Chapter 8

Development of homology model, docking protocol and machine-learning based scoring functions for identification of *Electrophorus electricus's* AChE inhibitors

8. Development of homology model, docking protocol and machine-learning based scoring functions for identification of *Electrophorus electricus*'s AChE inhibitors

8.1. Introduction

SBDD involves the use of the protein structure and features to design the ligands, which is accomplished through various *in silico* techniques, including molecular docking and MD simulation methods [252]. The use of SBDD methods depends upon the availability of a 3-dimensional protein structure. In the absence of protein structure, homology modelling remains an important technique for the prediction of structure. Molecular docking is a popular technique used for virtual screening. However, the validation of docking protocols is still a tedious process. One of the frequently available methods is to compare protein-ligand complex obtained from experimental data and docking to establish the pose reproducibility. The results are calculated in terms of RMSD between the docked and co-crystallised poses of the ligand [253]. The other method includes docking of a set of ligands with known experimental activity data and comparing the docking score with it using ranking, classification or regression-based methods [242, 254]. Another crucial factor involved in the success of molecular docking is the prediction accuracy of the scoring function (SF) available with the docking program. Sometimes, it is observed that the performance of problem-specific customised scoring is better than the general SFs. Further, the use of ML and deep learning techniques in the development of such SF have gained acceptance in recent years in providing validated results [176, 255-257].

AChE, obtained from *Electrophorus electricus* (ee), which shares structural homology with the human AChE, is widely used for the screening of AChE inhibitors. Since the 3-dimensional crystal structure of the enzyme from ee is not available and therefore, it is challenging to use the SBDD approach for the identification of inhibitors. In the present study, a homology model for eeAChE was developed and was subjected to its structural

refinement through energy minimisation. The docking protocol was also developed and validated by redocking a set of co-crystallised inhibitors from mouse AChE, and their interaction profiles were compared. The results indicated a poor performance of the Autodock SF. Hence, a batch of machine learning-based SFs was developed and validated. It also included the development of dedicated ML-based SFs for obtaining improved prediction results.

8.2. Materials and methods

8.2.1. Sequence alignment and analysis

The protein sequence for eeAChE was obtained from Uniport web database (accession code O42275). The sequence similarity search was carried out using blastp (BLASTP 2.9.0+) with the help of Blosum62 scoring matrix and gaps were allowed during sequence comparison [188, 258, 259].

8.2.2. Active site mapping

The active site cavity was mapped by using CavityPlus (http://www.pkumdl.cn/) web server, that provided a comprehensive analysis of cavities of the protein [260]. The cavity analysis for AChE obtained from *Tetronarce californica* (tc), *Mus musculus* (mm), and *Homo sapiens* (hs) was performed on chain A of the protein using the default parameters.

8.2.3. Homology modelling

The homology modelling for protein structure of eeAChE was performed. Various models were developed using SWISS-MODEL web server accessible via the ExPASy (https://swissmodel.expasy.org) [261]. The sequence of eeAChE (UniProt access code – O42275) was used for carrying out a template search. The templates obtained were used to build the eeAChE models. The quality of the obtained homology models were checked

through Global Model Quality Estimation (GMQE), QMEAN [262] and Ramachandran plot obtained from Molprobity [263] [197].

8.2.4. Homology model refinement and protein preparation

The selected homology models were refined using DOCKPREP utility of Chimera-1.4. The PDB was further uploaded to the PDB2PQR server (http://nbcr-222.ucsd.edu/pdb2pqr_2.1.1) to assign protonation states to the various amino acid residues of model at a physiological pH of 7.4 [264, 265]. The obtained PDB was subjected to energy minimisation using Amber18 with the ff14SB as a force field using method stated in section 4.2.9 [200, 266, 267].

8.2.5. Ligand preparation and grid generation

The ligand preparation was carried out using method reported in section 4.2.7. Autogrid-4.2.6 was used to calculate grid maps. The grid box size was set to 84 × 66 × 72 with a grid point spacing of 0.375 Å. The grid centre was placed at 51.05, 28.382 and 54.297, representing X, Y and Z coordinates, respectively.

8.2.6. Molecular docking and validation of docking protocol and scoring function.

The molecular docking was performed through Autodock-4.2.6. LGA along with Solis-Water local search to identify the various poses [196, 235]. The docking results were processed by a python script, i.e., vstools_v0.16. The post docking analysis and visualisation were performed by Discovery studio visualiser 2020. The docking protocol was validated through two approaches. The structures of the ligands that were co-crystallised with mmAChE were collected and docked on the eeAChE. The interactions obtained from docked protein-ligand complexes were compared with the respective native co-crystals. The other validation process involved docking of the pre-collected dataset of ligands with known IC_{50} against the eeAChE using the predefined protocol. The comparison of the obtained binding energies with their IC_{50} was carried out through three

approaches. The first approach used binary classification involving the division of datasets into two classes, i.e., active and moderately active. The area under the receiver operating characteristic curve (ROC) was determined assuming the docking as a classification-based problem. The other method involved treating the docking results as continuous data similar to the regression problem. It was validated using a coefficient of determination between actual pIC_{50} and predicted binding energies and pKi after using appropriate data pre-processing viz. logarithmic conversion of predicted Ki and experimental IC_{50} values.

8.2.7. Development of the scoring functions for eeAChE

SFs were developed through ML and deep learning techniques using various python libraries. A set of three SFs were developed using the principles of binary classification, multiclass classification and regression. The binary and multiclass classification SFs were developed using various ML algorithms viz. support vector classification (SVC), LR, KNN, RF, Naïve Bayesian and a variety of ensembled based techniques. A grid search was carried out to identify the best hyperparameter combinations. Initially, a hyperparameter tuning was performed for each algorithm and the selection of the best hyperparameters was carried out on the basis of mean accuracy obtained from 5-fold validation. The best model selected from each algorithm was tested on an independent test set to select the final model for formulating the SF. The final model was selected by comparing the various scores viz. confusion matrix, accuracy, precision, recall and F1-score. Further, the third set of ML and deep learning models were developed using regression-based algorithms viz. linear regression, ridge regression, elasticnet regression, lasso regression, support vector regression (SVR), RF regressor, Bayesian Ridge regression, Stochastic gradient regression and neural networks. These models were validated by using the coefficient of determination (r^2), which 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. The Q^2_{ext}-based metrics were used to ascertain model performance on the test data. The value of Q^2_{F1} and Q^2_{F2} should be greater than 0.5 for an acceptable quality of the model. The value less than the threshold indicated that though the model fits better on the training set; but have poor predictivity on a different dataset, i.e., overfitting. Further, MAE and Golbraikh and Tropsha's criteria were also used for the evaluation of the model.

8.3. Results and discussion

8.3.1. Sequence alignment and analysis

The eeAChE is 633 amino acids containing protein with a molecular weight of 71,815 Da. The first 23 amino acid residues are part of a signal peptide responsible for protein translocation, which is eventually cleaved. The blast search reflected that the top 500 reported hits were of two types of enzymes, i.e., Carboxylic ester hydrolases (CEHs) and AChE. Further, about 88.9 % of the hits were CEH of various organisms and, their 3D structures were not reported. The blast search also returned 6 % of hits as AChE with *Triplophysa tibetana* AChE being most similar to eeAChE (81.3 %). However, the significant queries returned had no 3D protein structures reported (**Figure 8.1**). The tcAChE, mmAChE and hsAChE obtained from blast search have sequence identities of 64.5, 59.8 and 60.8 %, respectively with eeAChE. The PDB structures of AChE of the three organisms were available in the PDB databank (https://www.rcsb.org/). AChE crystal structure of *Drosophila melanogaster* was also available but had a low sequence identity of 35.91 % with eeAChE. In the protein sequence analysis, the residues having physiochemically similar properties were treated as similar residues. Similar residues present in eeAChE sequence were found in a range of 12-15 %, when compared with the three organisms. The three selected organisms had a high homology of more than 70 %,

except for *drosophila* that shared a low homology with eeAChE (**Figure 8.1**). The sequence alignment also showed that all the major residues of various sites of AChE were identical and highly conserved among the organisms. The conserved residues of major sites were highlighted in **Figure 8.2,** representing the aligned sequences.

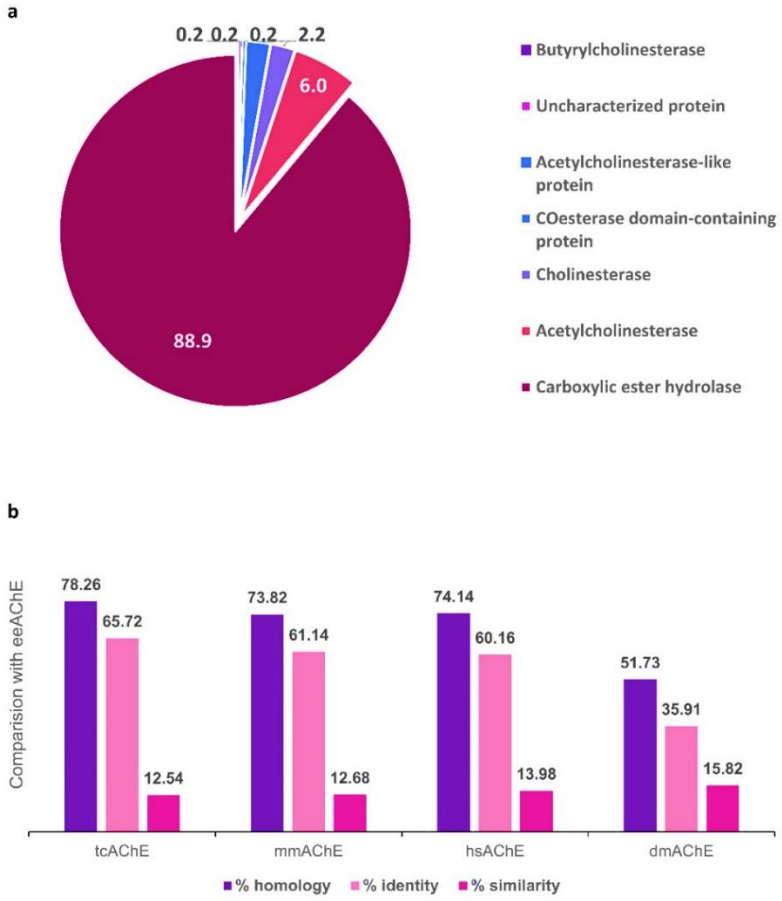

Figure 8.1 (a) Hits returned from a protein blast search using eeAChE as a query. (b) Comparison of eeAChE sequence with other organisms.

8.3.2. Active site mapping

CavityPlus helps in identifying various cavities along with their druggability scores. Thus, the mapping of active site and tunnel were performed for tcAChE, mmAChE and hsAChE using PDB codes 1EA5, 5DTI and 4EY4, respectively, that could serve as potential templates for homology model development for eeAChE.

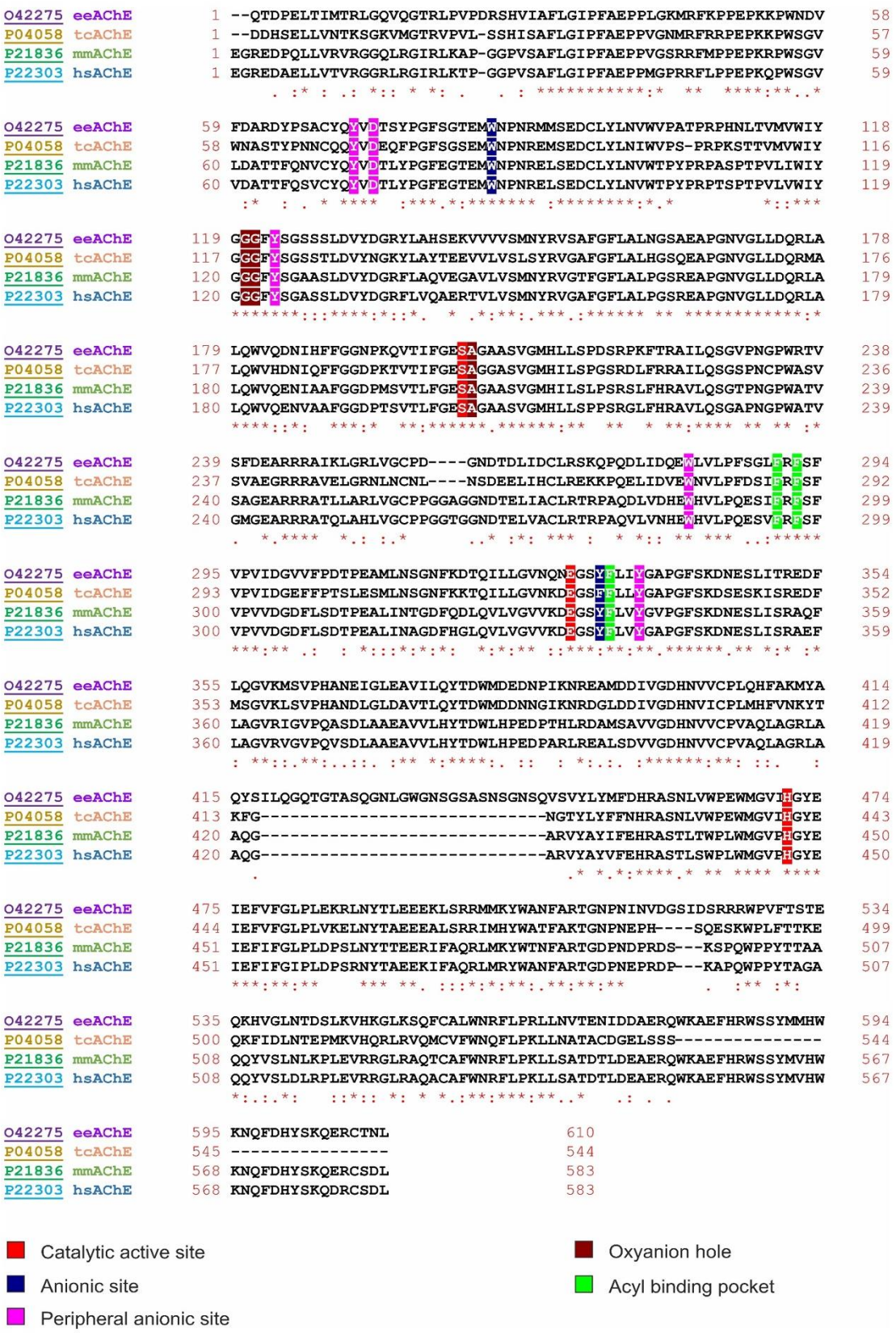

Figure 8.2 Sequence alignment of AChE Electrophorus electricus with *Tetronarce californica*, *Mus musculus* and *Homo sapiens*.

a

```
O42275 eeAChE      70 QYVDTSY 76  81 GTEMWNP 87  116 W  118 YGGGFYSGS  129 L  131 VY 132
P04058 tcAChE      69 QYVDEQF 75  80 GSEMWNP 86  114 W  116 YGGGFYSGS  127 L  129 VY 130
                      **** .:      *******        *      *********      *      **
P21836 mmAChE      71 QYVDTLY 77  82 GTEMWNP 88         119 YGGGFYSGA  130 L  132 VY 133
                      ***** *      *******                ********:      *      **
P22303 hsAChE      71 QYVDTLY 77  82 GTEMWNP 88         119 YGGGFYSGA  130 L  132 VY 133
                      ***** *      *******                ********:      *      **

O42275 eeAChE     201 ESAG 204  235 W  277 ID 278  280 EWLVLPFSGLFRFS 293  329 EGSYFLIYGA 338
P04058 tcAChE     199 ESAG 202  233 W  275 ID 276  278 EWNVLPFDSIFRFS 291  327 EGSFFLLYGA 336
                      ****         *      **          ** ****.:****         ***:**:***
P21836 mmAChE     202 ESAG 205  236 W  282 VD 283  285 EWHVLPQESIFRFS 298  334 EGSYFLVYGV 344
                      ****         *     :*           ** *** .:****         ******:**.
P22303 hsAChE     202 ESA  204                       285 EWHVLPQESVFRFS 298  334 EGSYFLVYGA 344
                      ***                            ** *** .:****         ******:***

O42275 eeAChE     360 M  463 W  470 IHGY 473  474 EI 475
P04058 tcAChE     358 L  432 W  439 IHGY 442       I 444
                      :      *      ****           *
P21836 mmAChE     365 I  439 W  446 PHGY 449  450 EI 451
                      :      *      ***            **
P22303 hsAChE     365 V  439 W  446 PHGY 449       I 451
                      :      *      ***            *
```

b

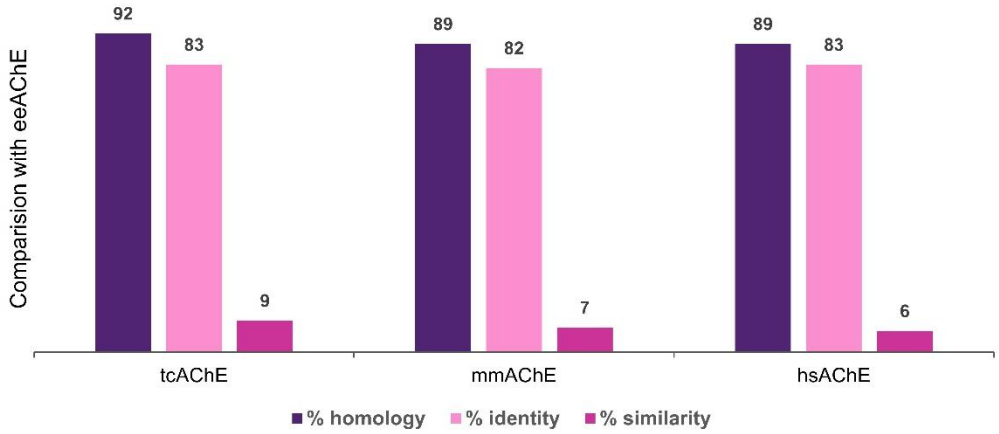

Figure 8.3 (a) Sequence comparison (b) % homology, identity and similarity of active site and tunnel of eeAChE with other organisms.

Sixteen cavities were identified from tcAChE, but only one druggable site consisting of an active catalytic site and the leading tunnel was predicted. Similarly, twelve and fourteen cavities are identified for mmAChE and hsAChE, respectively, with one cavity each containing the tunnel and catalytic site residues Sequence analysis was performed from the obtained data to identify the similarity of the active site and tunnel residues between eeAChE and the organisms. The results of the sequence comparison are presented in **Figure 8.3**. Interestingly, the homology of the three selected proteins with eeAChE sequence was approximately 90 % or more. In case of tcAChE, it was observed

that 65 residues made up the active site with two residues had radical replacements, five and three residues displayed conservative and semi-conservative replacements, respectively, when compared with corresponding eeAChE residues. Further, 65 and 59 residues were involved in the active site formation for mmAChE and hsAChE, respectively. The mmAChE have 5, 3 and 5 non-conservative, semi-conservative and conservative replacement respectively, in comparison to the corresponding eeAChE sequence. While, hsAChE has 3 radical replacements and 3 and 5 semi-conservative and conservative replacements, respectively, in relation to eeAChE (**Figure 8.3**) [268].

8.3.3. Homology modelling

The template search on SWISS MODEL retrieved 48 templates. Three PDB of AChE from *Tetronarce californica, Mus musculus,* and *Homo sapiens*, having higher resolution, were selected for homology modelling. It was ensured that none of these PDB had any mutations and were XRD-derived protein structures due to better atomic resolution. GMQE scores of all the models were in an acceptable range of 0.70 – 0.74. Three out of nine homology models were selected for model refinement that displayed better QMEAN scores. It was observed that all of the developed models had 91 – 93% residues in the favoured region and 0 – 2% outliers. The details of the homology models, PDB templates and their quality are indicated in **Table 8.1**.

8.3.4. Homology model refinement and protein preparation

Three homology models, one each from *Tetronarce californica*, *Mus musculus*, *and Homo sapiens*, were selected based on the QMEAN scores. The selected models were subjected to protein preparation. The correct protonation states of various amino acid residues were assigned using PDB2PQR server at physiological pH of 7.4. The energy minimisation was performed to reduce the inconsistency in the prepared model and was carried out using Amber18. The protein structure at each stage of energy minimisation

was assessed by the structure assessment tool available on the SWISS-MODEL. A thorough comparison of models at every minimisation stage on the various parameters are presented in **Figure 8.4**.

Table 8.1 Validation data of the developed homology models of eeAChE from crystal structures of various organisms.

S. No.	Template PDB	Resolution (Å)	Organism	GMQE	QMEAN	Ramachandran Favoured (%)	Ramachandran Outliers (%)	Model code
1	**5FPQ**	**2.4**	***Homo sapiens***	**0.74**	**-0.78**	**92.76**	**0.88**	**eeAChE_5FPQ**
2	**6FQN**	**2.3**	***Torpedo californica***	**0.70**	**-0.85**	**92.02**	**0.35**	**eeAChE_6FQN**
3	**2C0P**	**2.5**	***Mus musculus***	**0.73**	**-0.94**	**91.71**	**0.53**	**eeAChE_2C0P**
4	1KU6	2.5	*Mus musculus*	0.73	-1.32	91.07	1.77	eeAChE_1KU6
5	5HF5	2.2	*Homo sapiens*	0.73	-1.42	92.58	0.71	eeAChE_5HF5
6	4PQE	2.9	*Homo sapiens*	0.73	-1.46	92.38	1.06	eeAChE_4PQE
7	6EUE	2.0	*Torpedo californica*	0.71	-1.47	92.20	0.89	eeAChE_6EUE
8	2WHP	2.2	*Mus musculus*	0.73	-1.48	92.58	1.59	eeAChE_2WHP
9	5EHX	2.1	*Torpedo californica*	0.70	-2.06	91.84	1.42	eeAChE_5EHX

The coordinates obtained from stages 6, 6 and 5 of the energy minimisation of homology models obtained from templates of *Mus musculus* (mmAChE: 2C0P), *Tetronarce californica* (tcAChE: 6FQN) and *Homo sapiens* (hsAChE: 5FPQ), respectively, displayed optimum features. In the case of eeAChE_2C0P, the clash score decreased along with rotamer outliers and bad angles. However, there was a slight increase in Ramachandran outliers. The model eeAChE_6FQN displayed a decrease in the bad angles, rotamer outliers and clash score. The protein model obtained from *Homo sapiens* showed low Ramachandran outliers, low rotamer outliers and bad angles. Three optimised protein models were then compared, and the 3D structure obtained from *Mus musculus* showed a minimum deviation (**Table 8.2**). The final PDB (eeAChE_2C0P_min6) and its

Ramachandran plot are represented in **Figure 8.4**. The obtained PDB was converted into pdbqt format using Autodock Tools-1.5.6 [269].

Table 8.2 Comparison of validation parameters of selected homology models obtained from energy minimisation.

Model	MolProbity Score	Clash Score	Ramachandran Favoured (%)	Ramachandran Outliers (%)	Rotamer Outliers (%)	C-Beta Deviations	Bad Bonds	Bad Angles
eeAChE_2C0P	1.25	0.57	92.37	0.6	1.28	0	0	16
eeAChE_5FPQ	1.38	1.03	91.6	0.4	1.28	0	0	18
eeAChE_6FNQ	1.47	2.29	92.73	0.61	0.86	0	0	20

8.3.5. Ligand preparation and grid generation

The 3-dimensional coordinates of molecules were generated from SMILES strings which were subjected to energy minimisation using the MMFF94s force field. Subsequently, the minimised structures were converted to pdbqt files using AutoDockTools-1.5.6. The grid box used in the study covered the whole active site and the tunnel, which had been mapped and is indicated in **Figure 8.2**.

8.3.6. Molecular docking and validation of docking protocol and scoring function.

Autodock-4.2.6 uses a semi-empirical force field in its SF to evaluate various generated poses. The binding free energy is calculated as:

$\Delta G = (V_{bound}{}^{L\text{-}L} - V_{unbound}{}^{L\text{-}L}) + (V_{bound}{}^{P\text{-}P} - V_{unbound}{}^{P\text{-}P}) + (V_{bound}{}^{P\text{-}L} - V_{unbound}{}^{P\text{-}L} + \Delta S_{conf})$

where, L and P refers to the ligand and protein and the conformational entropy lost upon binding (ΔS_{conf}). The pairwise energy term (V) represents the contributions of van der Waals, electrostatic, hydrogen bond, de-solvation and torsional penalty involved in protein-ligand binding.

$$V = W_{vdw}\sum_{i,j}\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} + W_{hbond}\sum_{i,j}E(t)\left\{\frac{C_{ij}}{r_{ij}^{12}} + \frac{D_{ij}}{r_{ij}^{10}}\right\} + W_{ele}\sum_{i,j}\frac{q_i q_j}{e(r_{ij})\,r_{ij}} + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{\left(-r_{ij}^2/2\sigma^2\right)} + W_{tor}N_{tor}$$

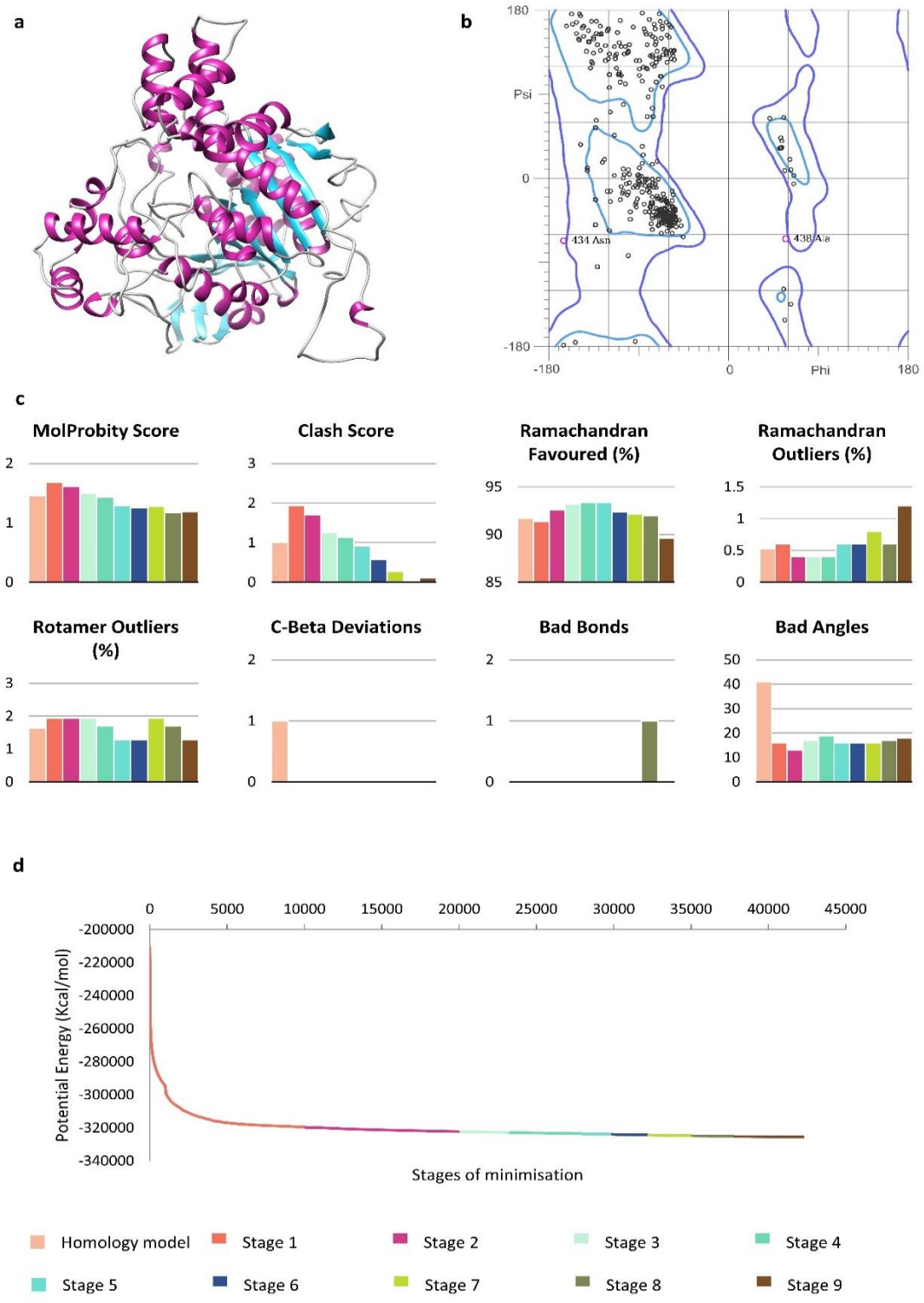

Figure 8.4 (a) Homology model of eeAChE, (b) Ramachandran plot of homology model after energy minimisation, (c) Validation score of protein structure at various stages of energy minimisation, (d) Potential energy (Kcal/mol) of the protein model during energy minimisation.

Conventionally, the ligands that were co-crystalised with protein are redocked and an RMSD is determined between the docked and co-crystallised poses for validation of the docking protocols [197, 267]. However, in this case, a homology model of eeAChE was built and hence no co-crystallised ligand could be obtained. Hence, a set of fifteen co-crystallised ligands with mmAChE were selected from the PDB databank. These ligands were docked with eeAChE and the poses with the lowest binding energy were selected. The interactions of the selected pose of the ligand with eeAChE were compared with the mmAChE interactions, as both the enzymes, had active site homology of about 88 %. The results are presented in **Figures 8.5, 8.6** and **8.7**. The ligand 5gz, when docked on eeAChE retained interactions with Trp85, Trp281, Phe290, Tyr332, Phe333, Tyr336 and His471; while A36 retained binding with three out of four interacting residues, i.e., Trp85, Tyr123 and Tyr332, when compared with mmAChE. The ligands B2V and B3W displayed interactions with four and three out of five and six residues, respectively, on comparison with mouse AChE.

B32 retained interactions with Trp85, Tyr123, Tyr332 and His471 and displayed an additional interaction with Tyr473 in comparison to mmAChE. The ligand C56 displayed retention of five out of seven interacting residues when docked to eeAChE, while DUC retained one out of two interacting residues along with extra interactions with Tyr71, Asp73, Leu284, Phe290, Arg291 and Tyr336 with eeAChE when compared to co-crystallised mouse AChE. E5H, E5K and GC8 displayed three, three and two out of three, five and five interactions, respectively with mmAChE when docked with eeAChE. E5H retained interactions with Trp85, Trp281, Tyr332, and E5K with Trp85, Trp281 and Tyr336. Similarly, N2K, Q4Q, SOF, Z5K and ZN4 retained five, four, three, four and three interactions with eeAChE out of eight, seven, five, six and six interactions present in mmAChE.

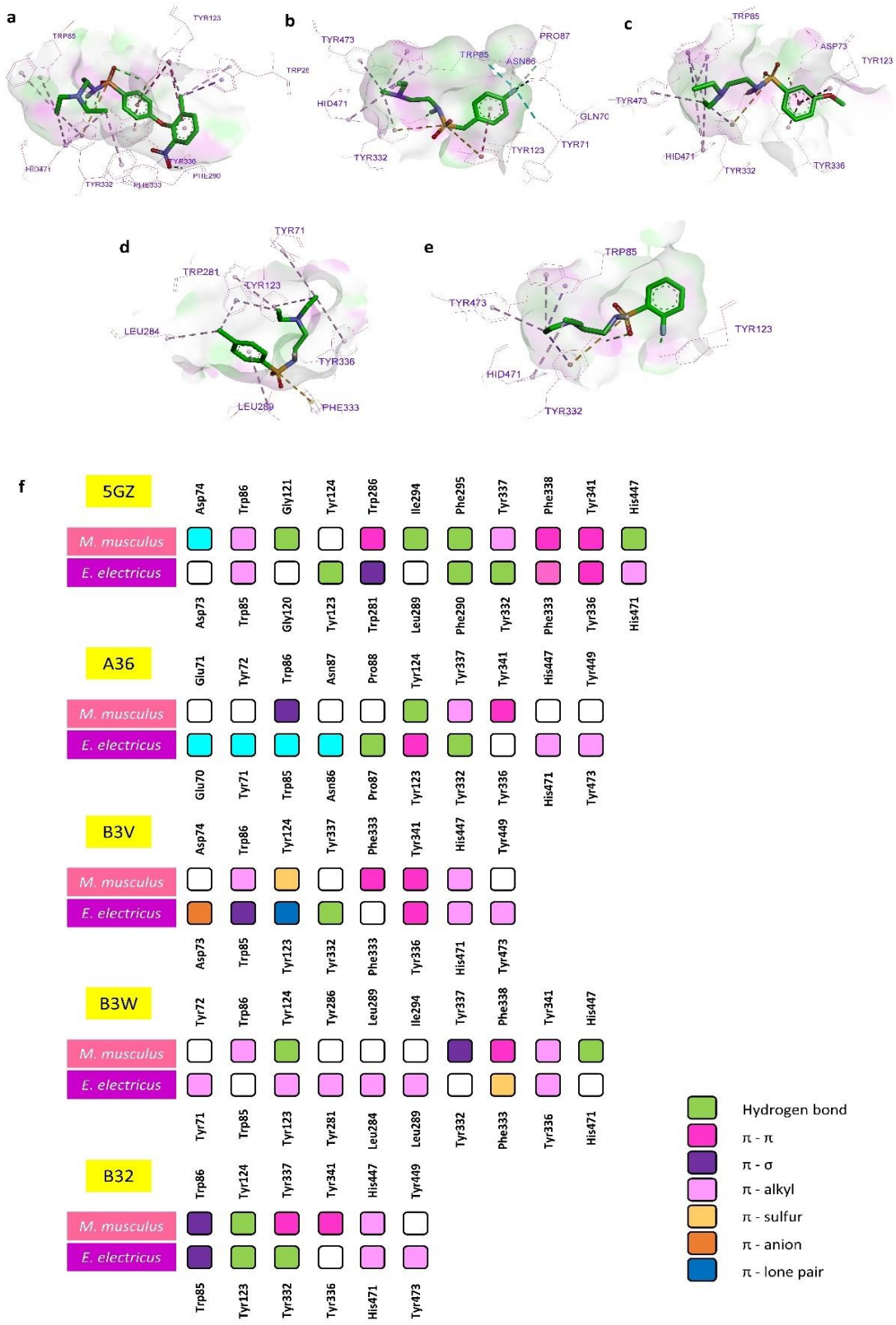

Figure 8.5 (a, b, c, d, e) 3D interaction diagrams of 5GZ, A36, B3V, B3W and B32 with eeAChE, (f) Comparison of the interaction profile of 5GZ, A36, B3V, B3W and B32 with mmAChE and eeAChE.

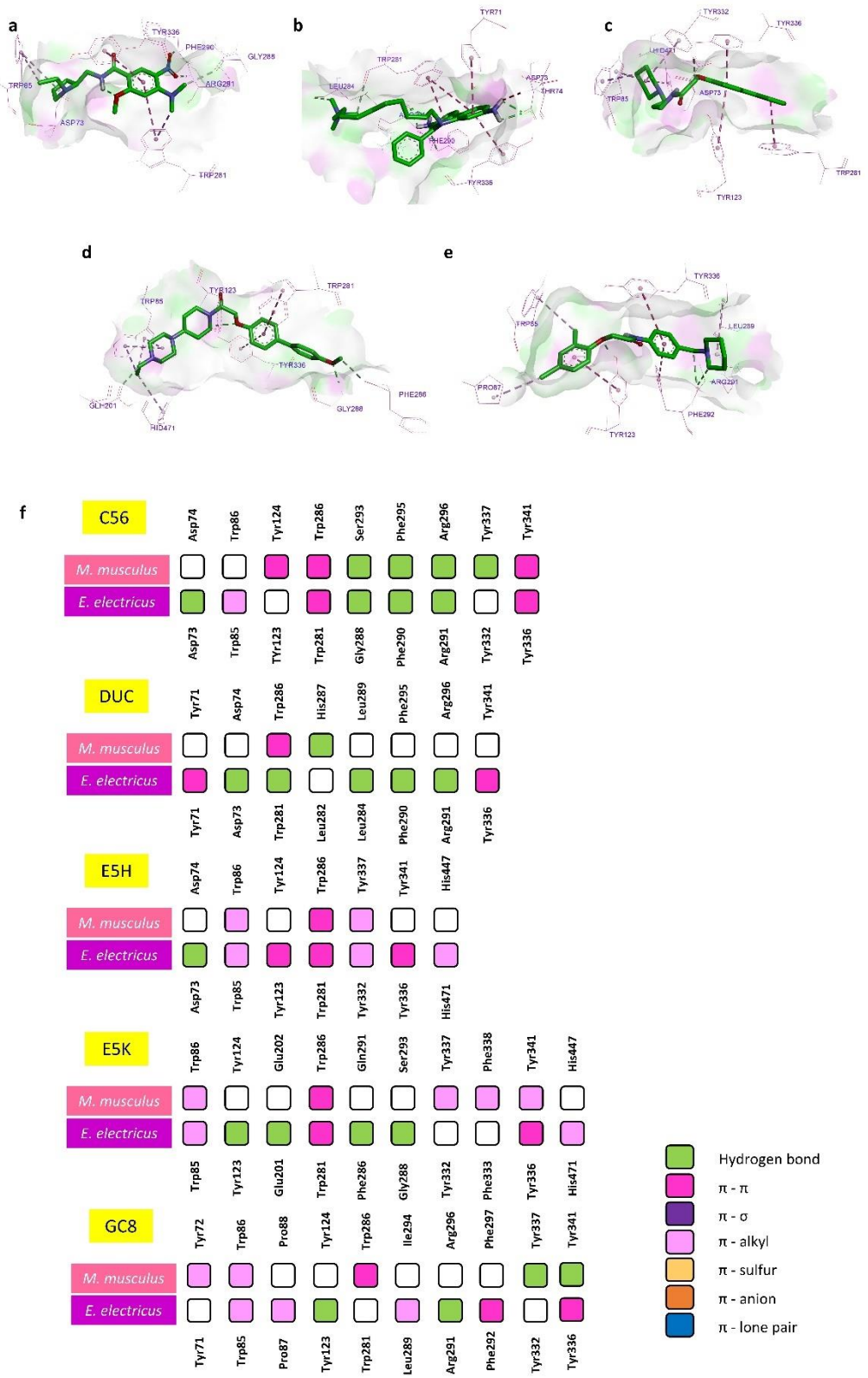

Figure 8.6 (a, b, c, d, e) 3D interaction diagrams of C56, DUC, E5H, E5K and GC8 with eeAChE, (f) Comparison of the interaction profile of C56, DUC, E5H, E5K and GC8 with mmAChE and eeAChE.

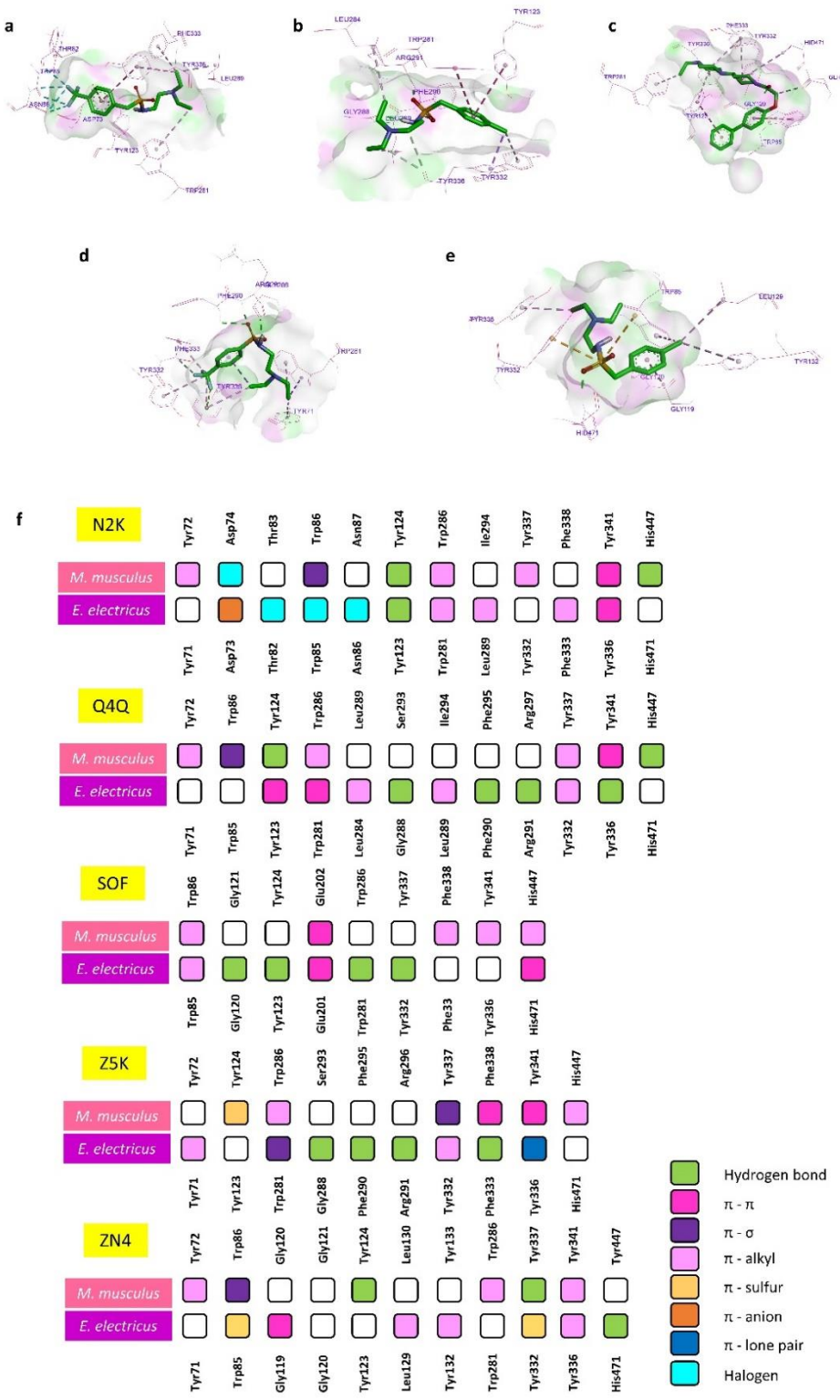

Figure 8.7 (a, b, c, d, e) 3D interaction diagrams of N2K, Q4Q, SOF, Z5K and ZN4 with eeAChE, (f) Comparison of the interaction profile of N2K, Q4Q, SOF, Z5K and ZN4 with mmAChE and eeAChE.

Overall, it was observed that the ligands interacted with all the major residues upon docking with eeAChE similar to mmAChE. Thus, it indicated that the ligand-bound in a

similar manner in both enzymes and validated the docking protocol. The quantitative aspect of the validation process was performed by generating a dataset of eeAChE inhibitors from the Bindingdb database (http://bindingdb.org/) [270]. Fifteen ligands, co-crystallised with the enzyme, were used for the identification of crucial fingerprints responsible for AChE inhibition (**Figure S14 in appendix**). The fingerprint search was performed with the help of RDkit, which resulted in the identification of CCN(C)C, a SMILES string, present in all mmAChE inhibitors and ACh. The *N,N-dimethylethanamine* group was used as a substructure filter to obtain a dataset of 1507 compounds from a total of 4,846 eeAChE inhibitors. Further, it was ensured that the selected compounds should have less than 16 rotatable bonds as desired for Autodock-4.2.6. The SMILES strings were converted to the tripos mol format using pybel module of open babel and stripped off the salt molecules. The obtained set was then converted to pdbqt format followed by docking on eeAChE using Autodock-4.2.6 through an in-house build python script with multiprocessing capability to speed up the process. The best poses for the ligands with the lowest binding energies were obtained from output files using a python script, vstools_v0.16. The obtained binding energies were used for the validation process along with the reported IC_{50} values. The cut-off values of 100, 1000, 2000 and 10000 nM were selected for assigning the ligands with binary classification labels, i.e., active and inactive. ROC curves were plotted for all the selected cut-off values. ROC is a plot between the true positive and false positive rates (TPR and FPR) obtained from a classification model.

$$TPR\ /\ Recall = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where TP: true positive, FP: False positive, TN: True negative and FN: False Negative

The area under curve (AUC) represents the degree of the separability of various classes of a classification model. The value of the AUC ranges from 0 to 1, the higher the AUC more accurate is the prediction. The value of the cut-off IC_{50}(s) and ROC-AUC are presented in **Figure 8.8**. The results indicated that the cut-off value of 10000 nM has the highest AUC value of 0.6355 but had an imbalanced class. Further, the regression-based analysis was carried out after the transformation of the IC_{50} values due to their wide range of distribution from 0.12 to 1233105 nM. The scatter plot of binding energy vs pIC_{50} displayed a poor coefficient of determination of 0.007. Similarly, the plot between predicted pki vs pIC_{50} has also shown the coefficient of determination to be 0.007 (**Figure 8.8**). However, any mathematical model with a good predictivity should have this value as high as possible, and for the best fit model, it tends to be 1.

8.3.7. Development and validation of the scoring functions for eeAChE

In the present study, various machine and deep learning models were developed and validated using three strategies, i.e., binary classification, multiclass classification and regression.

8.3.7.1. Dataset preparation

The poses, obtained after docking of 1507 compounds, were used to calculate 102 molecular descriptors using RDkit. Protein-ligand interaction profiler (https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index) was used to identify the interactions between protein and ligand along with the number of different types of interactions. Finally, the molecular descriptors, interaction profiles and binding energies were combined to form the final dataset of 677 features.

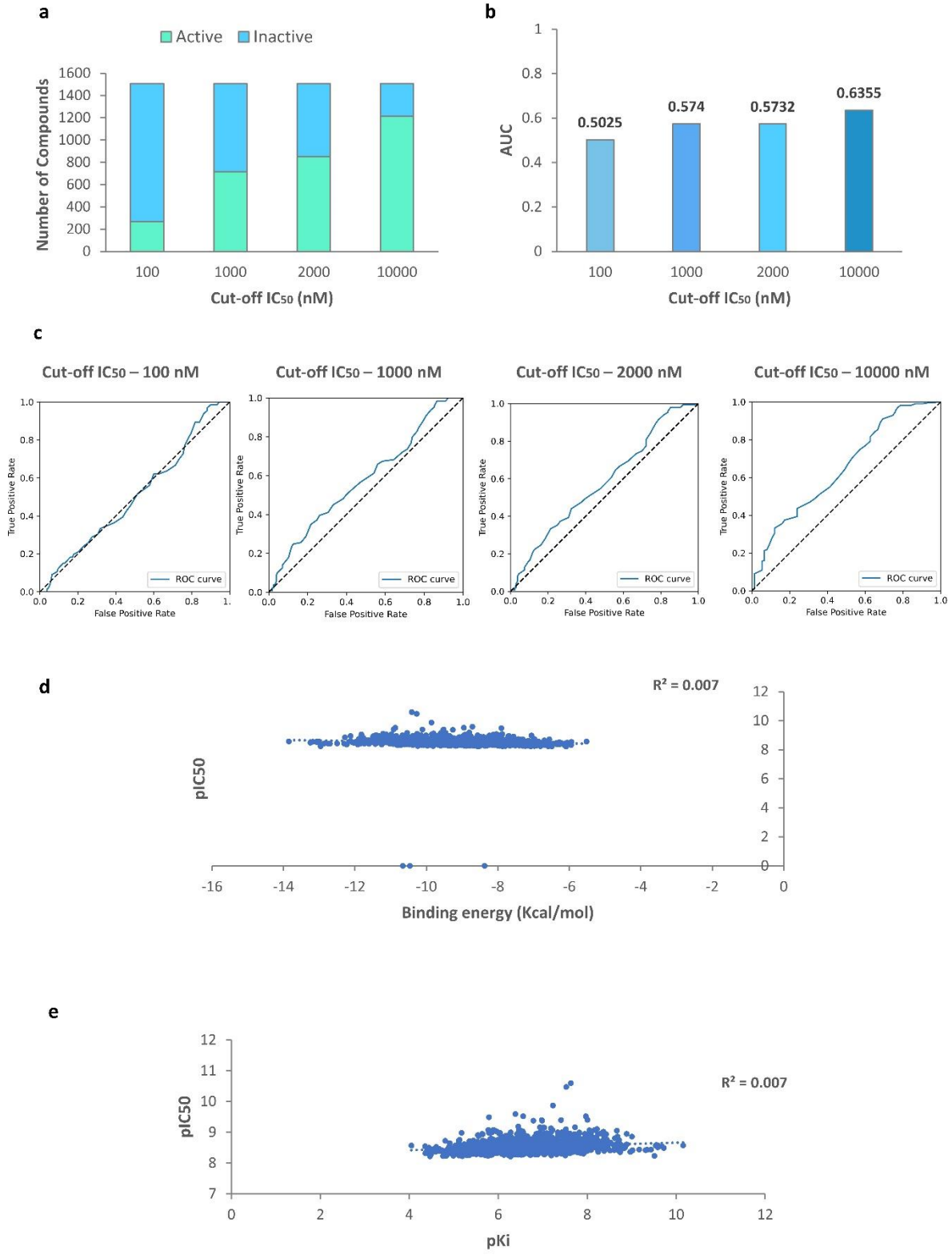

Figure 8.8 (a, b, c) IC_{50} cut-off, area under the curve and receiver operating characteristic of the docking validation set for validation of the Autodock scoring function using binary classification. (d) Scatter plot showing the relationship between binding energy and pIC_{50}. (e) Scatter plot showing the relationship between pK_i and pIC_{50}.

8.3.7.2. Development of scoring function based on binary classification models

IC_{50} of the compounds i.e., independent feature, was converted to two classes, active (IC_{50} < 1000 nM) and inactive, with a cut-off value of 1000 nM for the development of binary models **(Figure 8.9)**. Seventeen machine learning algorithms were employed for model development using various combinations of hyperparameters and five-fold validation. Logistic regression, SVM and KNN were trained on the scaled dataset. The standardisation and normalisation techniques were employed.

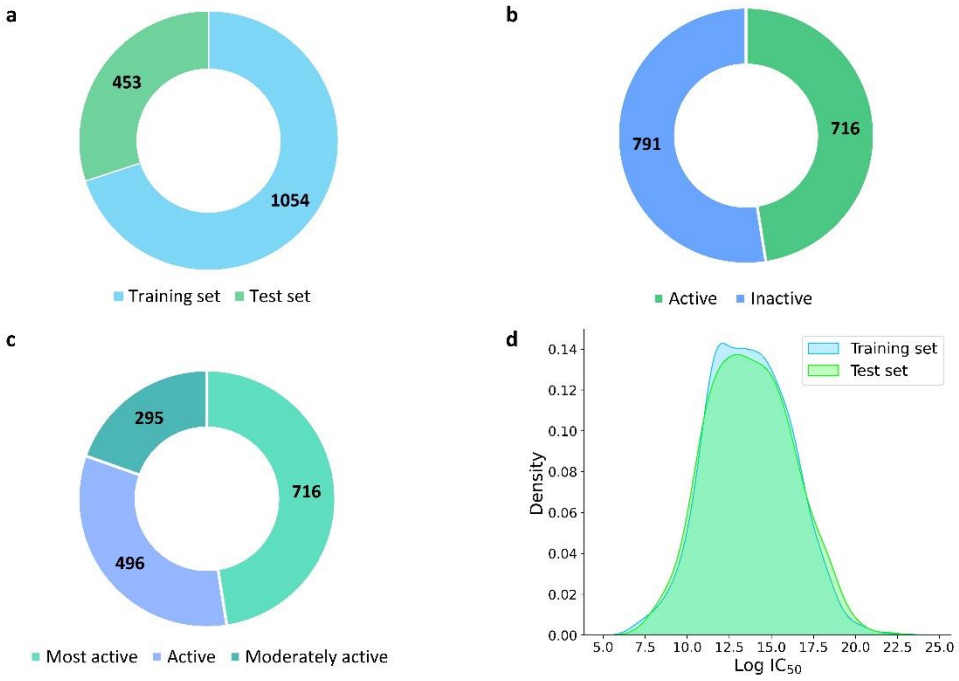

Figure 8.9 (a) Proportion of compounds used in training and test sets for the ML, (b) Compounds labelled as active and inactive at IC_{50} cut-off value of 1000 nM for binary classification, (c) Compounds labelled as most active, active and moderately active at IC_{50} cut-off values of 1000 and 10000 nM for multi-class classification, (d) Distribution of the compounds in training and test sets on a log IC_{50} scale for regression modelling.

The LR yielded similar accuracy on training and test sets obtained from both data pre-processing techniques. However, SVM and KNN showed better results with a normalised dataset on both training and test sets. The probability-based Naïve Bayesian classifiers have poor results with the lowest accuracies among all the algorithms. The multinomial and complement based Naïve Bayesian classifiers were unable to fit on the dataset.

Further, a decision tree has a flow chart based architecture that has branches and leaves representing the outcome of the test and a class label, respectively, while the RF is a cluster of various decision trees [271]. It was observed that both methods performed equally well on the training set. Still, the test set accuracy of the decision tree was higher than the random forest during binary classification. The opted techniques provided better accuracy than the native Autodock-4.2.6 SF but the accuracy above 80% is greatly appreciated for any ML model. Ensembled methods are essential techniques that improve predictivity by combining various models together. The ensemble models available with scikit-learn library are ada-boost, bagging, extra trees, gradient boosting and histgradient boosting classifiers. It was indicated that except the ada-boost classifier, all other classifiers improved the prediction accuracy on both training and test sets compared with the previous algorithms. Hence, the most appropriate ensembled classifier was identified through a comparison of precision, recall and F1 scores. Precision is usually defined as the number of TPs identified by an ML model from all the predicted positives, while recall identifies the fraction of true positives identified from all actual positives. As the precision of the model improves, the value of recall typically falls and vice versa. Usually, the selection of a model depends upon the trade-off between precision and recall by identifying a sweet spot where both values are high. Thus, besides the two parameters, F1 score, which is a harmonic average of precision and recall, plays an important role in the model selection.

The precision and recall scores indicated that the bagging classifier has better predictivity for identifying TPs, i.e., the active compounds, which is the primary function of any SF. Further, the precision also indicated that the classifier had reduced the number of FPs, which was crucial for the performance. It was clearly indicated that the bagging classifier outperformed other classifiers with an F1 score of 83.7 % (**Table 8.3**).

Table 8.3 Validation scores of various machine learning algorithms used to develop binary scoring function.

			Training dataset (k = 5)	Test dataset			
Algorithm	**Parameters**	**Feature scaling**	**Accuracy (%) ***	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**
Logistic Regression	C = 500, max iter = 10000, penalty = l1, solver = liblinear	Standardisation	74.19 ± 2.86	75.27	72.39	75.82	75.27
Logistic Regression	C = 10, max iter = 10000, penalty = l1, solver = saga	Normalisation	74 ± 2.84	75.05	74.48	69.85	75.05
Support Vector Classifier	C = 1, gamma = 1, kernel = poly	Standardisation	78.08 ± 1.96	71.96	67.82	74.64	71.07
Support Vector Classifier	C = 10, gamma = scale, kernel = rbf	Normalisation	75.5 ± 2.37	75.71	74.87	71.29	73.03
K-nearest neighbour	metric = manhattan, n neighbors = 3	Standardisation	81.21 ± 1.62	76.82	73.63	77.51	75.52
K-nearest neighbour	metric = manhattan, weights = distance, n neighbors = 5	Normalisation	78.93 ± 1.39	80.57	76.88	82.77	79.72
Naïve Bayes	Guassian	-	48.2 ± 3.6	47.9	47.41	99.06	64.13
Naïve Bayes	Bernoulli	-	65.3 ± 1.6	63.57	61.32	61.03	61.17
Random Forest	max depth = 15, max leaf nodes = 10, min samples leaf = 2, n estimators = 100	-	75.23 ± 1.34	72.62	71.29	69.95	70.61
Decision tree	criterion = gini, max depth = 8, min samples leaf = 2, min samples split = 8	-	75.61 ± 2.29	77.92	75.56	78.4	76.95
AdaBoost Classifier	algorithm = SAMME.R, learning rate = 0.1, n estimators = 1000	-	77.04 ± 3.23	79.02	74.1	81.77	79.02
Bagging Classifier	**max features = 500, n estimators = 5000**	**-**	**81.68 ± 1.73**	**83.66**	**81.81**	**85.52**	**83.26**
Extra Trees Classifier	criterion = gini, max features = sqrt, n estimators = 1000	-	81.96 ± 4.05	81.23	81.42	78.8	81.23
Gradient Boosting Classifier	criterion = mse, learning rate = 1, loss = exponential, max features = auto, n estimators = 10000	-	83.1 ± 2.69	81.67	80.73	81.1	81.67
Hist Gradient Boosting Classifier	learning rate = 0.01, max iter = 100		81.2 ± 2.61	81.89	82.71	79.72	81.89

* Expressed as Mean ± SD

8.3.7.3. Development of scoring function based on multiclass classification models

A multiclass based SF was also developed by using IC_{50} cut-off value of 1000 and 10000 nM, that led to the division of dataset into three classes, i.e., most active (IC_{50} < 1000 nM), active (IC_{50} > = 1000 and IC_{50} <10000 nM) and moderately active (IC_{50} >= 10000 nM) **(Figure 8.9)**. Various classification algorithms were tested to obtain the best model. A wide range and combinations of hyperparameters were tested through the grid search method along with five-fold validation. LR, linear SVC, KNN and ridge classifier were tested on both standardised and normalised data. It was observed that all the four algorithms performed poorly with the accuracy of less than 70 % on training as well as test sets, except KNN which showed training accuracies of 70.64 % and 71.08 % on standardised and normalised datasets, respectively (**Table 8.4**). The probability-based classifiers, i.e., Gaussian and Bernoulli Naïve Bayesian poorly performed. Further, the decision tree and extra tree classifiers had also performed poorly, when compared to their binary classification performance. The ensemble algorithms again outperformed others and have better accuracies among all. Generally, the bagging classifier performs data fitting on smaller sets of training dataset and then take the average of predictions made by all individual models to reach a final decision [272]. In contrast, extra tree classifier generates decision trees with all data available in the training set, followed by creating subsets of the features randomly to generate models [273]. Both algorithms performed well on the training and test datasets, with almost equal accuracies. In multiclass problems, we used micro-averaging that led to an equal value of precision, recall and f1-scores. Micro-averaging did not distinguish between different classes, but averaged their metric scores and hence performed well on unequally distributed classes, as in the case. The bagging classifier performed slightly better with AUC for ROC of 87.0% compared to extra tree classifier (**Figure 8.10 (b)**).

Table 8.4 Validation scores of various machine learning algorithms used to develop multiclass scoring function.

			Training dataset (k = 5)	Test dataset			
Algorithm	**Parameters**	**Feature scaling**	**Accuracy***	**Accuracy**	**Precision**	**Recall**	**F1 score**
Logistic Regression	C = 1000, max iter = 10000, multi class = multinomial, penalty = l2, solver = newton-cg	Standardisation	58.92 ± 1.98	59.6	59.6	59.6	59.6
Logistic Regression	C = 10, max iter = 10000, multi class = multinomial, penalty = l1, solver = saga	Normalisation	57.02 ± 2.47	62.69	62.69	62.69	62.69
Linear SVC	C = 1, class weight = balanced, loss = squared hinge, multi class = ovr, penalty = l2	Standardisation	57.49 ± 1.58	60.92	60.92	60.92	60.92
Linear SVC	C = 10, class weight = balanced, loss = hinge, multi class = ovr, penalty = l2	normalisation	57.87 ± 1.94	62.47	62.47	62.47	62.47
K-Neighbors Classifier	algorithm = auto, metric = manhattan, n neighbors = 3, weights = distance	Standardisation	65.08 ± 2.65	70.64	70.64	70.64	70.64
K-Neighbors Classifier	algorithm = auto, metric = manhattan, n neighbors = 5, weights = distance	Normalisation	64.32 ± 2.78	71.08	71.08	71.08	71.08
Ridge Classifier	class weight = balanced, solver = auto	Standardisation	57.21 ± 1.85	60.92	60.92	60.92	60.92
Ridge Classifier	class weight = balanced, solver = auto	Normalisation	56.73 ± 2.66	60.48	60.48	60.48	60.48
Gaussian NB			46.9 ± 2.9	48.56	48.56	48.56	48.56
Bernoulli NB			54.3 ± 1.8	49.66	49.66	49.66	49.66
Decision Tree Classifier	criterion = gini, max depth = 8, min samples leaf = 1, min samples split = 4		63.47 ± 2.52	62.91	62.91	62.91	62.91
AdaBoost Classifier	algorithm = SAMME, learning rate = 1, n estimators = 1000	-	78.17 ± 2.19	78.8	78.8	78.8	78.8
Bagging Classifier	**max features = 500, n estimators = 10000**		**82.92 ± 3.05**	**83**	**83**	**83**	**83**
Extra Trees Classifier	criterion = entropy, max features = sqrt, n estimators = 1000		82.92 ± 1.55	82.27	82.27	82.27	82.27
Gradient Boosting Classifier	criterion = mse, learning rate = 1, loss = exponential, max features = sqrt, n estimators = 1000		80.93 ± 2.64	82.56	82.56	82.56	82.56
HistGradient Boosting Classifier	classifier learning rate = 0.01, classifier max iter = 1000		67.65 ± 2.13	61.45	61.45	61.45	61.45

* Expressed as Mean ± SD

8.3.7.4. Development of scoring functions based on regression-based models

The quantitative output would provide much better information about the order of the predicted activity, which is crucial for any virtual screening. Hence, a wide range of regression-based models were developed in order to ascertain the activity in terms of the predicted IC_{50} using various ML and deep learning algorithms. The IC_{50} was the dependent feature and was processed on a logarithmic scale **(Figure 8.9 (d))**. It was observed that all the selected algorithms did not show any difference in model fitting on normalised and standardised data. The r^2 values were better for random forest regressor and neural networks. At the same time, all other algorithms displayed an underfitting, which was also indicated from the mean absolute and mean squared error values of training and test datasets. The Q^2_{F1} and Q^2_{F2} values for every algorithm were almost equal, indicating that the mean of training set lies in closeness with the test set mean and the test set covers the complete response domain of the model. Both neural networks and RF regressor have these values higher than 0.5, which was the acceptable threshold. The criteria proposed by Golbraikh and Tropsha assisted in determining the external predictiveness of the model [274]. It accounted for the observed (Y_{obs}), predicted (Y_{pred}) activities, and squared correlation coefficients (r^2_0 or $r`^2_0$).

$$k = \frac{\Sigma Y_{obs} * Y_{pred}}{\Sigma Y^2_{pred}}$$

$$k` = \frac{\Sigma Y_{obs} * Y_{pred}}{\Sigma Y^2_{obs}}$$

$$r^2_0 = 1 - \frac{\Sigma (Y_{obs} - k * Y_{pred})^2}{\Sigma (Y_{obs} - \bar{Y}_{obs})^2}$$

$$r`^2_0 = 1 - \frac{\Sigma (Y_{obs} - k` * Y_{pred})^2}{\Sigma (Y_{pred} - \bar{Y}_{pred})^2}$$

According to these criteria, models are considered satisfactory if all of the following conditions are satisfied:

i. $Q^2_{training} > 0.5$

ii. $R^2_{test} > 0.6$

iii. $1 - r_0^2 / r^2 < 0.1 \; and \; 0.85 \leq k \leq 1.15 \; or \; 1 - r`^2_0 / r^2 < 0.1 \; and \; 0.85 \leq k` \leq 1.15$

iv. $|r_0^2 - r`^2_0| < 0.3$

It was observed that only the neural network and RF satisfied the criteria. These criteria are especially important for non-linear methods such as neural networks as in such cases, r^2 is always high and tend to be close to 1, but only fewer of such models perform well on the external test sets. The models showing better prediction on both training and test sets, could only pass the above criteria. Finally, the models were also tested for MAE criteria which states that:

i. Good prediction: MAE ≤ 0.1 X training set range AND MAE ± 3 X σ ≤ 0.2 X training set range.
ii. Bad prediction: MAE > 0.15 X training set range OR MAE ± 3 X σ > 0.25 X training set range

the σ value denotes the standard deviation of the absolute error values for the test dataset. If both of the criteria (i) and (ii) are not satisfied, then the model is considered of moderate quality. However, if the test set is large enough, then it is allowed to drop 5% of samples with high absolute errors, as such data points are outliers and heavily penalise the predictivity of the model. The RF regression-based models were only satisfying the MAE criteria on 95 % of test dataset and were selected. Although, the performance of RF regression algorithm was almost similar on both standardised and normalised data with a slightly better performance on standardised data and hence this model was selected for the development of an SF (**Table 8.5**).

Table 8.5 Validation scores of various machine learning algorithms used to develop regression-based scoring function.

Model	Feature scaling	Coefficient of determination (train)	MAE Test	MAE Train	MAE criteria	MAE criteria (95%)	MSE Test	MSE Train	Q^2_{F1}	Q^2_{F2}	TROPSA criteria
Linear regression	normalisation	0.544	2439.997	1.286	Bad	Bad	2693373901	2.793	-420204120	-420426134	Fail
Linear regression	standardisation	0.543	65669877079	1.285	Bad	Bad	8.24E+23	2.797	-1.29E+23	-1.29E+23	Fail
Ridge regression	normalisation	0.439	1.657	1.425	Bad	Bad	4.421	3.434	0.31	0.31	Fail
Ridge regression	standardisation	0.492	1.668	1.36	Bad	Bad	4.587	3.11	0.284	0.284	Fail
Lasso regression	normalisation	0	2.077	2.016	Bad	Bad	6.41	6.124	0	-0.001	Fail
Lasso regression	standardisation	0	2.077	2.016	Bad	Bad	6.41	6.124	0	-0.001	Fail
Elasticnet regression	normalisation	0.048	2.02	1.97	Bad	Bad	6.32	6.01	0.02	0.026	Fail
Elasticnet regression	standardisation	0.048	2.021	1.962	Bad	Bad	6.094	5.83	0.049	0.049	Fail
SVR regression	normalisation	0.406	1.645	1.42	Bad	Bad	4.365	3.641	0.319	0.319	Fail
SVR regression	standardisation	0.554	1.566	1.143	Bad	Bad	4.047	2.731	0.369	0.368	Fail
Bayesian Ridge regression	normalisation	0.408	1.669	1.468	Bad	Bad	4.429	3.626	0.309	0.309	Fail
Bayesian Ridge regression	standardisation	0.403	1.68	1.475	Bad	Bad	4.474	3.658	0.302	0.302	Fail
RandomForest regression	normalisation	0.94	1.157	0.443	Bad	Good	2.38	0.362	0.629	0.629	Pass
RandomForest regression	**standardisation**	**0.94**	**1.148**	**0.444**	**Bad**	**Good**	**2.342**	**0.367**	**0.635**	**0.634**	**Pass**
SGD regression	normalisation	0.373	1.689	1.522	Bad	Bad	4.474	3.84	0.302	0.302	Fail
SGD regression	standardisation	0.407	1.791	1.487	Bad	Bad	6.145	3.63	0.041	0.041	Fail
Neural network *	normalisation	0.992	1.431	0.17	Bad	Bad	3.357	0.048	0.39300001	0.389	Fail
Neural network **	standardisation	0.977	1.357	0.309	Bad	Moderate	3.14	0.146	0.491	0.491	Fail

* Batch size = 32, loss = Mean squared error, Activation function = relu, Optimiser = Adamax and number of hidden layers = 3

** Batch size = 16, loss = Huber loss, Activation function = relu, Optimiser = Adamax and number of hidden layers =

8.3.8. Improved scoring function

When the AUC of ROC of the Autodock SF were compared with selected binary and multiclass models, a significant improvement in model predictivity was observed **(Figure. 8.10)**. Further, the selected regression model also displayed improved results when compared to the native SF. Hence, the selected models were compiled and converted to a python-based package, Protein-Ligand Scoring Function (PLSF), which could be obtained through the website (http://:www.drugdesign.in/tools).

8.3.9. Applicability domain

The applicability domain is another important aspect of any prediction process, which defines a chemical space used for building ML model. It is generally described in terms of chemical descriptors and features for which the prediction results could be reliable. Any compound that falls out of this space would have an unreliable prediction result. In this study, the applicability domain is defined with two conditions, i.e., presence of a *N,N-dimethylethanamine* group and number of the rotatable bonds less than sixteen.

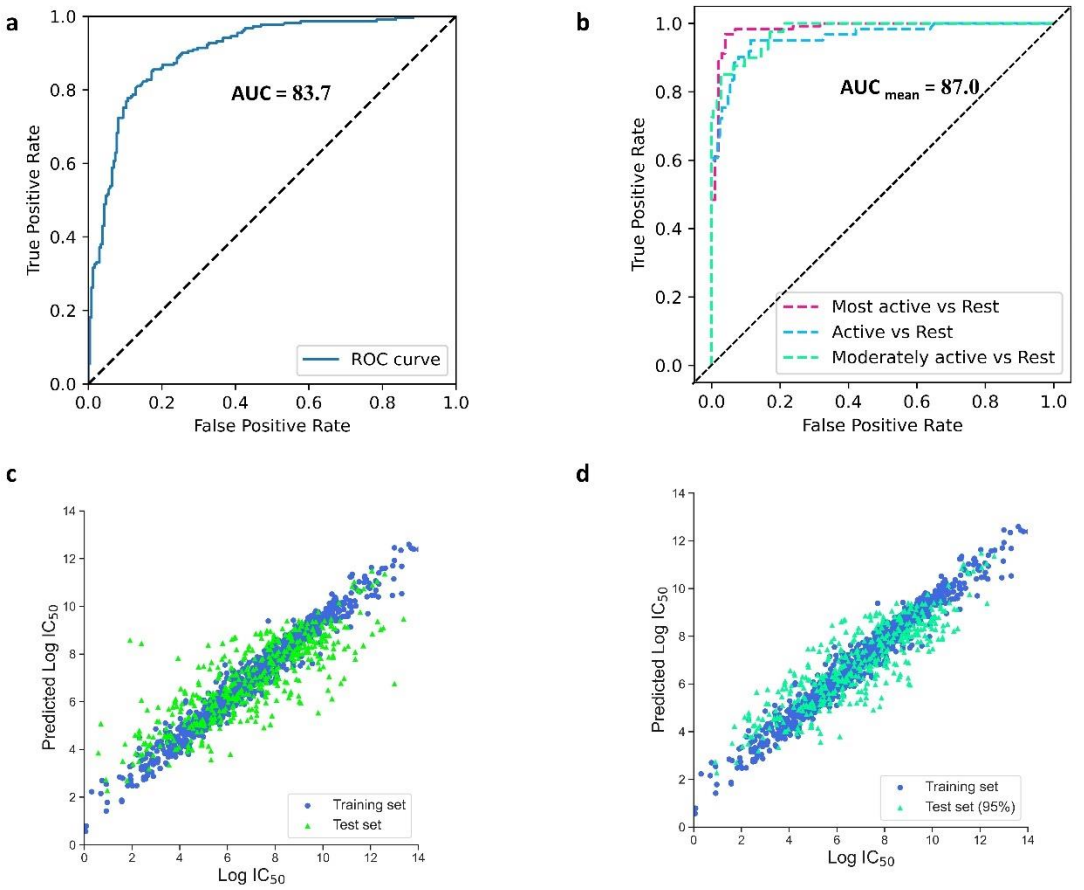

Figure 8.10 (a, b) Receiver operating characteristic of binary and multiclass scoring function, respectively for the bagging classifier. (c) Plot between log IC_{50} and predicted $LogIC_{50}$ of the training and test sets for RF regressor. (d) Plot between log IC_{50} and predicted Log IC_{50} of the training and and 95% of the test set after removing outliers for RF regressor.