Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking

Saw Simeon[1], Nuttapat Anuwongcharoen[1], Watshara Shoombuatong[1], Aijaz Ahmad Malik[1], Virapong Prachayasittikul[2], Jarl E. S. Wikberg[3], and Chanin Nantasenamat*[1]

[1]**Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand**
[2]**Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand**
[3]**Department of Pharmaceutical Biosciences, Uppsala University, Uppsala 751 24, Sweden**

ABSTRACT

Alzheimer's disease (AD) is a chronic neurodegenerative disease which leads to the gradual loss of neuronal cells. Several hypotheses for AD exists (e.g. cholinergic, amyloid, tau hypotheses, etc.). As per the cholinergic hypothesis, the deficiency of choline is responsible for AD therefore the inhibition of AChE is a lucrative therapeutic strategy for the treatment of AD. Acetylcholinesterase (AChE) is an enzyme that catalyzes the breakdown of the neurotransmitter acetylcholine that is essential for cognition and memory. A large non-redundant data set of 2,570 compounds with reported IC_{50} values against AChE was obtained from ChEMBL and employed in quantitative structure-activity relationship (QSAR) study so as to gain insights on their origin of bioactivity. AChE inhibitors were described by a set of 12 fingerprint descriptors and predictive models were constructed from 100 different data splits using random forest. Generated models afforded R^2, Q^2_{CV} and Q^2_{Ext} values in ranges of 0.66–0.93, 0.55–0.79 and 0.56–0.81 for the training set, 10-fold cross-validated set and external set, respectively. The best model built using the substructure count was selected according to the OECD guidelines and it afforded R^2, Q^2_{CV} and Q^2_{Ext} values of 0.92±0.01, 0.78±0.06 and 0.78±0.05, respectively. Furthermore, Y-scrambling was applied to evaluate the possibility of chance correlation of the predictive model. Subsequently, a thorough analysis of the substructure fingerprint count was conducted to provide informative insights on the inhibitory activity of AChE inhibitors. Moreover, Kennard-Stone sampling of the actives were applied to select 30 diverse compounds for further molecular docking studies in order to gain structural insights on the origin of AChE inhibition. Site-moiety mapping of compounds from the diversity set revealed three binding anchors encompassing both hydrogen-bonding and van der Waals interaction. Molecular docking revealed that compound **13** exhibited the lowest binding energy of –12.2 kcal/mol against human AChE, which is modulated by hydrogen bonding, π-π stacking and hydrophobic interaction inside the binding pocket. These information may be used as guidelines for the design of novel and robust AChE inhibitors.

Keywords: Acetylcholinesterase, Acetylcholinesterase inhibitor, Alzheimer's disease, Quantitative structure–activity relationship, QSAR, Data mining

INTRODUCTION

Neurodegenerative diseases is caused by the progressive loss of neural cells thereby leading to nervous system dysfunction (Beal, 1995; Kuca et al., 2016). In particular, Alzheimer's disease (AD) is a debilitating illness that is expected to triple by the year 2050 (Brookmeyer et al., 2007). AD is characterized by gradual cognitive impairment, memory loss and decline in speech, behavioral abnormality and eventually death. The pathological changes in the AD are mainly because of the dramatic loss of neurons in many areas of

*Corresponding author. E-mail: chanin.nan@mahidol.ac.th

the central nervous system accompanied by a great reduction in the levels of neurotransmitters. Thus, a promising therapeutic approach is to maintain the level of the acetylcholine (ACh) neurotransmitter by inhibiting the enzyme that is responsible for its breakdown.

Acetylcholinesterases (AChE) are mainly found in the joints of neuromuscular junction in which it not only serves as a synaptic transmission but also as an enzyme that catalyzes the breakdown of ACh to choline and acetic acid (Quinn, 1987). The structure of AChE showed that it is comprised of two active sites consisting of anionic and esteratic sites (Bourne et al., 1995). The anionic site is involved in the binding of the positive quaternary amine of ACh (Ordentlich et al., 1993). The substrate interacts with the 14 aromatic residues that forms the active site. Of these 14 aromatic residues, Trp84 is important for the enzyme activity because when it is replaced by alanine, the activity of the enzyme decreased by 3000-fold (Tougu, 2001). In contrast, the esteratic site hydrolyzed ACh to acetate and choline in which the existence of a catalytic triad (i.e. Ser200, His440 and Glu327) resembled that of chymotrypsin and other serine proteases (Harel et al., 1993). The mechanism of the hydrolysis starts from the carboxyl ester leads to the formation of an acyl-enzyme and choline. Finally, the acyl-enzyme undergoes nucleophilic attack by water molecules thereby regenerating the enzyme (Tougu, 2001).

AChE inhibitors form one of the most actively investigated classes of compounds having been labeled as a potential agent for the treatment of AD by inhibiting AChE from hydrolyzing ACh as to increase the level of ACh and maintain homeostasis (Birks, 2006). Generally, they can be classified into reversible and irreversible inhibitors. Reversible inhibitor bind to the enzyme at allosteric sites as to reduce the activity of the enzyme whether or not the enzyme has already bind the substrate or not. For example, tacrine was synthesized nearly five decades ago and in 1993 it has become the first drug to be marketed as a form of treatment for AD with approval from the U.S. Food and Drug Administration (Racchi et al., 2004). On the other hand, irreversible inhibitors such as metrifonate (Morris et al., 1998) bind to the target enzymes and dissociates very slowly from the enzyme via either covalent or non-covalent interactions (Kitz and Wilson, 1962).

Quantitative structure-activity relationship (QSAR) is a paradigm that enables the prediction of biological activities for compounds of interest as a function of their descriptors through the use of statistical or machine learning methods (Nantasenamat et al., 2009). Aside from the ability to predict the activity, QSAR models have been instrumental in enabling understanding on the origin of these biological activities by means of interpreting the descriptors used in building such models.

Historically, the first QSAR investigation of AChE inhibitors was reported by Mundy et al. (1978) almost 40 years ago in which the $\log(1/LD_{50}$ for a series of twelve substituted 0,0-dimethyl 0-(*p*-nitrophenyl) phosphorothioates and 0-analogs was predicted as a function of the octanol/water partition coefficient. Analysis of the literature of QSAR studies of AChE revealed that much of the early studies are classical QSAR models (i.e. Hansch and Free-Wilson approach) that are based on small congeneric compound set and primarily aimed at predicting AChE inhibition as to investigate the toxic effect of pesticides of various chemotypes belonging to either organophosphates (Mager, 1983; Aaviksaar, 1990) or carbamates (Su and Lien, 1980; Goldblum et al., 1981; Walters and Hopfinger, 1986). Recent QSAR studies are based on the use of large and heterogeneous data sets comprising of structurally diverse chemotypes. This include the study from Yan and Wang (2012) where they predicted AChE inhibition for a large set of 404 compounds using multiple linear regression and support vector machine. Furthermore, Lee and Barron (2016) performed a 3D-QSAR investigation on a large set of 341 compounds comprising of organophosphates and carbamates. Moreover, Veselinović et al. (2015) compiled a set of 278 organophosphates for which they developed QSAR models for predicting AChE inhibition using SMILES-based descriptors.

Research in this field had experienced two distinct transitions when viewed from biological and computational viewpoints. Biologically, early QSAR studies treat AChE as a biomarker of toxicity from pesticides while investigations from later years had shifted the focus by viewing AChE as a therapeutic target for the treatment of AD. In regards to the latter point, viewpoint on targeting AChE as a single target for treating AD is starting to be replaced by the multi-target concept in which the treatment for AD can be approached by a panel of key targets (Fang et al., 2015; Huang et al., 2011). Computationally, early studies are predominantly based on simple 2D-QSAR (Mundy et al., 1978; Su and Lien, 1980) while later years started to use more sophisticated approach for understanding AChE inhibition encompassing 3D-QSAR (Deb et al., 2012; Lee and Barron, 2016; Prado-Prado et al., 2012), molecular dynamics (Shen et al., 2002), molecular docking (Lu et al., 2011; Deb et al., 2012; Giacoppo et al., 2015), pharmacophore

modeling Lu et al. (2011); Gupta and Mohan (2014) and statistical molecular design (Andersson et al., 2014; J Prado-Prado et al., 2013).

Herein, we propose the first large-scale QSAR investigation for predicting AChE inhibition, which to the best of our knowledge represents the largest collection of 2,570 non-redundant compounds. QSAR models were built using interpretable learning methods (e.g. random forest) and descriptors (i.e. molecular fingerprints) as to unravel the underlying AChE inhibitory activity, which was performed in accordance with guidelines of the Organisation for Economic Cooperation and Development (OECD). Molecular docking was also performed on a chemically diverse set of compounds selected from active AChE inhibitors. Together, the ligand and structure-based approach employed in this study is anticipated to be useful in the design and development of robust AChE inhibitors.

MATERIALS AND METHODS

A summary of the workflow of this study is provided in Figure 1. Briefly, this included a large-scale QSAR model for predicting and analyzing the AChE inhibition, which was performed in accordance with the OECD guidelines as follows: (i) a data set with a defined endpoint; (ii) an unambiguous learning algorithm; (iii) a defined applicability domain of the QSAR model; (iv) using appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation of the QSAR model. Furthermore, molecular docking was also performed on a chemically diverse data set as to elucidate the underlying binding mechanism. To facilitate the reproducibility of the research work performed herein, the data set and codes used to perform the multivariate analysis as well as the results tables and figures are made freely available on GitHub at https://github.com/sawsimeon/AChE.

Data set

A data set of inhibitors against human AChE (Target ID CHEMBL220) were compiled from the ChEMBL 20 database (Gaulton et al., 2012) that is comprised of a total number of 9,242 bioactivity data points from 5,049 compounds. SMILES notations of the compounds were curated with ChemAxon's Standardizer (**?**) using the same parameter settings as described in our previous study (Simeon et al., 2016). The initial data set was assembled from several bioactivity measurement units including (in order of decreasing data size) IC_{50}, K_i, % activity, % inhibition, MIC, EC_{50}, etc. IC_{50} was selected for further investigation as they constituted the largest subset with 4,910 compounds. A closer look revealed that 1,301 compounds had no reported IC_{50} values or had lesser/greater than signs which were subjected to removal to produce in 3,609 compounds. Only data points having nM as the bioactivity unit were selected for further study, which produced 3,596 compounds. Furthermore, redundant compounds having different activity values were kept if the standard deviation of IC_{50} was less than 2 and this resulted in 2,571 compounds. Moreover, some compounds were found to have no SMILES notation associated with it and were thus removed. A final data set comprising of 2,570 compounds was obtained.

Description of inhibitors

AChE inhibitors were encoded by a vector of fingerprint descriptors accounting for its molecular constituents. Prior to calculating descriptors, salts were removed and tautomers were standardized using the built-in function of the PaDEL-Descriptor software (Yap, 2011).

Although fingerprint descriptors are able to capture the feature space of chemical compounds, however their ability to be used as descriptors for bioactivity modeling can vary. In fact, performance differences existing amongst the different fingerprint type has been the subject of several investigations into its utilization for bioactivity modeling. Riniker and Landrum (2013) benchmarked and assessed the performance of predictive models constructed from 2D fingerprint descriptors obtained from RDKit.

In this study, the suitability of 12 different fingerprint descriptors for predicting the bioactivity of AChE inhibitors was investigated. Table 1 summarizes the employed fingerprints along with their corresponding size, description and reference.

Additionally, the four molecular descriptors that are used to define the Lipinski's rule-of-five comprising of molecular weight (MW), logarithm of the octanol/water partition coefficient (ALogP), number of hydrogen bond donor (nHBDon) and number of hydrogen bond acceptor (nHBAcc) were also computed by the PaDEL-Descriptor software.

Feature selection

Collinearity is a condition where descriptor pairs are known to have intercorrelation, which not only add complexity to the model but could potentially give rise to bias. To remedy this, the *cor* function from the R package *stats* was used to find the pairwise correlation among descriptors, and descriptors in a pair with a Pearson's correlation coefficient greater than the threshold of 0.7 was filtered out using the *findCorrelation* function from the R package *caret* to obtain a smaller subset of descriptors (Kuhn, 2008).

Data splitting

To avoid the possibility of bias that may arise from a single data split when building predictive models (Puzyn et al., 2011), predictive models were constructed from 100 independent data splits and the mean and standard deviation values of statistical parameters were reported. The data set was split into internal and external sets in which the former comprises 80% whereas the latter constitutes 20% of the initial data set. The *sample* function from the R *base* package was used to split the data.

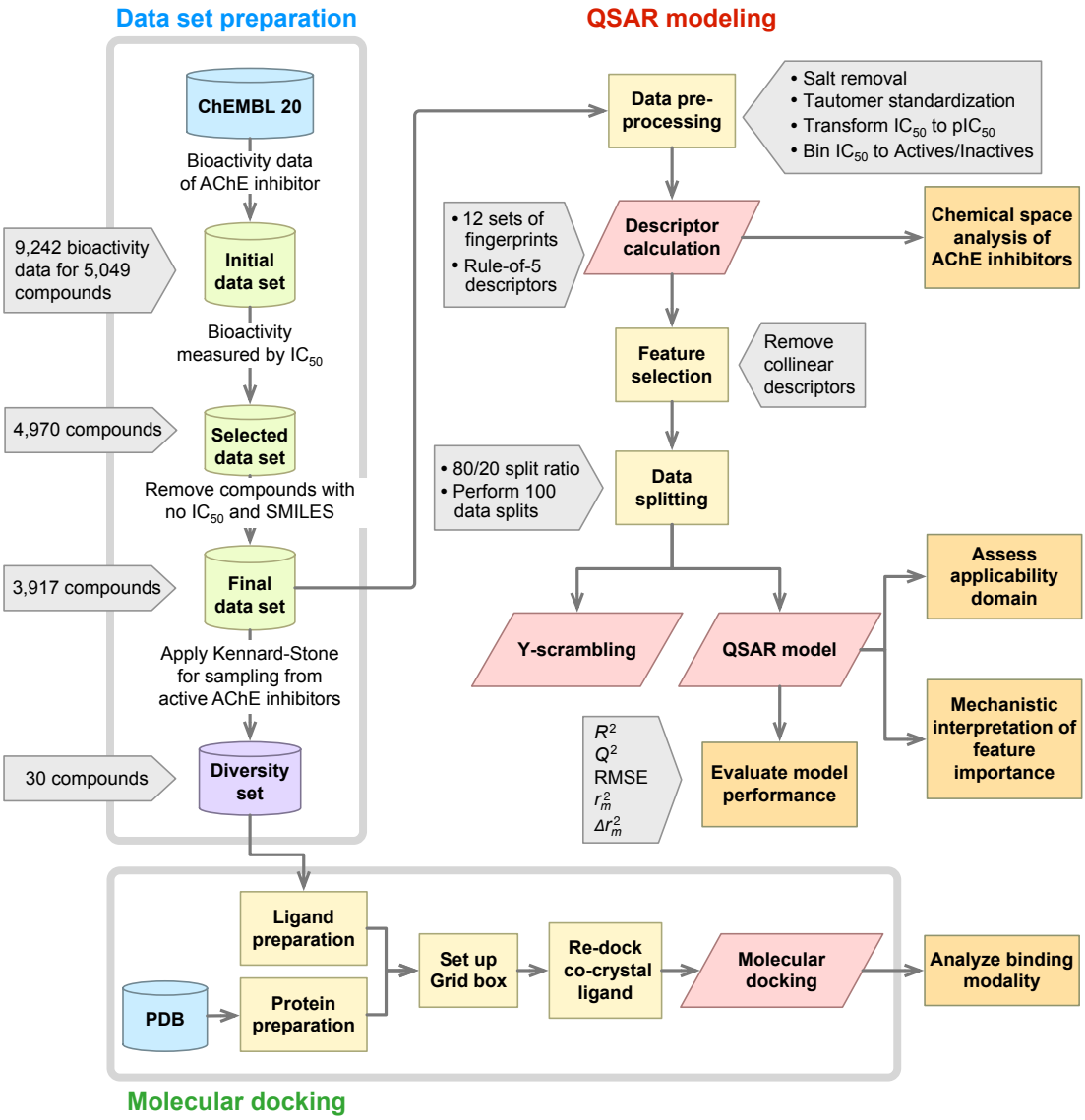

Figure 1. Workflow of QSAR modeling and molecular docking for investigating AChE inhibitory activity.

Table 1. Summary of 12 sets of fingerprint descriptors employed in this study.

No.	Fingerprint	Number	Description	Reference
1	CDK	1024	Fingerprint of length 1024 and search depth of 8	Steinbeck et al. (2003)
2	CDK extended	1024	Extends the fingerprint with additional bits describing ring features	Steinbeck et al. (2003)
3	CDK graph only	1024	A special version that considers only the connectivity and not bond order	Steinbeck et al. (2003)
4	E-state	79	Electrotopological state atom types	Hall and Kier (1995)
5	MACCS	166	Binary representation of chemical features defined by MACCS keys	Durant et al. (2002)
6	PubChem	881	Binary representation of substructures defined by PubChem	NCBI (2009)
7	Substructure	307	Presence of SMARTS patterns for functional groups	Laggner (2005)
8	Substructure count	307	Count of SMARTS patterns for functional groups	Laggner (2005)
9	Klekota-Roth	4860	Presence of chemical substructures	Klekota and Roth (2008)
10	Klekota-Roth count	4860	Count of chemical substructures	Klekota and Roth (2008)
11	2D atom pairs	780	Presence of atom pairs at various topological distances	Carhart et al. (1985)
12	2D atom pairs count	780	Count of atom pairs at various topological distances	Carhart et al. (1985)

Multivariate analysis

Supervised learning is to learn a model from labeled training data which can be used to make prediction about unseen or future data (James et al., 2013). This study constructs regression models, which affords the prediction of the continuous response variable (i.e. pIC_{50}) as a function of predictors (i.e. fingerprint descriptors).

Random forest (RF) is an ensemble classifier that is composed of several decision trees (Breiman, 2001). Briefly, the main idea behind RF is that instead of building a deep decision tree with an ever-growing number of nodes, which may be at risk for overfitting and overtraining of the data, rather multiple trees are generated as to minimize the variance instead of maximizing the accuracy. As such, the results will be more noisier when compared to a well-trained decision tree, yet, these results are usually reliable and robust. The *ranger* function from the R package *ranger*, which is a fast implementation of the RF algorithm that was used for constructing the models (Wright and Ziegler, 2015).

Validation of QSAR models

Model validation is an important process, which should be performed to ensure that a fitted model can accurately predict responses for future or unknown subjects. Two statistical parameters were used to evaluate the performance of the QSAR models consisting of Pearson's correlation coefficient (r) and root mean squared error (RMSE). The r value is a commonly used metric to represent the degree of relationship between two variables of interest. It can range from –1 to +1 in which negative values are indicative of negative correlation between two variables and vice versa. RMSE is a commonly used parameter to assess the relative error of the predictive model. The predictive performance of the QSAR models was verified by 10-fold cross-validation, external validation and Y-scrambling test.

The 10-fold cross-validation technique does not used the entire data set to build predictive model. Instead, it splits the data into training and testing data set by allowing model that is built with training data set us allow to assess the performance of the model on the testing data set. By performing repeats of the 10-fold validation, the average accuracies can be used to truly assess the performance of the predictive model.

Y-scrambling test was used to ensure the robustness of the predictive model not only to rule out the possibility of chance correlations but also to assess the statistical significance of R^2 and Q^2, ensuring the generalizability of QSAR model. The true Y-dependent variable (i.e. pIC_{50}) was randomly scrambled and the statistical assessment parameters are recalculated. Performance of the Y-scrambling test can be deduced from the regression line of the plot of R^2 versus Q^2. Intercept values for R^2 and Q^2 as denoted by iR^2 and iQ^2, respectively, were calculated. Negative iQ^2 is indicative of an acceptable QSAR model and that there is no chance correlation from the real model (Eriksson et al., 2003). Furthermore, r_m^2 and Δr_m^2 metrics introduced as introduced by Roy et al. (2013) were used to verify the robustness of the proposed QSAR model in which an acceptable QSAR model should give $r_m^2 > 0.5$ and $\Delta r_m^2 < 0.2$.

Applicability domain analysis

The applicability domain (AD) estimates the likelihood of reliable prediction for compounds on the basis of how similar they are to compounds used to build the model. Thus, compounds falling outside the AD may lead to unreliable predictions. The most common approach for determining AD is described by Gramatica (2007) and Tropsha et al. (2003), which is to compute the leverage values for each compound. The leverage value allows one to identify whether new compounds will lie within or outside the domain. Leverage values for all compounds are calculated via adjustment of X to give the hat matrix H:

$$H = X(X^T X)^{-1} X^T \tag{1}$$

where X is a two-dimensional matrix comprising of n compounds and m descriptors while X^T is the transpose of X. Meanwhile, the leverage value of the i^{th} compound (h_i) is the i^{th} diagonal element of H:

$$h_i = x_i^T (X^T X)^{-1} x_i \tag{2}$$

where x_i is the descriptor row-vector of the i^{th} compound. The warning leverage h^* is calculated by:

$$h^* = 3(p+1)/n \tag{3}$$

Practically, the leverage value along with the plot is often used to assess the AD of QSAR models. The plot is constructed by depicting the standardized residuals versus the leverage value for each compound's h_i. If the i^{th} compound has $h_i > h^*$ then it means that the i^{th} compound exerts a great influence on the QSAR model and may be excluded from the AD. In spite of this, it does not appear to be an outlier because its standardized residual may be small.

Molecular docking

The co-crystal structure of human AChE with donepezil (PDB ID: 4EY7) was retrieved from the Protein Data Bank and initially prepared by removing alternative side chains and water molecules. The protein was prepared via the rebuilding of bonds and the addition of missing hydrogen atoms. Subsequently, the protein was cleaned by merging the atomic charge and removing lone pair atoms, non-polar hydrogen atoms and non-standard amino acid residues. Grid box was set up to provide coverage of the active site of human AChE with a dimension of 40 × 30 × 40 Å (X, Y and Z axes of -13.987, -41.668 and 27.109, respectively). Molecular docking was consequently performed with AutoDock Vina (Trott and Olson, 2010) using default parameters. The docking protocol was validated in order to ensure its reliability for subsequent analysis of the studied compounds. This was performed by extracting the co-crystal ligand, donepezil, from the PDB file and re-docked to the co-crystal human AChE protein. The root mean squared deviation (RMSD) of the atomic position between the original orientation of the co-crystal ligand and the re-docked ligand is computed and is deemed acceptable if the RMSD value is less than or equal to 2.0 Å.

A set of 30 representative and chemically diverse compounds, which will be referred hereafter as the diversity set, were extracted from the full set of active AChE inhibitors (i.e. $IC_{50} < 1\ \mu M$) using the Kennard-Stone algorithm (Kennard and Stone, 1969). These compounds were used as ligands for molecular docking against the human AChE. The binding energy (kcal/mol) of AChE inhibitors were calculated according to the built-in scoring function of Autodock Vina and conformers providing the lowest binding energy were selected for further analysis of the binding mode. Furthermore, key-interacting residues and their moiety preferences were analyzed using LigPlot+ (Wallace et al., 1995) and the SiMMap web server (Bollback, 2006). Finally, three-dimensional structure of protein-ligand interaction was created and visualized using Pymol (Schrödinger, LLC, 2015a).

RESULTS AND DISCUSSION

Chemical space of AChE inhibitors

Navigation of the chemical space of AChE inhibitors was performed to gain insights into the structure-activity relationship by analyzing the Lipinski's rule-of-five descriptors. Chemical space analysis may provide important knowledge on the general character of compounds governing inhibitory properties of compounds. Exploratory data analysis was performed using the Lipinski's rule-of-five descriptors comprising of MW, ALogP, nHBDon and nHBAcc. MW represents the molecular size of a compound that is commonly used because of it can be easily interpreted and calculated as well as appropriate size of a compound is important for its passage via lipid membrane. ALogP is a widely used parameter for determining the lipophilicity of a compound and used in calculating the membrane penetration and permeability of compounds. nHBDon and nHBAcc describe the number of hydrogen bond donors and hydrogen bond acceptors, respectively, which is used to measuring hydrogen bonding capacity. Visualization of the chemical space of ALogP as a function of MW is shown in Figure 2, as to investigate the chemical space of AChE inhibitors. A dense distribution of inhibitors was observed within the space of MW starting from approximately 300 to 600 Da and within the space of ALogP ranging from approximately -2.5 to 2.5. In addition, the box plot of the Lipinski's descriptors is shown in Fig. 3. Compounds with negative ALogP values approximately of closer to 0.0 can be found in inactive inhibitors whereas most of the active inhibitors tend to possess approximately lower values in average of ALogP values.

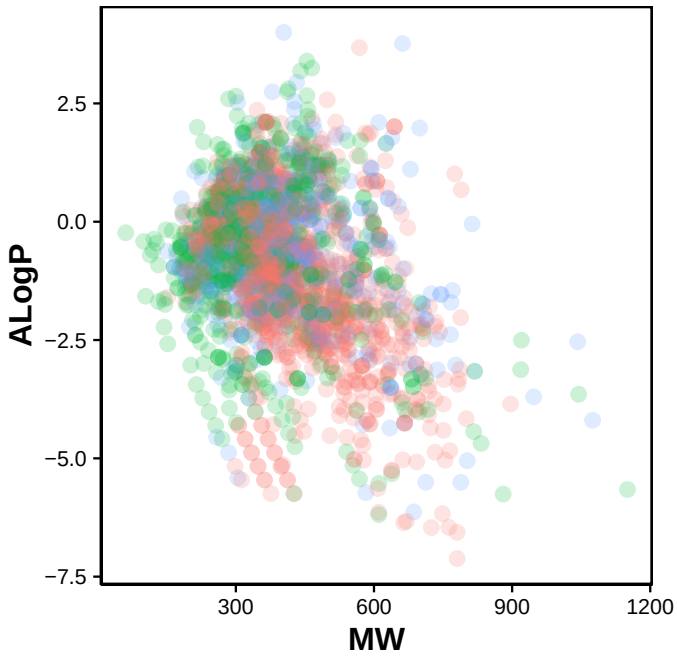

Figure 2. Chemical space of AChE inhibitors are shown as actives (green), inactives (red) and intermediates (blue)

Visual representation of the overall distribution of data values of Lipinski's descriptors is shown as box plots in Fig. 3 in which the ALogP, nHBAcc, nHBDon and MW are shown in the top-left, top-right, bottom-left and bottom-right corner, respectively. Analysis of the box plots revealed that there were no differences amongst the three bioactivity classes for nHBAcc and nHBAcc as deduced from the boundaries of the boxes (i.e. representing the first and third quartiles). ALogP and MW were found to display differences amongst the bioactivity classes. Particularly, ALogP values for actives were the lowest while negligible differences were observed for the other two classes. Furthermore, MW for actives were

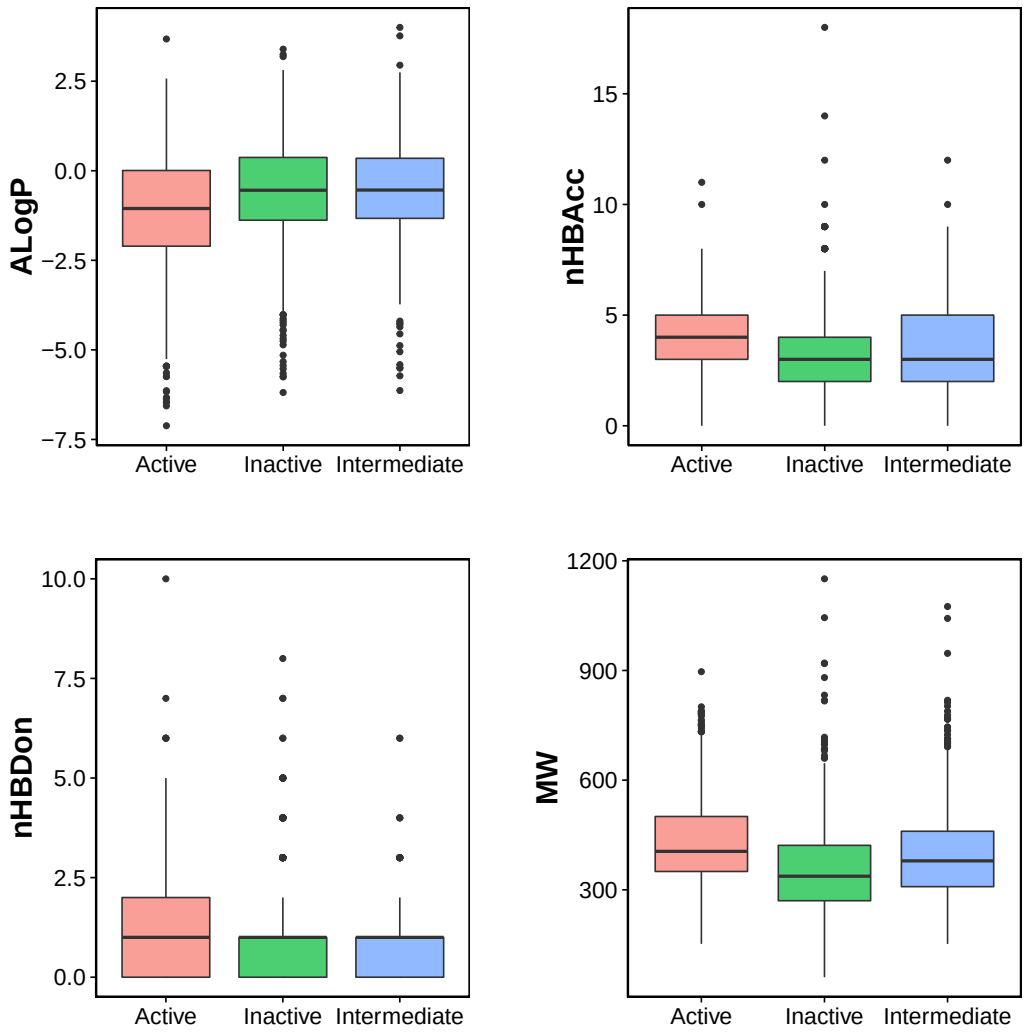

Figure 3. Box plot of AChE inhibitors using Lipinski's rule-of-five descriptors (e.g. ALogP, nHBAcc, nHBDon and MW)

the largest amongst the three bioactivity classes, which is followed by the intermediates while inactives were smallest.

QSAR model for predicting AChE inhibitory activity

A data set comprising of 2,570 compounds were used for construction of QSAR models. Particularly, twelve sets of fingerprint descriptors were benchmarked in order to find the best performing set. Prior to modeling, feature selection was applied to remove collinear descriptors. Each of the twelve models were then built using a data split ratio of 80/20 in which 80% of the data set was used as the internal set and 20% as the external set. This procedure was iteratively performed in which each of the 100 independent data splits were used for model construction and the performance results given in Table 2 are the mean and standard deviation values derived from these runs.

It can be observed that all twelve models are capable of capturing the inhibitory activity space of AChE inhibitors as they provided R^2 and Q^2 (i.e. both 10-fold CV and external sets) greater than the threshold values proposed by Golbraikh and Tropsha (2002) of 0.6 and 0.5, respectively, which is indicative of robust model performance. The possibility of chance correlation can be assessed from the R^2–Q^2 margin as described by Eriksson and Johansson (1996) in which values < 0.2–0.3 are indicative of predictive and

reliable models while values > 0.2–0.3 suggests possible chance correlation or the presence of outliers in the data set. Furthermore, observation of the Q^2_{CV}–Q^2_{Ext} margin revealed that the difference was negligible with values in the range of 0 and 0.01.

Generally, it can be seen that models with larger descriptor size, namely CDK and CDK extended, afforded the best performance with Q^2_{CV} of 0.79±0.07 and 0.79±0.06, respectively, and Q^2_{Ext} of 0.80±0.04 and 0.81±0.04, respectively. The opposite also holds true as the model with the least number of descriptors were also found to perform the worst amongst the other fingerprints with Q^2_{CV} of 0.55±0.09 and Q^2_{Ext} of 0.56±0.05. In a nutshell, the model performance in order of decreasing value is as follows: CDK extended > CDK > MACCS ≈ Substructure count ≈ Klekota-Rota count > PubChem > Klekota-Roth ≈ 2D atom pairs count > CDK graph only > Substructure > 2D atom pairs > E-state.

The best performing model is not necessarily the best choice considering the fact that the descriptor size for the best models were quite high and is consequently prone to overfitting. It was found that the substructure count provided reasonably good predictive performance (i.e. Q^2_{CV} and Q^2_{Ext} of 0.78±0.06 and 0.78±0.05, respectively) with the advantage of making use of a small set of 26 descriptors. Therefore, this fingerprint was selected for further interpretation of the feature importance.

To further check the reliability and validity of the selected model, Y-scrambling test was performed for 100 iterations. Table demonstrates that QSAR models built using substructure count has a low Q^2 (–0.0013), which rules out the possibility of chance correlation. Furthermore, model afforded an r^2_m value of 0.61±0.06 thereby revealing its robustness. It is observed in Table that the value of Δr^2_m is greater that 0.2 but also close to 0.20.

As shown in Fig. 4, it can also be seen that scatter plots of experimental versus predicted pIC_{50} of panels A, C and F displayed narrower variance of the data points than the other methods as assessed via 10-fold cross-validation and external set.

Applicability domain

The AD of the proposed QSAR model was defined as provided by the Williams plot shown in Fig. The employed data set consisting of 2,570 compounds was randomly split to two separate subset in which the first subset constituting 80% of the data set was used as an internal set while the second subset constituting the remaining 20% were used as an external set. Compounds representing the internal set (blue dots) and external set (red dots) are shown in the Williams plot and it can be clearly seen that almost all of the 2,570 compounds were located within the boundaries of applicability domain, which indicated that our proposed QSAR model had a well-defined AD.

As can be seen in Fig. very few compounds indeed fall outside the ±3 standardized residual range. This consisted of 6 compounds (**997**, **1829**, **62**, **1096**, **13**, **677**) from the internal set and 11 compounds (**2116**, **2120**, **2388**, **2117**, **2392**, **2323**, **2423**, **2424**, **2507**, **2219**, **2422**) from the external set that had standardized residual higher than 3. On the other hand, 7 compounds (**1567**, **576**, **1644**, **1098**, **2022**, **1447** and **322**) from the internal set and 8 compounds (**2486**, **2353**, **2130**, **2553**, **2389**, **2072**, **2103**, **2125**) from the external set had standardized residual lower than -3. The corresponding chemical structures are provided in Table S1.

Mechanistic interpretation of feature importance

Feature importance analysis help reveal features that are important toward bioactivity. There are essentially two parameters for evaluating the relative importance of features used in models using the RF algorithm: (i) accuracy and (ii) Gini index (i.e. variance of the responses). The latter was selected as a metric for ranking important features (i.e. mean decrease of the Gini index) for predicting the pIC_{50} of AChE inhibitors (Fig. 6). Table 4 lists the substructure fingerprints along with their respective descriptions.

As can be seen in Fig. 6, the top ranking feature is secondary carbon (SubFPC2), which is a carbon atom with two carbon neighbors. In the context of drug design, such central carbon atom may be more difficult to be accessed and metabolized by cytochrome P450 (Uetrecht and Trager, 2007) and therefore are more metabolically stable.

The second most important feature is the rotatable bond (SubFPC302). Based on the rule of three for defining lead-like compounds, a compound may have a lead-like characters if it does have rotatable bonds of no more than 3. On the other hand, Veber et al. (2002) noted that the the upper limit of a orally bioavailable drugs is of seven rotatable bonds. Nevertheless, it has been found that number of rotatable bonds provide better discrimination between compounds that are orally active and those that are not. Kryger et al. (1999) claimed that E2020 (i.e. also known as donepezil and marketed as Aricept) needs at

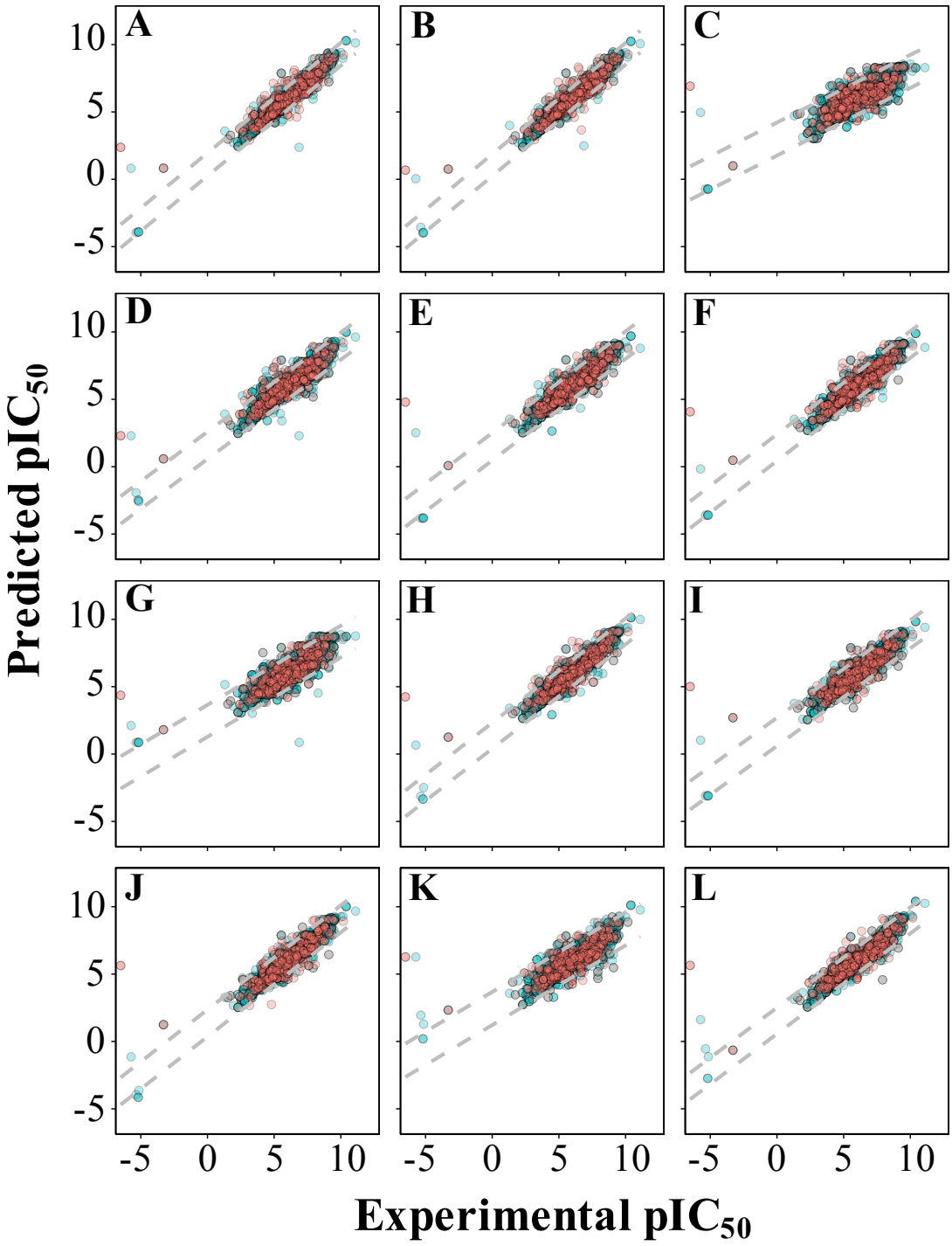

Figure 4. Plot of experimental versus predicted pIC_{50} values for models constructed with 12 different fingerprint descriptors. Shown are models built with CDK fingerprint (A), CDK extended fingerprint (B), E-State fingerprint (C), CDK graph only fingerprint (D), MACCS fingerprint (E), PubChem fingerprint (F), substructure fingerprint (G), substructure fingerprint count (H), Klekota-Roth fingerprint (I), Klekota-Roth fingerprint count (J), 2D atom pairs (K) and 2D atom pairs count (L).

least two rotatable bonds on each side of the piperidine in which two aromatic moieties of E2020 interact with Trp86 and Trp286 (human AChE numbering), suggesting that links between aromatic systems of the inhibitor against its AChE counterparts are essential to yield high affinity.

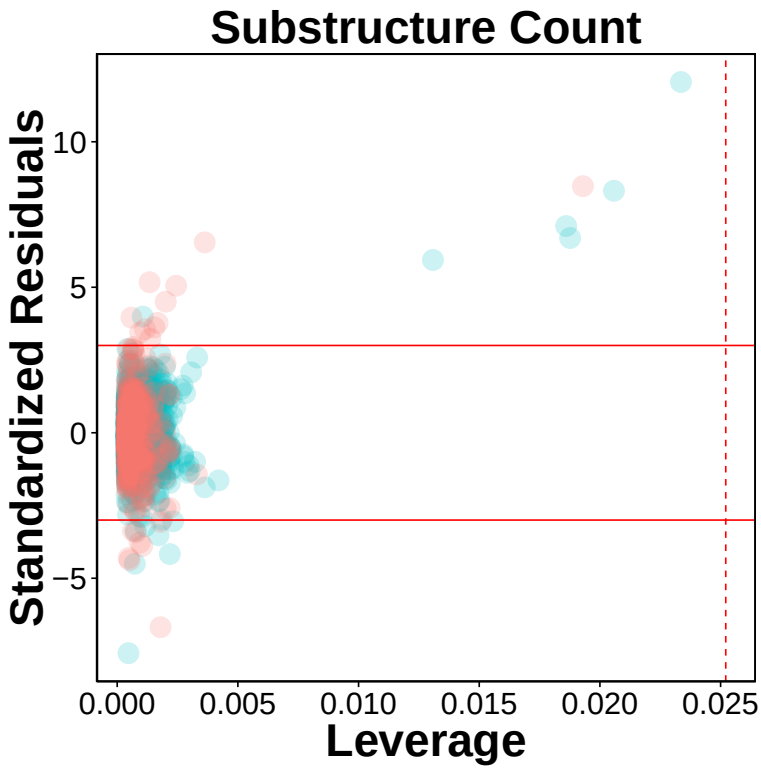

Figure 5. William plot for the internal (blue dots) and external (red dots) sets for QSAR model built using substructure fingerprint count. The solid and dashed lines correspond to the ±3 standardized residual and the warning leverage value ($h^* = 0.025$), respectively.

The third important substructure is the aromatic ring (SubFPC274). Findings from X-ray crystallographic study showed that in the binding site of the co-crystal structure of AChE with tacrine, the aromatic ring of acridine engages in a π-π stacking interaction with the indole of Trp86 (human AChE numbering) thereby indicating the importance of the aromatic ring for AChE inhibition. Chen et al. (2012)

The fourth important feature is C ONS bond (SubFPC295), which is defined as the presence of any carbon connected with either oxygen, nitrogen or sulfur atom in a molecule. These atoms contribute electrostatic charge, which deemed to increase the polarity

The fifth important substructure is secondary mixed amine (SubFPC32). The importance of the moiety was demonstrated in the work of Bembenek et al. (2008) in which a structure-based approach was used to reveal that in the catalytic triad, Trp86 interacts with the quaternary amine of ACh through a cation-π interaction. Furthermore, in the 'anionic' site Trp286 appears to attract the amine moiety via cation and/or hydrophobic interactions.

The sixth, seventh and eighth important substructures are heterocyclic rings (SubFPC275), tertiary carbon (SubFPC3) and primary carbon (SubFPC1). Heterocycles are of high relevance in the design of AChE inhibitors as it allows π-π stacking interaction with key amino acid residues in the binding site of AChE. It is observed that the binding site of the AChE are highly hydrophobic in nature. Particularly, the Trp286 (human AChE numbering) which is a part of the peripheral anionic site of the AChE is involved in the π-π interaction with heterocycles of AChE inhibitors (Lu et al., 2011). The aforementioned explanation made for the secondary carbon is also applicable for the tertiary carbon in which the higher number of carbon neighbors would also confer high stability against cytochrome P450 metabolism.

The ninth and tenth important substructures are Hetero N nonbasic (SubFPC181) and tertiary aliphatic amine (SubFPC26) in which former is defined as aromatic nitrogen. Present of Nitrogen atom, which is cation moiety deemed to interact with aromatic residues through π-cation interaction as observed in E2020 agianst *Torpedo californica* AChE (tcAChE). The charge of nitrogen atom located in piperidine ring provide π-cation interaction with the side chain of Phe337 (Phe330 in tcAChE numbering) (Jianxin Guo et al., 2004). Since the active site gorge of AChE conprised of several aromatic residues, also known as

aromatic patch, adding cation moiety possibly increase the binding affinity when the ligand arranged in suitable conformation against aromatic side cahin of residue the active site.

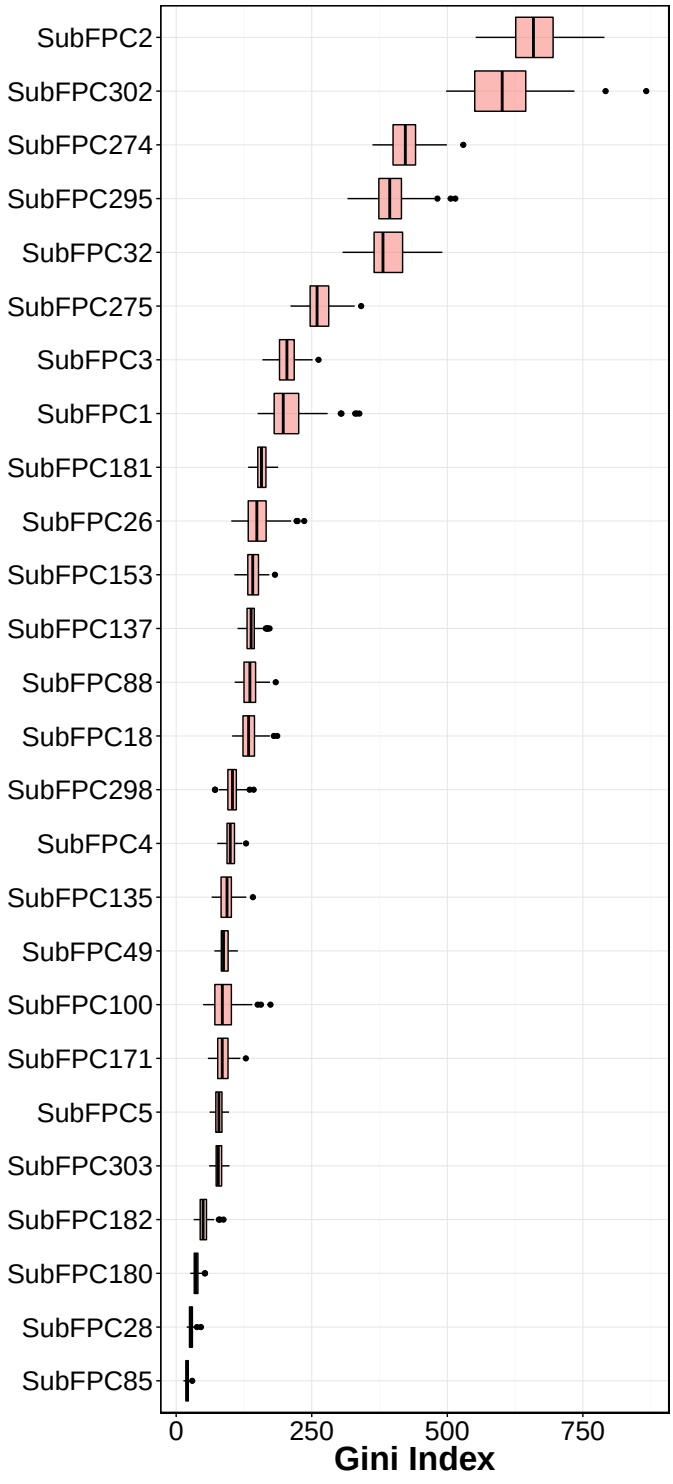

Figure 6. **Plot of feature importance as exemplified by the Gini index.**

Table 2. Performance summary of QSAR models for predicting pIC_{50}.

Descriptor class	N	Training set		10–fold CV set		External set		R^2–Q^2_{CV}	R^2–Q^2_{Ext}
		R^2	$RMSE_{Tr}$	Q^2_{CV}	$RMSE_{CV}$	Q^2_{Ext}	$RMSE_{Ext}$		
CDK	960	0.93±0.01	0.44±0.04	0.79±0.07	0.76±0.15	0.80±0.04	0.73±0.09	0.14	0.13
CDK extended	948	0.94±0.01	0.42±0.03	0.79±0.06	0.76±0.12	0.81±0.04	0.72±0.08	0.15	0.13
CDK graph only	198	0.87±0.01	0.61±0.03	0.72±0.06	0.87±0.13	0.72±0.05	0.87±0.09	0.15	0.15
E-State	21	0.66±0.03	1.00±0.05	0.55±0.09	1.11±0.13	0.56±0.05	1.10±0.08	0.11	0.10
MACCS	77	0.89±0.01	0.56±0.03	0.77±0.07	0.81±0.15	0.77±0.04	0.80±0.09	0.12	0.12
PubChem	103	0.90±0.01	0.55±0.03	0.76±0.05	0.80±0.11	0.78±0.03	0.79±0.08	0.14	0.12
Substructure	30	0.75±0.01	0.85±0.03	0.64±0.06	1.00±0.13	0.65±0.05	0.98±0.08	0.11	0.10
Substructure count	26	0.92±0.01	0.50±0.02	0.78±0.06	0.77±0.14	0.78±0.05	0.77±0.10	0.14	0.14
Klekota-Roth	111	0.89±0.01	0.59±0.03	0.74±0.07	0.85±0.14	0.76±0.05	0.81±0.10	0.15	0.13
Klekota-Roth count	72	0.91±0.01	0.52±0.03	0.78±0.06	0.79±0.14	0.78±0.05	0.77±0.11	0.13	0.13
2D atom pairs	42	0.75±0.03	0.85±0.06	0.61±0.08	1.03±0.15	0.60±0.06	1.05±0.12	0.14	0.15
2D atom pairs count	38	0.92±0.01	0.51±0.02	0.74±0.07	0.84±0.15	0.76±0.05	0.82±0.10	0.18	0.16

Table 3. Performance summary of QSAR models assessed using iR^2, iQ^2

, r^2_m and Δr^2_m metrics.

Descriptor class	N	Training set		10–fold CV set		External set		iR^2	iQ^2
		r^2_m	Δr^2_m	r^2_m	Δr^2_m	r^2_m	Δr^2_m		
CDK	960	0.82±0.02	0.07±0.01	0.62±0.09	0.20±0.06	0.64±0.05	0.19±0.03	0.0003	–0.0003
CDK extended	948	0.83±0.01	0.07±0.01	0.62±0.08	0.20±0.05	0.65±0.05	0.18±0.03	0.0006	–0.0005
CDK graph only	198	0.70±0.02	0.14±0.01	0.51±0.07	0.27±0.05	0.53±0.05	0.26±0.03	0.0007	–0.0006
E-State	21	0.35±0.03	0.38±0.02	0.27±0.07	0.40±0.04	0.28±0.05	0.41±0.03	0.0011	–0.0009
MACCS	77	0.73±0.01	0.12±0.01	0.57±0.09	0.23±0.05	0.58±0.05	0.23±0.03	0.0005	–0.0004
PubChem	103	0.74±0.02	0.12±0.01	0.57±0.07	0.23±0.04	0.59±0.05	0.22±0.03	0.0006	–0.0005
Substructure	30	0.50±0.02	0.28±0.01	0.39±0.07	0.34±0.05	0.41±0.05	0.33±0.03	0.0033	–0.0027
Substructure count	26	0.77±0.01	0.10±0.01	0.60±0.08	0.22±0.05	0.61±0.06	0.21±0.04	0.0015	–0.0013
Klekota-Roth	111	0.71±0.02	0.14±0.01	0.54±0.08	0.25±0.05	0.56±0.06	0.24±0.03	0.0006	–0.0004
Klekota-Roth count	72	0.76±0.02	0.11±0.01	0.60±0.08	0.22±0.05	0.61±0.07	0.21±0.04	0.0006	–0.0005
2D atom pairs	42	0.49±0.03	0.28±0.02	0.35±0.08	0.36±0.04	0.35±0.06	0.36±0.03	0.0010	–0.0008
2D atom pairs count	38	0.75±0.01	0.10±0.01	0.52±0.05	0.26±0.05	0.54±0.06	0.25±0.04	0.0006	–0.0005

Table 4. List of top substructure fingerprints and their corresponding description.

Fingerprints	Description
SubFPC1	Primary carbon
SubFPC2	Secondary carbon
SubFPC3	Tertiary carbon
SubFPC5	Alkene
SubFPC18	Alkylaryether
SubFPC23	Amine
SubFPC26	Tertiary aliphatic amine
SubFPC28	Primary aromatic amine
SubFPC32	Secondary mixed amine
SubFPC35	Ammonium
SubFPC49	Ketone
SubFPC88	Carboxylic acid derivative
SubFPC100	Secondary amide
SubFPC135	Carbonyl derivative
SubFPC137	Vinylogous ester
SubFPC143	Carbonic acid diester
SubFPC153	Urethan
SubFPC171	Arylchloride
SubFPC180	Hetero N basic H
SubFPC181	Hetero N nonbasic
SubFPC182	Hetero O
SubFPC184	Heteroaromatic ring
SubFPC274	Aromatic ring
SubFPC275	Heterocyclic ring
SubFPC276	Epoxide
SubFPC287	Conjugated double bond
SubFPC295	C ONS bond
SubFPC296	Charged
SubFPC298	Cation
SubFPC300	1,3-Tautomerizable
SubFPC301	1,5-Tautomerizable
SubFPC302	Rotatable bond
SubFPC303	Michael acceptor
SubFPC307	Chiral center specified

MOLECULAR DOCKING OF ACHE INHIBITORS

To gain a further understanding on the non-covalent interaction between AChE and their inhibitors, a chemically diverse set of 30 representative compounds was extracted from active AChE inhibitors (i.e. having $IC_{50} < 1\ \mu M$) using the Kennard-Stone algorithm and subjected to an investigation on its binding modality against the active site of AChE. Figure 7 shows the distribution of the selected subset of compounds in the context of the full set of actives, which was found to provide a full coverage of the original chemical space. The chemical structures of these compounds are shown in Fig. 8.

The active site of this enzyme is buried inside a narrow gorge of 20 Ådeep, which permits multiple enzyme-substrate interaction thereby facilitating the formation of the transition state of ACh (Silman and Sussman, 2008; Zhou et al., 2010; Cheung et al., 2012a). The entry of the active site gorge is lined up by peripheral anionic site (PAS), which is composed of Tyr72, Tyr124, Trp286 and Try341. The function of PAS is to trap the substrate via π-cation interaction and proceed through the constriction residues Tyr124 and Phe338 and onto the catalytic site (Silman and Sussman, 2008; Dvir et al., 2010). As a serine hydrolase, AChE contains Ser203, Glu334 and His447 in the catalytic triad that catalyzed the acylation and deacylation of ACh. The catalytic triad is surrounded by the catalytic anionic site (CAS) (i.e. contains Trp86, Glu202 and Tyr337, oxyanion hole (i.e. comprising of Glu121, Glu122 and Ala204) and the acyl pocket (i.e. comprised of Phe295 and Phe297). These sites help to tie up the ACh and make the substrate arranged in a suitable orientation for interacting with the catalytic triad as well as providing proton transfer that is essential for nucleophilic substituion during the catalytic reaction (Zhou et al., 2010). As a result of the acylation process, the nucleophilic attack from the O atom of Ser203 to the carbonyl group of ACh affords the proton transfer from Ser203 to His447, which consequently breaks down the choline moiety and forms a covalent acylenzyme complex between Ser203 and the acetyl group. This complex consequently proceeds with deacylation, which follows a similar mechanism with the acylation stage. The protonation of Glu202 provides a water molecule and the proton transfer from His447 to the water molecule leads to the nucleophilic attack against the acetyl group of the complex. Finally, this results in the breaking down of the complex thereby restoring wild-type AChE and causing the release of acetic acid from the active site (Zhou et al., 2010).

Prior to carrying out the molecular docking calculations, the docking protocol was validated by re-docking the co-crystal ligand and protein. It was found that the re-docked ligand exhibited negligible deviation from the co-crystal conformation with an RMSD value of 0.963 Å, which was deemed to be suitable for further molecular docking investigation and its subsequent interpretation. Consequently, the binding modality was analyzed in order to gain understanding on the contribution of key residues in interacting with the investigated set of 30 compounds. This was performed using the SiMMap web server, which revealed three major binding anchors: Hbond1, vdW1 and vdW2 along with their site-moiety preferences. The first anchor site involves hydrogen bond interaction between Tyr124 (i.e. an important residue in the PAS that is spatially located as a bottleneck between the peripheral region and the catalytic site of AChE) and the following ligand moieties: secondary amide, secondary amine, nitrogen moiety in aromatic ring, ketone and ester. Such interaction can be observed in the co-crystal structure of huperzine A with human AChE (Cheung et al., 2012b). Interestingly, analysis of the important features from QSAR models also revealed the importance of "C ONS bond" , "secondary mixed amine", "heterocyclic" and "hetero N non-basic" as they were found to be in the top ten important substructures and is therefore crucial for forming hydrogen bonds. Furthermore, the other anchor sites involve van der Waals interaction in which members of the first van der Waals interaction site (vdW1) are comprised of Tyr124, Phe338 and Tyr341, which has a preference to interact with heterocyclic, aromatic, phenol and other non-polar moieties from representative inhibitors. The second van der Waals' interaction site (vdW2) is consisted of Trp86 and Gly121 with preference for the following ligand moieties: aromatic ring, heterocyclic ring, aliphatic moiety with alkene linkage and phenol moiety. These residues contain either bulky aromatic ring or non-polar moiety as their side chain to provide the van der Waal's surface contact against non-polar moiety from ligands. Notably, aromatic and heterocyclic substructures were also observed in the top ten important substructures for predicting the inhibitory activity of human AChE inhibitors.

Furthermore, analysis of the binding energy from the 30 representative compounds revealed that compounds **13**, **5** and **28** exhibited the lowest binding energy of –12.2, –12.0 and –12.0 kcal/mol, respectively, when interacting with the human AChE binding site, which is comparable to donepezil (–12.2 kcal/mol) as indicated in Figure 8. Key interacting residues and their moiety preference was deduced from the protein-ligand interaction diagram generated by LigPlot+ (Wallace et al., 1995) in combination

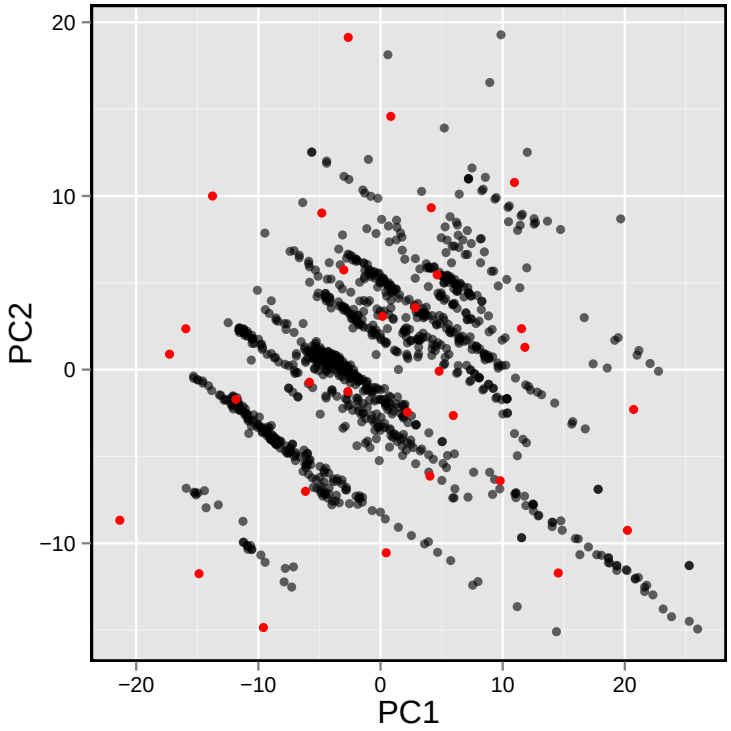

Figure 7. Plot showing the distribution of active AChE inhibitors (gray circles) and the diversity set (red circles) selected for molecular docking.

with Maestro (Schrödinger, LLC, 2015b) and their three-dimensional structure was visualized by PyMOL (Schrödinger, LLC, 2015a).

Figures 9A-B revealed three major interaction sites for **13** inside the binding pocket of human AChE. The first interaction site (Fig. 9C) is formed by residues from the PAS consisting of Trp286 and Tyr341 both of which engages in π-π stacking interaction where the terminal benzene and the attached pyrazole moiety of the ligand interacts with the former residue while the pyrazole moiety of **13** provide contact with the phenol moiety of the latter residue, which are deemed to increase the binding fitness against the active site of human AChE. The second interaction site (Fig. 9D) is dominated by Tyr124 and Phe338 in which the former interacts with the ketone moiety of **13** by means of a hydrogen bond, which tends to increase the binding affinity of this compound. The side chain of Phe338 is involved in hydrophobic interaction with carbon atoms from the piperidine ring and the central aliphatic chain of **13**. It can be noted that these residues are members of constriction site, which arranged as bottleneck of active site. The third interaction site (Fig. 9E) is moderated by π-π stacking between Trp86 and the terminal benzene with an attached piperidine moiety as well as hydrophobic interaction between Tyr337 and the piperidine moiety of **13** in which both residues belonged to the CAS. In addition, Gly121 of oxyanion hole also interact with terminal benzene of the ligand through hydrophobic contact increasing binding affinity against catalytic site of AChE.

Analysis of the binding modality of compound **5** revealed interactions with all subsites of the AChE active site gorge. PAS was the first subsite dominated by π-π stacking between Trp286 and the benzisoxazole moiety of compound **5**, which is essential for stabilizing the binding affinity of the ligand against entry into the gorge. The second interaction site was observed at the constriction site in which the piperidine moiety makes contact with Phe338 via hydrophobic interaction thereby increasing the binding fitness against the bottleneck region of the active site. Similar hydrophobic interaction was also observed in the binding pocket in which Tyr337 from CAS interacts with the piperidine moiety and Gly121 of oxyanion pocket interacts with the terminal benzene of compound **5**. The π-π interaction between the terminal benzene and Trp86 from the CAS was deemed to increase the binding fitness with the catalytic site of AChE. In addition, H-bond interaction is facilitated by the N atom from Phe295 at the acyl pocket

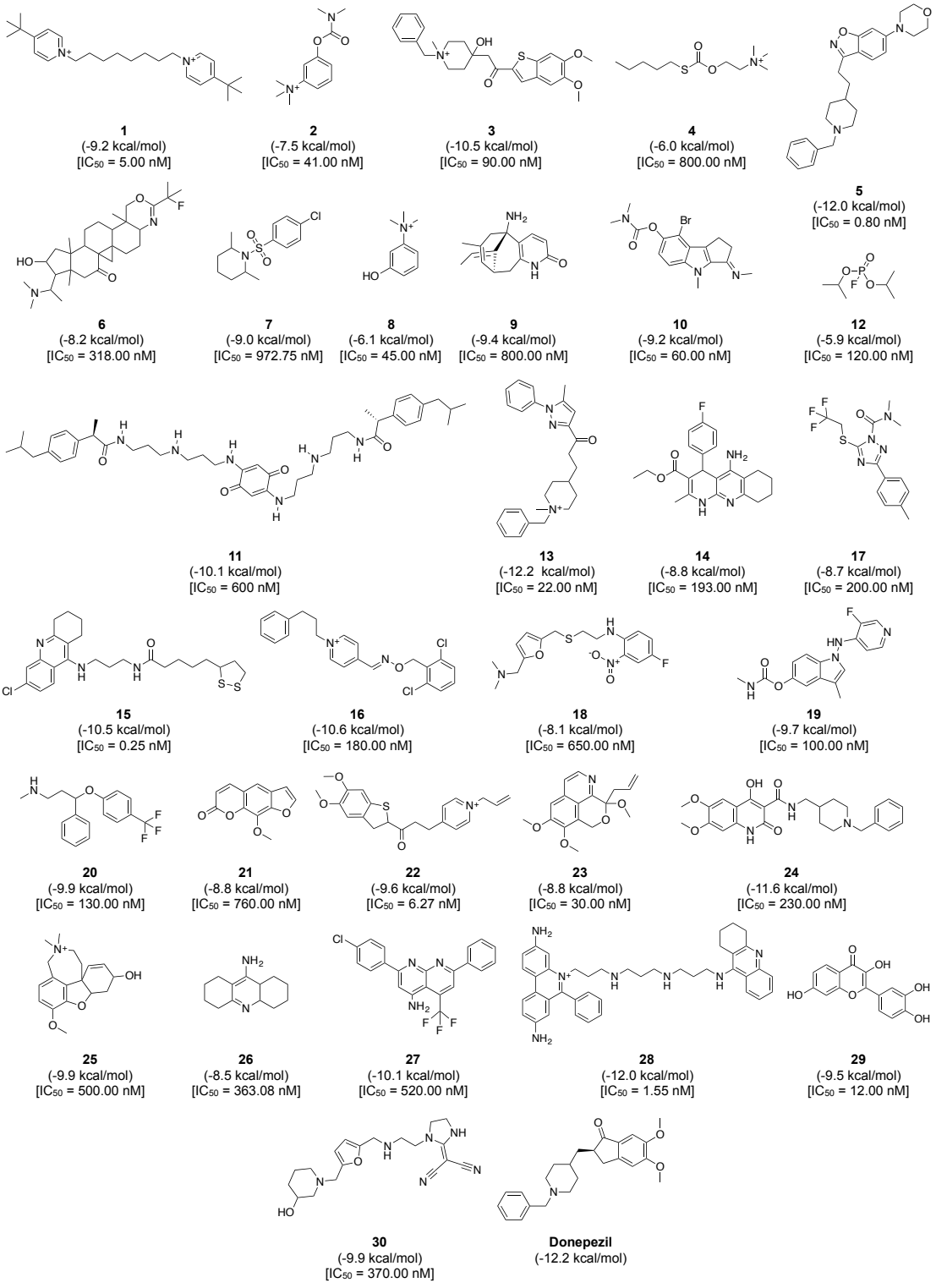

Figure 8. Chemical structures, binding energy and bioactivity of the diversity set consisting of 30 representative compounds from active AChE inhibitors.

to the O atom from the benzisoxazole moiety. Notably, all sites from the active site gorge are snugly bound by compound **5**, which is deemed to exhibit strong intermolecular interaction with human AChE.

The binding energy of compound **5** was similar to that of compound **28**. These compounds possessed

several aromatic rings at both terminal, which are favourable for interacting with aromatic residues lining up the surface of the gorge and these are known as the aromatic patch. The 5,7-dihydrophenanthridine moiety facilitates π-π stacking with the side chain of Trp286 from PAS at the gorge opening. Meanwhile, this moiety also engages in hydrophobic interaction with Phe295, which tends to increase the binding fitness for the acyl pocket. Aside from the former moiety, 1,2,3,4-tetrahydroacridine at the opposite terminal provides π-π interaction with Trp86 and hydrophobic contact with Tyr337 where both of which are members of CAS in the catalytic site. Furthermore, long aliphatic chain linking the two aromatic moieties provide hydrophobic contact with several aromatic residues in the aromatic patch consisting of Tyr72, Tyr124, Trp286, Tyr337, Phe338 and Tyr341 (i.e. these residues are the members of PAS, CAS and constriction site of the gorge). Moreover, this chain contain several N atoms, which can act as H-bond donor to Tyr124 and Tyr337 from PAS and CAS, respectively. This would tighten the binding between compound **28** and the active site gorge.

It should be noted that compounds exhibiting strong binding fitness against AChE are those that interact with residues from both PAS and CAS at the entry and inner pocket of the gorge, respectively, as dual-binding site inhibitor through either π-π stacking or π-cation interaction together with hydrophobic contact. These compounds competes with the natural substrate in interacting with these residues. For non-covalent inhibitors, the aromatic moiety is preferred for occupying the interaction sites while hydrophobic moieties are preferred for making contact with the aromatic residues surrounding the catalytic site. H-bond donor moieties such as secondary amine and heterocyclic ring can be employed for interacting with the O atom on the side chain of Tyr residues. Interestingly, this finding is corroborated by the feature importance results obtained from the QSAR model as shown in Fig. 6 in which the aromatic moiety, C ONS bond, secondary mixed amine, heterocyclic ring and the hetero N non-basic moiety were found amongst the top ten important substructures that are essential for the bioactivity of AChE inhibitor.

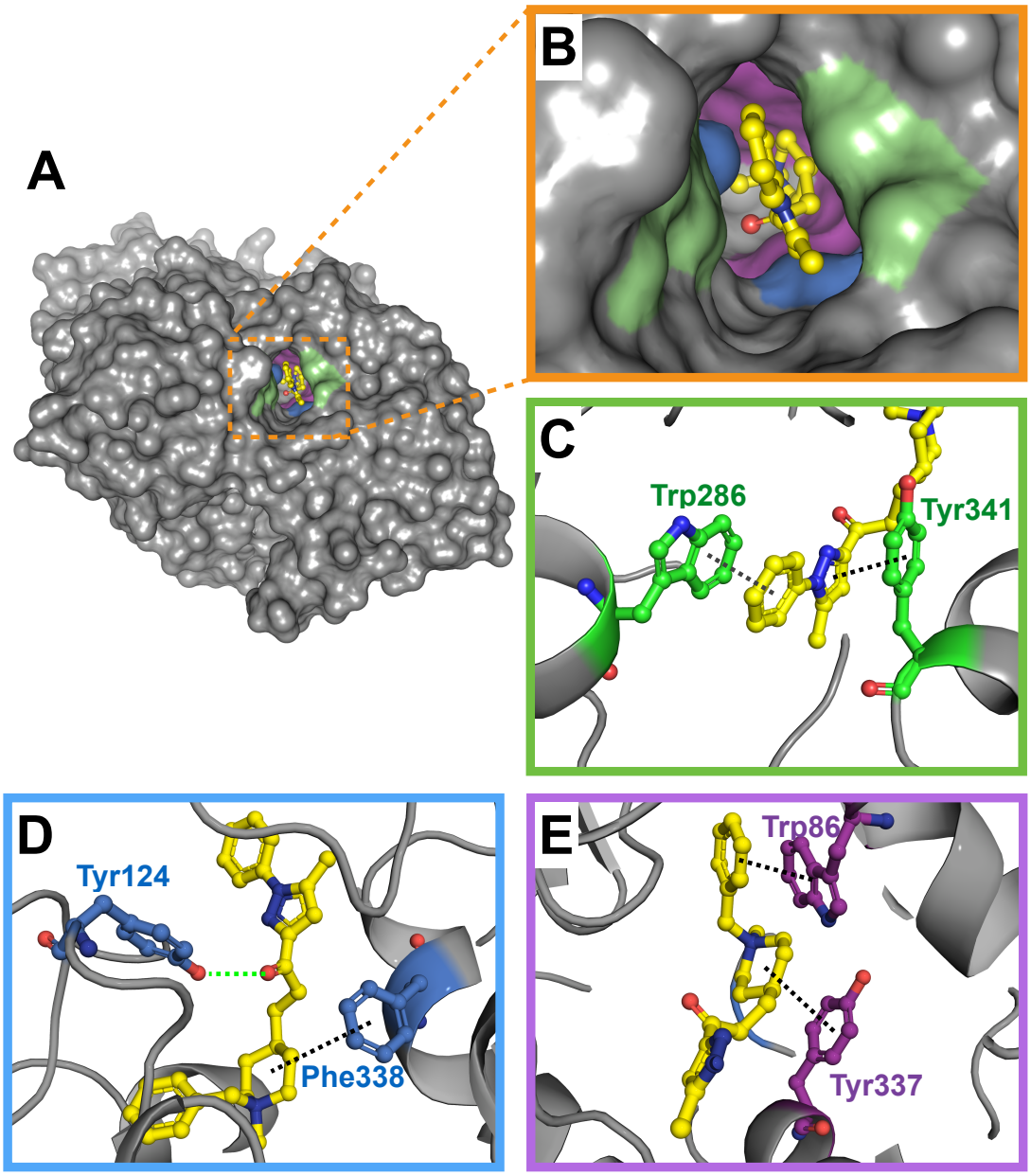

Figure 9. Molecular docking of the representative compound 13 against AChE. Illustration of the binding between AChE (gray colored surface representation) and **13** (yellow colored ball-and-stick) (A) and a zoom-in view showing the interaction of **13** with three spatially juxtaposed areas (blue, green and purple) of the AChE binding pocket (B). The first interaction site (green shaded area) situates at the entrance of the binding pocket and primarily engages in π-π stacking interaction (C), the second interaction site (blue shaded area) constitutes hydrogen bond interaction between Tyr124 and ketone moiety while hydrophobic interaction between Phe338 and piperidine ring plus central aliphatic chain of **13** (D) and the third interaction site (purple shaded area) constitutes π-π stacking between Trp86 and terminal aromatic ring while also affording hydrophobic interaction between Tyr337 and the piperidine moiety of **13** (E).

CONCLUSION

In conclusion, twelve sets of fingerprint descriptors were used for constructing QSAR models and their performances were comparatively evaluated. It was observed that several fingerprint descriptors afforded

good performance for the constructed models indicating that they could capture the feature space of AChE inhibitors. By taking advantage of the built-in feature importance estimator from RF known as the Gini index, the following important features that are critical for AChE inhibition were identified: secondary carbon (SubFPC2), rotatable bond (SubFPC302), aromatic (SubFPC274), C ONS bond (SubFPC295), secondary mixed amine (SubFPC32) and heterocyclic (SubFPC275). Results from molecular docking also support the aforementioned findings from the QSAR models in which the aromatic, heteroaromatic and heterocyclic rings were preferable moieties for interacting with the hydrophobic pocket of AChE. These can be used as a guide for designing novel inhibitors.

ACKNOWLEDGEMENTS

This research is supported by a Research Career Development Grant (No. RSA5780031) from the Thailand Research Fund and the Swedish Research Links program (No. C0610701) from the Swedish Research Council.

REFERENCES

Aaviksaar, A. (1990). QSAR in reactions of organophosphorus inhibitors with acetylcholinesterase. *Phosphorus, Sulfur, and Silicon and the Related Elements*, 51(1-4):47–50.

Andersson, C. D., Hillgren, J. M., Lindgren, C., Qian, W., Akfur, C., Berg, L., Ekström, F., and Linusson, A. (2014). Benefits of statistical molecular design, covariance analysis, and reference models in QSAR: a case study on acetylcholinesterase. *Journal of Computer-Aided Molecular Design*, 29(3):199–215.

Beal, M. F. (1995). Aging, energy, and oxidative stress in neurodegenerative diseases. *Annals of Neurology*, 38(3):357–366.

Bembenek, S. D., Keith, J. M., Letavic, M. A., Apodaca, R., Barbier, A. J., Dvorak, L., Aluisio, L., Miller, K. L., Lovenberg, T. W., and Carruthers, N. I. (2008). Lead identification of acetylcholinesterase inhibitors-histamine H3 receptor antagonists from molecular modeling. *Bioorganic & Medicinal Chemistry*, 16(6):2968–2973.

Birks, J. (2006). Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database of Systematic Reviews*, (1):CD005593.

Bollback, J. P. (2006). SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7:88.

Bourne, Y., Taylor, P., and Marchot, P. (1995). Acetylcholinesterase inhibition by fasciculin: crystal structure of the complex. *Cell*, 83(3):503–512.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia*, 3(3):186–191.

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73.

Chen, Y., Fang, L., Peng, S., Liao, H., Lehmann, J., and Zhang, Y. (2012). Discovery of a novel acetylcholinesterase inhibitor by structure-based virtual screening techniques. *Bioorganic & Medicinal Chemistry Letters*, 22(9):3181–3187.

Cheung, J., Rudolph, M. J., Burshteyn, F., Cassidy, M. S., Gary, E. N., Love, J., Franklin, M. C., and Height, J. J. (2012a). Structures of human acetylcholinesterase in complex with pharmacologically important ligands. *Journal of Medicinal Chemistry*, 55(22):10282–10286. PMID: 23035744.

Cheung, J., Rudolph, M. J., Burshteyn, F., Cassidy, M. S., Gary, E. N., Love, J., Franklin, M. C., and Height, J. J. (2012b). Structures of human acetylcholinesterase in complex with pharmacologically important ligands. *Journal of Medicinal Chemistry*, 55(22):10282–10286.

Deb, P. K., Sharma, A., Piplani, P., and Akkinepally, R. R. (2012). Molecular docking and receptor-specific 3D-QSAR studies of acetylcholinesterase inhibitors. *Molecular Diversity*, 16(4):803–823.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280.

Dvir, H., Silman, I., Harel, M., Rosenberry, T. L., and Sussman, J. L. (2010). Acetylcholinesterase: From 3d structure to function. *Chemico-Biological Interactions*, 187(1–3):10 – 22. 10th International Meeting on Cholinesterases.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., and Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based qsars. *Environmental Health Perspectives*, 111(10):1361.

Eriksson, L. and Johansson, E. (1996). Multivariate design and modeling in QSAR. *Chemometrics and Intelligent Laboratory Systems*, 34(1):1–19.

Fang, J., Li, Y., Liu, R., Pang, X., Li, C., Yang, R., He, Y., Lian, W., Liu, A.-L., and Du, G.-H. (2015). Discovery of multitarget-directed ligands against Alzheimer's disease through systematic prediction of chemical–protein interactions. *Journal of Chemical Information and Modeling*, 55(1):149–164.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107.

Giacoppo, J. O., CC França, T., Kuča, K., Cunha, E. F. d., Abagyan, R., Mancini, D. T., and Ramalho, T. C. (2015). Molecular modeling and in vitro reactivation study between the oxime bi-6 and acetylcholinesterase inhibited by different nerve agents. *Journal of Biomolecular Structure and Dynamics*, 33(9):2048–2058.

Golbraikh, A. and Tropsha, A. (2002). Beware of q^2! *Journal of Molecular Graphics and Modelling*, 20(4):269–276.

Goldblum, A., Yoshimoto, M., and Hansch, C. (1981). Quantitative structure-activity relationship of phenyl N-methylcarbamate inhibition of acetylcholinesterase. *Journal of Agricultural and Food Chemistry*, 29(2):277–288.

Gramatica, P. (2007). Principles of qsar models validation: internal and external. *QSAR & Combinatorial Science*, 26(5):694–701.

Gupta, S. and Mohan, C. G. (2014). Dual binding site and selective acetylcholinesterase inhibitors derived from integrated pharmacophore models and sequential virtual screening. *BioMed Research International*, 2014:291214.

Hall, L. H. and Kier, L. B. (1995). Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045.

Harel, M., Schalk, I., Ehret-Sabatier, L., Bouet, F., Goeldner, M., Hirth, C., Axelsen, P., Silman, I., and Sussman, J. (1993). Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. *Proceedings of the National Academy of Sciences of the United States of America*, 90(19):9031–9035.

Huang, W., Tang, L., Shi, Y., Huang, S., Xu, L., Sheng, R., Wu, P., Li, J., Zhou, N., and Hu, Y. (2011). Searching for the multi-target-directed ligands against Alzheimer's disease: Discovery of quinoxaline-based hybrid compounds with AChE, H3R and BACE 1 inhibitory activities. *Bioorganic & Medicinal Chemistry*, 19(23):7158–7167.

J Prado-Prado, F., Escobar, M., and Garcia-Mera, X. (2013). Review of bioinformatics and theoretical studies of acetylcholinesterase inhibitors. *Current Bioinformatics*, 8(4):496–510.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: With Applications in R*. Springer, New York.

Jianxin Guo, Margaret M. Hurley, Jeffery B. Wright, , , and Gerald H. Lushington*, 2004). A docking score function for estimating ligandprotein interactions: application to acetylcholinesterase inhibition. *Journal of Medicinal Chemistry*, 47(22):5492–5500. PMID: 15481986.

Kennard, R. W. and Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1):137–148.

Kitz, R. and Wilson, I. B. (1962). Esters of methanesulfonic acid as irreversible inhibitors of acetylcholinesterase. *Journal of Biological Chemistry*, 237(10):3245–3249.

Klekota, J. and Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525.

Kryger, G., Silman, I., and Sussman, J. L. (1999). Structure of acetylcholinesterase complexed with E2020 (aricept®): implications for the design of new anti-Alzheimer drugs. *Structure*, 7(3):297–307.

Kuca, K., Soukup, O., Maresova, P., Korabecny, J., Nepovimova, E., Klimova, B., Honegr, J., Ramalho, T. C., and França, T. C. (2016). Current approaches against alzheimer's disease in clinical trials. *Journal of the Brazilian Chemical Society*, 27(4):641–649.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*,

28(5):1–26.

Laggner, C. (2005). Smarts patterns for functional group classification. *Inte: Ligand Software-Entwicklungs und Consulting GmbH, Maria Enzersdorf, Austria*.

Lee, S. and Barron, M. G. (2016). A mechanism-based 3D-QSAR approach for classification and prediction of acetylcholinesterase inhibitory potency of organophosphate and carbamate analogs. *Journal of Computer-Aided Molecular Design*, 30(4):347–363.

Lu, S.-H., Wu, J. W., Liu, H.-L., Zhao, J.-H., Liu, K.-T., Chuang, C.-K., Lin, H.-Y., Tsai, W.-B., and Ho, Y. (2011). The discovery of potential acetylcholinesterase inhibitors: a combination of pharmacophore modeling, virtual screening, and molecular docking studies. *Journal of Biomedical Science*, 18(8):b22.

Mager, P. (1983). QSAR applied to aging of phosphylated acetylcholinesterase. *Pharmazie*, 38(4):271–272.

Morris, J. C., Cyrus, P. A., Orazem, J., Mas, J., Bieber, F., Ruzicka, B. B., and Gulanski, B. (1998). Metrifonate benefits cognitive, behavioral, and global function in patients with Alzheimer's disease. *Neurology*, 50(5):1222–1230.

Mundy, R., Bowman, M., Farmer, J., and Haley, T. (1978). Quantitative structure activity study of a series of substituted 0,0-dimethyl 0-(p-nitrophenyl) phosphorothioates and 0-analogs. *Archives of Toxicology*, 41(2):111–123.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., and Prachayasittikul, V. (2009). A practical overview of quantitative structure-activity relationship. *EXCLI J*, 8(7):74–88.

NCBI (2009). *PubChem Substructure Fingerprint, version 1.3*. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt.

Ordentlich, A., Barak, D., Kronman, C., Flashner, Y., Leitner, M., Segall, Y., Ariel, N., Cohen, S., Velan, B., and Shafferman, A. (1993). Dissection of the human acetylcholinesterase active center determinants of substrate specificity. Identification of residues constituting the anionic site, the hydrophobic site, and the acyl pocket. *Journal of Biological Chemistry*, 268(23):17083–17095.

Prado-Prado, F., Garcia-Mera, X., Escobar, M., Alonso, N., Caamano, O., Yanez, M., and Gonzalez-Diaz, H. (2012). 3d mi-dragon: new model for the reconstruction of us fda drug-target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for ache. *Current Topics in Medicinal Chemistry*, 12(16):1843–1865.

Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A., Skrzyński, M., and Worth, A. P. (2011). Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Structural Chemistry*, 22(4):795–804.

Quinn, D. M. (1987). Acetylcholinesterase: enzyme structure, reaction dynamics, and virtual transition states. *Chemical Reviews*, 87(5):955–979.

Racchi, M., Mazzucchelli, M., Porrello, E., Lanni, C., and Govoni, S. (2004). Acetylcholinesterase inhibitors: novel activities of old molecules. *Pharmacological Research*, 50(4):441–451.

Riniker, S. and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):1–17.

Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., and Das, R. N. (2013). Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *Journal of Computational Chemistry*, 34(12):1071–1082.

Schrödinger, LLC (2015a). PyMOL, version 1.7.6.3. New York, NY.

Schrödinger, LLC (2015b). Schrödinger Release 2015-4: Maestro, version 10.4. New York, NY.

Shen, T., Tai, K., Henchman, R. H., , and McCammon, J. A. (2002). Molecular dynamics of acetylcholinesterase. *Accounts of Chemical Research*, 35(6):332–340.

Silman, I. and Sussman, J. L. (2008). Acetylcholinesterase: How is structure related to function? *Chemico-Biological Interactions*, 175(1–3):3 – 10. Proceedings of the {IX} International Meeting on Cholinesterases.

Simeon, S., Möller, R., Almgren, D., Li, H., Phanus-umporn, C., Prachayasittikul, V., Bülow, L., and Nantasenamat, C. (2016). Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via qsar modeling. *Chemometrics and Intelligent Laboratory Systems*, 151:51–60.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500.

Su, C. and Lien, E. (1980). QSAR of acetylcholinesterase inhibitors: A reexamination of the role of charge-transfer. *Research Communications in Chemical Pathology and Pharmacology*, 29(3):403–415.

Tougu, V. (2001). Acetylcholinesterase: mechanism of catalysis and inhibition. *Current Medicinal Chemistry-Central Nervous System Agents*, 1(2):155–170.

Tropsha, A., Gramatica, P., and Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of qspr models. *QSAR & Combinatorial Science*, 22(1):69–77.

Trott, O. and Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461.

Uetrecht, J. and Trager, W. (2007). *Drug Metabolism: Chemical and Enzymatic Aspects*. CRC Press, Boca Raton, Florida.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623.

Veselinović, J., Nikolić, G., Trutić, N., Živković, J., and Veselinović, A. (2015). Monte Carlo QSAR models for predicting organophosphate inhibition of acetycholinesterase. *SAR and QSAR in Environmental Research*, 26(6):449–460.

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering*, 8(2):127–134.

Walters, D. and Hopfinger, A. (1986). Case studies of the application of molecular shape analysis to elucidate drug action. *Journal of Molecular Structure: THEOCHEM*, 134(3-4):317–323.

Wright, M. N. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv e-print*. http://arxiv.org/abs/1508.04409.

Yan, A. and Wang, K. (2012). Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 22(9):3336–3342.

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474.

Zhou, Y., Wang, S., and Zhang, Y. (2010). Catalytic reaction mechanism of acetylcholinesterase determined by bornoppenheimer ab initio qm/mm molecular dynamics simulations. *Journal of Physical Chemistry B*, 114(26):8817–8825. PMID: 20550161.