Chapter 4

FRAGMENT BASED DRUG DISCOVERY FOR THE DESIGN OF SELECTIVE BACE1 INHIBITORS – INSIGHTS FROM FB-QSAR, FB-QSSR, MULTI OBJECTIVE (MO-QSPR) AND MIF STUDIES
FRAGMENT BASED DRUG DISCOVERY FOR THE DESIGN OF SELECTIVE BACE1 INHIBITORS – INSIGHTS FROM FB-QSAR, FB-QSSR, MULTI OBJECTIVE (MO-QSPR) AND MIF STUDIES

4.1 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 Linear Free Energy Relationship (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 (Nuclear Magnetic Resonance (NMR), X-Ray, Surface Plasmon Resonance (SPR) and, Mass Spectrometry (MS)) and by few structure based computational techniques like Multi-Copy Simultaneous Search (MCSS) [4-6], GRID [7]. Fragment screening, conducted using biochemical assay or a biophysical technique, are 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 quantifying weak binding. Jencks concept of additivity phenomena [8] that laid the foundation for Fragment Based Drug Design (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, a hybrid Fragment Based-Quantitative Structure Activity Relationship (FB-QSAR) modeling approach was developed that incorporates 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.
To underpin the concept of FB-QSAR in a pragmatic fashion, in this study the attention was turned towards BACE1, an extensively studied aspartic protease, involved in etiopathogenesis and progression of Alzheimer’s disease (AD). From a therapeutic point of view, selectivity towards BACE1 over other human aspartic proteases like BACE2 and Cat-D (Cathepsin D) [11] is expected to be important, as promiscuous binding would bring about undesirable side effects. Targeting BACE1 without inhibiting ~50% identical enzyme termed BACE2 [12], which is hypothesized to be physiologically important protein, and Cat-D, a ubiquitous protein, which shares ~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, disclosed by GlaxoSmithKline pharmaceuticals was considered in the study [13-16]. 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 peptido-mimetic inhibitors exhibit high potency in cell free and cell based assays, but fail to turn up as promising drug candidates due to poor pharmacokinetic 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 [17, 18].

To address the issue of target selectivity, Fragment Based-Quantitative Structure Selectivity Relationship (FB-QSSR) models were developed using activity values reported for two promiscuous targets namely BACE2 and CAT-D [13-16]. Conventional 2D QSAR and 3D QSAR approaches focus on a single objective and overlooks the importance of multiple objectives required for drug-like behavior [19]. As an adaptive response to the changing scenario, Multi Objective Optimization-Quantitative Structure Property Relationship (MO-QSPR) study was carried out, using Derringer and Suich algorithm [20] 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 the MO-QSPR approach was used in Inverse(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 [21].

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 herald the approach of revitalizing QSAR to address the demands of modern day drug discovery.

### 4.1.1 Selectivity in drug design

Selective inhibition of drug target is very important to achieve desired therapeutic effect. The withdrawal or discontinuing of most of the drugs in the advance drug development stage is mainly due to the undesirable side effects. The environment in which drugs act is complex, with many potential interaction partners. Proteins, DNA, RNA, lipids, sugars, metabolites, and other small molecules all have the potential to interact with a drug, and in many cases these unexpected interactions lead to undesired and often severe side effects. A drug's selectivity can often be explained by how selectively it binds to receptors. Selectivity in drug action is related to the structural specificity of drug binding to receptors. The ideal drug would interact only with a molecular target that causes the desired therapeutic effect but not with molecular targets
that cause unwanted adverse effects. Although no such drug has yet been discovered (i.e., all drugs currently in clinical use have the potential to cause adverse effects as well as therapeutic effects), several determinants of drug selectivity can be considered in an attempt to reach the goal. All drugs have multiple effects, both desirable (beneficial) and undesirable (adverse effects, or “side effects”). Selectivity is partly intrinsic to the nature of the drug-receptor interaction. Selectivity of drug action can be conferred by at least two classes of mechanisms, that include (i) the cell-type specificity of receptor subtypes, and (ii) the cell-type specificity of receptor-effector coupling [22].

Proper tuning of binding selectivity is a primary objective in the discovery and optimization of a compound on the path toward developing a drug. Designing a drug with the appropriate balance of avoidance of undesirable targets (narrow selectivity) is a continual drug development challenge. This objective can be attained through the rational approaches that can guide the tuning of selectivity, and a number of generalizable strategies. The general principles that underlie to drive selectivity should allow for more efficient design of compounds with desirable selectivity profiles. Conceptually, the problem of designing for a particular selectivity profile is significantly more complex than designing for high affinity to a single target. The underlying problem is challenging because it is necessary to evaluate energy differences for each ligand binding to a panel of targets and decoys rather than to a single desirable target. In very simple yet useful terms, achieving narrow selectivity involves recognizing and exploiting differences between targets and decoys.

There are six selective drug designing strategies, based on five principles (shape, electrostatics, flexibility, hydration, and allostery) of binding and complementarity that fall under structure-based approaches and can be employed to gain binding selectivity for a given target. These are, (i) optimization of ligand charges specifically for the target and against the decoy; (ii) displacement of a high-energy water molecule in the target that is not present in the decoy; (iii) binding to an allosteric pocket in the target that is not present in the decoy; (iv) creating a clash with the decoy receptor but not the target receptor, where the decoy is unable to alleviate the clash by structural rearrangement; (v) binding to a receptor conformation that is accessible in the target but inaccessible in the decoy; (vi) creating an interaction with the target receptor but
not the decoy receptor, where the decoy is unable to form the interaction by structural rearrangement [23, 24]. Nowadays there are many drug designing software available to model the five principles, which can be exploited in selective drug designing.

4.1.2 Bioisosteric replacement

In drug discovery, when lead compound that targets a particular disease is discovered, it often lacks the required potency and pharmacokinetic properties suitable for the development of a viable clinical candidate. Hence, to improve these properties, a very well known conservative medicinal chemistry technique is employed, which involves the replacement of one fragment in a bioactive molecule with another fragment that is known to closely mimic the original fragment or moiety. This is a usual process in drug discovery, while moving from hit to lead or from lead to clinical candidate. Replacement or modification of functional groups with other groups, having similar properties is known as isosteric or bioisosteric [25] replacement. Bioisosterism represents the rational modification of lead compounds into safer and more clinically effective agents. Bioisosterism recognizes that certain functional groups have similar biological, chemical and physical properties. The concept of bioisosterism is often considered to be qualitative and intuitive.

Bioisosteres have been classified as either classical or nonclassical. Isosteres, which obey Grimm’s hydride displacement law and Erlenmeyer’s [26] definition of isosteres, are called as classical bioisosteres. Classical bioisosteres represent the results of an early appreciation of the concept and encompass structurally simple, mono-, di-, and trivalent atoms or groups and ring equivalents. In contrast, non classical bioisosteres extend the concept to structural elements that offer a more subtle and sophisticated form of biochemical mimicry, relying upon functionality that can differ quite substantially in electronic, physicochemical, steric, and topological representation from that being emulated.

Bioisosteric replacement frequently introduces structural changes that can be beneficial or harmful depending on the context, with size, shape, electronic distribution, polarizability, dipole, polarity, lipophilicity, and pKₐ, potentially playing key contributing roles in molecular recognition and mimicry. In the contemporary practice of medicinal chemistry, the
development and application of bioisosteres have been adopted as a fundamental tactical approach useful to address a number of aspects associated with the design and development of drug candidates. The established utility of bioisosteres is broad in nature, extending to improving potency, enhancing selectivity, altering physical properties, understanding and optimizing drug-target interactions and specificity, improving drug permeability, reducing or redirecting metabolism, eliminating or modifying toxicophores, and acquiring novel intellectual property.

As an established and powerful concept in medicinal chemistry, the application of bioisosteric replacement plays an important role in drug discovery [27, 28]. In addition to chemical experience and intuition, computational methods have also been utilized to identify or predict bioisosteres.

4.1.3 2D-QSAR

Quantitative Structure-Activity Relationship (QSAR) is based on the general principle of medicinal chemistry that the biological activity of a ligand or compound is related to its molecular structure or properties, and structurally similar molecules may have similar biological activities. Such structural information is encoded in molecular descriptors and a QSAR model defines mathematical relationships between descriptors and biological activities of known ligands to predict unknown ligands’ activities. Developing QSAR using various physicochemical parameters has been an important task in lead optimization. A wide variety of molecular descriptors are available and descriptor selection is an integral process in QSAR modelling. 2D QSAR models are generated using descriptors derived from the two-dimensional graphical representation of a molecule. The correlation of physicochemical properties to activity is generally carried out using multivariable regression methods. Regression Analysis models the activity of molecules through an equation constructed using a combination of physicochemical properties. The coefficient for each variable in the equation can, consequently, be examined to determine the extent to which each property contributes towards the activity of the molecule. QSAR finds immense applicability and accelerates the drug discovery process by (i) Forecasting the activity of the molecules yet to be synthesized.
(ii) Helping to identify the pertinent features of the molecules that play a decisive role in driving biological activity. (iii) As a cost effective strategy, this can be retrospectively used for designing compounds using inverse QSAR paradigm. 2D-QSAR is attractive because predicting molecular properties and activities based on 2D molecular structures is simple, fast and robust. 2D-QSAR methods allow modelling of a wide variety of ligands or compounds including cases where 3D structures of target (receptor) are not available. Several modifications and new additions have been made to the 2D-QSAR method in which fragment based-QSAR is an effective approach, capable of assisting in lead optimization [29].

4.1.4 FB-QSAR and FB-QSSR

In standard 2D QSAR, the molecular descriptors representing all physicochemical properties are for a whole molecule. However, in the modern drug discovery practice, chemists are often focusing only on a few substitutes in a molecule. Therefore, in the QSAR study chemists want to know how the small structural changes in substitutes affect the bioactivity of the drug candidates. The physicochemical properties of substitutes may be more important and more sensitive than the properties of a whole molecule in QSAR studies. This demand led the computational chemist to develop a QSAR method that only uses the substituent property to correlate with the activity. This QSAR approach is called as FB-QSAR. In recent times, FB-QSAR has emerged as a versatile tool to explore the chemical and biological space of data sets of compounds. FB-QSAR approaches have evolved from classical use in the generation of standard QSAR models into advanced drug design tools for database mining, pharmacokinetic property prediction and optimization of multiple parameters. FB-QSAR method is the ideal approach for the purpose of lead optimization [30].

In the FB-QSAR model, the framework of a molecular family is divided into several fragments and the total binding free energy between ligand and its receptor is considered as the summation of contributions, from all fragments. In two-dimensional FB-QSAR model building, the only requirements for model generation are the 2D structures of the data set compounds and the corresponding property values. The fragment structure pattern change in the training set compounds can be related to their corresponding experimental biological
parameters using Multiple Linear Regression (MLR) or Partial Least Squares (PLS) [31] analysis in order to generate FB-QSAR models. The only difference between FB-QSAR and FB-QSSR is that the former uses the biological response like inhibitor activity to relate with compounds physicochemical property, while the later uses the affinity differences (ΔpIC50) obtained between the two different biological responses, as dependent variable, to model the selectivity.

FB-QSAR, either alone or in combination with conventional 3D-QSAR, docking or virtual screening strategies, is being successfully applied to identify hits or lead candidates against a variety of target proteins. The FB-QSAR models can be applied to search large chemical databases to identify novel active compounds for a given biological target. The fragment-based method is also applied to simplify the computational analysis of ligand binding and to map out different structural and chemical elements required for binding affinity and biological activity. The concept of this approach involves the principle that each unique interaction in the binding site represented by different molecular fragments should produce an optimized compound with binding affinity that is the sum of the individual essential interactions [32].

The early known Fragment based-QSAR model was developed by Free-Wilson [9] in which the biological activity values were correlated with the presence and absence of molecular substituents. This method focuses on the effect of substituent changes on the overall biological activity. The Free-Wilson approach addresses structure-activity relationship in a congeneric series as described in eq. (1) [9].

\[ BA = \sum A_i X_i + \mu \]  

(1).

where, BA is the biological activity, \( \mu \) 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 FBDD, 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 FBDD. To weigh the R-group’s contribution for activity/selectivity, Fujita-Ban approach is employed as a method for estimating de novo group contributions, which is expressed in eq. (2) [10].

\[ \text{Log BA} = \sum G_i X_i + \mu \]  

where, \( \mu \) 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.

4.1.5 Multi Objective Optimization (MOOP) in drug discovery

Modern drug discovery involves simultaneous optimization of many physicochemical and biological properties that transcend the traditional focus on bioactivity alone. Its success or failure depends on the simultaneous control of numerous, often conflicting, molecular and pharmacological properties. In recent times, the awareness has increased that successful drug discovery increasingly requires more than just finding a molecule that is highly potent at the target, it needs to be as close to optimal for other desired physicochemical properties too. This is perhaps important for compounds that have activity against multiple targets or are promiscuous, which has enabled molecule repurposing in some cases. The process of resolving many requirements is termed ‘multi-objective optimization’. Multi-objective optimization strategies represent a new approach to capture the occurrence of varying optimal solutions based on trade-offs among the objectives taken into account. In view of this, multi-objective optimization aims to discover a set of satisfactory compromises that may in turn be used to find the global optimal solution by optimizing numerous dependent properties simultaneously. In drug discovery, the need for a simultaneous, multi-objective optimization of various molecular properties with efficacy data is well understood. Similarly, the need for optimizing the absorption, distribution, metabolism and excretion (ADME), toxicity and selectivity data that are generated for compounds [33] much earlier in the process is also strongly acknowledged. Drug discovery increasingly requires the simultaneous optimization of many measured and calculated properties.
Optimization problems can be divided into two broad categories, single-objective or multi-objective, depending on the number of criteria that their objective function encodes. The solution to a single-objective optimization (SOOP) problem involves finding the optimum to a one-objective function, whereas a MOOP problem requires the more difficult task of finding solutions that satisfy a whole spectrum of objectives. Multi-objective problems are often characterized by vast, complex search spaces with various local optima that are difficult to explore exhaustively. When the objectives are in competition, the task of finding a solution is further complicated because there is no single best solution that outperforms all the other solutions in all criteria [19].

The straightforward approach to find compromise solutions when numerous objectives are present is to transform the problem to a single-objective one by combining the multiple objectives. An example of this approach is the weighted-sum-of-objective-functions method [34]. According to this method, a weight is associated with each objective function and the weighted sum of the functions is taken as the new composite (or fitness) function, as defined by the following eq. (3).

\[
f(n) = w_1(Objective_1) + w_2(Objective_2) + \ldots + w_n(Objective_n)
\]  

(3)

where \(f(n)\) is the fitness function, and \(w_1, w_2, \ldots\) are the user defined weights.

Desirability functions and Pareto optimization [35] are the two methods used for solving a typical Multi-objective problem. These methods have been applied to numerous problems, including those in the area of compound and library optimization. One challenge is that as the number of properties to be optimized increases, the required calculation time greatly increases, hence, efforts to combine properties into composites (e.g. as desirability functions) should be made to decrease the number of properties undergoing Pareto optimization [36].

These MOOP methods are implemented in commercial software packages. One commercially available set of software tools for performing Pareto optimization of compounds and compound libraries is found in the Accelrys Pipeline Pilot™ and Discovery Studio [37].
programs. These can be used to optimize a set of compound libraries to be both maximally drug-like and maximally diverse. Another commercial implementation of Pareto optimization is found in SAS [38]. Desirability-based multi-objective optimization is implemented in the commercial packages JMP [39], Minitab [40], STATISTICA [41] and Stat-Ease [42].

4.2 Materials (Software used in this study)
The molecular modeling and statistical software’s used in this study are Hyperchem [43], MOE [44], SPSS [45], TSAR [46], STATISTICA [41].

4.2.1 Brood
Brood [47] is commercial software developed by Openeye Incs. It is designed to search databases of chemical fragments to identify and select fragments with similarities to the query fragment and can perform bioisosteric replacements to develop new leads. Brood also generates analogs to leads by assembling and replacing different fragments based on shape, electrostatics and molecular properties. Brood is accompanied by CHOMP which serves for fragmentation of a molecule and MERGE for fragment merging. Additionally, if a crystal structure for the target protein is known, Brood can use information from the protein structure to eliminate fragments which will not fit in the binding pocket and will verify protein-ligand close contacts. The fragment replacement approaches have been applied using a variety of molecular similarity descriptors with the Brood program pioneering the technique from the perspective of shape-based template comparison [48]. Shape-based fragment-similarity tools automate and extend this common practice. Rather than sampling thousands of fragments that might come into a chemist’s mind, such tools can search millions of fragments and winnow the list, so a modeler or chemist can assess the viability of just a few compounds, knowing that all of the fragment replacements have the same shape as the original fragment, balancing the need for change and the need for stability.
4.3 Computational Methods

4.3.1 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 [13-16]. 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 [43]. Upon obtaining reasonable starting geometries, batch optimization of all the modeled compounds were carried out using MMFF94 force field, implemented in MOE [44]. 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 [45] for the rational division of training and test series [49]. Clustering was carried out using more than one hundred 2D whole molecule descriptors representing topology, structural information and group counts, calculated using MOE [44] and TSAR [46]. 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 provided vide Table 1 of appendix I and the common scaffold is shown in Fig 1.

![Fig 1. The common core scaffold considered for FB-QSAR/FB-QSSR.](image-url)
4.3.2 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 R₁, R₂ and R₃ as shown in Fig 1. The fragmentation of a molecule assumes that the fragments follow a modular binding mode and exert an additivity phenomenon of the foundational principles, which FBDD relies on [50]. 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.

4.3.4 FB-QSAR/FB-QSSR and MO-QSPR methods

The 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 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; \( \bar{y}_i \) obtained from forward QSAR approaches (FB-QSAR/FB-QSSR) in to a dimensionless partial 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 [20]. 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 in eq. (4).

\[
d_i = \begin{cases} 
0 & \text{if } \hat{y}_i \leq L_i \\
\left( \frac{\hat{y}_i - L_i}{T_i - L_i} \right)^s & \text{if } L_i < \hat{y}_i < T_i \\
1 & \text{if } \hat{y}_i \geq T_i = U_i 
\end{cases} \quad (4).
\]

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. This can be achieved through eq. (5).

\[
D = (d_1 \times d_2 \times \ldots \times d_k)^{\frac{1}{k}} \quad (5).
\]
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 $R^2_D$ as proposed by Cruz-Monteagudo et al [51] and provided in eq. (6). $R^2_D$ is analogous to the determination coefficient $R^2$.

$$R^2_D = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum (D_{yi} - D_{fi})^2}{\sum (D_{yi} - \bar{D}_{yi})^2} \tag{6}$$

where, $D_{yi}$ is the actual desirability, $D_{fi}$ 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. $R^2_D$ 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 $R^2 (Q^2)$ and given in eq. (7).

$$Q^2_D = 1 - \frac{SSE_{LOO-CV}}{SSTO} = 1 - \frac{\sum (D_{yi} - D_{fi}(LOO-CV))^2}{\sum (D_{yi} - \bar{D}_{yi})^2} \tag{7}$$

where, $Q^2_D$ is the overall predictive desirability, $SSE_{LOO-CV}$ and $D_{yi}(LOO-CV)$ are the leave one out cross validation error sum of square and the predicted desirability by LOO-CV, respectively.

The response values ($Y_{pred}$) 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 [52]. 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.

4.3.5 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.

4.4 Results & Discussion

The chemical structures of the compounds considered in the present study are shown in Table 1 of appendix I and their observed and predicted values are enlisted in Table 1. The best FB-QSAR, FB-QSSR models obtained are presented as equations (8, 9 and 10) and the corresponding statistical values are provided as foot note below the equations. The descriptors obtained in these equations and their properties are provided vide Table 2 of appendix I.

Table 1. Observed and predicted affinity/selectivity of BACE1, BACE2, CAT-D and their affinity differences are indicated by Δ.

<table>
<thead>
<tr>
<th>NAME</th>
<th>BACE1 (pIC50) (Obs)</th>
<th>BACE1 (pIC50) (Pred)</th>
<th>BACE1 - BACE2 (Obs)</th>
<th>BACE1 - BACE2 (Pred)</th>
<th>BACE1 - CAT-D (Obs)</th>
<th>BACE1 - CAT-D (Pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.16</td>
<td>7.13</td>
<td>1.8</td>
<td>1.68</td>
<td>1.53</td>
<td>1.64</td>
</tr>
<tr>
<td>2*</td>
<td>7.52</td>
<td>7.86</td>
<td>2.42</td>
<td>2.32</td>
<td>2.4</td>
<td>2.57</td>
</tr>
<tr>
<td>3</td>
<td>6.92</td>
<td>6.43</td>
<td>2.21</td>
<td>2.22</td>
<td>2.57</td>
<td>2.4</td>
</tr>
<tr>
<td>4</td>
<td>6.13</td>
<td>6.69</td>
<td>1.58</td>
<td>1.85</td>
<td>2.4</td>
<td>1.92</td>
</tr>
<tr>
<td>5</td>
<td>7.23</td>
<td>6.69</td>
<td>1.94</td>
<td>1.66</td>
<td>3.1</td>
<td>1.76</td>
</tr>
<tr>
<td>6</td>
<td>6.92</td>
<td>6.69</td>
<td>1.6</td>
<td>1.46</td>
<td>2.01</td>
<td>1.85</td>
</tr>
<tr>
<td>7</td>
<td>7.49</td>
<td>6.69</td>
<td>1.1</td>
<td>1.07</td>
<td>1.13</td>
<td>1.57</td>
</tr>
<tr>
<td>8*</td>
<td>5.84</td>
<td>5.62</td>
<td>0.95</td>
<td>0.83</td>
<td>0.63</td>
<td>0.85</td>
</tr>
<tr>
<td>9</td>
<td>6.22</td>
<td>6.69</td>
<td>1.13</td>
<td>1.46</td>
<td>1.89</td>
<td>1.75</td>
</tr>
<tr>
<td>10</td>
<td>7.89</td>
<td>7.23</td>
<td>2.14</td>
<td>1.69</td>
<td>2.32</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>7.41</td>
<td>7.23</td>
<td>1.33</td>
<td>1.48</td>
<td>2.14</td>
<td>1.81</td>
</tr>
<tr>
<td>12*</td>
<td>6.74</td>
<td>7.23</td>
<td>0.56</td>
<td>1.29</td>
<td>1.23</td>
<td>1.42</td>
</tr>
<tr>
<td>13</td>
<td>7.14</td>
<td>7.23</td>
<td>1.34</td>
<td>1.37</td>
<td>1.8</td>
<td>1.78</td>
</tr>
<tr>
<td>14</td>
<td>5.23</td>
<td>6.29</td>
<td>0.99</td>
<td>1.45</td>
<td>0.31</td>
<td>1.1</td>
</tr>
<tr>
<td>15</td>
<td>7.48</td>
<td>6.79</td>
<td>2.25</td>
<td>1.66</td>
<td>1.88</td>
<td>1.37</td>
</tr>
</tbody>
</table>

123
### 4.4.1 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. (8).

\[
pIC_{50} = 3.59668 + 0.016594 \times \text{MM (R3)} + 1.07651 \times \text{NHBA (RQ)} - 0.541331 \times \text{NHBA (R2)}
\]  

(8)

<table>
<thead>
<tr>
<th>Compound</th>
<th>pIC50</th>
<th>MM (R3)</th>
<th>NHBA (RQ)</th>
<th>NHBA (R2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>6.52</td>
<td>1.07</td>
<td>1.61</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>6.26</td>
<td>1.74</td>
<td>1.75</td>
<td>1.29</td>
</tr>
<tr>
<td>18*</td>
<td>6.1</td>
<td>1.94</td>
<td>1.8</td>
<td>1.45</td>
</tr>
<tr>
<td>19</td>
<td>7.4</td>
<td>2.01</td>
<td>1.9</td>
<td>2.16</td>
</tr>
<tr>
<td>20</td>
<td>7.37</td>
<td>2.25</td>
<td>2.12</td>
<td>2.07</td>
</tr>
<tr>
<td>21</td>
<td>6.12</td>
<td>1.92</td>
<td>1.86</td>
<td>1.4</td>
</tr>
<tr>
<td>22</td>
<td>7.77</td>
<td>1.55</td>
<td>1.73</td>
<td>1.74</td>
</tr>
<tr>
<td>23</td>
<td>5.87</td>
<td>1.52</td>
<td>1.76</td>
<td>0.86</td>
</tr>
<tr>
<td>24*</td>
<td>5.77</td>
<td>1.1</td>
<td>1.32</td>
<td>0.44</td>
</tr>
<tr>
<td>25</td>
<td>6.22</td>
<td>0.75</td>
<td>1.02</td>
<td>0.47</td>
</tr>
<tr>
<td>26</td>
<td>6.8</td>
<td>1.37</td>
<td>1.47</td>
<td>1.52</td>
</tr>
<tr>
<td>27</td>
<td>7.85</td>
<td>1.42</td>
<td>1.29</td>
<td>1.43</td>
</tr>
<tr>
<td>28</td>
<td>7.7</td>
<td>2</td>
<td>1.99</td>
<td>2.25</td>
</tr>
<tr>
<td>29</td>
<td>7.08</td>
<td>2.26</td>
<td>1.7</td>
<td>1.83</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>1.84</td>
<td>2.01</td>
<td>1.77</td>
</tr>
<tr>
<td>31</td>
<td>6.48</td>
<td>2</td>
<td>2.08</td>
<td>1.9</td>
</tr>
<tr>
<td>32*</td>
<td>6.64</td>
<td>2.16</td>
<td>1.99</td>
<td>2.31</td>
</tr>
<tr>
<td>33</td>
<td>6.74</td>
<td>1.44</td>
<td>1.1</td>
<td>0.64</td>
</tr>
<tr>
<td>34</td>
<td>5.95</td>
<td>1.03</td>
<td>1.05</td>
<td>0.33</td>
</tr>
<tr>
<td>35</td>
<td>5.61</td>
<td>1.23</td>
<td>1.2</td>
<td>0.65</td>
</tr>
<tr>
<td>36</td>
<td>6.19</td>
<td>0.96</td>
<td>0.97</td>
<td>0.38</td>
</tr>
<tr>
<td>37</td>
<td>6.42</td>
<td>1.85</td>
<td>1.92</td>
<td>1.96</td>
</tr>
<tr>
<td>38</td>
<td>8.3</td>
<td>2.18</td>
<td>2.4</td>
<td>3.09</td>
</tr>
<tr>
<td>39*</td>
<td>8.3</td>
<td>2.08</td>
<td>2.38</td>
<td>2.71</td>
</tr>
<tr>
<td>40</td>
<td>6.92</td>
<td>2.43</td>
<td>2.42</td>
<td>2.34</td>
</tr>
<tr>
<td>41</td>
<td>7.47</td>
<td>1.83</td>
<td>2.09</td>
<td>1.96</td>
</tr>
<tr>
<td>42</td>
<td>7.74</td>
<td>2.21</td>
<td>2.2</td>
<td>1.77</td>
</tr>
<tr>
<td>43</td>
<td>7.92</td>
<td>2.76</td>
<td>2.89</td>
<td>3.67</td>
</tr>
<tr>
<td>44</td>
<td>8.1</td>
<td>2.73</td>
<td>2.35</td>
<td>2.93</td>
</tr>
<tr>
<td>45</td>
<td>6.6</td>
<td>1.22</td>
<td>1.2</td>
<td>0.85</td>
</tr>
<tr>
<td>46</td>
<td>6.55</td>
<td>1.04</td>
<td>0.95</td>
<td>1.02</td>
</tr>
<tr>
<td>47</td>
<td>6.8</td>
<td>0.82</td>
<td>0.89</td>
<td>1.1</td>
</tr>
<tr>
<td>48*</td>
<td>7.3</td>
<td>1.97</td>
<td>2.6</td>
<td>2.23</td>
</tr>
<tr>
<td>49*</td>
<td>8.22</td>
<td>2.23</td>
<td>2.51</td>
<td>2.99</td>
</tr>
<tr>
<td>50</td>
<td>8.52</td>
<td>2.68</td>
<td>2.54</td>
<td>3.11</td>
</tr>
<tr>
<td>51</td>
<td>8.3</td>
<td>2.46</td>
<td>2.31</td>
<td>2.98</td>
</tr>
<tr>
<td>52</td>
<td>7.96</td>
<td>2.18</td>
<td>2.35</td>
<td>3.36</td>
</tr>
</tbody>
</table>

*Compound marked as * indicates test set compound.
The BACE1 (SMLR) activity 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 R₃ appears in the equation with a positive coefficient, which suggests that a bulky substituent is acceptable at this position. The presence of NHBA (R₁) 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 (R₂) in this equation with a negative coefficient indicates that hydrogen bonding acceptor groups at R₂ 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) [13]. This module of MOE is very close in spirit to knowledge-based tools like X-Site [53] and SuperStar [54]. 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.
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.

Visual interpretation of the map displays a large hydrophobic patch around the core moiety and around the R₃ 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 trifluoro methyl, with a potential to accommodate even bulkier group. The occurrence of molecular mass for the R₃ fragment with a positive coefficient as shown in eq. (8) 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 R₃ fragment possessing an amide linker binds to the S₂' site, in contrast R₃ 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 does not take 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 structures 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 R₁ fragment housed in the S2 site. Based on the crystallographic information (2VIZ) [14], it is evident that the carbonyl oxygen of lactam mediates a strong hydrogen bond (2.87 Å) with Asn 294. The modeling hypothesis is in consonance with this finding as it reveals that NHBA of R₁ 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 R₁ fragment. Occurrence of NHBA (R₂) in this equation with a negative coefficient implies that hydrogen bonding acceptor groups at R₂ position are unfavorable for biological activity. Non bonded iso contour preference map points out that the R₂ 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 Å 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 Thr 293, 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) R₂ Fragment), in contrast O linked derivatives forms a hydrogen bond of the type (Thr293-O-H (Don)---O-Ligand (Acc) R₂ Fragment), which clearly highlights that the former mediates a stronger hydrogen bonding interaction in comparison to the later, as evident from literatures [55-58]. 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.

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

The FB-QSSR model 2 (eq. (9)), which evaluates the selectivity criteria for BACE1 over BACE2 had an explained variance of 77.3% and a predicted variance of 65.2%.

\[
\Delta(p\text{IC}_{50 \text{BACE1}} - p\text{IC}_{50 \text{BACE2}}) = 0.280745 + 0.224262 \times \text{VB4} (R_3) - 0.256772 \times \log P (R_1) - 0.031452 \times \text{BL} (R_3) - 0.296178 \times \frac{1}{P^3} (R_3) + 0.606211 \times \text{lip_acc} (R_1) - 0.012459 \times \text{vsa_hyd} (R_2)
\]

\[N = 43, R^2 = 0.805, R^2_{\text{adj}} = 0.773, P_{\text{adj}} = 0.806, Q^2 (LOO) = 0.739, \text{Randomized } R^2 = 0.352, F_{\text{test}} = 24.786, \text{PRESS } = 3.264, R^2_{\text{pred}} = 0.652, (R^2 - R^2_0) / R^2 = 0, \text{ and } k = 1.\]

In this study, the FB-QSSR modeling suggests that (Verloop) VB4 [59] at R3 position, which signifies the steric influence of the substituent, is a positive contributor for selectivity over BACE2. The in silico modeling results agree well with SAR findings wherein, compounds 15, 19, 20 and 29 possessing linear and bulky fragments at R3 position are more selective compared to compounds 34, 36 and 14 which possess only small R3 fragments. The appearance of bond lipole with a negative coefficient at R3 position suggests that less lipophilic groups are favorable for selective inhibition of BACE1 over BACE2. This modeling hypothesis can be substantiated by comparing the R3 substituent of selective compounds 19, 21 (less lipophilic R3-groups) over compounds 16, 17 (highly lipophilic R3-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 [60]. The appearance of connectivity index \(\frac{1}{P^3}\) with negative coefficient at R3 position suggests that, the branching of fragments at R3 position is detrimental for selectivity. This computational modeling result can
be further substantiated by SAR upon comparing compound 20, (decreased $\chi_\rho^2$ index and high selectivity) with compound 36, (increased $\chi_\rho^2$ index and non selective). A holistic picture of the nature of $R_3$ fragment can be drawn from the above results, which imply that steric fragments with less number of branching and less lipophilic property are favorable for $R_3$ fragments. Since chain branching and lipophilicity are competing properties, here it is presumed that lipophilicity dictates pharmacokinetic property and branching dictates the pharmacodynamic property. The appearance of logP [61] 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).

4.4.3 FB-QSSR model for BACE1 selectivity over CAT-D (Model 3)
The FB-QSSR model 3 (eq. (10)), which evaluates the selectivity criteria for BACE1 over CAT-D had an explained variance of 81.2% and a predicted variance of 77.6%.

$$\Delta (pIC_{50BACE1} - pIC_{50CAT-D}) = -1.8697 - 0.131048 \times VL (R_3) + 0.377365 \times VB4 (R_3) + 0.363482 \times LYC (R_2) + 0.966115 \times lip\_acc (R_1)$$ (10).

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

The 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 R₂ 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 R₁ 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 R₁ 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, but here it is unconvincing to corroborate the FB-QSSR finding by a hypothetical evidence, as docking would not serve as a proof of concept in the present scenario.

4.4.4 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 (R²). Since the value of R² tends to be inflated as the number of terms in the equation increases, R²adj was also calculated for expository reasons. The near value of R² and R²adj, for all the obtained models, as evident from the footnotes provided with equations (8, 9, and 10) ensures the absence of over fitting. Correlation plot between the experimental pIC₅₀ and predicted pIC₅₀ 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 R²cv (q²) 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 [62].

\[
R^2_{\text{pred}} = \frac{SD - \text{PRESS}}{SD} \tag{11}
\]

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 [62].

\[
(R^2 - R^2_o)/R^2 < 0.1 \text{ or } (R^2 - R^{2'}_o)/R^2 < 0.1 \tag{12}
\]

And

\[
k \text{ or } k' \text{ close to 1} \tag{13}
\]

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.
Fig 3. Correlation coefficient between the experimental pIC₅₀ and predicted pIC₅₀ values for the training set (Blue-round) and the test set compounds (Red-diamond) at 95% confidence level. BACE1 (A), BACE1-BACE2 (B), and BACE1-CAT-D (C).

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 [63]. 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:
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:

$$h^* = \frac{3 \times p'}{n}$$

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, and therefore, the predicted value must be used with great care. When the leverage value of a compound is lower than the warning 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, allow a graphical depiction of both outliers and “out of domain” compounds for all the three (BACE1, BACE1-BACE2, BACE1-CAT-D) models.
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.

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 4 A, 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 with great care. In FB-QSSR model (Model3) three compounds (14, 32, and 48) were detected as statistical outliers and no compound was 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 [64]. 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 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 equations 8, 9 and 10 reveal that both BACE1 and BACE1-BACE2, 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 $R^2_{pred} > 0.5$) [65].

4.4.5 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 R₁, R₂ and R₃ fragments toward its intended target BACE1 and its off targets BACE2 and CAT-D.

The FB-QSAR models were also constructed using the experimentally reported IC₅₀ values for BACE2 and CAT-D. The FB-QSAR models for BACE2 and CAT-D along with the statistical parameters are provided in appendix I, as their lone objective was to aid in the construction of affinity heat maps [66] 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 compound R-group /target combination is assigned a color ranging from green (low) to red (high).

**Fig 5.** Heat map showing R-group affinity (A), selectivity (B) profile of BACE1, BACE2, and CAT-D.

### 4.4.6 Multi objective QSPR (MO-QSPR)

The 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 (\(Y_j\)) at a maximum combination of levels (\(\dagger\)activity and \(\dagger\)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. The FB-QSAR/FB-QSSR modelling highlights MM, VB4, VL, \( \chi_p^v \), 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 equations (8, 9, and 10) 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 \( R^2_D \) value of 0.95 and \( Q^2_D \) 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 appendix I. The optimal level for the independent variables obtained from MO-QSPR is shown in Table 2 and the derringer desirability plot obtained are provided in appendix I.

**Table 2. Results of the Multi objective QSPR (MO-QSPR)**

<table>
<thead>
<tr>
<th>S. No</th>
<th>Descriptors</th>
<th>Optimum Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MM(R3)</td>
<td>132.27</td>
</tr>
<tr>
<td>2</td>
<td>NHBA(R1)</td>
<td>2.1388</td>
</tr>
<tr>
<td>3</td>
<td>NHBA(R2)</td>
<td>-0.1948</td>
</tr>
<tr>
<td>4</td>
<td>VL(R3)</td>
<td>7.6091</td>
</tr>
<tr>
<td>5</td>
<td>VB4(R3)</td>
<td>8.8932</td>
</tr>
<tr>
<td>6</td>
<td>logP (R1)</td>
<td>0.56037</td>
</tr>
<tr>
<td>7</td>
<td>Bond Lipole (R3)</td>
<td>10.902</td>
</tr>
<tr>
<td>8</td>
<td>((\chi_p^v) / KCV3PI (R3))</td>
<td>2.7438</td>
</tr>
<tr>
<td>9</td>
<td>lip_acc (R1)</td>
<td>3.2776</td>
</tr>
<tr>
<td>10</td>
<td>vsa_hyd (R2)</td>
<td>74.051</td>
</tr>
<tr>
<td>11</td>
<td>LYC (R2)</td>
<td>0.80075</td>
</tr>
</tbody>
</table>

### 4.4.7 Inverse QSAR based virtual combinatorial library design

The 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 3 having descriptor values near the optimal level (vide...
Table 2) 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 bioisosteric database containing fragments of synthetic tractability using the Brood [43] software. The isosteres identified for the defined R-group fragments were filtered based on the desirable levels evident from Table 2. 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, some promising molecules with an overall desirability level of 0.91 (Table 4) were obtained 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 in to the AD of the trained model.

Table 3. R-group fragments considered for bioisosteric replacement along with the descriptor values.

<table>
<thead>
<tr>
<th>S. N</th>
<th>Fragments</th>
<th>R-</th>
<th>Descriptor Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Gr out M</td>
<td>NH</td>
<td>M BA BA 0</td>
</tr>
<tr>
<td>o</td>
<td>P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>-0.2</td>
<td>3 NA NA NA</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>50</td>
<td>50 0.10</td>
</tr>
<tr>
<td>3</td>
<td>7.2</td>
<td>12</td>
<td>12 1.56</td>
</tr>
</tbody>
</table>

*NA= Not applicable
Table 4. Hit molecules with predicted pIC50, desirability for BACE1, BACE1-BACE2, and BACE1-CAT-D and predicted Overall desirability

<table>
<thead>
<tr>
<th>S. No</th>
<th>Compound Structure</th>
<th>Hits ID</th>
<th>pIC50</th>
<th>Pred BACE1</th>
<th>Pred BACE1-BACE2</th>
<th>Pred BACE1-CAT-D</th>
<th>Pred Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Compound Structure" /></td>
<td>1007</td>
<td>8.29</td>
<td>1.75</td>
<td>1.75</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2.png" alt="Compound Structure" /></td>
<td>1008</td>
<td>8.29</td>
<td>1.74</td>
<td>0.84</td>
<td>0.81</td>
<td>0.88</td>
</tr>
</tbody>
</table>

4.5 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 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. Through this study, it is apparent that FB-QSAR/FB-QSSR and MO-QSPR methodology will drive fragnomics in identifying potential fragments for an exploratory chemical synthesis.
4.6 References


40. http://www.minitab.com
41. http://www.statsoft.com
42. http://www.statease.com


44. Molecular Operating Environment (MOE), version 2009.10, Chemical Computing Group, Montreal, Quebec, Canada, 2009.

45. SPSS, version 15.0, SPSS Inc., Chicago IL, 2009.


47. BROOD, version 1.1.1, OpenEye Scientific Software, Santa Fe, NM, 2006.


