Chapter 9

Development of homology model, docking protocol and machine-learning based scoring functions for identification of *Equus caballus*'s BChE inhibitors

9. Development of homology model, docking protocol and machine-learning based scoring functions for identification of *Equus caballus's* BChE inhibitors

9.1. Introduction

Horse BChE is widely used for the screening of BChE inhibitors and shares structural homology with the human BChE. The 3-dimensional crystal structure of the enzyme from horse is not available, which makes it challenging to use the SBDD approach for the identification of inhibitors. A homology model was developed for the horse BChE in the present study. It was further subjected to its structural refinement through energy minimisation. The docking protocol was developed and validated by redocking a set of co-crystallised inhibitors, obtained from human BChE and their interaction profiles were compared. The performance of the Autodock SF was poor and therefore, a ML-based SF was developed and validated.

9.2. Material and methods

9.2.1. Homology modelling

SWISS-MODEL (https://swissmodel.expasy.org/) was used to generate a 3-dimensional protein structure for *Equus caballus*'s BChE (ecBChE) by homology modelling [188, 261]. UniProt accession code - P81908 (https://www.uniprot.org) was employed to obtain various templates to build ecBChE models[275]. The developed models were evaluated using GMQE, QMEAN Ramachandran plot, and parameters obtained from MolProbity [197, 262, 263].

9.2.2. Protein model refinement and preparation

The selected homology models were refined by DOCKPREP utility of Chimera-1.4 to repair incomplete side chains, add hydrogens and charges [276, 277]. Further, PDB2PQR server (http://nbcr-222.ucsd.edu/pdb2pqr_2.1.1) was used to assign protonation states to various amino acid residues, at a physiological pH of 7.4 [264, 265]. Each model developed was then subjected to energy minimisation by using Amber18 using protocol described in section 4.2.9 (**Table T1 of appendix**) [197, 266, 267].

9.2.3. Ligand preparation and grid generation

The SMILES strings of the ligands were PDBQT protocol described in section 4.2.7. The grid box size of 78 × 90 × 86 was selected for grid generation with the grid centre placed at 55.007, 53.553 and 43.81 representing X, Y and Z coordinates, respectively.

9.2.4. Validation of docking protocol and scoring function

Autodock-4.2.6 was used to perform molecular docking. The conformational search was carried out by LGA and Solis-Water local search [112, 196, 235]. The docking results were processed by a python script, i.e., vstools_v0.16 and post docking analysis and visualisation were performed by Discovery studio visualiser 2020.

BChE inhibitors co-crystallised with human BChE were docked on the homology model and the interaction profiles of the inhibitor, with both the enzymes, were compared to evaluate the interaction similarity. The validation of binding energy, obtained from Autodock SF, was performed by docking BChE inhibitors on horse BChE by the above-mentioned protocol. The binding energies of the inhibitors and the experimentally determined IC_{50} values were used for validation. The cut-off of IC_{50} values viz. 100, 500, 1000, 5000 and 10000 nM were used to classify the compounds in two classes (0, if IC_{50} <= cut-off value and 1, if IC_{50} > cut-off value). Similarly, various threshold cut-off values of binding energies were selected to perform binary classification of the compounds. The above obtained binary datasets were used for the development of a classification matrix, calculation of TPR and FPR. A ROC curve was plotted and the AUC was calculated. The other validation method involved the conversion of IC_{50} value using logarithm and were compared with the binding energies of ligands to obtain correlation coefficient. The IC_{50}

values were also standardised about the mean and standard deviation and were correlated with binding energies [112].

9.2.5. Development and validation of the scoring function

9.2.5.1. Preparation of datasets

BChE inhibitors docked on the homology model were utilised to prepare a dataset for the development of SF. The binding energy obtained from the native SF was used as one of the features. The best ligand poses, selected by native SF, were used to calculate the protein-ligand interactions through PLIP package [278]. Further, the ligand SMILES strings were used to calculate 2D descriptors from RDkit. The features were pre-processed for the training of certain ML algorithms, through standardisation about the mean and standard deviation value of the feature or were normalised in a range of 0 – 1.

The IC_{50} values were converted into a binary categorical variable, using a cut-off value of 10000 nM, for the development of the binary classification-based SF. On the other hand, IC_{50} values were converted to a logarithmic value for development of regression-based models. Finally, the dataset was divided into training and test subsets (85:15). Five-fold validation was used for the training of models on the training set. The training set was divided into five equal parts and one part was kept for validation, while training was performed on the remaining four parts iteratively. The test set was used for testing the final selected models.

9.2.5.2. Development of scoring function using binary classification algorithms

Various ML algorithms were employed to develop binary classification models by using Sci-kit learn, a python library. The binary classification models were developed by using machine learning algorithms viz. logistic regression, SVM, KNN, Naïve Bayesian, discriminant analysis and a variety of ensembled based and semi-supervised techniques. The models developed were identified from a combination of hyperparameters using grid search methods for each algorithm. The developed models were validated using the confusion matrix with the use of accuracy, precision, recall and F1-score on the independent test dataset [112].

9.2.5.3. Development of scoring function using regression algorithms

Two sets of regression-based models were developed on the basis of IC_{50} values. If the selected binary classification model, developed in the previous section, predicted that a compound was belonging to a class with IC_{50} less than or equal to 10000 nM; then the IC_{50} prediction would be carried out by using the model trained on a dataset containing compounds with IC_{50} below 10000 nM. Otherwise, the prediction would be carried by using the second model developed on a dataset with IC_{50} above 10000 nM.

The regression algorithms viz. linear regression, ridge regression, lasso regression, elastic net regression, SVR, RF regressor, Bayesian Ridge regression, stochastic gradient regression and neural networks were used in the study.

The coefficient of determination (r^2) defines the dependence of one variable on another and ranges between 0 to 1. A higher value of r^2 represents a better fitting of the line or manifold plain on the data. Mean absolute error (MAE) and mean squared error (MSE) are other essential parameters and should be as low as possible. The Q^2ext-based metrics, i.e., Q^2_{F1} and Q^2_{F2} should be greater than 0.5. The lower value indicates that the model fits better on the training set but have poor predictivity on an independent test dataset, i.e., overfitting. Further, Golbraikh and Tropsha's criterion were also used for the evaluation of the model, which takes into account observed (Y_{obs}), predicted (Y_{pred}) activities, and squared correlation coefficients (r^2_0 or $r`_0^2$) [279].

9.3. Results and discussion

9.3.1. Homology modelling

Homology modelling utilised a template structure having a high degree of sequence similarity for protein model development . The selection of the template was performed through sequence alignment. The template search identified twelve PDB, including human BChE and rat AChE, and were used for model building. The homology model developed by using crystal structures of human BChE were found to be superior in quality than mouse AChE due to high sequence similarity. The X-ray diffraction (XRD) derived protein structures were better than cryo-electron microscopy generated structures as a template for model building. QMEAN score is a linear combination of Cβ, all-atom, solvation and torsion potentials and is expected to be in the range of -4 – 1 [280]. QMEAN scores for Model_6EMI, Model_4TPK and Model_5LKR were -0.46, -0.67 and -0.87, respectively and were better than the remaining models. The selected models had good GMQE scores of above 0.8. Interestingly, the templates of the three selected models were derived from XRD and were human BChE with high sequence similarity. The Ramachandran plot exhibited that more than 93% of residues were in favoured region, and less than 1 % in outlier region for the selected models. Model_6EMI had Asn54, Model_4TPK had Asn54 and Asp283 and Model_5LKR had Glu158 as outlier residues. The details of the homology models that includes PDB templates and their validation parameters are included in table (**Table 9.1)**.

9.3.2. Protein model refinement and preparation

The missing and incorrect side chains of the selected models were refined through the assignment of partial charges and appending them with the help of the Dunbrack rotamer library [281]. The hydrogens and protonation states of the side chains of amino acids were assigned by using the PDB2PQR server at a physiological pH of 7.4 and considering de-solvation, hydrogen bonding, and charge-charge interactions [282]. The other parameters were standardised by using AMBER forcefields. The energy minimisation was carried out on the protein soaked in a box of TI3P water molecules. The disulphide bonds providing structural stability were assigned between Cys64 – Cys92, Cys252 – Cys263 and Cys400 – Cys519 prior to the energy minimisation. The minimisation process allowed a gradual decrease in the energy by reducing the restrain on the protein structure and increasing the flexible movement in order to reach minima **(Figure 9.1**).

The coordinates obtained at the end of each minimisation stage was evaluated by using a structure assessment tool accessible through the SWISS-MODEL web server. The comparison between the quality of protein structures obtained after various stages of energy minimisation of the selected models are presented in figure **(Figure 9.1)**. The energy minimisation of Model_4TPK, Model_5LKR and Model_6EMI ceased with the potential energies of -297340, -309150, -308320 Kcal/mol, respectively, at the end-stage. The models throughout energy minimisation displayed Pro102 residue in the cis conformation, similar to human BChE. The MolProbity and clash scores gradually decreased with minimisation for Model_5LKR and Model_6EMI. On the contrary, Model_4TPK showed an initial increase, followed by a decrease in their values with the progress of energy minimisation process. In the case of Ramachandran favoured and outlier residues, there was an increase and decrease in their corresponding values, respectively, during energy minimisation, except the end phase. The fifth minimisation stage of the Model_4TPK displayed an optimal trade-off between the clashes and beneficial factors with zero C-β deviation, low clash score, rotamer and Ramachandran outliers. Although the Model_6EMI displayed slightly better, Ramachandran favoured residue scores than others but had the highest clash score throughout the minimisation process. The three optimised protein models were selected and compared (**Table 9.2**). The final selected model (**Model_4TPK, stage 5**) was converted into PDBQT format using Autodock Tools-1.5.6, along with the assignment of the Gasteiger partial charges and atom types for each atom.

Table 9.1 Homology models developed for ecBChE from crystal structures of various organisms.

S.No.	Template PDB	Resolution (Å)	Organism	Model code	GMQE	QMEAN	Ramachandran Favoured (%)	Ramachandran Outliers (%)
1	**6EMI(Chain A)**	**2.48**	***Homo sapiens***	**Model_6EMI**	**0.83**	**-0.46**	**95.81**	**0.38**
2	**4TPK(Chain A)**	**2.7**	***Homo sapiens***	**Model_4TPK**	**0.86**	**-0.67**	**94.67**	**0.95**
3	**5LKR(Chain A)**	**2.52**	***Homo sapiens***	**Model_5LKR**	**0.86**	**-0.87**	**93.51**	**0.57**
4	6I2T(Chain B)*	5.7	*Homo sapiens*	Model_6I2T_B	0.87	-1.37	94.24	0.9
5	2JGE(Chain A)	2.6	*Mus musculus*	Model_2JGE	0.76	-1.39	93.21	0.75
6	2C0P(Chain A)	2.5	*Mus musculus*	Model_2C0P	0.77	-1.39	92.84	0.94
7	1MAA(Chain A)	2.9	*Mus musculus*	Model_1MAA	0.76	-1.44	93.79	0.94
8	3O9M(Chain A)	2.98	*Homo sapiens*	Model_3O9M	0.86	-1.47	92.44	1.13
9	1Q83(Chain A)	2.65	*Mus musculus*	Model_1Q83	0.76	-1.55	94.9	0.57
10	6I2T(Chain A)*	5.7	*Homo sapiens*	Model_6I2T_A	0.87	-1.61	94.64	1.61
11	1KU6(Chain A)	2.5	*Mus musculus*	Model_1KU6	0.76	-1.63	93.76	1.13
12	2WHP(Chain A)	2.2	*Mus musculus*	Model_2WHP	0.77	-1.73	95.85	0.57

* Structure derived through cryo-EM

Table 9.2 Homology models developed for ecBChE from crystal structures of various organisms.

Model	Stage of minimisation	MolProbity Score	Clash Score	Ramachandran Favoured (%)	Ramachandran Outliers (%)	Rotamer Outliers (%)	C-Beta Deviations	Bad Bonds	Bad Angles
Model_4TPK	**5**	**0.88**	**0.12**	**95.3**	**0.41**	**0.45**	**0**	**0**	**12**
Model_5LKR	7	0.9	0	93.98	0.21	0.68	1	0	9
Model_6EMI	6	0.96	0.36	95.24	0	0.45	1	0	16

Models indicated in bold were selected for further study.

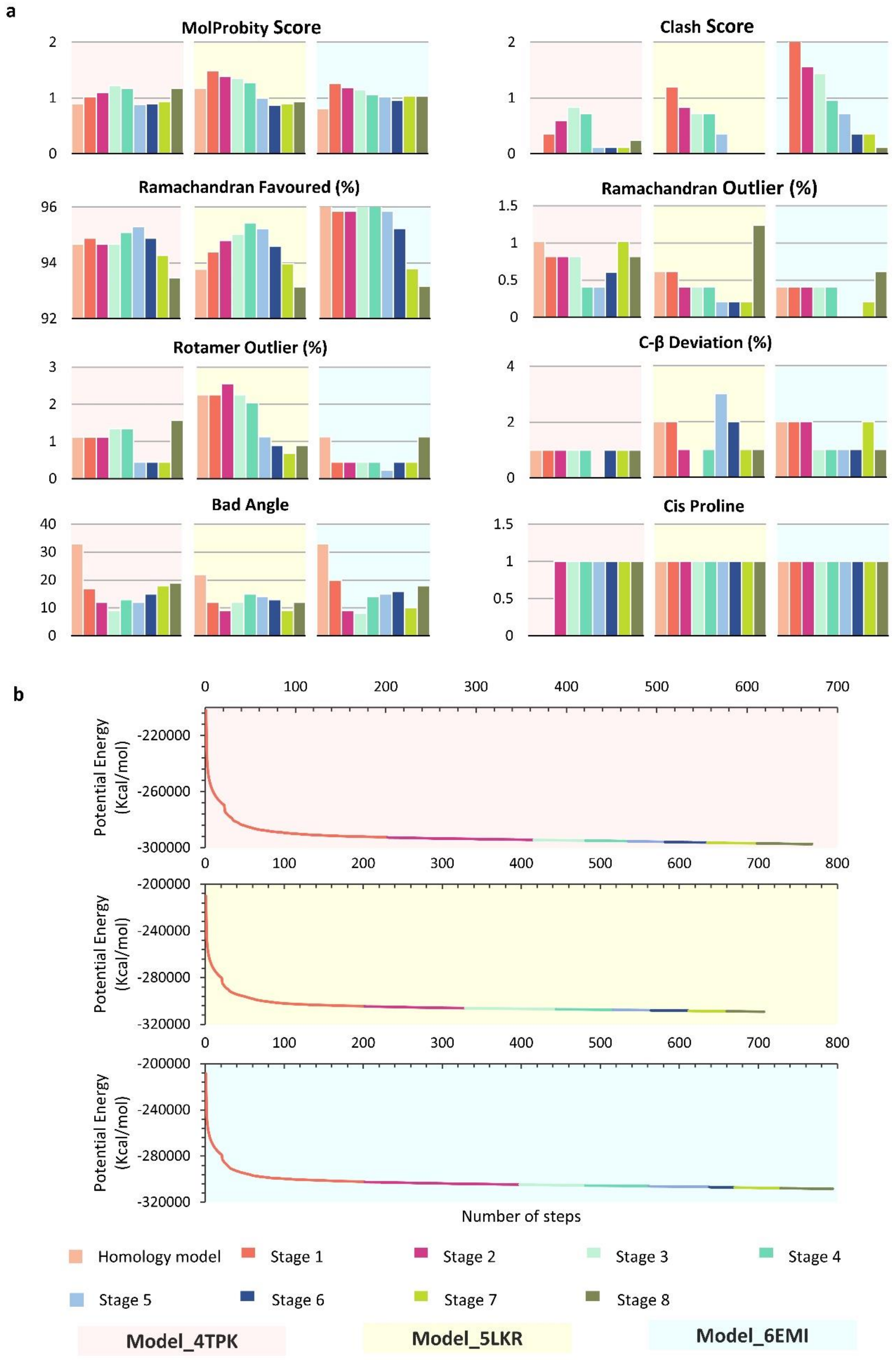
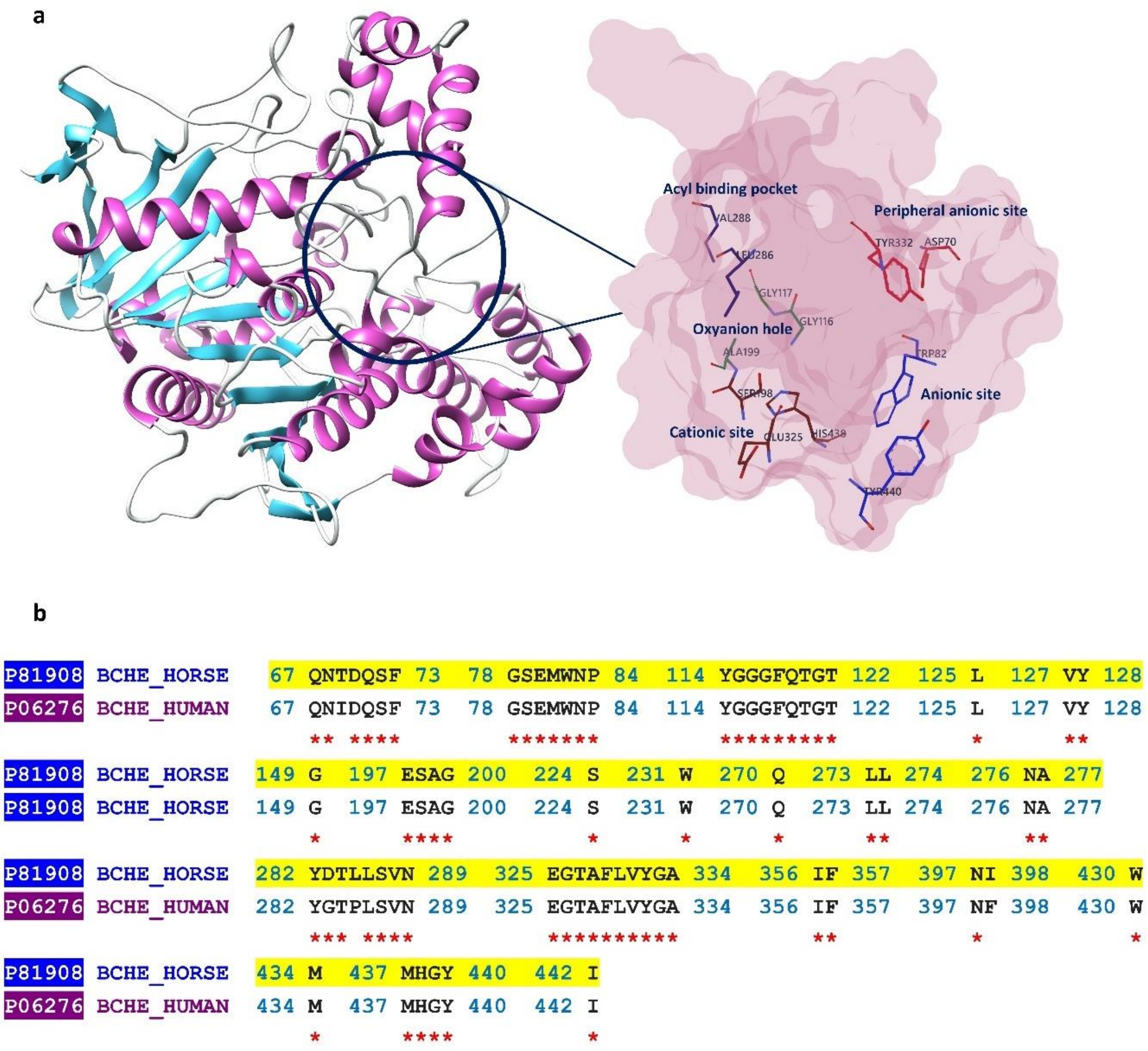


Figure 9.1 (a) Validation score of protein structures at various energy minimisation stages, (d) Potential energy (Kcal/mol) of the protein models during energy minimisation.



```
P81908 BCHE_HORSE   67 QNTDQSF 73   78 GSEMWNP 84   114 YGGGFQTGT 122   125 L   127 VY 128
P06276 BCHE_HUMAN   67 QNIDQSF 73   78 GSEMWNP 84   114 YGGGFQTGT 122   125 L   127 VY 128
                       ** ****         *******          *********          *         **

P81908 BCHE_HORSE  149 G   197 ESAG 200   224 S   231 W   270 Q   273 LL 274   276 NA 277
P81908 BCHE_HORSE  149 G   197 ESAG 200   224 S   231 W   270 Q   273 LL 274   276 NA 277
                       *       ****          *       *       *        **           **

P81908 BCHE_HORSE  282 YDTLLSVN 289   325 EGTAFLVYGA 334   356 IF 357   397 NI 398   430 W
P06276 BCHE_HUMAN  282 YGTPLSVN 289   325 EGTAFLVYGA 334   356 IF 357   397 NF 398   430 W
                       *** ****           **********           **           *            *

P81908 BCHE_HORSE  434 M   437 MHGY 440   442 I
P06276 BCHE_HUMAN  434 M   437 MHGY 440   442 I
                       *       ****          *
```

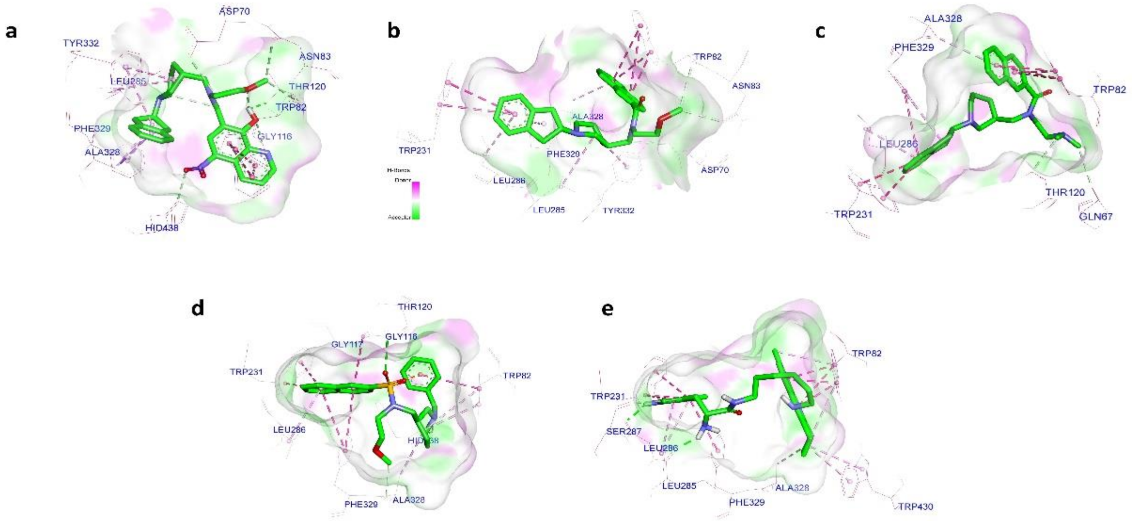
Figure 9.2 (a) Active site of human BChE (b) Sequence comparison (active site and tunnel residues) of ecBChE with human.

9.3.3. Ligand preparation and grid generation

The rdkit was used to generate 3D structures which were then converted to PDBQT format with the assignment of Gastiger partial charges to each atom and an Autodock atom type. The grid maps generated using grid box covered the entire active site and the tunnel approaching it, which consisted of sixty-six amino acid residues, as indicated in the figure (**Figure 9.2**).

9.3.4. Validation of docking protocol and scoring function

Molecular docking was performed by Autodock-4.2.6, using a force field-based scoring function and de-solvation term. The usual validation process involved the redocking of the co-crystallised ligand, followed by calculating the root mean square deviation (RMSD). However,



f

40V	Ile69	Asp70	Trp82	Asn83	Gly116	Thr120	Glu197	Ser198	Trp231	Leu285	Ala328	Phe329	Tyr332	His438
H. sapiens	🟥	🟧	🟪	□	🟩	□	🟩	🟩	🟥	□	□	🟥	▨	🟩
E. caballus	□	🟩	🟩	🟩	🟩	🟩	□	□	□	▨	🟪	▨	▨	🟩
	Ile69	Asp70	Trp82	Asn83	Gly116	Thr120	Glu197	Ser198	Trp231	Leu285	Ala328	Phe329	Tyr332	His438

3F9	Ile69	Asp70	Trp82	Asn83	Gly116	Glu197	Trp231	Leu285	Leu286	Ala328	Phe329	Tyr332
H. sapiens	▨	🟧	□	□	🟩	🟩	🟥	□	▨	▨	▨	▨
E. caballus	□	🟩	🟩	🟩	□	□	🟥	▨	▨	□	🟥	▨
	Ile69	Asp70	Trp82	Asn83	Gly116	Glu197	Trp231	Leu285	Leu286	Ala328	Phe329	Tyr332

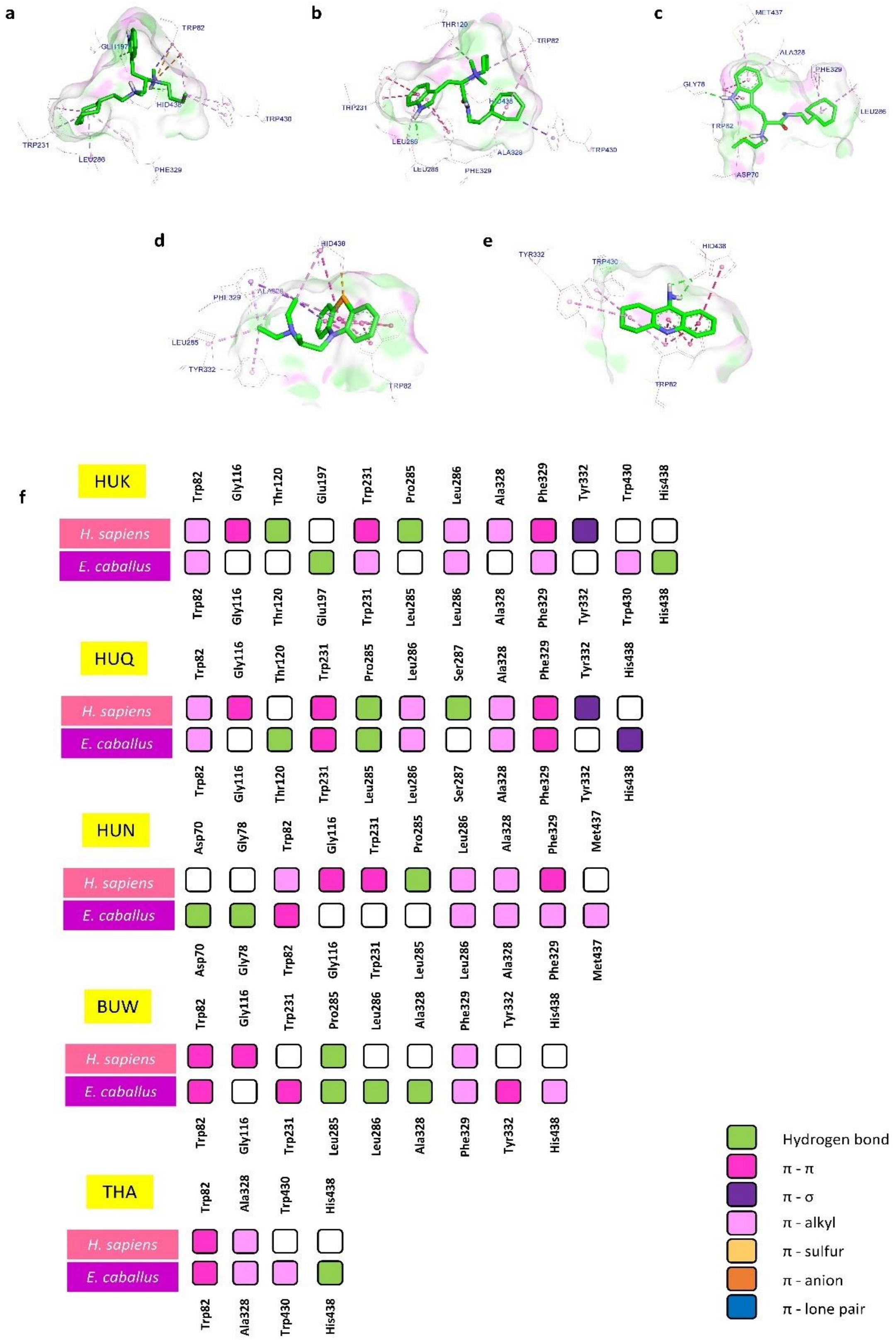
92H	Glu67	Asp70	Trp82	Gly116	Thr120	Glu197	Trp231	Leu286	Ala328	Phe329	Tyr332
H. sapiens	▨	🟧	□	□	🟩	🟩	🟥	□	▨	▨	▨
E. caballus	□	🟩	🟩	🟩	□	□	🟥	▨	▨	□	🟥
	Glu67	Asp70	Trp82	Gly116	Thr120	Glu197	Trp231	Leu286	Ala328	Phe329	Tyr332

5HF	Asn68	Trp82	Gly116	Thr120	Trp231	Leu286	Ala328	Phe329	Tyr332	His438
H. sapiens	🟩	🟥	🟥	🟩	🟥	▨	▨	🟥	▨	□
E. caballus	□	🟥	🟥	🟩	🟥	▨	🟩	🟥	□	🟩
	Asn68	Trp82	Gly116	Thr120	Trp231	Leu286	Ala328	Phe329	Tyr332	His438

9A5	Trp82	Gly116	Trp231	Pro285	Leu286	Ser287	Ala328	Phe329	Trp430
H. sapiens	🟥	🟥	□	🟩	□	□	▨	□	□
E. caballus	🟥	□	🟥	🟩	🟩	🟩	▨	🟥	▨
	Trp82	Gly116	Trp231	Leu285	Leu286	Ser287	Ala328	Phe329	Trp430

🟩	Hydrogen bond
🟥	π - π
🟪	π - σ
▨	π - alkyl
🟨	π - sulfur
🟧	π - anion
🟦	π - lone pair

Figure 9.3 (a, b, c, d, e) 3D interaction diagrams of 40V, 3F9, 92H, 5HF and 9A5 with ecBChE, (f) Comparison of the interaction profile of 40V, 3F9, 92H, 5HF and 9A5 among human and horse BChE.



HUK	Trp82	Gly116	Thr120	Glu197	Trp231	Pro285	Leu286	Ala328	Phe329	Tyr332	Trp430	His438
H. sapiens	▨	🟥	🟩	□	🟥	🟩	▨	▨	🟥	🟪	□	□
E. caballus	▨	□	□	🟩	▨	□	▨	□	▨	□	▨	🟩
	Trp82	Gly116	Thr120	Glu197	Trp231	Leu285	Leu286	Ala328	Phe329	Tyr332	Trp430	His438

HUQ	Trp82	Gly116	Thr120	Trp231	Pro285	Leu286	Ser287	Ala328	Phe329	Tyr332	His438
H. sapiens	▨	🟥	□	🟥	🟩	▨	🟩	▨	🟥	🟪	□
E. caballus	▨	□	🟩	🟥	🟩	▨	□	▨	🟥	□	🟪
	Trp82	Gly116	Thr120	Trp231	Leu285	Leu286	Ser287	Ala328	Phe329	Tyr332	His438

HUN	Asp70	Gly78	Trp82	Gly116	Trp231	Pro285	Leu286	Ala328	Phe329	Met437
H. sapiens	□	□	▨	🟥	🟥	🟩	▨	▨	🟥	□
E. caballus	🟩	🟩	🟥	□	□	□	▨	▨	▨	▨
	Asp70	Gly78	Trp82	Gly116	Trp231	Leu285	Leu286	Ala328	Phe329	Met437

BUW	Trp82	Gly116	Trp231	Pro285	Leu286	Ala328	Phe329	Tyr332	His438
H. sapiens	🟥	🟥	□	🟩	□	□	▨	□	□
E. caballus	🟥	□	🟥	🟩	🟩	🟩	▨	🟥	▨
	Trp82	Gly116	Trp231	Leu285	Leu286	Ala328	Phe329	Tyr332	His438

THA	Trp82	Ala328	Trp430	His438
H. sapiens	🟥	▨	□	□
E. caballus	🟥	▨	▨	🟩
	Trp82	Ala328	Trp430	His438

🟩 Hydrogen bond
🟥 π - π
🟪 π - σ
▨ π - alkyl
🟨 π - sulfur
🟧 π - anion
🟦 π - lone pair

Figure 9.4 (a, b, c, d, e) 3D interaction diagrams of HUK, HUQ, HUN, BUW and THA with ecBChE, (f) Comparison of the interaction profile of HUK, HUQ, HUN, BUW and THA among human and horse BChE.

the docking was performed on a homology model of horse BChE with no co-crystallised ligands. Hence, ten ligands co-crystallised with human BChE were docked on the homology model. The comparison of their interactions of these lignads with the human and horse BChEs are presented in the figures **(Figures 9.3** and **9.4)**. Most of the ligands docked with horse BChE displayed interactions similar to human BChE. The redocking also indicated that the interactions with the major sites of the enzyme were retained when compared to human BChE. The validation study showed that the protocol developed for docking produced poses similar to human BChE.

Ten inhibitors co-crystallised with human BChE, used for pose validation, were analysed by using molecular fingerprints to identify essential features for BChE inhibition. It was observed that 149 fingerprints were present in the compounds and were crucial for the enzyme inhibition (**Table T33 in appendix**). About 1110 BChE inhibitors were acquired from the Bindingdb database (http://bindingdb.org/). The inhibitors were filtered to remove duplicate compounds or that have missing IC_{50} values to obtain 795 compounds [270]. Further, the compounds were subjected to a fingerprint filter with a criterion of at least 30% selected fingerprints should be present in a molecule, resulting in 755 inhibitors. The selected inhibitors with their known IC_{50} values were used as a dataset for validation of SF of Autodock-4.2.6 and the development of new ML-based SF. The binding energies obtained after docking of these inhibitors on the homology model were correlated with experimentally obtained IC_{50} values to validate the SF. ROC is a plot between the specificity and 1-sensitivity (TPR and FPR). Its AUC represents the degree of separability of various classes of a classification model. A randomly generated predictor has an expected AUC of 0.5. In contrast, an ideal predictor has an AUC value of 1 for the ROC curve. The AUC of the SF was found to be between 0.4 – 0.65 for various IC_{50} cut-off values, indicating poor prediction by Autodock SF (**Figure 9.5**).

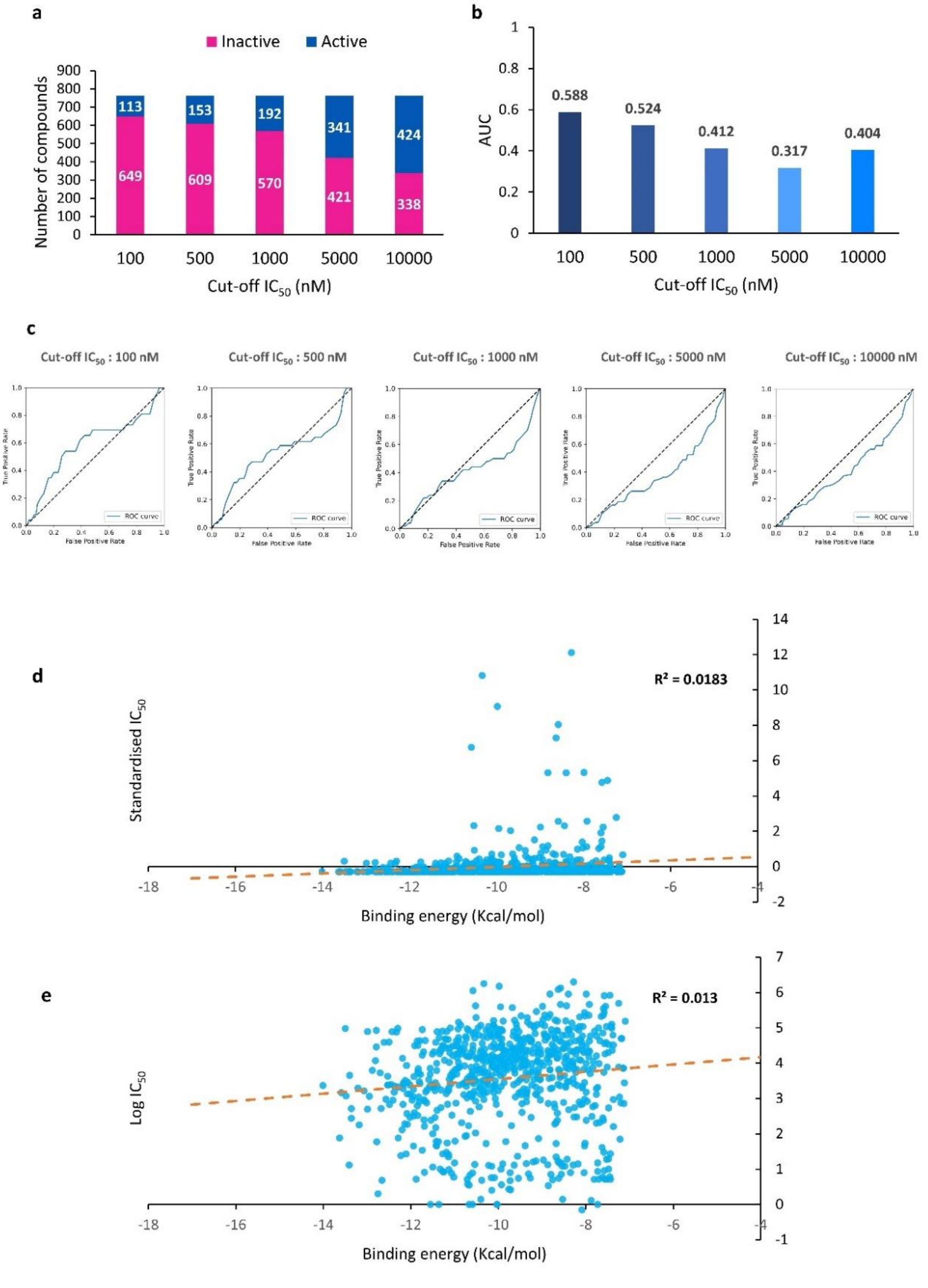


Figure 9.5 (a, b, c) IC_{50} cut-off, area under the curve and receiver operating characteristic of the docking validation set for validation of the Autodock scoring function using binary classification. (d and e) Scatter plot showing the relationship between binding energy with standardised IC_{50} and Log IC_{50}, respectively.

The second validation method assumed both binding energies and IC_{50} values as continuous variables, and a relationship between the two was identified. The range of IC_{50} values of the

BChE inhibitors (0 – 1980000 nM) was quite extensive. Hence, the IC_{50} values were scaled through logarithmic conversion and standardisation processes. The plots between the binding energies and the IC_{50} values are included in the figure **(Figure 9.5)**. The correlation coefficients were found to be 0.0183, between standardised IC_{50} values and binding energies and 0.013, between log IC_{50} values and binding energies. The regression coefficients displayed a poor correlation between the two variables, indicating the unsuitable performance of the SF of Autodock-4.2.6.

9.3.5. Development and validation of the scoring function

9.3.5.1. Preparation of datasets

A molecular library of 755 compounds used for docking validation studies was employed. The predicted binding energy, ligand interactions with the 64 residues of the active site and 102 2D-descriptors of ligands were used as predictors. The IC_{50} values of the compounds were used to assign labels as active (1) and moderately active (2), using a cut-off IC_{50} value of 10000 nM for building binary classification models. The dataset of the BChE inhibitors was divided into training and test set in a ratio of 85:15 (**Figure 9.6**). The log IC_{50} values of the training and test sets overlapped in both the regression models, indicating the correct division of the datasets.

9.3.5.2. Development of scoring function using binary classification algorithms

Sixteen algorithms were used for the training of the binary classification model. Shapiro-Wilk test for normality was performed and the features displaying normal distribution were standardised about their mean and standard deviation, while the other features were scaled in the range of 0 – 1. The pre-processed dataset was employed for the training of models for LR, SVC, KNN classifier, ridge classifier, perceptron and MLP algorithms. The features with different ranges create bias in model training towards any specific feature with large values for the algorithms. LR, SVC and KNN performed well with training accuracy of 80% and above, but the test accuracy was poor, thus indicating underfitting. The perceptron and MLP models performed poorly on training and test sets, with an accuracy below 80 % in each case. Similar results were also observed with the models developed from LDA, QDA, Bernoulli Naïve Bayesian, label propagation and label spreading and were not suitable.

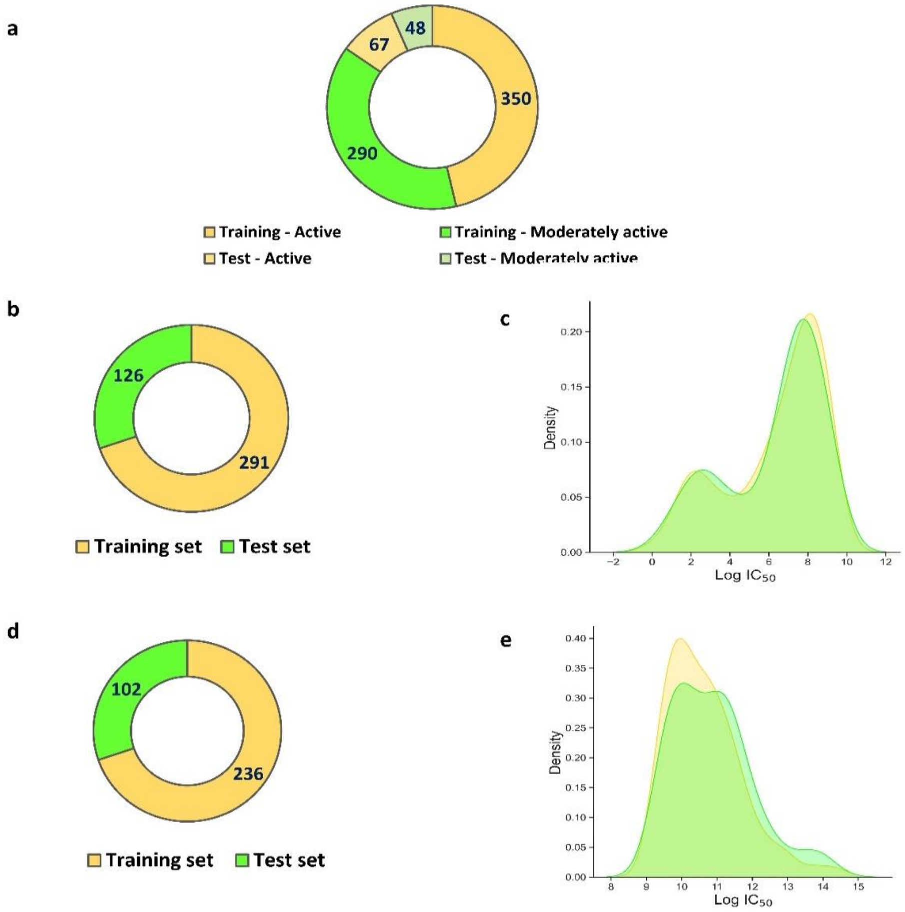


Figure 9.6 Data distribution for (a) classification models, (b and d) regression models dataset with IC_{50} below and above 10000 nM, (c and e) Log IC_{50} distribution of training and test sets for regression models dataset with IC_{50} below and above 10000 nM

The decision tree displayed low accuracies on the training and test datasets. However, ensemble techniques that used various weak decision trees improved the performance, which was evident in the case of RF and extra tree classifiers. The gradient and adaptive boosting techniques performed well, like the RF and extra tree classifiers. Further, the extra tree classifier was better than other ensemble techniques due to the higher precision value. Molecular docking and virtual screening require the correct prediction of active compounds to yield good results in the *in vitro* testing. Hence, the precision value was the most important criterion for selecting the final model to develop as SF. The AUC for the extra tree classifier was 0.911, indicating good quality. Finally, the extra trees classifier was selected for SF. The validation scores and AUC are presented in **Table 9.3** and **Figure 9.7,** respectively.

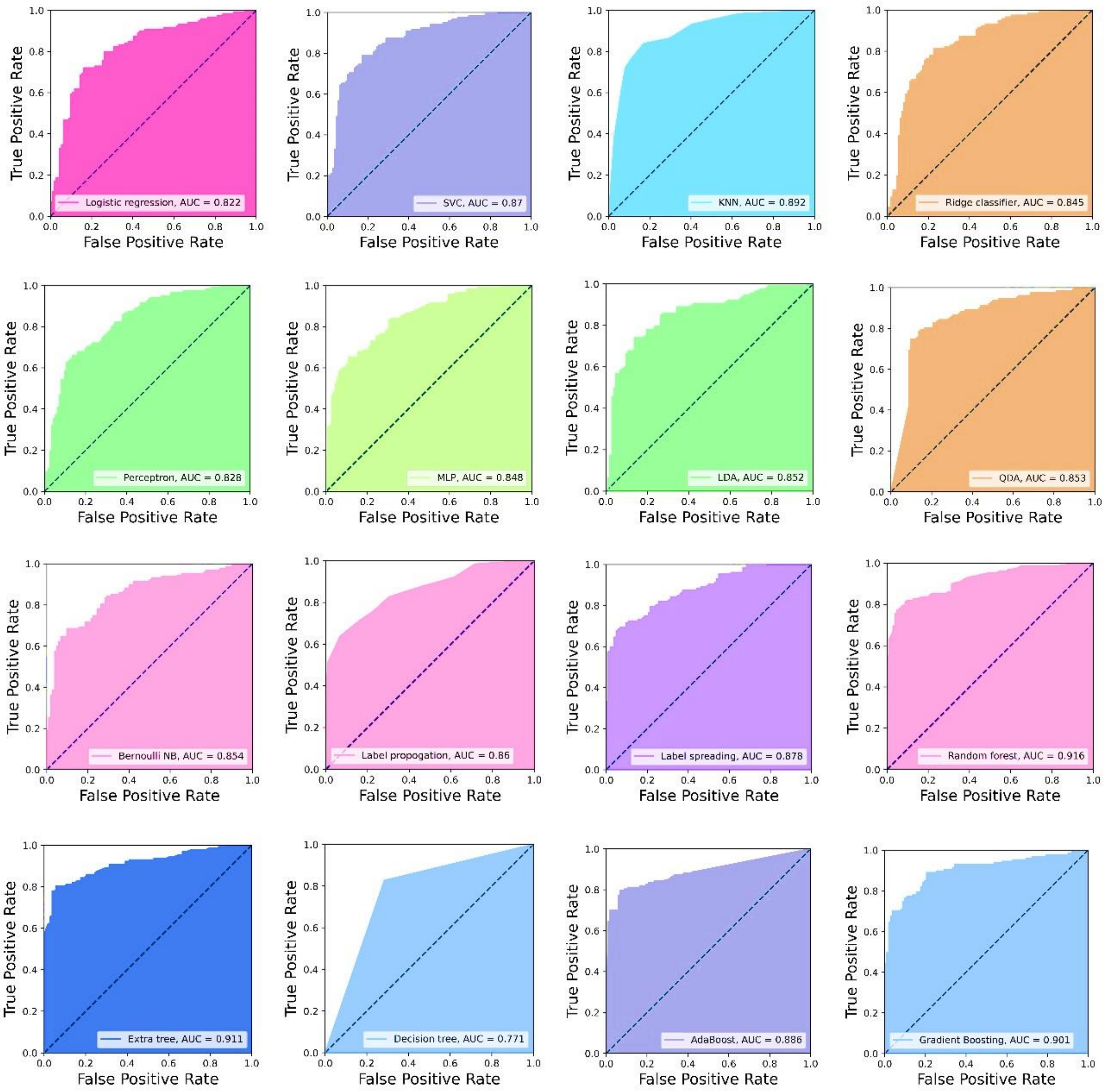


Figure 9.7 ROC of various algorithms used for the generation of binary classification models.

Table 9.3 Validation scores of machine learning algorithms employed in the development of binary classification models.

			Training set	**Test set**			
Algorithm	**Parameters**	**Data Preprocessing**	**Accuracy (Mean ± SD)**	**Accuracy**	**Precision**	**Recall**	**F1 score**
Logistic regression	C = 50, class_weight = balanced, max_iter = 10000, penalty = l2, solver = saga	Yes	80.28 ± 6.57	75.55	73.85	81.36	77.42
SVC	C = 10, gamma = 0.1, kernel = rbf, probability = 1	Yes	81.63 ± 5.05	78.6	78.51	80.51	79.5
K Neighbors Classifier	metric = manhattan, n_neighbors = 9, weights = uniform	Yes	82.54± 6.22	83.41	83.9	83.9	83.9
Ridge classifier	alpha = 0.001, class_weight = balanced, solver = sag	Yes	81.08 ± 6.32	77.73	76.8	81.36	79.01
Perceptron	alpha = 0.001, class_weight = balanced, max_iter = 10000	Yes	72.04 ± 8.11	72.05	87.5	53.39	66.32
MLP	activation = relu, alpha = 0.01, batch_size = 8, learning_rate = adaptive, solver = adam	Yes	79.95 ± 4.32	74.24	73.98	77.12	75.52
Linear Discriminant Analysis	solver = svd	No	77.12 ± 3.85	78.17	80.8	79.53	80.16
Quadratic Discriminant Analysis		No	74.15 ± 8.15	74.24	89.53	60.63	72.3
Bernoulli Naïve Bayesian		No	73.77 ± 5.39	75.98	78.57	77.95	78.26
Label Propagation	kernel = knn, n_neighbors = 5	No	78.82 ± 4.57	76.42	80.67	75.59	78.05
Label Spreading	kernel = knn, n_neighbors = 5	No	81.62 ± 4.29	78.17	80.8	79.53	80.16
Random forest classifier	bootstrap = True, criterion = gini, max_features = auto, n_estimators = 100	No	83.48 ± 5.75	83.84	87.5	82.68	85.02
Extra trees classifier	bootstrap = False, criterion = gini, max_features = auto, n_estimators = 200	No	84.07 ± 3.61	83.41	87.39	81.89	84.55
Decision tree classifier	class_weight = balanced, criterion = gini, max_features = auto, splitter = best	No	78.42 ± 6.14	77.73	78.36	82.68	80.46
AdaBoost classifier	n_estimators=100, algorithm = SAMME.R	No	84.63 ± 2.78	82.97	87.29	81.1	84.08
Gradient boosting classifier	loss = exponential, n_estimators = 200	No	82.18 ± 5.19	83.41	83.97	86.61	85.27

9.3.5.3. Development of scoring function using regression algorithms

The first set of regression-based models were developed for predicting IC_{50} of the 'active compounds' as predicted by the binary classification-based model in the previous section. This set of models would predict IC_{50} below 10000 nM and are designated as 1.01 and so on **(Table 9.4)**. A set of 417 compounds, having IC_{50} below 10000 nM, were used to develop these models. The pre-processing, either by standardisation or normalisation, was performed depending on the distribution of data in a feature determined using Shapiro-Wilk test for normality for each feature. The dataset was divided into training and test sets in a ratio of 75:25. Various regression models, neural networks and ensemble algorithms were employed for model training. Linear regression displayed a good r^2 of 0.893 on the training set but had poor MAE and MSE. The lasso and elastic net regression showed poor r^2 values on the training set. Further, Q^2_{F1} and Q^2_{F2} indicated a correlation between the actual and predicted activity of the test set with respect to mean activity of training and test sets, respectively. A value of above 0.5 was an indicator of a good model. A poor value of Q^2_{F1} and Q^2_{F2} with a good r^2 indicated overfitting of the model on the training set. In contrast, good Q^2_{F1} and Q^2_{F2} scores above the desired threshold with a poor r^2 indicated underfitting. Linear regression showed overfitting, while the lasso and elastic net regression indicated poor fitting. Bayesian ridge regression, used naïve Bayes probability for prediction, performed quite good with an r^2 value of 0.75 and acceptable Q^2_{F1} and Q^2_{F2} scores.

Similarly, ridge, SVR, SGD and ARD regression models performed well with the r^2, Q^2_{F1} and Q^2_{F2} values comparable to the model developed from Bayesian ridge regression. Among the other regression methods, MLP and Huber regression displayed good r^2. Q^2_{F1} and Q^2_{F2} were in the acceptable range for MLP but were poor in the case of Huber regression. However, ensemble methods have outperformed all other algorithms with the lowest MAE and MSE on training and test datasets. The RF and extra tree regression have the best Q^2_{F1} and Q^2_{F2} among all the regression models. The two algorithms were able to pass Tropsa criteria among the developed models. Further, RF regression performed better among the two and was selected. The details of the validation parameters for all the generated models are tabulated (**Table 9.4**) and presented in figure **(Figure 9.8)**.

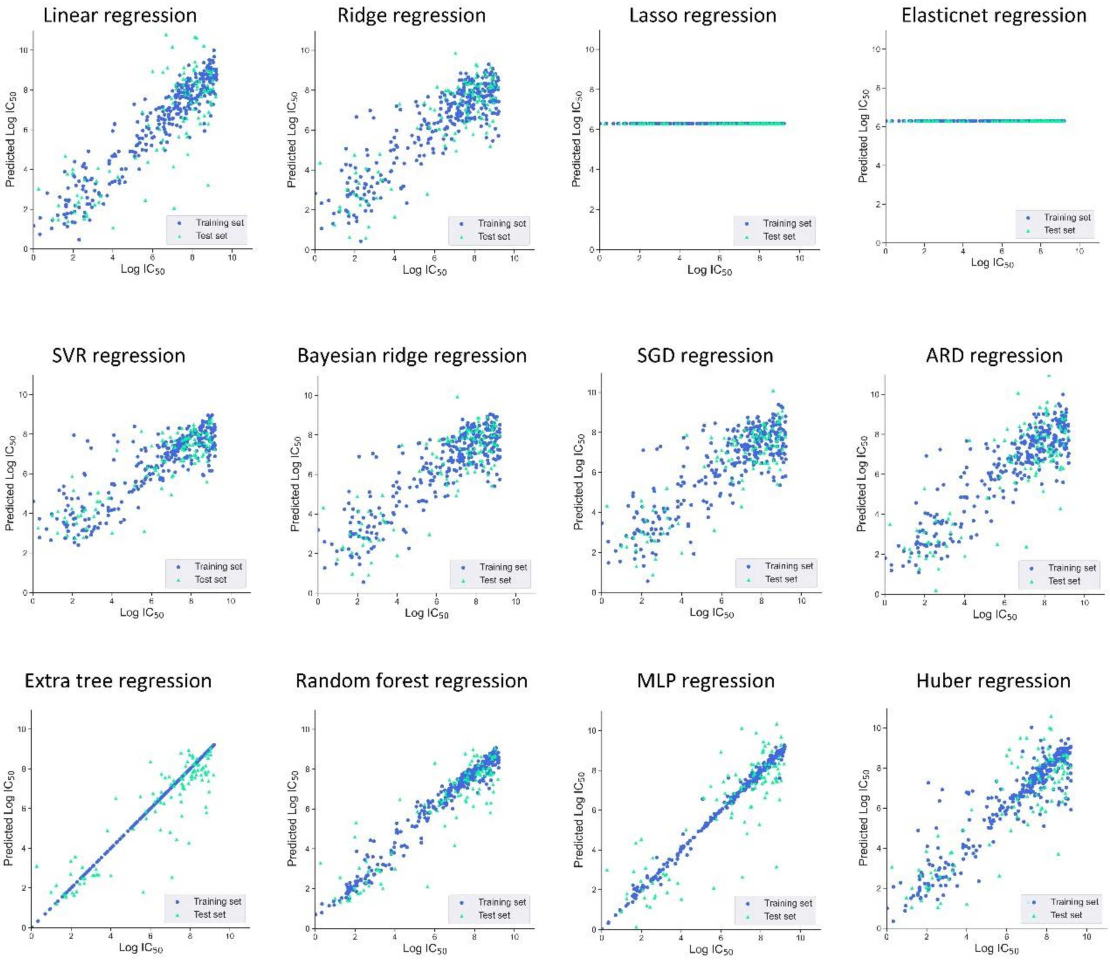
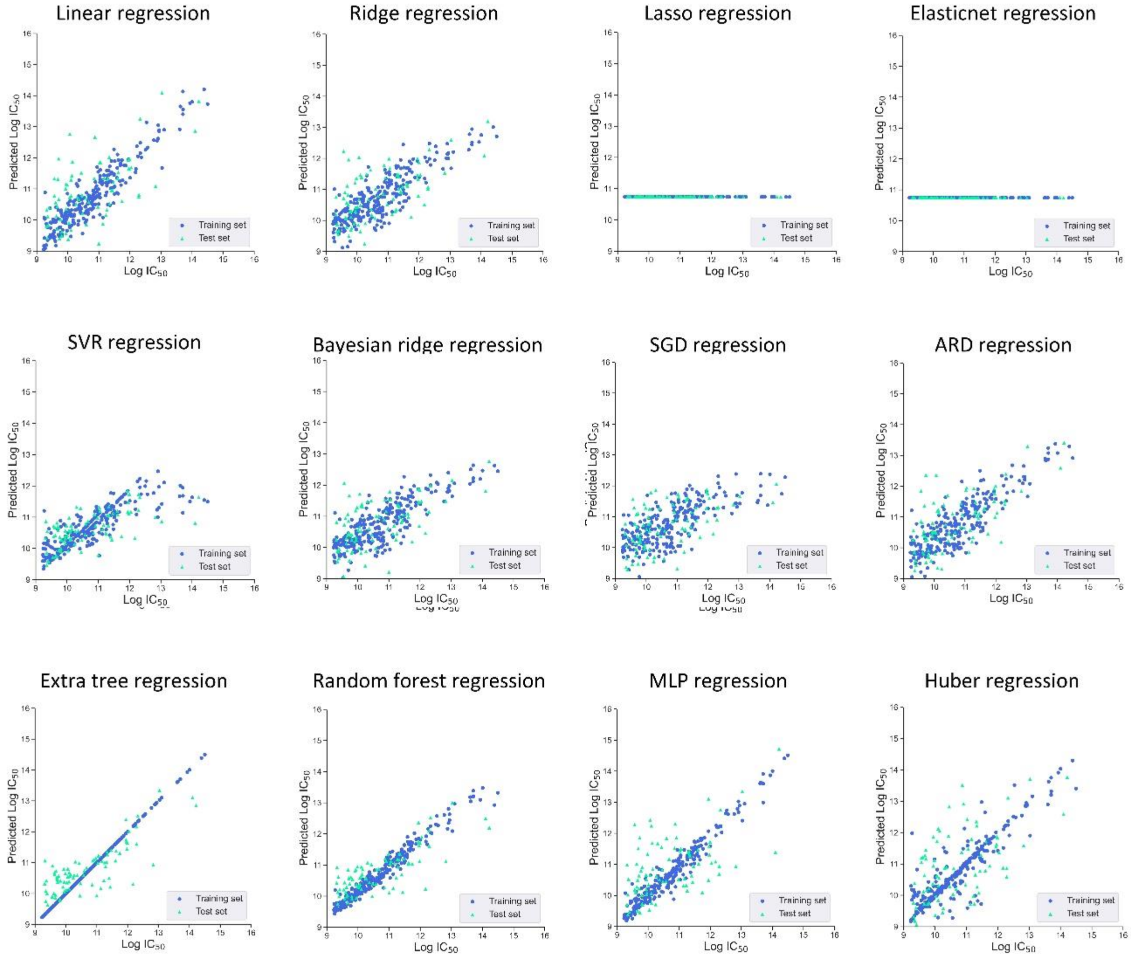


Figure 9.8 Scatter plot between predicted and experimental IC_{50} obtained from various algorithms for the development of regression-based models for the compounds with IC_{50} below 10000 nM (active).

Table 9.4 Validation scores of machine learning algorithms employed in the regression-based models for prediction of IC_{50} below 10000 nM.

Model	**Algorithm**	**Coefficient of determination(train)**	**MAE Test**	**MAE Train**	**MSE Test**	**MSE Train**	Q^2_{F1}	Q^2_{F2}	**TROPSA criteria**
1.01	Linear regression	0.893	1460421	0.617	2.24E+14	0.643	-3.6E+13	-3.6E+13	FAIL
1.02	Ridge regression	0.774	1.03	0.889	1.749	1.365	0.719	0.719	FAIL
1.03	Lasso regression	0	2.119	2.071	6.228	6.03	0	-0.002	FAIL
1.04	ElasticNet regression	0	2.119	2.071	6.228	6.03	0	-0.002	FAIL
1.05	SVR regression	0.72	1.037	0.862	1.796	1.691	0.712	0.711	FAIL
1.06	BayesianRidge regression	0.75	1.034	0.93	1.717	1.507	0.724	0.724	FAIL
1.07	SGD regression	0.712	1.046	0.999	1.866	1.738	0.7	0.7	FAIL
1.08	ARD regression	0.798	1.305	0.833	3.741	1.217	0.399	0.398	FAIL
1.09	ExtraTrees regression	1	0.795	0	1.391	0	0.777	0.776	PASS
1.10	**RandomForest regression**	**0.964**	**0.771**	**0.331**	**1.227**	**0.214**	**0.803**	**0.803**	**PASS**
1.11	MLP regression	0.992	1.084	0.121	2.248	0.05	0.639	0.638	PASS
1.12	Huber regression	0.837	1.34	0.605	5.923	0.986	0.049	0.047	FAIL

Similarly, the other set of models were developed from a dataset of compounds having IC_{50} above 10000 nM. The dataset had 338 compounds with 167 features. The pre-processing and dataset bifurcation was carried out as described previously.



Linear regression

Ridge regression

Lasso regression

Elasticnet regression

SVR regression

Bayesian ridge regression

SGD regression

ARD regression

Extra tree regression

Random forest regression

MLP regression

Huber regression

Figure 9.9 Scatter plot between predicted and experimental IC_{50} obtained from various algorithms for developing regression-based models for compounds with IC_{50} above 10000 nM (moderately active).

Linear and ridge regression models performed well on the training dataset with r^2 values of 0.865 and 0.715, respectively. However, the test set evaluation indicated poor correlation between the predicted and observed activities with poor Q^2_{F1} and Q^2_{F2}. Lasso and elastic net regression performed poorly on both datasets. The models developed from SVR, SGD, ARD and Bayesian ridge regressions performed quite better than lasso regression, but still had a suboptimal fitting on both the datasets. Interestingly, the MLP and Huber regression models were performing well on the training dataset with high values for r^2, but were underfitted on the test set. The ensemble-based extra trees regressions algorithms performed well in all aspects and fulfilled Tropsa criteria with acceptable values of Q^2_{F1} and Q^2_{F2}. The extra trees regression model was finally selected to develop the SF due to better performance than the random forest algorithm on the test dataset. The validation report for all the developed models is tabulated **(Table 9.5)** and presented in the figure **(Figure 9.9)**.

9.3.5.4. Applicability domain for the scoring function

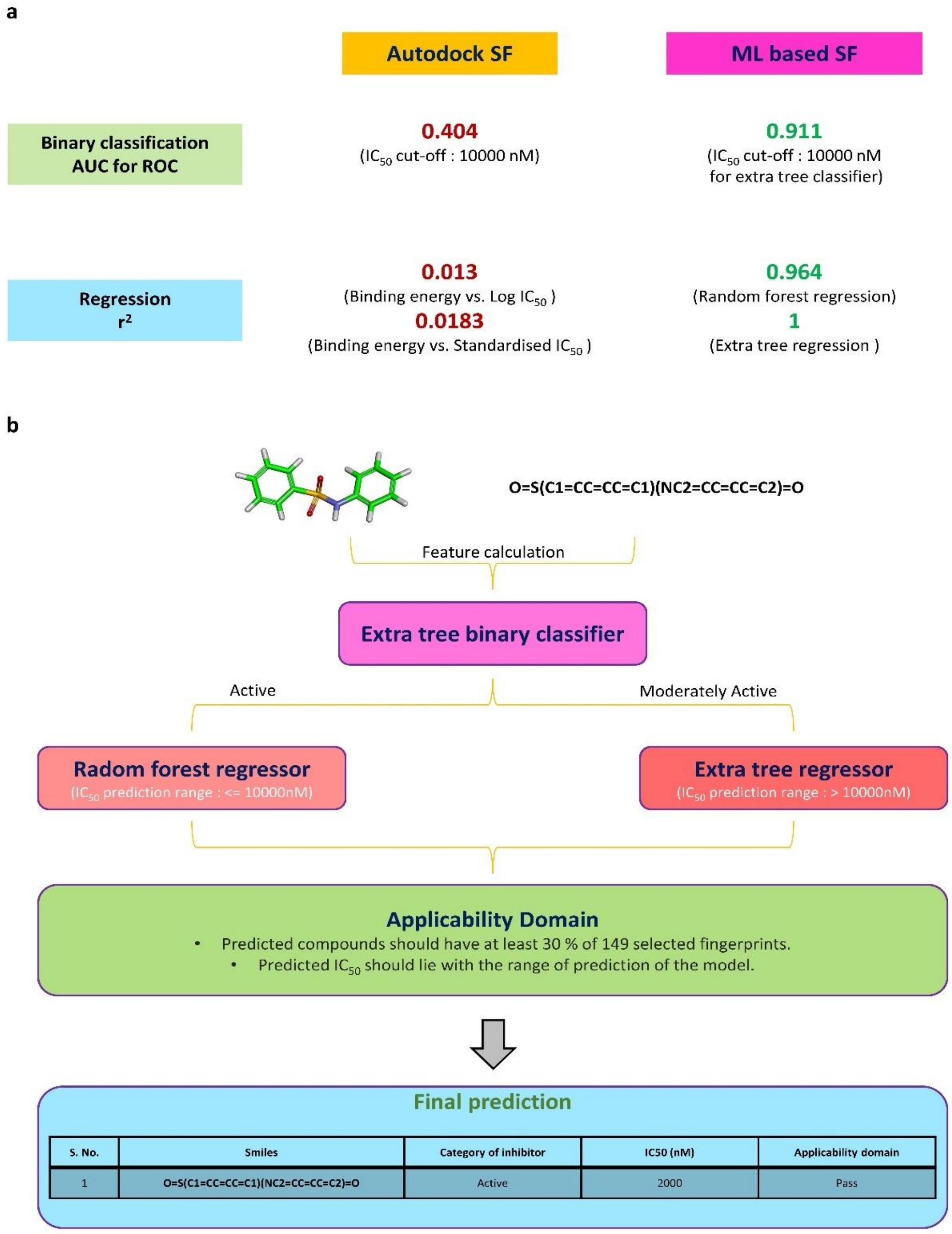
The goal of the applicability domain is to define the assumptions in the development of a ML model and to assess that the assumptions are satisfied during the prediction in order to ascertain confidence over the obtained results. Usually, the applicability domain is the chemical space used to build or train the ML or a QSAR model and the compounds that fall within the applicability domain space can be predicted with certainty. It depends on the similarity of the compound to that of the training set used to build the model. In the present study, the applicability domain for the developed SF was also defined. The training and test datasets were populated with the compounds having similarities with ten inhibitors co-crystallised with human BChE. The 149 previously identified fingerprints from the BChE inhibitors were used to define applicability domain. The molecules having at least 30% of the fingerprints in their chemical structure were only selected. The other criterion was the predicted activity of the compound, which should be within the range of the observed activity of the training set (i.e., 0 – 1980000 nM). The molecules satisfying both the criteria of the applicability domain were selected.

Table 9.5 Validation scores of machine learning algorithms employed in the regression-based model for prediction of IC_{50} above 10000 nM.

Model	**Algorithm**	**Coefficient of determination (train)**	**MAE Train**	**MAE Test**	**MSE Train**	**MSE Test**	Q^2_{F1}	Q^2_{F2}	**TROPSA criteria**
2.01	Linear regression	0.865	38.082	0.304	24968.49	0.158	-21178.6	-21300.7	FAIL
2.02	Ridge regression	0.715	0.73	0.462	0.858	0.334	0.272	0.268	FAIL
2.03	Lasso regression	0	0.912	0.84	1.179	1.17	0	-0.006	FAIL
2.04	ElasticNet regression	0	0.912	0.84	1.179	1.17	0	-0.006	FAIL
2.05	SVR regression	0.651	0.628	0.395	0.723	0.409	0.387	0.383	FAIL
2.06	BayesianRidge regression	0.658	0.718	0.5	0.806	0.4	0.317	0.313	FAIL
2.07	SGD regression	0.5	0.74	0.592	0.865	0.585	0.266	0.262	FAIL
2.08	ARD regression	0.737	0.699	0.445	0.874	0.308	0.259	0.254	FAIL
2.09	**ExtraTrees regression**	**1**	**0.512**	**0**	**0.439**	**0**	**0.627**	**0.625**	**PASS**
2.10	RandomForest regression	0.933	0.555	0.205	0.5	0.078	0.576	0.574	FAIL
2.11	MLP regression	0.95	0.826	0.168	1.186	0.058	-0.006	-0.011	FAIL
2.12	Huber regression	0.813	0.923	0.26	1.347	0.218	-0.142	-0.149	FAIL

9.3.5.5. Formulation of the scoring function

The selected binary classification as well as the regression-based models, performed better than the native SF of the Autodock for BChE inhibitors **(Figure 9.10**). Various selected models were compiled as SF along with the applicability domain and termed as Protein-Ligand Scoring Function (PLSF). SF accepted two inputs, i.e., the pose of the ligand obtained from docking and its SMILES string. It was followed by a calculation of the various features. The features were fed to the Extra tree binary classification model to produce one of the outputs, i.e., active or moderately active. If the compound was predicted as active, the IC_{50} prediction was carried out by using a RF based regression model; otherwise, the extra tree regression model was employed. Further, the input was also tested for the applicability domain. The developed SF was compiled as a python library and made available to users (https://www.drugdesign.in/tools).


a
Autodock SF
ML based SF
Binary classification AUC for ROC
0.404
(IC_{50} cut-off : 10000 nM)
0.911
(IC_{50} cut-off : 10000 nM for extra tree classifier)
Regression r^2
0.013
(Binding energy vs. Log IC_{50})
0.0183
(Binding energy vs. Standardised IC_{50})
0.964
(Random forest regression)
1
(Extra tree regression)
b
O=S(C1=CC=CC=C1)(NC2=CC=CC=C2)=O
Feature calculation
Extra tree binary classifier
Active
Moderately Active
Radom forest regressor
(IC_{50} prediction range : <= 10000nM)
Extra tree regressor
(IC_{50} prediction range : > 10000nM)
Applicability Domain
• Predicted compounds should have at least 30 % of 149 selected fingerprints.
• Predicted IC_{50} should lie with the range of prediction of the model.
Final prediction

S. No.	Smiles	Category of inhibitor	IC50 (nM)	Applicability domain
1	O=S(C1=CC=CC=C1)(NC2=CC=CC=C2)=O	Active	2000	Pass

Figure 9.10 (a) Comparison between the Autodock SF and selected ML models, (b) Schematic representation of the formulated scoring function.