, share a common ideological credo. Given the ability of Fujita-Ban [10] QSAR modelling in weighing the R-group contribution for activity, we developed a hybrid FB-QSAR modeling approach, incorporating the essential elements of classical Free-Wilson [9] model and Fujita-Ban [10] model. The method envisaged herein would enable lead optimization using a novel approach where QSAR confluence with FBDD (Fragment Based Drug design).

To underpin the concept of FB-QSAR in a pragmatic fashion, we turned our attention towards BACE1, an extensively studied aspartic protease, involved in etiopathogenesis and progression of Alzheimer's disease (AD). BACE1, also termed as β-secretase or memapsin 2, plays a central role in the cleavage of amyloid precursor protein (APP) that results in the formation of the amyloid-β (Aβ) peptide [11, 12], which is the principal component of neuritic plaques, found in the brains of Alzheimer patients. The amyloid hypothesis holds that the neuronal dysfunction and clinical manifestation of AD are a consequence of the long-term deposition and accumulation of 40/42 amino acid long amyloid-β (Aβ) peptide [13]. Hence inhibition of the enzyme BACE1 offers a promising strategy that would aid in the management and delay the progression of AD.

From a therapeutic point of view, selectivity towards BACE1 over other human aspartic proteases like BACE2 and Cat-D (Cathepsin D) [14] is expected to be important, as promiscuous binding would bring about undesirable side effects. Targeting BACE1 without inhibiting an $\sim 50\%$ identical enzyme termed BACE2 [15], which is hypothesized to be physiologically important protein along with Cat-D, a ubiquitous protein which shares $\sim 35\%$ sequence similarity, obviously reminds the need to consider selectivity aspects in the early part of the investigation. First generation BACE1 inhibitor, based on a hydroxyl ethylamine (HEA) core, recently disclosed by *GlaxoSmithKline* Pharmaceuticals was considered in the study [16–19]. The vast majority of the BACE1 inhibitors reported to date are notably tetrahedral intermediate isosteres that mimic the scissile amide bond of the substrate. These peptidomimetic inhibitors exhibit high potency in cell free and cell based assays, but fail to turn up as promising drug candidates due to poor pharmacokinetics property and Blood Brain Barrier (BBB) penetration. For this tormenting reason many HTS campaigns aimed at identifying good quality nonpeptidic hits have been pursued, which turned up to be unsuccessful. Of notable exception is the successful application of FBDD by Astex Therapeutics, which led to the identification of aminopyridines and cyclic amidines which are currently been optimized by Astra

Zeneca. This success clearly highlights that FBDD detected fragments are highly recommended candidates for BACE1 [20, 21].

To address the issue of target selectivity, we developed FB-QSSR models using activity values reported for two promiscuous targets namely BACE2 and CAT-D [16–19]. Conventional 2D QSAR and 3D QSAR approaches focus on a single objective and overlooks the importance of multiple objectives required for drug-like behavior [22]. As an adaptive response to the changing scenario, we carried out MO-QSPR studies using Derringer and Such algorithm [23] as a multi-objective optimization technique to duly consider the multiple objectives of activity and selectivity in arriving at a balanced model. The information gleaned from our MO-QSPR approach was used in a (I)-QSAR manner to identify ideal bioisosteric fragments, which in turn were employed to enumerate a focused virtual combinatorial library that was subsequently screened using the forward FB-QSAR/FB-QSSR models.

The key strategic advantages of the present study are its ability to carry out fragment based drug design even in the absence of experimentally resolved structure of the receptor and carry out affinity prediction for fragments even in the absence of specialized assay. Classical 2D QSAR modeling builds a mathematical relationship between physicochemical properties of the molecule and its biological activity; on the contrary FB-QSAR paradigm relates sub structural fragments present in the data set compounds to their corresponding biological activity using standard chemometric methods. Based on the obtained FB-QSAR model, fragment weights could be assigned according to their individual coefficient value and descriptor value. The key advantage of undertaking FB-QSAR is the ease in optimizing fragments that could be tethered eventually to yield optimized drug like molecules rather than embarking on optimizing a whole molecule that often tends to reduce the drug likeliness property [24].

Further, FB-QSAR offers the unique advantage of mitigating the risk of sharing QSAR data that can be confronted by concealing the core structure without jeopardizing the intellectual rights and commercialization plans especially when data are not under patent cover. FB-QSAR/FB-QSSR and the MO-QSPR methods implemented herein heralds the approach of revitalizing QSAR to address the demands of modern day drug discovery.

Materials and methods

Data modeling and K-means clustering

An internally consistent data set in terms of activity range, distribution, assay method and experimental conditions

were taken from four literature sources reported by the same group [16–19]. The skewness in the dataset was removed by converting IC_{50} (nM) values to pIC_{50} (nM) using a simple logarithmic transformation log $(1/IC_{50})$. Data set compounds were modeled using HyperChem version 7.52 [25]. Upon obtaining reasonable starting geometries, batch optimization of all the modeled compounds were carried out using MMFF94 force field, implemented in MOE, version 2008.10 [26]. Development of predictive QSAR models relies on a multitude of factors and one such important prerequisite is the rational selection of training and a test set with adequate representation. A non-hierarchical method called K-means clustering was performed using SPSS 15.0 [27] for the rational division of training and test series [28]. Clustering was carried out using more than one hundred 2D whole molecule descriptors representing topology, structural information and group counts, calculated using MOE [26] and TSAR v3.3 [29]. Manual selection of compounds was done from each cluster so as to ensure training set and test set had adequate coverage in terms of activity range and "chemical diversity". Accordingly, 43 compounds served as training set and 9 compounds served as test set. FB-QSSR studies were carried out using ΔpIC_{50} values, which reflect the computed pairwise selectivity differences of BACE1 with respect to BACE2 and Cat-D. The structures of the dataset compounds are shown in Table 1 and the common scaffold is shown in Fig. 1.

Fragmentation and descriptor selection

Being a congeneric data set, the fragmentation rule was defined based on the substitutions present around the defined common HEA derivative core. Accordingly, the congeneric dataset was fragmented as R_1, R_2 and R_3 as shown in Fig. 1. The fragmentation of a molecule assumes that the fragments follow a modular binding mode and exert an additivity phenomenon, the key principle, which FBDD relies on [30]. Upon marking of the fragments, established 2D descriptors were calculated for all the R-group fragments, using MOE and TSAR. Hence, changes in descriptor values directly reflect the changes at the sub structural level in an intuitive manner. This procedure is akin to the traditional Free–Wilson QSAR approach of using indicator variables to quantify the presence/absence of a particular fragment. The risk of over correlation among descriptors would be high in FB-QSAR as too many variables would be involved. The crux of variable reduction is to select a subset of descriptors that retains maximal information by pruning non-informative and correlated descriptors. A feature selection routine operated on such a pruned descriptor space offers a better QSAR model with a reduced risk of chance correlation and over fitting. Hence,

a two-stage descriptor pruning was carried out by performing a simple pair wise correlation analysis initially with the activity/property, followed by a complete pair wise correlation analysis within the descriptors. This procedure ensures elimination of descriptors not/less correlated to property of interest and the elimination of highly inter correlated variables.

FB-QSAR/FB-QSSR and MO-QSPR theory

Free-Wilson approach addresses structure–activity relationship in a congeneric series as described in Eq. 1 [9].

$$BA = \sum A_i X_i + \mu \tag{1}$$

where, BA is the biological activity, μ is the average contribution of the parent molecule (unsubstituted compound), and A_i is the contribution of each structural feature; X_i denotes the presence $(X_i = 1)$ or absence $(X_i = 0)$ of a particular structural fragment. Free-Wilson approach of using indicator variable (I_0, I_1) has been replaced by using R-group based descriptors in the present study, to assist fragment based drug discovery as indicator variables are prone to overlook even bioisosteric fragments. Hence, this modified approach is termed herein as Fragment Based QSAR (FB-QSAR) approach, tuned to address the present day needs of fragment based drug discovery. To weigh the R-group's contribution for activity/selectivity, we employed the Fujita-Ban approach as a method for estimating de novo group contributions, which is expressed by the following equation [10].

$$Log\ BA = \sum G_i X_i + \mu \tag{2}$$

where, μ is the average contribution of the parent molecule, G_i represents the biological activity contribution of the substituent. Individual R-group (substituent) contributions were estimated, assuming that activity contribution from each R-group follows an additive principle on lines with the Free-Wilson theory.

SMLR, a variant of MLR approach that uses a combination of forward and backward MLR, was used as a chemometric method for variable selection and statistical fitting. SMLR identifies an initial model and proceeds by altering the model by adding and removing explanatory variable in accordance with a F criterion, which controls the inclusion or exclusion of explanatory variables until an optimum model is found. SMLR analyses were performed using SPSS with F to include set at 4, and F to exclude set at 3.5 for BACE1 and BACE1-CAT-D model, and F to include set at 3 and F to exclude set at 2 for BACE1-BACE2 model. The F value signifies the square of the t value of the regression coefficient of the variable being included or excluded. The above mentioned FB-QSAR/FB-QSSR

Table 1 R-group fragments considered for the generation of the FB-QSAR Models
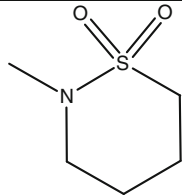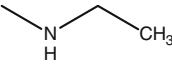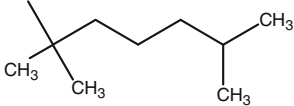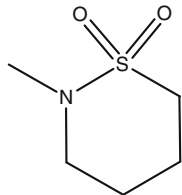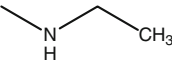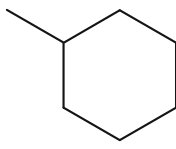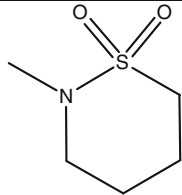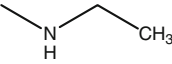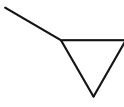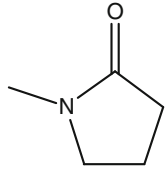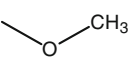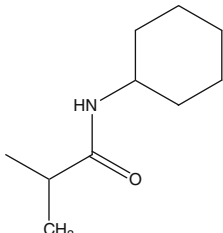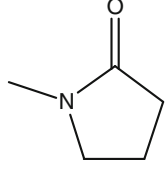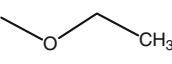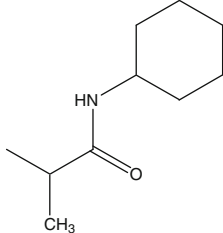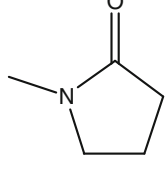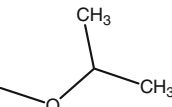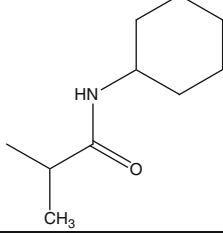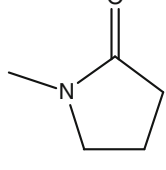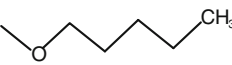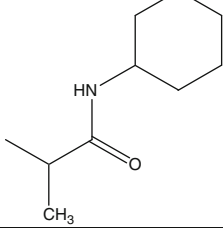
NAME	R 1	R 2	R 3
1			
2*			
3			
4			
5			
6			
7			

Table 1 continued
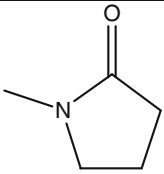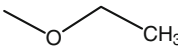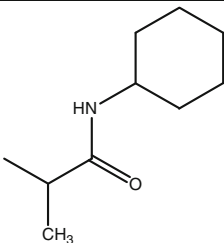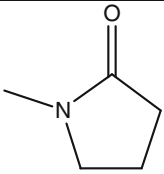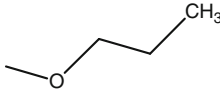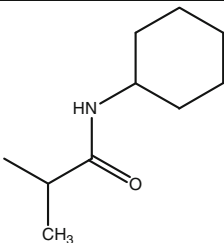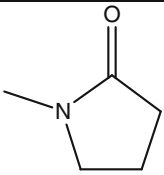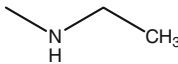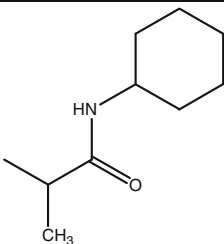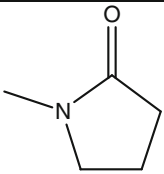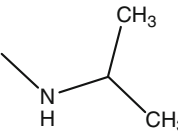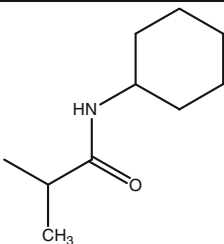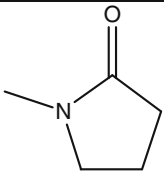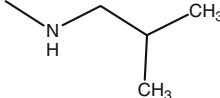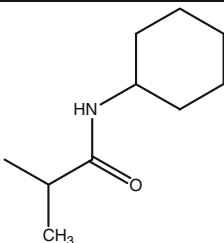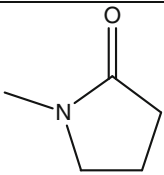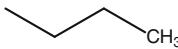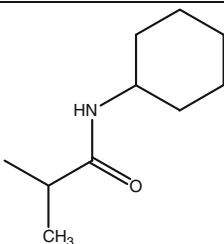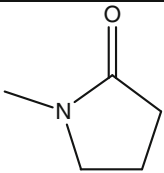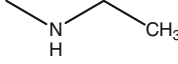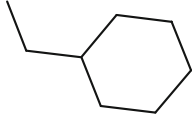
8*			
9			
10			
11			
12*			
13			
14			

Table 1 continued
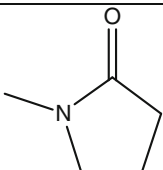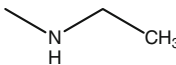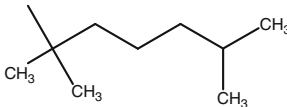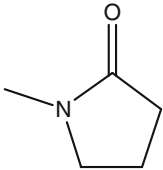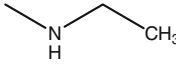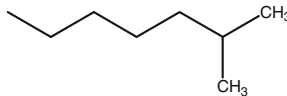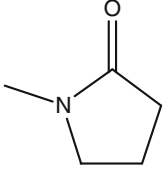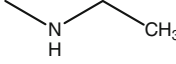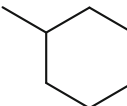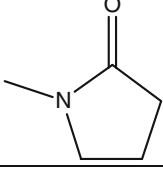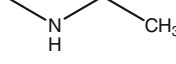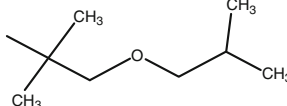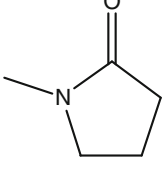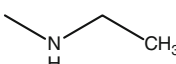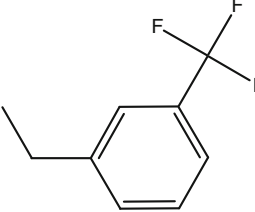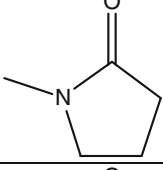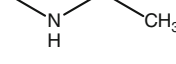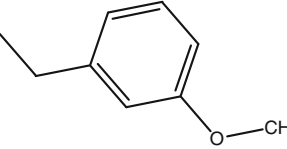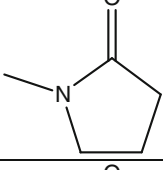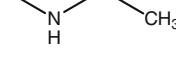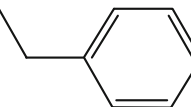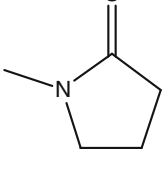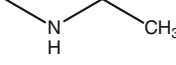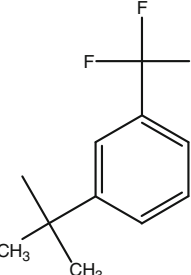
15			
16			
17			
18*			
19			
20			
21			
22			

Table 1 continued
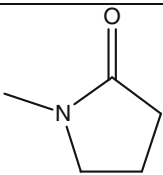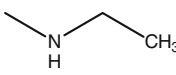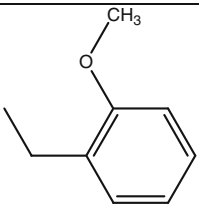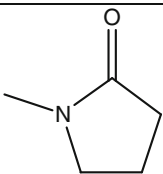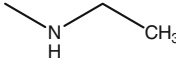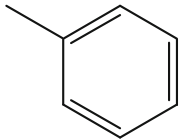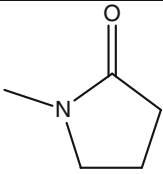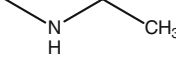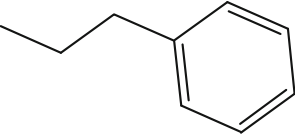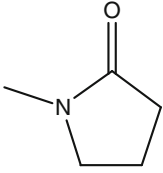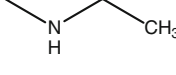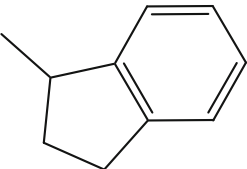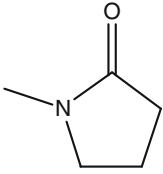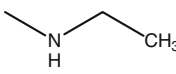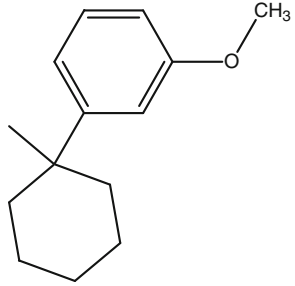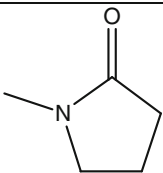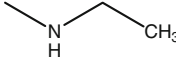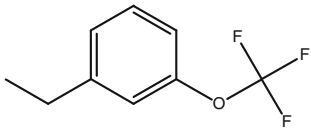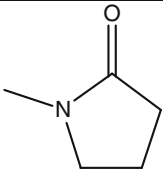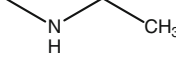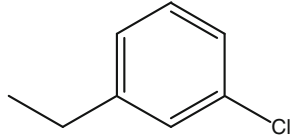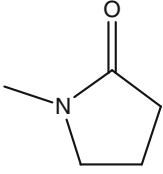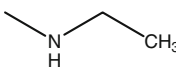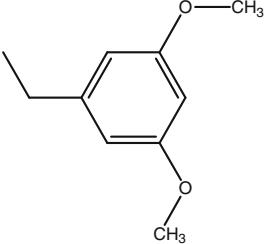
23			
24*			
25			
26			
27			
28			
29			
30			

Table 1 continued
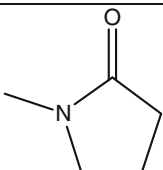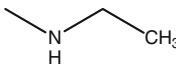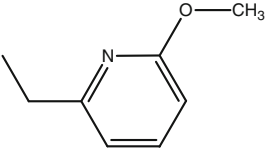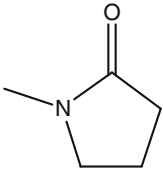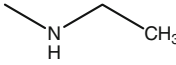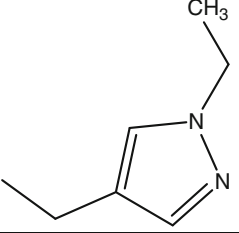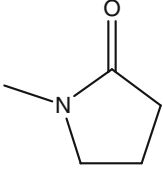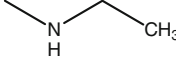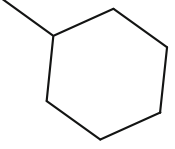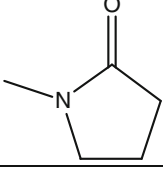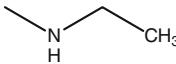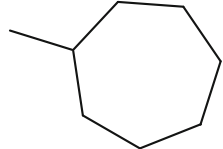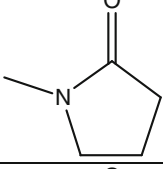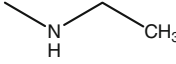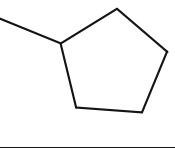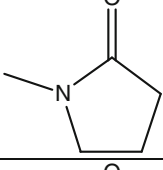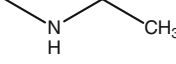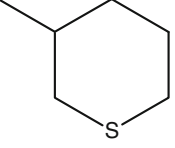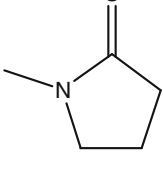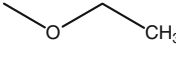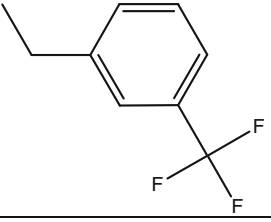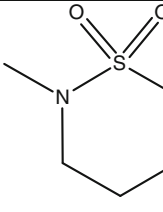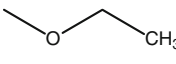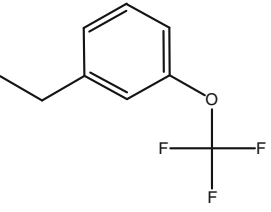
31			
32*			
33			
34			
35			
36			
37			
38			

Table 1 continued
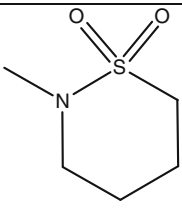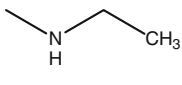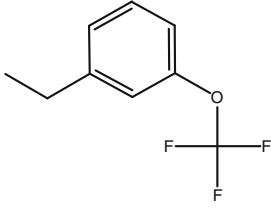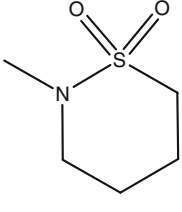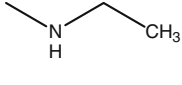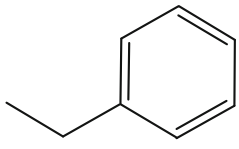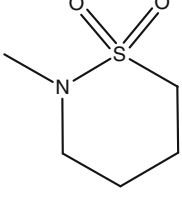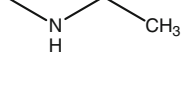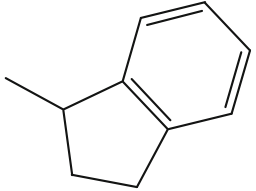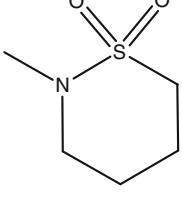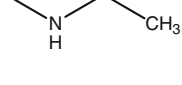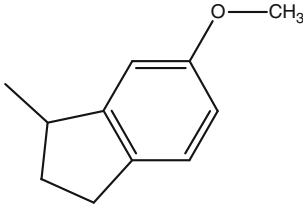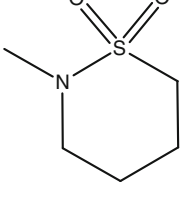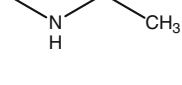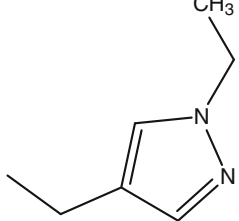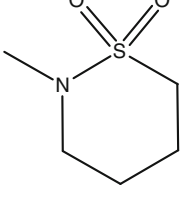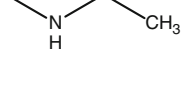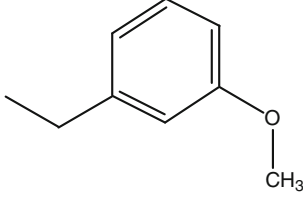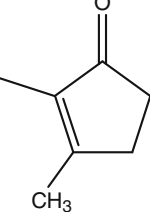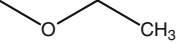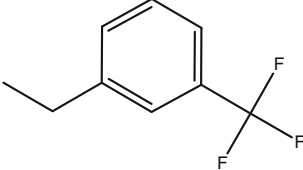
39*			
40			
41			
42			
43			
44			
45			

Table 1 continued

46			
47			
48*			
49*			
50			
51			
52			

Compound marked as *
indicates test set compound

approach focuses on a lone objective (Single Object Optimization-SOO) and produces only one single optimal solution. Since lead optimization demands optimization of multiple properties, that are often conflicting in nature, we employed Derringer and Suich desirability function method to carry out a multi objective (MO-QSPR) study. MO-QSPR can be termed as an inverse (*i*-QSAR) approach which involves the transformation of each predicted response; (*Yi*) obtained from forward QSAR approaches (FB-QSAR/FB-QSSR) into a dimensionless partial
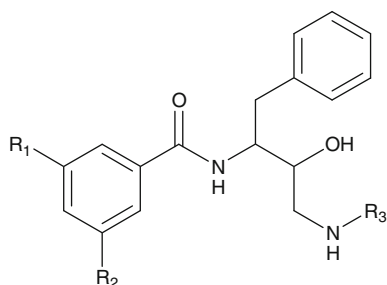
Fig. 1 The common core scaffold considered for FB-QSAR/FB-QSSR

desirability function, d_i. This transformation depends on the researcher's priority that suits the optimization procedure, depending on whether a particular response needs to be maximized, minimized or to retain the allotted target value [23]. In the present study since both the dependent properties (affinity and selectivity) need to be maximized, one-sided transformation was applied. Accordingly, the individual desirability function was defined as

$$di = \begin{cases} 0 & \text{if } \hat{Y}i \leq Li \\ \left[\frac{\hat{Y}i - Li}{Ti - Li}\right]^s & \text{if } Li < \hat{Y}i < Ti \\ 1 & \text{if } \hat{Y}i \geq Ti = Ui \end{cases} \quad (3)$$

where L_i, U_i, and T_i are the lower, upper, and target values, respectively. The value of d_i will vary non-linearly while approaching the desired value; hence the exponent term(s) was assigned a value of 1 to make the desirability function linear. For the estimation of d_i, the lower value L_i was assigned to the least active and the least nonselective compounds and the upper value U_i was assigned to the most active and the most selective compounds.

The normalized individual desirability functions (d_i) thus obtained were combined into a single composite response termed global desirability function (D), obtained as geometric mean of different d_i values.

$$D = (d1 \times d2 \times \ldots \times dk)^{\frac{1}{k}} \quad (4)$$

where, k denotes the number of responses.

This process ensures simultaneous optimization by taking into account the relative importance of each response. Irrespective of the modelling technique employed, model validation and predictivity estimation are mandatory for any QSAR/MO-QSPR model. The goodness of fit for the MO-QSPR model was quantified using a metric R_D^2 as proposed by Cruz-Monteagudo et al. [31]. R_D^2 is analogous to the determination coefficient R^2.

$$R_D^2 = 1 - \frac{\text{SSE}}{\text{SSTO}} = 1 - \frac{\sum (D_{Yi} - D_{\hat{Y}i})^2}{\sum (D_{Yi} - \bar{D}_{Yi})^2} \quad (5)$$

where, D_{Yi} is the actual desirability, $D\hat{Y}i$ is the predicted desirability and $\bar{D}_{Yi}$ is mean of the actual desirability.

SSTO is the total sum of squares, and SSE is the sum of squared error. R_D^2 is not reflective of the predictive ability of MO-QSPR model and it is only a measure of the goodness of fit. Hence the predictive ability was estimated using the LOO-CV defined in a manner analogous to conventional cross validated R^2 (Q^2).

$$Q_D^2 = 1 - \frac{\text{SSE}_{\text{LOO-CV}}}{\text{SSTO}} = 1 - \frac{\sum (D_{Yi} - D_{\hat{Y}i}(\text{LOO} - \text{CV}))^2}{\sum (D_{Yi} - \bar{D}_{Yi})^2} \quad (6)$$

where, Q_D^2 is the overall predictive desirability, $\text{SSE}_{\text{LOO-CV}}$ and $D_{Yi}(\text{LOO-CV})$ are the leave one out cross validation error sum of square and the predicted desirability by LOO-CV, respectively.

The response values ($Y_{\text{predicted}}$) and the predictor variables (X variables) obtained from individual FB-QSAR/FB-QSSR models were used to generate Derringer desirability plots using the "Response surface regression" tool available in STATISTICA [32]. The desirability plot, thus obtained, highlights the optimal value for the independent X variables for achieving maximal selectivity and activity. This methodology is akin to the engineering concept of Design of Experiments (DOE) in industrial settings, which could be extrapolated for drug design scenario using inverse QSAR paradigms. Such approach has the potential to deliver quality pre-clinical molecules that strike a subtle balance between activity and selectivity.

Statistical validation

The generated FB-QSAR/FB-QSSR models were assessed for four important qualities namely goodness of fit, model stability, predictive ability and domain applicability. Routine standard metrics like explained variance, LOO-CV and predicted variance, F statistics were used to judge the quality of the models.

Results and discussion

The chemical structures of the compounds considered in the present study are shown in Table 1 and their observed and predicted values are enlisted in Table 2. The best FB-QSAR, FB-QSSR models obtained are presented as Eqs. (7, 8 and 9) and its corresponding statistical values are provided as foot note below the equations.

Activity based FB-QSAR for BACE1 (Model 1)

The best QSAR equation obtained in terms of statistical quality for our activity based FB-QSAR model 1 is shown in Eq. 7.

Table 2 Observed and predicted affinity/selectivity of BACE1, BACE2, CAT-D and their affinity differences is indicated by Δ

NAME	BACE1 (pIC$_{50}$) (Obs)	BACE1 (pIC$_{50}$) (Pred)	(BACE1-BACE2) Δ pIC$_{50}$ (Obs)	(BACE1-BACE2) Δ pIC$_{50}$ (Pred)	(BACE1-CAT-D) Δ pIC$_{50}$ (Obs)	(BACE1-CAT-D) Δ pIC$_{50}$ (Pred)
1	7.16	7.13	1.8	1.68	1.53	1.64
2*	7.52	7.86	2.42	2.32	2.24	2.57
3	6.92	6.43	2.21	2.22	2.4	1.92
4	6.13	6.69	1.58	1.85	1.61	1.84
5	7.23	6.69	1.94	1.66	2.13	1.76
6	6.92	6.69	1.6	1.46	2.01	1.85
7	7.49	6.69	1.1	1.07	1.13	1.57
8*	5.84	5.62	0.95	0.83	0.63	0.85
9	6.22	6.69	1.13	1.46	1.89	1.75
10	7.89	7.23	2.14	1.69	2.32	2
11	7.41	7.23	1.33	1.48	2.14	1.81
12*	6.74	7.23	0.56	1.29	1.23	1.42
13	7.14	7.23	1.34	1.37	1.8	1.78
14	5.23	6.29	0.99	1.45	0.31	1.1
15	7.48	6.79	2.25	1.66	1.88	1.37
16	6.52	6.32	1.07	1.61	0.5	1.07
17	6.26	6.55	1.74	1.75	1.29	1.49
18*	6.1	6.82	1.94	1.8	1.45	1.24
19	7.4	7.31	2.01	1.9	2.16	1.71
20	7.37	6.68	2.25	2.12	2.07	1.87
21	6.12	6.19	1.92	1.86	1.4	1.33
22	7.77	7.78	1.55	1.73	1.74	1.79
23	5.87	6.68	1.52	1.76	0.86	1.37
24*	5.77	5.95	1.1	1.32	0.44	0.19
25	6.22	6.42	0.75	1.02	0.47	−0.1
26	6.8	6.62	1.37	1.47	1.52	1.28
27	7.85	7.81	1.42	1.29	1.43	1.54
28	7.7	7.58	2	1.99	2.25	1.88
29	7.08	6.76	2.26	1.7	1.83	1.36
30	7	7.18	1.84	2.01	1.77	1.77
31	6.48	6.72	2	2.08	1.9	1.91
32*	6.64	6.5	2.16	1.99	2.31	1.55
33	6.74	6.05	1.44	1.1	0.64	0.86
34	5.95	6.29	1.03	1.05	0.33	0.95
35	5.61	5.82	1.23	1.2	0.65	0.5
36	6.19	6.35	0.96	0.97	0.38	0.22
37	6.42	6.77	1.85	1.92	1.96	2.03
38	8.3	8.11	2.18	2.4	3.09	2.85
39*	8.3	8.66	2.08	2.38	2.71	2.83
40	6.92	7.26	2.43	2.42	2.34	2.44
41	7.47	7.69	1.83	2.09	1.96	2.38
42	7.74	8.19	2.21	2.2	1.77	2.6
43	7.92	7.56	2.76	2.89	3.67	3.61
44	8.1	7.76	2.73	2.35	2.93	2.74
45	6.6	6.77	1.22	1.2	0.85	1.06
46	6.55	6.77	1.04	0.95	1.02	1.04
47	6.8	6.77	0.82	0.89	1.1	1.21
48*	7.3	7.85	1.97	2.6	2.23	3

Table 2 continued

NAME	BACE1 (pIC$_{50}$) (Obs)	BACE1 (pIC$_{50}$) (Pred)	(BACE1-BACE2) Δ pIC$_{50}$ (Obs)	(BACE1-BACE2) Δ pIC$_{50}$ (Pred)	(BACE1-CAT-D) Δ pIC$_{50}$ (Obs)	(BACE1-CAT-D) Δ pIC$_{50}$ (Pred)
49*	8.22	7.85	2.23	2.51	2.99	2.91
50	8.52	8.39	2.68	2.54	3.11	3.06
51	8.3	8.39	2.46	2.31	2.98	3.25
52	7.96	8.39	2.18	2.35	3.36	3.02

Compound marked as * indicates test set compound

$$pIC_{50} = 3.59668 + 0.016594 * MM(R_3) + 1.07651 \\ * NHBA(R_1) - 0.541331 * NHBA(R_2) \quad (7)$$

$N = 43$, $R^2 = 0.725$, $R^2_{adj} = 0.704$, $BsR^2 = 0.726$, $Q^2_{(LOO)} = 0.671$, Randomized $R^2 = 0.041$, Ftest = 34.310, PRESS = 8.489, $R^2_{Pred} = 0.793$, $(R^2 - R^2o)/R^2 = 0$, and $k = 1$.

The BACE1 (SMLR) model was derived using some of the simple descriptors, which offer ease for interpretation and chemical translatability. The best FB-QSAR model obtained for activity (Model 1) had an explained variance of 70.4% and a predicted variance of 79.3%. Molecular Mass (MM) at R_3 appears in the equation with a positive coefficient, which suggests that a bulky substituent is acceptable at this position. The presence of NHBA (R_1) in this equation with positive coefficient suggests that groups with hydrogen bond acceptor functionalities are favorable at this position for enhanced biological activity. Occurrence of NHBA (R_2) in this equation with a negative coefficient indicates that hydrogen bonding acceptor groups at R_2 position are unfavorable for biological activity. Significantly, SAR also reveals that O linked derivatives, which have a HBA functional moiety have reduced activity, whereas N linked derivatives possessing HBD groups had increased activity. The availability of the crystal structure of BACE1 in complex with an HEA derivative allowed us to further examine and reinforce the conclusion drawn from FB-QSAR approach. In the light of some recent sardonic comments about QSAR modeling stating that correlation doesn't infer causation, a sanity assessment of the FB-QSAR models was undertaken by corroborating the ligand based findings with those of structure based approach. Though very few QSAR modeling papers highlight such a type of assessment, we strongly believe that cross correlating QSAR findings using approaches like Molecular Interaction Field (MIF), in cases when the 3D structure of the receptor is available, would greatly substantiate the interpretation of the FB-QSAR models in terms of their chemical and biological significance. The surface and maps utility, present in MOE, was used to generate iso-contour non bonded, contact preference maps for the co-crystallized complex (2VNM) [16].
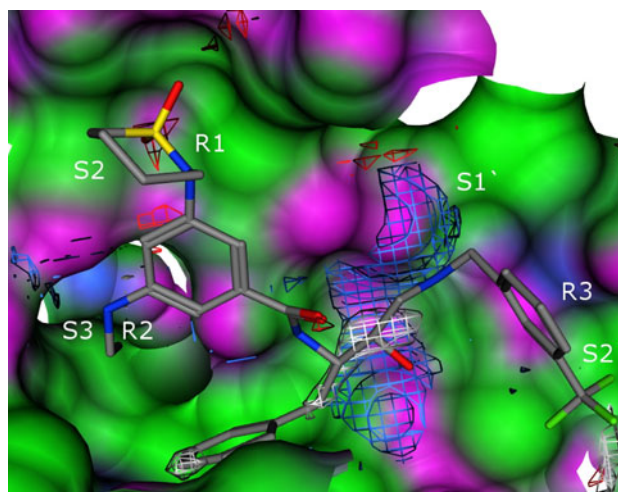
Fig. 2 The activity contour map of BACE1. The Interaction potential contour shown as *red* color mesh represents HBA ($E = -2.5$ kcal). *Blue* color mesh represents HBD ($E = -2.5$ kcal). *White* color mesh represents DRY ($E = -3.3$ kcal) (Hydrophobic) probe which models Hydrophobicity

This module of MOE is very close in spirit to knowledge-based tools like X-Site [33] and SuperStar [34]. The iso-contour map depicted in Fig. 2 highlights the preferred location for hydrophobic and hydrophilic (HBA/HBD) ligand atoms, based on the probabilistic non bonded contact preference values, obtained using a knowledge based approach.

Visual interpretation of the map displays a large hydrophobic patch around the core moiety and around the R_3 fragment Fig. 2. The S2′ site, termed as the prime site, reveals the presence of energetically favorable hydrophobic contours in the vicinity of meta substituted trifluro methyl, with a potential to accommodate even bulkier group. The occurrence of molecular mass for the R_3 fragment with a positive coefficient as shown in Eq. 7 is in accordance with this finding. It should be noted that a close scrutiny of the crystal structure of BACE1 in complex with its inhibitors (compound 10, PDB ID 2VJ6) reveals that the R_3 fragment possessing an amide linker binds to the S2′ site, in contrast R_3 fragment lacking the amide linker (compound 17, PDB ID 2VJ9) binds to the S1′ site. Since

the FB-QSAR method implemented herein is 2D in nature it doesn't takes into consideration the realistic or a probable binding mode for the fragments. Hence no information pertaining to site specific binding of the fragments could be inferred from our FB-QSAR modeling. This observed difference in the binding mode as evident from crystallographic could in no way influence the FB-QSAR models as it is largely ligand based and doesn't consider protein–ligand interaction cross terms during model building. A very large acceptor iso contour is evident around the R_1 fragment housed in the S2 site. Based on the crystallographic information (2VIZ) [17], it is evident that the carbonyl oxygen of lactam mediates a strong hydrogen bond (2.87 Å) with Asn 294. Our modeling hypothesis is in consonance with this finding as it reveals that NHBA of R_1 fragment as a positive contributor for activity. Accordingly, the increased activity evinced in sultam derivatives could be attributed to the bifurcated H bonding, evident in the crystal structure of 2VNM due to the presence of two sulfonyl oxygen in the R_1 fragment. Occurrence of NHBA (R_2) in this equation with a negative coefficient implies that hydrogen bonding acceptor groups at R_2 position are unfavorable for biological activity. Non bonded iso contour preference map points out that the R_2 fragments anchored in the S3 site is predominately occupied by HBD patches. So it becomes obvious that a HBA at this position has a detrimental effect on activity. Though FB-QSAR modeling was not able to shed information on this requirement, it could be fairly envisaged if one corroborates SAR information in tandem with QSAR models. From SAR it is obvious that O linked derivatives which posses HBA functionality have reduced activity, whereas N linked derivatives having HBD functionality have increased activity. Examination of the binding modes of these two classes of ligands whose complex structures are available in PDB, reveals that the NH group of N linked derivate (2VNM) and Oxygen of the O linked derivate (2VIZ) are involved in hydrogen bonding distance of 2.72 and 3.31 Å respectively with the side chain hydroxyl of THR 293.

Though both O linked and N linked derivatives mediate hydrogen bonding interaction with THR293, the nature of the bonding is strikingly different. N linked derivatives mediate a hydrogen bonding of the type (THR293-O–H (Acc)—HN-Ligand (Don) R_2 Fragment), in contrast O linked derivatives forms a hydrogen bond of the type (THR293-O–H (Don)—O-Ligand (Acc) R_2 Fragment), which clearly highlights that the former mediates a stronger hydrogen bonding interaction in comparison to the later, as evident from literatures [35–38]. This finding justifies the appearance of NHBA with a negative coefficient in the FB-QSAR equation. Corroboration of FB-QSAR with crystallographic findings vindicates that the FB-QSAR modeling hypothesis is congruent, and proves that the

findings from FB-QSAR are truly causative rather than merely being correlative.

FB-QSSR model for BACE1 selectivity over BACE2 (Model 2)

FB-QSSR model 2 (Eq. 8), which evaluates the selectivity criteria for BACE1 over BACE2 model had an explained variance of 77.3% and a predicted variance of 65.2%.

$$
\begin{aligned}
\Delta(\mathbf{pIC_{50BACE1}} - \mathbf{pIC_{50\,BACE2}}) &= 0.280745 + 0.224262 \\
&\quad * VB4(R_3) - 0.256772 * logP(R_1) - 0.031452 \\
&\quad * BL(R_3) - 0.296178 *{}^3\chi_p^v(R_3) + 0.606211 \\
&\quad * lip\_acc(R_1) - 0.012459 * vsa\_hyd(R_2)
\end{aligned}
$$

$$(8)$$

$N = 43$, $R^2 = 0.805$, $R_{adj}^2 = 0.773$, $BsR^2 = 0.806$, Q^2 (LOO) = 0.739, Randomized $R^2 = 0.352$, Ftest = 24.786, PRESS = 3.264, $R_{Pred}^2 = 0.652$, $(R^2 - R^2o)/R^2 = 0$, and $k = 1$.

Our FB-QSSR modeling suggests that (Verloop) VB4 [39] at R3 position, which signifies the steric influence of the substituent, is a positive contributor for selectivity over BACE2. Our Insilico modeling results agree well with SAR findings wherein, compounds 15, 19, 20 and 29 possessing linear and bulky fragments at R_3 position are more selective compared to compounds 34, 36 and 14 which possess only small R_3 fragments. The appearance of bond lipole with a negative coefficient at R_3 position suggests that less lipophilic groups are favorable for selective inhibition of BACE1 over BACE2. This modeling hypothesis can be substantiated by comparing the R_3 substituent of selective compounds 19, 21 (less lipohilic R_3-groups) over compounds 16, 17 (highly lipohilic R_3-groups) which are relatively non selective.

The kier connectivity index descriptors which appeared in FB-QSSR model provides information on the skeletal variation of the fragments, including degree of branching, types of branching (three- and four-way branches), ring structure, as well as various patterns of branching, adjacency of branch points and also about the valance state [40]. The appearance of connectivity index ${}^3\chi_p^v$ with negative coefficient at R_3 position suggests that, the branching of fragments at R_3 position is favorable for selectivity. This computational modelling result can be further substantiated by SAR upon comparing compound 20, (decreased ${}^3\chi_p^v$ index and high selectivity) with compound 36, (increased ${}^3\chi_p^v$ index and non selective). A holistic picture of the nature of R_3 fragment can be drawn from the above results, which implies that steric fragments with more number of branching and less liphophilic property are favorable for R_3 fragments. Since chain branching and liphophilicity are

complementary in this model, we presume that liphophilicity dictates pharmacokinetic property and branching dictates the pharmacodynamic property. The appearance of logP [41] with negative contribution and lip_acc (Number of O and N atoms) with a positive contribution collectively indicates the preference for groups with acceptor functionality over lipophilic groups at R_1 position for the selective inhibition of BACE1. The negative contribution of vsa_hyd for R_2 fragment suggests that hydrophobic fragments are not tolerated for achieving BACE1 selectivity. This can be correlated with the experimental SAR results which reveal that the O and N linked derivatives were selective than their carbon analogues, that are more hydrophobic (compare 49 and 50 with 51).

FB-QSSR model for BACE1 selectivity over CAT-D (model 3)

FB-QSSR model 3 (Eq. 9), which evaluates the selectivity criteria for BACE1 over CAT-D had an explained variance of 81.2% and a predicted variance of 77.6%.

$$\Delta(\mathbf{pIC_{50\,BACE1}-pIC_{50\,CAT-D}}) = -1.8697 - 0.131048$$
$$* VL(R_3) + 0.377365 * VB4(R_3) + 0.363482$$
$$* LYC(R_2) + 0.966115 * lip\_acc(R_1)$$

$$(9)$$

$N = 43$, $R^2 = 0.830$ $R^2_{adj} = 0.812$, Bs $R^2 = 0.830$, Q^2 $(LOO) = 0.766$, Randomized $R^2 = 0.241$, Ftest = 46.230, PRESS = 7.049, $R^2_{Pred} = 0.776$, $(R^2 - R^2o)/R^2 = 0$, and $k = 1$.

Verloop length (VL) parameter, which describes the steric nature of the fragment, appears with a negative coefficient for the R_3 fragment. This implies that long fragments are not tolerated for the R_3 fragment for attaining selectivity. The appearance of VB4 at R_3 position with positive coefficient suggests that the width of the substituent is essential for the selective inhibition of BACE1. These findings together, lead to a conclusion that nonlinear and bulky groups are ideal fragments for R_3 position. This is also evident if one compares the selectivity profile and the nature of the R_3 fragment for compounds 34, 16 (which posses low VL and high VB4 values along with high selectivity) with 19, 20 (high VL and low VB4 values with low selectivity). The Lipole Y component (LYC) is the measure of lipophilic distribution of the fragment. Large lipole Y values indicate a large distribution of lipophilic groups distant from the point of attachment. The appearance of LYC at R_2 position with positive coefficient suggests that Lipophilic distribution at this site is essential for the selective inhibition of BACE1 over CAT-D. This result can be substantiated by comparing compound 10, (which possess large LYC value and good selectivity) with

compounds 12 (which possess low LYC value and low selectivity). The appearance of lip_acc with positive coefficient for the R_1 fragment suggests that the acceptor functionality plays an important role in selective inhibition of BACE1 over CAT-D. This is also clearly evident upon analyzing the nature of R_1 substituent of compounds 45, 46 and 47 which exhibit low selectivity, compared to sultams and lactams that exhibit high selectivity. In the absence of co-crystalized complexes of HEA derivatives with BACE2 and CAT-D; we are constrained to limit our model validation within the framework of a ligand based approach. Though it sounds plausible that a docking study undertaken on BACE2 and CAT-D would enable one to carry out validation on similar lines like BACE1 model, we were skeptical of corroborating our FB-QSSR finding by a hypothetical evidence, as docking would not serve as a proof of concept in the present scenario.

Statistical validation of FB-QSAR/QSSR models

Several statistical parameters were considered to ensure the quality of models with emphasis on four important qualities namely goodness of fit, model stability, predictive ability and domain applicability. Goodness of fit was judged using square correlation coefficient (R^2). Since the value of R^2 tends to be inflated as the number of terms in the equation increases, R^2_{adj} was also calculated for expository reasons. The near value of R^2 and R^2_{adj}, for all the obtained models, as evident from the footnotes provided with Eqs. (7, 8 and 9) ensures the absence of over fitting. Correlation plot between the experimental pIC_{50} and predicted pIC_{50} values for the training set and the test set compounds fitted at 95% confidence level for all three models are shown in Fig. 3. Internal predictivity of the models was validated using two re-sampling techniques namely leave –one –out (LOO) cross validation and bootstrapping. Obtained R^2cv (q^2) values of 0.671 (BACE1), 0.739 (BACE1-BACE2) 0.766 (BACE1-CAT-D) assure the internal predictive ability of the models. Model stability and chance correlation were evaluated by subjecting the developed models to a Y-randomization procedure that scrambles the dependent variable set, and rebuilds a new QSAR model based on the permuted response. Randomized R^2 values of 0.041 (BACE1), 0.352 (BACE1-BACE2) and 0.241 (BACE1-CAT-D) apparently signify the stability of these models.

The above mentioned metrics are only indicative of the interpolative ability of the models and do not reflect the extrapolative ability of the models. The predictive ability and extensibility of the models was hence established using a reserved test set consisting of 9 compounds that was not considered during the model generation process.

The predictive power of the models was calculated using the following formula [42].
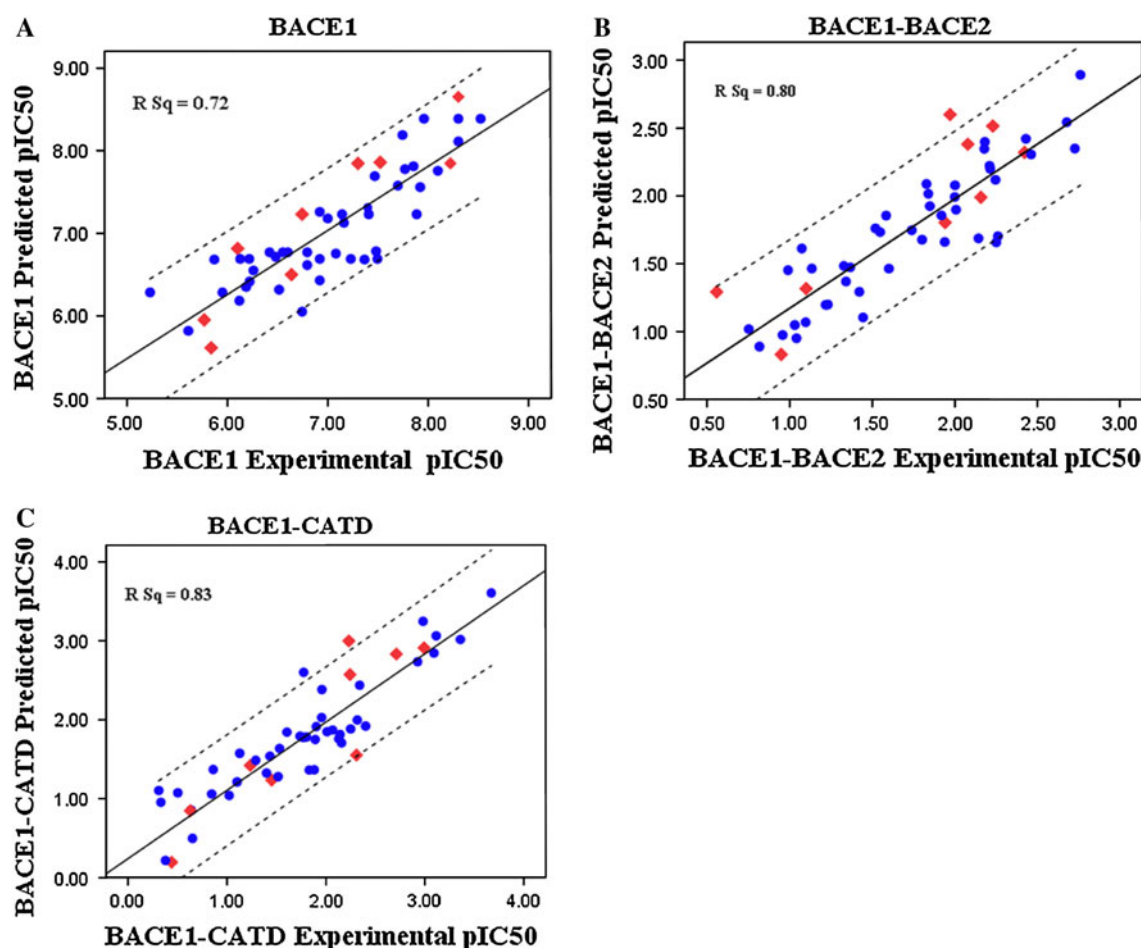
Fig. 3 Correlation coefficient between the experimental pIC_{50} and predicted pIC_{50} values for the training set (*Blue-round*) and the test set compounds (*Red-diamond*) at 95% confidence level

$$R^2_{pred} = \frac{SD - PRESS}{SD} \qquad (10)$$

where SD is the sum of square deviations between the biological activities of each molecule in the test and the mean activity of the training set molecules and PRESS is the sum of square deviations between the predicted and the actual activities of molecules in the test set. All the obtained models have acceptable levels of predictive quality. More rigorous validation was carried out using the other parameters as proposed by Tropsha et al. [42].

$$(R^2 - R^2_O)/R^2 < 0.1 \text{ or } (R^2 - R^2_O)/R^2 < 0.1 \qquad (11)$$

and

$$k \text{ or } k' \text{ close to } 1 \qquad (12)$$

where, R^2 represents the square correlation coefficient between the predicted and observed activities. R^2_O and $R^{2'}_O$ are quantities characterizing the square correlation between predicted vs observed and observed vs predicted activity

respectively, with Y-intercept set to zero and k, k' are their corresponding slopes.

When embarking on an extrapolative adventure using FB-QSAR/FB-QSSR models, assessment of applicability domain with reference to the calibrated model is of immense importance to obtain predictions with confidence. In consonance with the OECD (Organisation for Economic Co-operation and Development) regulatory guidelines for model development, Applicability Domain (AD) estimations were also carried out to ascertain the reliability of the predictions [43]. A plot of standardized residual against leverage (h) values, termed as Williams Plot was used to investigate the applicability domain of the FB-QSAR/FB-QSSR models. Leverage (h) of a compound in the original variable space, which measures its influence on the model is defined as:

$$h_i = x_i^T (X^T X)^{-1} \times (i = 1, \ldots, n) \qquad (13)$$

where, x_i is the descriptor vector of the considered compound and X is the model matrix derived from the

training set descriptor values. The warning leverage h* is defined as follows:

$$h^* = 3 \times p'/n \qquad (14)$$

where n is the number of training compounds and p′ is the number of model variable plus one parameters. A leverage greater than the warning leverage h* (h > h*) means that the compound predicted response can be extrapolated from the model, but, the predicted value must be used with great care. When the leverage value of a compound is lower than the leverage h*, the probability of accordance between predicted and actual values is as high as that for the training set chemicals.

Compounds whose residual value exceeds twice the standard deviation were considered as statistical outliers. Williams plots shown in Fig. 4 allows a graphical depiction of both outliers and "out of domain" compounds for all the three (BACE1, BACE1-BACE2, BACE1-CAT-D) models.

In FB-QSAR model (model1), compound 14 was the only compound with a value of standardized residual (−2.488) higher than twice the standard deviations from the mean value of pIC_{50}. Compounds with h > h*

(h* = 0.279) are out of the model's applicability domain. As observed in Fig. 4A all the compounds in training set lie within the model's applicability domain. Specifically, compound 14 (h = 0.004) although in the outlier zone, but placed perfectly in the applicability domain. In FB-QSSR model (Model2) four compounds were detected as statistical outliers (12, 29, 15, and 48). One compound (47) has leverage value greater than the warning leverage h* which indicates that the predicted value of that compound must be used for model development with great care. In FB-QSSR model (Model3) three compounds (14, 32, and 48) were detected as statistical outliers and no compounds were identified to be out of applicability domain. Removal of outliers from QSAR models remains a contentious issue, as it is often termed as method of polishing R^2 value, unless obvious evidence supports the uniqueness of the outlier compound [44]. In light of the good statistical quality index, and the relatively low number of outliers, evident for all the models, the models were considered inclusive of outliers.

Descriptor-based K-Means clustering performed earlier for the rational division of training set and test set reveals

Fig. 4 Williams plot for the obtained model BACE1 (**A**), BACE1-BACE2 (**B**), BACE1-CATD (**C**): Test set compounds are denoted as *red triangle*, Training set compounds are denoted as *blue diamond*

that outlier compounds share the same descriptor domain space as the rest in the training set, further the activity distribution (Y) follows a normal distribution pattern; hence the question of they being an X or a Y dependent outlier was ruled out. In light of this observation, we are convinced that these compounds should be X/Y dependent outliers having a relationship that is unable to fit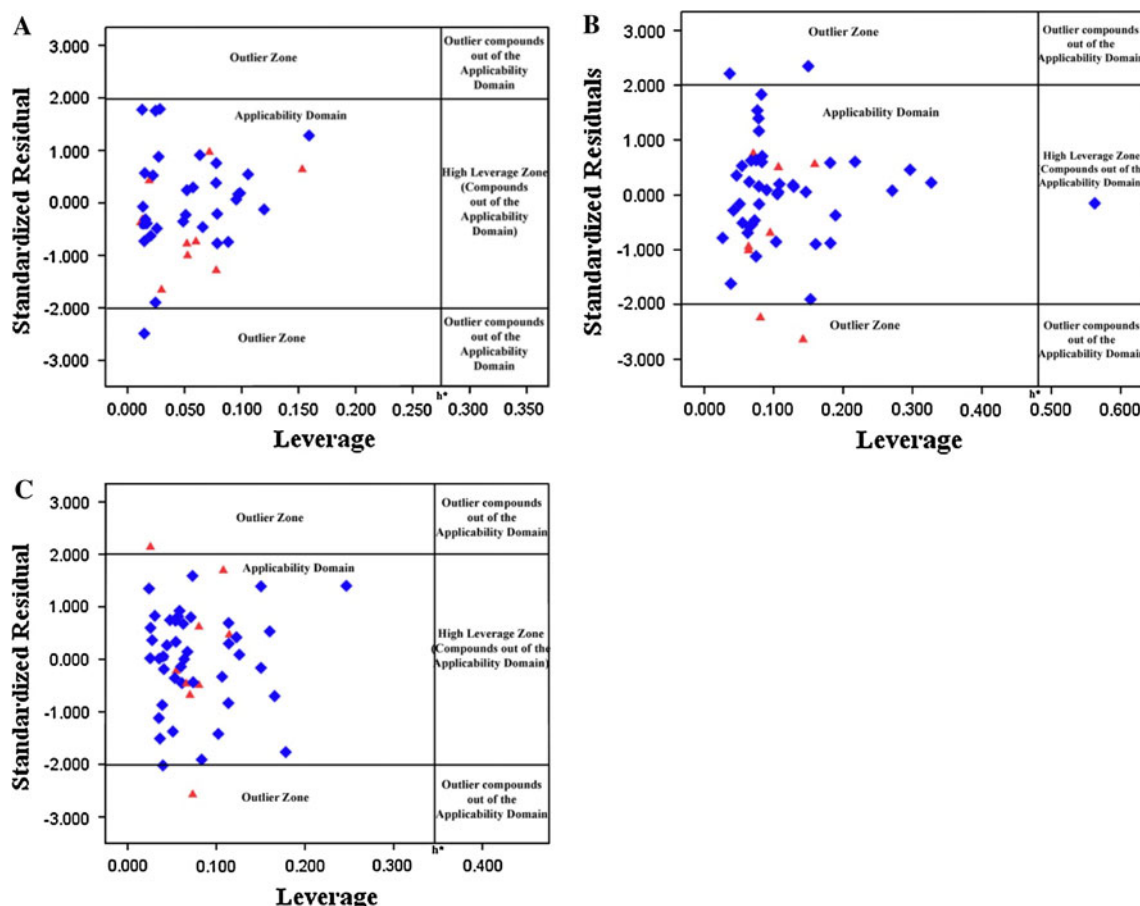 with the current equation. The overall statistical quality index as evident from the footnotes provided in Eqs. 7, 8 and 9 reveal that BACE1,BACE1-BACE2 and BACE1-CAT-D models satisfy the acceptability criteria for a valid QSAR model ($R^2 > 0.6$, $q^2 > 0.5$, $R^2 - q^2 < 0.3$ and $RPred^2 > 0.5$) [45].

Identification of superior R-Group fragments

To identify superior R-group fragments that could help in optimizing potency and selectivity a systematic study was undertaken to construct affinity and selectivity heat maps for the marked R_1, R_2 and R_3 fragments toward its intended target BACE1 and its off targets BACE2 and CAT-D.

FB-QSAR models were also constructed using the experimentally reported IC_{50} values for BACE2 and CAT-D. The FB-QSAR models for BACE2 and CAT-D along with the statistical parameters are provided in the supplementary information as their lone objective was to aid in the construction of affinity heat maps [46] shown in Fig. 5. Since FB-QSAR modeling involves the use of conventional 2D descriptors for the marked R-groups and those being employed as independent variables in constructing regression based QSAR models, the derived model coefficients can be treated as a quantitative estimate of the activity contribution of each R-group.

To construct heat maps, R-group coefficients obtained from the FB-QSAR/FB-QSSR models were transformed into Affinity/Selectivity maps, wherein each compounds R-group/target combination is assigned a color ranging from green (low) to red (high).

Multi objective QSPR (MO-QSPR)

Response/desirability profiling was carried out to inspect the response surface produced by fitting the predicted responses (Y_i) using a quadratic equation model based on levels of the independent variables. The *Profiler* option was used to examine the predicted values for the dependent variables (Y_i) at a maximum combination of levels (↑activity and ↑selectivity) of the independent variables. The objective of MO-QSPR is to identify the levels for the Independent variables that will maximize the desirability of the responses on the Dependent variable. The optimal value of the independent variables that produce the most desirable response was identified. Our FB-QSAR/FB-QSSR modelling highlights MM, VB4, VL, $^3\chi_p^v$, logP, BL, LYC, vsa_hyd, lip_acc as decisive descriptors in influencing the activity and selectivity of the compounds in question. Hence, to attain a trade-off between activity and selectivity, MO-QSPR was carried out by fitting the predicted responses obtained from Eqs. (7, 8, 9) using the response/desirability profiling option present in STATISTICA. The values of L_i, U_i, and T_i were assigned as stated earlier. The curvature parameters s and t were set at 1 to obtain linearity. Optimum desirability at exact grid points option, which performs an exhaustive search, was used in the study. It is evident from the statistical results obtained, that a R_D^2 value of 0.95 and Q_D^2 value of 0.75 assures the statistical significance of the obtained model. The predicted and actual desirability values for all the three models are provided in supplementary information. The optimal level for the independent variables obtained from MO-QSPR is

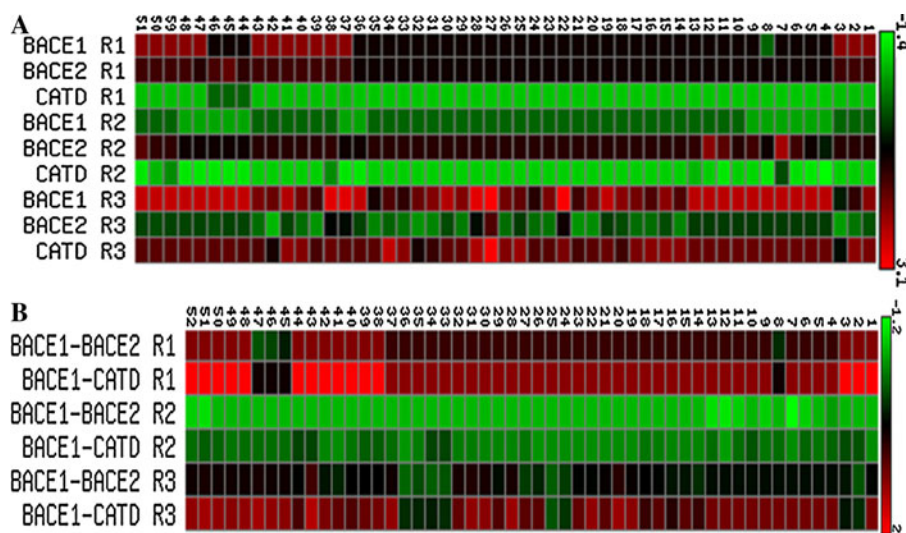**Fig. 5** Heat Map showing R-group Affinity (**A**), Selectivity (**B**) profile of BACE1, BACE2, and CATD

Table 3 Results of the multi objective QSPR (MO-QSPR)

S. No	Descriptors	Optimum level
1	MM(R3)	132.27
2	NHBA(R1)	2.1388
3	NHBA(R2)	−0.1948
4	VL(R3)	7.6091
5	VB4(R3)	8.8932
6	logP (R1)	0.56037
7	Bond Lipole (R3)	10.902
8	$(^3\chi_p^v)$/KCV3PI (R3)	2.7438
9	lip_acc (R1)	3.2776
10	vsa_hyd (R2)	74.051
11	LYC (R2)	0.80075

shown in Table 3 and the derringer desirability plot obtained are provided in supplementary information.

Inverse QSAR based virtual combinatorial library design
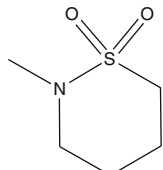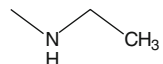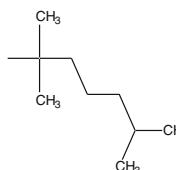
Ideal R-group fragments that help in achieving a subtle balance between activity and selectivity were identified from the desirable values obtained from MO-QSPR study. Accordingly fragments shown in Table 4 having descriptor values near the optimal level (vide table: 3) were subsequently used as the key R-group motifs for analog design. Bioisosteric analogs that are ostensibly known to yield molecular entities imparting similar biological properties

were considered as ideal replacements for carrying out analog design around the defined core scaffold. Bioisosteric analogs were identified from a biosisosteric database containing fragments of synthetic tractability using the Brood [47] software. The isosters identified for the defined R-group fragments were filtered based on the desirable levels evident from Table 3. The filtered R-groups thus obtained were used to enumerate a virtual library of all possible products that could be combinatorially obtained by using the QuaSARCombiGen module in MOE. In a retrospective fashion the obtained FB-QSAR/QSSR models were used to predict the activity and selectivity of the enumerated compounds. On the whole we were able to obtain some promising molecules with an overall desirability level of (0.91 and 0.88 as shown in Table 5) against the predicted level of 1. Further, the reliability of the prediction for the designed molecules was ascertained by ensuring whether the designed molecules fitted into the AD of the trained model.

Conclusion
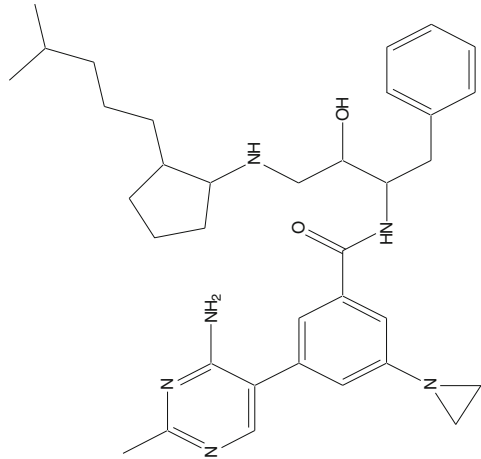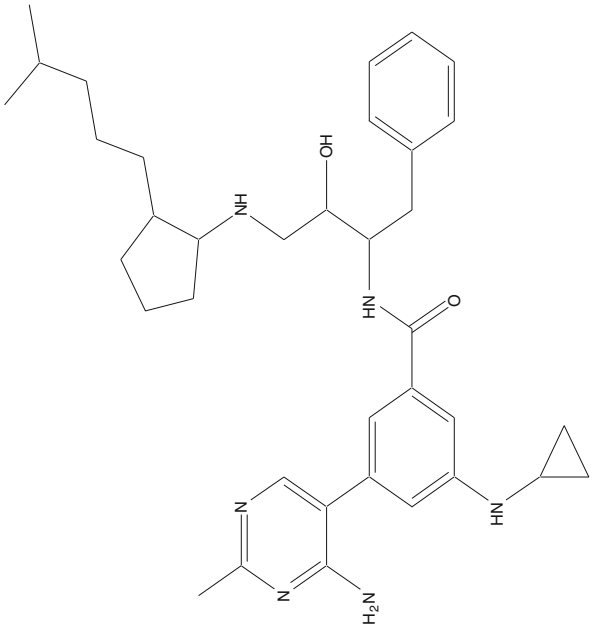
The work reported herein provides a pragmatic solution to rationalize fragment based drug discovery using a novel *In-silico* approach that employs FB-QSAR in combination with multi objective approach. FB-QSAR offers the advantage to weigh the fragment contributions to molecular activity/selectivity, which in turn provides a head start for the identification of fragments with enhanced activity

Table 4 R-group fragments considered for Isosteric replacement along with the descriptor values

S. No	Fragments	R-Group	Descriptor levels										
			MM	NHBA	NHBA	VL	VB4	logP	BL	KCV3PI $(^3\chi_p^v)$	lip_acc	vsa_hyd	LYC
1		R_1	NA	2	NA	NA	NA	−0.2	NA	NA	3	NA	NA
2		R_2	NA	NA	0	NA	NA	NA	NA	NA	NA	50	0.107
3		R_3	127.2	NA	NA	7.95	6.9	NA	12.29	1.56	NA	NA	NA

NA not applicable

Table 5 Hit molecules with predicted pIC$_{50}$, desirability for BACE1, BACE1-BACE2, and BACE1-CAT-D and predicted Overall desirability

S. No	Compound structure	Hits ID	Pred BACE pIC$_{50}$	Pred D BACE1	Pred BACE-BACE2 pIC$_{50}$	Pred D BACE1-BACE2	Pred BACE1-CATD pIC$_{50}$	Pred D BACE1-CATD	Pred D
1		1007	8.29	1	1.75	0.83	2.99	0.90	0.91
2		1008	8.29	1	1.74	0.84	2.74	0.81	0.88

and selectivity. It is particularly intriguing to note that the method employed herein has a unique advantage of handling multi objectives; a long standing issue that is often overlooked in QSAR modeling. We are optimistic that our FB-QSAR/FB-QSSR and MO-QSPR methodology will drive fragnomics in identifying potential fragments for an exploratory chemical synthesis.

References

1. Hansch C, Fujita T (1964) ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626

2. Zartler ER, Shapiro MJ (2005) Fragonomics: Fragment-based drug discovery. Curr Opin Chem Biol 9:366–370

3. Verdonk ML, Hartshorn MJ (2004) Structure-guided fragment screening for lead discovery. Curr Opin Drug Discov Dev 7:404–410

4. Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins 11:29–34

5. Caflisch A, Miranker A, Karplus M (1993) Multiple copy simultaneous search and construction of ligands in binding sites. J Med Chem 36:2142–2167

6. Evensen E, Joseph-McCarthy D, Karplus M (1997) MCSSV2. Harvard University, Cambridge

7. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28:849–857

8. Jencks WP (1981) On the attribution and additivity of binding energies. Proc Natl Acad Sci USA 78:4046–4050

9. Free SM, Wilson JW (1964) A mathematical contribution to structure- activity studies. J Med Chem 7:395–399

10. Fujita T, Ban T (1971) Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. J Med Chem 14:148–152

11. Vassar R, Bennett BD, Babu-Khan S, Kahn S, Mendiaz EA, Denis P, Teplow DB, Ross S, Amarante P, Loeloff R, Luo Y, Fisher S, Fuller J, Edenson S, Lile J, Jarosinski MA, Biere AL, Curran E, Burgess T, Louis JC, Collins F, Treanor J, Rogers G, Citron M (1999) β-Secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. Science 286:735–741

12. Yan R, Bienkowski MJ, Shuck ME, Miao H, Tory MC, Pauley AM, Brashier JR, Stratman NC, Mathews WR, Buhl AE, Carter DB, Tomasselli AG, Parodi LA, Heinrikson RL, Gurney ME (1999) Membrane-anchored aspartyl protease with Alzheimer's disease β-secretase activity. Nature 402:533–537

13. Selkoe DJ (1999) Translating cell biology into therapeutic advances in Alzheimer's disease. Nature 399:A23–A31

14. Baldwin ET, Bhat TN, Gulnik S, Hosur MV, Sowder RC 2nd, Cachau RE, Collins J, Silva AM, Erickson JW (1993) Crystal structures of native and inhibited forms of human cathepsin D: implications for lysosomal targeting and drug design. Proc Natl Acad Sci USA 90:6796–6800

15. Saunders AJ, Kim TW, Tanzi RE (1999) BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate down syndrome region of chromosome 21. Science 286:1255a

16. Charrier N, Clarke B, Cutler L, Demont E, Dingwall C, Dunsdon R, East P, Hawkins J, Howes C, Hussain I, Jeffrey P, Maile G, Matico R, Mosley J, Naylor A, O'Brien A, Redshaw S, Rowland P, Soleil V, Smith JK, Sweitzer S, Theobald P, Vesey D, Walter SD, Wayne G (2008) Second generation of hydroxyethylamine BACE-1 inhibitors: optimizing potency and oral bioavailability. J Med Chem 51:3313–3317

17. Clarke B, Demont E, Dingwall C, Dunsdon R, Faller A, Hawkins J, Hussain I, MacPherson D, Maile G, Matico R, Milner P, Mosley J, Naylor AA, O'Brien A, Redshaw S, Riddell D, Rowland P, Soleil V, Smith JK, Stanway S, Stemp G, Sweitzer S, Theobald P, Vesey D, Walter DS, Ward J, Wayne G (2008) BACE-1 inhibitors part 1: identification of novel hydroxy ethylamines (HEAs). Bioorg Med Chem Lett 18:1011–1016

18. Clarke B, Demont E, Dingwall C, Dunsdon R, Faller A, Hawkins J, Hussain I, MacPherson D, Maile G, Matico R, Milner P, Mosley J, Naylor AA, O'Brien A, Redshaw S, Riddell D, Rowland P, Soleil V, Smith JK, Stanway S, Stemp G, Sweitzer S, Theobald P, Vesey D, Walter DS, Ward J, Wayne G (2008) BACE-1 inhibitors part 2: Identification of hydroxyl ethylamines (HEAs) with reduced peptidic character. Bioorg Med Chem Lett 18:1017–1021

19. Beswick P, Charrier N, Clarke B, Demont E, Dingwall C, Dunsdon R, Faller A, Gleave R, Hawkins J, Hussain I, Johnson CN, MacPherson D, Maile G, Milner P, Naylor A, O'Brien A, Redshaw S, Riddell D, Rowland P, Skidmore J, Soleil V, Smith K, Stanway S, Stemp G, Stuart A, Theobald P, Vesey D, Walter DS, Ward J, Wayne G (2008) BACE-1 inhibitors. Part 3: identification of hydroxyl ethylamines (HEAs) with nanomolar potency in cells. Bioorg Med Chem Lett 18:1022–1026

20. Congreve M, Aharony D, Albert J, Callaghan O, Campbell J, Carr RA, Chessari G, Cowan S, Edwards PD, Frederickson M, McMenamin R, Murray CW, Patel S, Wallis N (2007) Application of fragment screening by X-ray crystallography to the discovery of aminopyridines as inhibitors of beta-secretase. J Med Chem 50:1124–1132

21. Murray CW, Callaghan O, Chessari G, Cleasby A, Congreve M, Frederickson M, Hartshorn JM, McMenamin R, Patel S, Wallis N (2007) Application of fragment screening by X-ray crystallography to the discovery of aminopyridines as inhibitors of beta-secretase. J Med Chem 50:1116–1123

22. Nicolaou AC, Brown N, Pattichis SC (2007) Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Devel 3:316–324

23. Derringer G, Suich R (1980) Simultaneous optimization of several response variables. J Quality Technol 12:214–219

24. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41:856–864

25. HyperChem 7.52 Hypercube Inc., Gainesville Fl 32608 USA

26. Molecular Operating EnVironment (MOE) (2009) Chemical computing group, Montreal

27. SPSS Version 15.0 (2009) SPSS Inc., Chicago

28. Vijayan RS, Ghoshal N (2008) Structural basis for ligand recognition at the benzodiazepine binding site of GABAA alpha 3 receptor, and pharmacophore-based virtual screening approach. J Mol Graph Model 27:286–298

29. TSAR Version 3.3 (2007) Accelrys Inc., San Diego

30. Hajduk PJ (2006) Puzzling through fragment-based drug design. Nat Chem Biol 12:658–659

31. Cruz-Monteagudo M, Borges F, Cordeiro MN (2008) Desirability-based multiobjective optimization for global QSAR studies: application to the design of novel NSAIDs with improved

analgesic, antiinflammatory, and ulcerogenic profiles. J Comput Chem 29:2445–2459

32. STATISTICA, 6.0 (2001) Statsoft_Inc

33. Laskowski RA, Thornton JM, Humblet C, Singh J (1996) X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. J Mol Biol 259:175–201

34. Nissink JWM, Verdonk ML, Klebe G (2000) Simple knowledge-based descriptors to predict protein-ligand interactions. methodology and validation. J Comput Aided Mol Des 14:787–803

35. Bissantz C, Kuhn B, Stahl M (2010) A medicinal chemist's guide to molecular interactions. J Med Chem 53:5061–5084

36. Panigrahi SK, Desiraju GR (2007) Strong and weak hydrogen bonds in the protein-ligand interface. Proteins 67:128–141

37. Desiraju GR, Sarkel S (2004) N–H …O, O–H …O, and C–H …O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. Proteins 54:247–259

38. Barratt E, Bronowska A, Vondrasek J, Cerny J, Bingham R, Phillips S, Homans SW (2006) Thermodynamic penalty arising from burial of a ligand polar group within a hydrophobic pocket of a protein receptor. J Mol Biol 362:994–1003

39. Verloop A, Hoogenstraaten W, Tipker J (1976) Development and application of new steric substituent parameters in drug design. In Drug Design Academic Press, New York

40. Hall LH, Kier LB (1992) The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling, Reviews in Computational Chemistry ed, Wiley

41. Hansch C, Leo AJ (1979) Substituent constants for correlation analysis in chemistry and biology. Wiley, New York

42. Golbraikh A, Tropsha A (2002) Beware of q2. J Mol Graph Model 20:269–276

43. http://www.oecd.org/dataoecd/33/37/37849783.pdf

44. Cronin MTD, Schultz TW (2003) Pitfalls in QSARs. THEO-CHEM 622:39–52

45. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression based QSARs. Environ Health Perspect 111: 1361–1375

46. Pavlidis P, Noble WS (2003) Matrix2png: a utility for visualizing matrix data. Bioinformatics 19:295–296

47. OpenEye (2006) OpenEye Scientific Software Santa Fe NM