28

Identification of lead inhibitors of TMPRSS2 isoform 1 of SARS-CoV-2 target using neural network, random forest, and molecular docking

Alakanse Suleiman Oluwaseun[3], Joel Ireoluwa Yinka[1], George Oche Ambrose[1], Adigun Temidayo Olamide[1], Sulaiman Faoziyat Adenike[1], Ohanaka Judith Nkechinyere[2], Idris Mukhtar[3], Yekeen Abeeb Abiodun[3], Olarewaju Ayodeji Durojaye[3]

[1]*DEPARTMENT OF BIOCHEMISTRY, FACULTY OF LIFE SCIENCES, UNIVERSITY OF ILORIN, ILORIN, KWARA, NIGERIA;* [2]*DEPARTMENT OF BIOCHEMISTRY, FACULTY OF NATURAL AND APPLIED SCIENCES, NILE UNIVERSITY OF NIGERIA, ABUJA, NIGERIA;* [3]*SCHOOL OF LIFE SCIENCES, UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA, HEFEI, ANHUI, PEOPLE'S REPUBLIC OF CHINA*

1. Introduction

Sever acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) have been linked to SARS coronavirus (SARS-CoV) and MERS-coronavirus (MERS-CoV), respectively [1,2]. As of September 1, 2021, a total of 218,946,836 laboratory-confirmed cases of SARS infection by SARS-CoV-2 was recorded; of which, 4,539,724 were recorded to have resulted in death [3].

SARS-CoV-2, a novel coronavirus detected late 2019 and closely related to SARS-CoV, has been etiologically implicated in the new pulmonary disease [4–6]. The viral endocytotic activities of coronaviruses and specifically those of SARS-CoV-2 in the targeted host cells are aided by the S1 unit of the spike protein (S), thus enhancing viral attachment and entry into the host cells via angiotensin-converting enzyme 2 (ACE2), which can be blocked via transmembrane protease serine 2 (TMPRSS2) inhibitors [7].

In addition, the priming of S protein for SARS-CoV-2 viral entry is also aided by TMPRSS2 [8–10] and endosomal cysteine proteases cathepsin B and L (CatB/L).

Data Science for COVID-19. https://doi.org/10.1016/B978-0-323-90769-9.00021-9

Nevertheless, the activity of CatB/L is negligible compared to the activity of TMPRSS2, which is necessary for viral pathogenesis and spread in the host cell [11–13]. The isoform 1 of TMPRSS2 has been shown to activate SARS-CoV spike protein for independent entry into the host cells via cathepsin L [14,61].

Currently, no broad spectrum of antiviral drugs are available for the treatment of highly virulent respiratory viruses, which include MERS and SARS. In this study therefore, we developed two machine learning models using random forest classifier (RFC) and neural networks (NNs) based on 2251 inhibitors of serine proteases to screen a database of 21,000,000 virtual compounds. We screened the hit compounds using absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties and finally docked the filtered compounds into the active site of TMPRSS2 to determine their corresponding binding affinity and plausible molecular interactions.

2. Materials and methods

2.1 Software and hardware

All analyses were carried out using Linux Ubuntu 18.04 system running on a 12-GB RAM, core i5, 4 core, 2.5-GHz hp Pavilion dm4 laptop.

All python packages were run on python 3.6 using Jupyter Lab 1.2.6. Python packages used include Scikit-learn 0.22.2, Tensorflow 2.1.0, Feature Selector, Pandas 1.0.3, Matplotlib 3.2.1, Seaborn 0.10.0, Numpy 1.18.1, Skater 1.0.2, and TPOT (Tree-Based Pipeline Optimization Tool) 0.11.1.

2.2 Data extraction and descriptor calculation

A total of 2253 inhibitors Simplified Molecular Input Line Entry System (SMILES) of TMPRSS2 isoform 1 was downloaded from the Chemical European Molecular Biology Laboratory (CHEMBL) database with their corresponding Half-maximal inhibitory concentration (IC50) values. A python script using the pandas module was written to clean up the downloaded data (i.e., remove SMILES with missing IC50 values and duplicates and extract only SMILES and their corresponding IC50).

Molecular Operating Environment (MOE) [15] was used to calculate 2D descriptors for the SMILES. A total of 206 descriptors were calculated.

2.3 Exploratory data analysis

Exploratory data analysis was carried out using a python script (Pandas, Feature Selector, Matplotlib, and Seaborn). IC50 values were converted to the negative log of the IC50 value when converted to molar (pIC50) using this formula: 9-log10[IC50]. The pIC50 values were converted into categorical values of active (1) and nonactive (0), and the activity threshold was set at 8.0 (active $\geq$ 8.0, nonactive $<$ 8.0). Correlation within descriptors was carried out using the python Feature Selector module, and the correlation threshold was set at 0.75. All intercorrelated descriptors were removed.

2.4 Tree-based pipeline optimization tool analysis

A python script implementing TPOT analysis [16] was carried out to investigate the best machine learning algorithms and their corresponding hyperparameters for our modeling task. TPOT classifier class was used, early stopping parameter was set to true, verbose was set to 4, and cross-validation with shuffling was set to 10-fold. All other parameters were left at default values.

2.5 Model building

In order to reduce the number of training features, we used a python script implementing the Feature Selector module. The module uses the XGBoost algorithm [17] to identify zero- and low-importance features. The cumulative importance value was set at 0.95, the task value was set to the classification, and the evaluation metric used was l2. All other parameters were left at default. The number of features was reduced to 43. The 43 descriptors were normalized using Scikit-Learn RobustScaler [18].

The RFC was the best according to our TPOT analysis, hence an RFC algorithm was used to build a model based on our 43 descriptors using Scikit-Learn RFC class [18]. The following hyperparameters were used: bootstrap: False, criterion: entropy, max_features: 0.2, min_samples_leaf: 1, min_samples_split: 15, n_estimators: 100, and class weight was set at 0: 0.60, 1: 0.4. All other parameters were left at default.

A deep NN model was built using a python TensorFlow module [19] in order to find the best values for NN hyperparameters, namely, epoch, batch size, and optimization function. We wrote a python script implementing Scikit-learn Keras wrapper and Grid Search CV class [18]. Cross-validation was set to 10 verbose 4, an NN with three layers of 100 units each, ReLU as the activation function.

After finding the best values for the abovementioned hyperparameters the NN for our modeling task was constructed. The NN architecture had the following parameters: input dimension: 44, three dense layers with 100 neurons (units) each, an output dense layer with 2 units (2 units to accommodate for both active(1) and non-active(0)), each of the dense layers used ReLU activation function and an l2(0.01) regularizer, and the output layer used sigmoid activation function. Finally, the network was compiled using binary cross-entropy loss function, Adam optimizer, categorical accuracy as metric evaluation, early stopping and model check point as the callback, epoch: 20, batch size: 5, verbose: 2, and class weight: 0: 60, 1: 40. All other parameters were left at default.

2.6 Classification metric evaluation

The classification models were evaluated using Scikit learn metric module [18] to calculate various classification evaluation metrics as described by Abadi et al. [19].

2.7 Database screening

SCUBIDOO a database [20] with 21,000,000 compounds was screened against both models. SCUBIDOO has sample representations of the database: 9994 (S); 99,977 (M); and 999,794 (L) compounds. We used the M sample to screen against our models, and hits were filtered using two criteria: synthesis probability (≥ 0.9) and drug-likeness ($= 1$). This criterion was calculated using MOE software [15], while the filtration was executed using a python script. The hit compounds from this filtration process were further subjected to ADMET property verification and further reduced.

2.8 Sequence retrieval, alignment, and homology modeling of transmembrane protease serine 2 isoform 1 target protein

The crystal structure of TMPRSS2 of either isoform 1 or 2 was not available in the PDB database; thus the homology model was generated for this study. The protein FASTA query sequence of TMPRSS2 isoform 1 was retrieved from the NCBI database with the accession number NP_001128571.1. The template 5ce1.1.A was selected for the homology modeling using the SWISS-MODEL webserver [21].

2.9 Optimization and refinement of transmembrane protease serine 2 isoform 1 modeled protein

The generated homology model of TMPRSS2 isoform 1 was uploaded on the 3Drefine webserver [22]; this makes use of iterative optimization of hydrogen bonding network in addition to atomic-level energy minimization on the optimized protein model using a composite physics and knowledge-based force field for efficient protein structure refinement. The output of 3Drefine webserver was further refined using GalaxyRefine webserver [23]; this works by rebuilding side chains, repacking of side chain, and relaxing overall protein structure using molecular dynamic simulation (MDS) [23].

2.10 Validation and quality estimation of optimized transmembrane protease serine 2 isoform 1 modeled protein

The optimized TMPRSS2 isoform 1 model was validated using RAMPAGE and PROSESS [24]. Quality estimation of the modeled protein was carried out using SAVES server and ProSA-web [25]. Ramachandran plot quality estimation was done using WHAT IF webserver. The resolution of the optimized modeled protein was calculated using ResProx server [26] and visualized using Discovery Studio 3.0 [27].

2.11 Physiochemical characterization of transmembrane protease serine 2 isoform 1

Using the ProtParam webserver, the individual percentage of amino acid residues, molecular weight, theoretical pI, atomic composition, extinction coefficient, and instability index of the model protein were determined [28].

2.12 Binding site prediction of transmembrane protease serine 2 isoform 1 modeled protein

The recognition of protein-ligand binding sites is of great importance in drug discovery. The binding site of the proposed lead compound(s) was predicted using P2Rank, which is a template-free machine learning algorithm embedded in PrankWeb [29,30]. This works based on local chemical neighborhood ligandability unified on junctures placed on a solvent-accessible protein surface. Junctures or points with increased ligandability score are then clustered to form the resulting ligand binding sites [29].

2.13 Submission of the model in protein model database

The model generated for TMPRSS2 isoform 1 was successfully submitted in the Protein Model Database (PMDB) having PMDB ID: PM0083140.

2.14 Molecular docking using AutoDock Vina

For our analysis, we used the PyRx, AutoDock Vina exhaustiveness search docking function. After the minimization process, the grid box resolution of TMPRSS2 homology model protein was centered at $1.1075 \times -1.3338 \times 15.7311$ along the x, y, and z center axes, respectively, at a grid dimension of $70.919 \times 58.432 \times 58.519$ Å to define the binding site of the protein. Camostat was first docked within the binding site of TMPRSS2 and the resulting interaction was compared with those of the hits into the same active sites using the same grid box dimension.

3. Result and discussion

3.1 Model building and database screening

The use of machine learning algorithms to virtually screen databases of compounds with unknown biological activity for hit compounds has become an established protocol in drug discovery [31]. Machine learning algorithms can be classified into various categories, including logic-based algorithms (decision trees), statistical algorithms (support vector machine, Bayesian network, K-nearest neighbor), and perceptron-based algorithms (NNs) [32], but for the sake of this study, we classified them into two: NNs and non-NNs (logic-based and statistical algorithms).

In this study, we built models based on both types of machine learning algorithms and screened for hit compounds.

3.2 Data preprocessing

A total of 2251 inhibitors of protease inhibitors were extracted from CHEMBL, with pIC50 ranging from 3 to 11 (Fig. 28.1). 2D descriptors (206) were calculated using MOE software, descriptors with a correlation greater than 0.75 were removed (Fig. 28.1), and

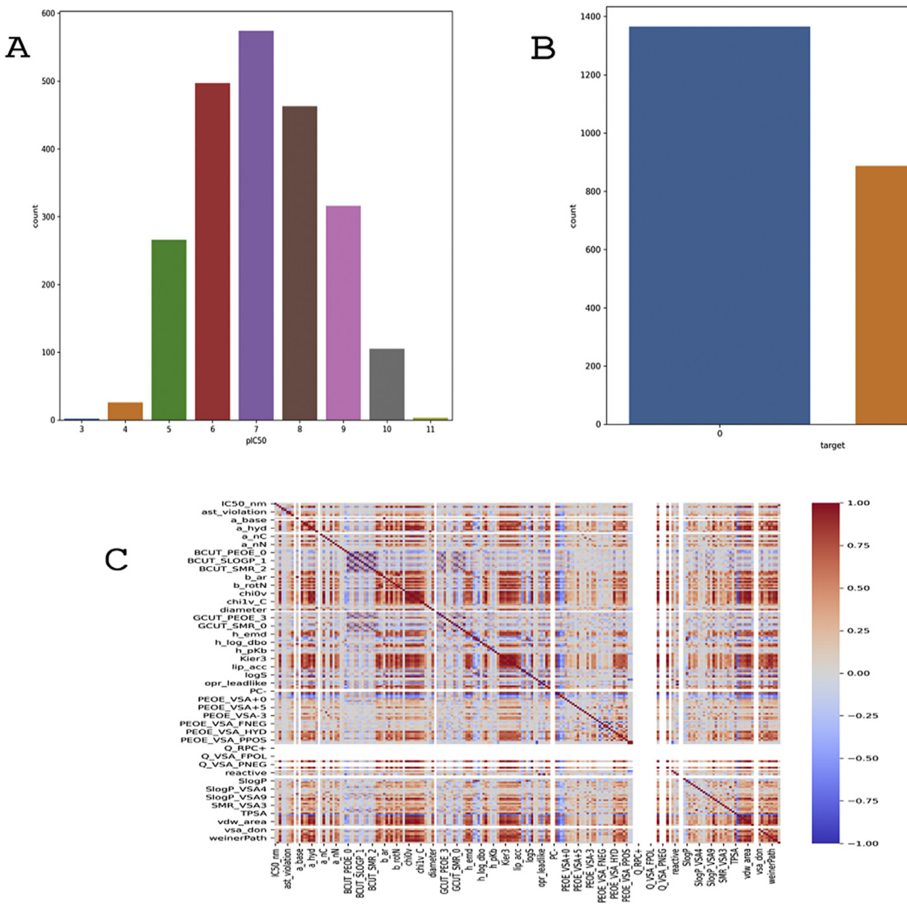


FIGURE 28.1 Distribution plot of (A) pIC50, (B) active and nonactive, and (C) correlation plot.

descriptors with zero or low importance were removed using the Feature Selector python package. A total of 44 descriptors were carried forward for modeling. The pIC50 values were converted into categorical data-type of active and nonactive using a python script, and the activity threshold was set to pIC50 8 and above.

3.3 Non-neural networks

There are huge arrays of non-NN algorithm, and each has its array of hyperparameters that must be tuned to achieve high-performance models. Selecting from these models could be a daunting task, therefore we used a python package, TPOT [16], to assist in this task.

TPOT is a genetic programming AutoML (Auto machine learning) protocol with the primary aim of maximizing classification or regression accuracy. TPOT classifier class was trained on the compounds and validated (training set, 1576; test set, 675). The analysis produced RFC as the best model for our modeling task with an accuracy of 0.802.

Table 28.1 The10-fold cross-validation evaluation of random forest classifier.

Metrics	Score
Precision	0.85
Recall (sensitivity)	0.77
Accuracy	0.85
Error rate	0.18
F1	0.80
ROC_AUC	0.84
Specificity	0.85
Balanced accuracy	0.81

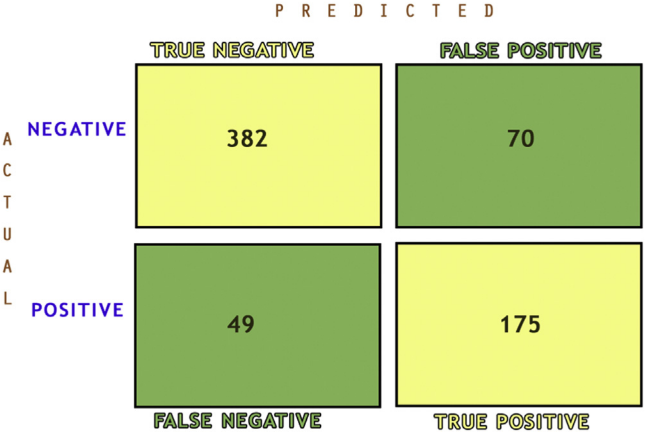


FIGURE 28.2 Random forest confusion matrix.

The RFC was trained on a training set of 1576 and validated on a test set of 675 compounds. The RFC model was subjected to a 10-fold cross-validation and was evaluated using various categorical classification evaluation metrics (Table 28.1). A confusion matrix (Fig. 28.2) was constructed from which the classification evaluation metrics were calculated.

3.4 Non-neural network performance evaluation

Precision metric measures the positive predictive value rate; that is, it measures how well the model identifies true positives as against false positives. The RFC model has a precision score of 0.85 (Table 28.1). The recall metric, also known as sensitivity score, is a measure of true positive rate, i.e., evaluates how well the model classifies true positives from false negatives. F1 is the harmonic mean of both precision and accuracy. Specificity measures the true negative rate; that is, it measures how well a model classifies true negatives correctly.

In order to get a full picture of how well the model is identifying true positive (sensitivity) and true negative (specificity), we calculate another metric call balanced accuracy (Table 28.1). Balanced accuracy is the mean of both sensitivity and specificity, and it enables us to evaluate the model's ability to identify active and nonactive inhibitors. The RFC model had a balanced accuracy of 0.808. The ROC_AUC score measures how often the model picks a true positive ahead of a false positive (Fig. 28.3). The RFC model had an ROC_AUC score of 0.84 (Note that all the metric values ranged from 0 to 1, with 1 being the best value and 0 the worst value possible).

Having evaluated this RFC model with the abovementioned metrics the results, therefore, suggest that the model is robust and reliable for the screening of compounds with unknown biological activity.

3.5 Model interpretation

Machine learning models provide an immerse opportunity in predicting unknowns, but they come with a challenge of interpretability. Most of the best performing algorithms

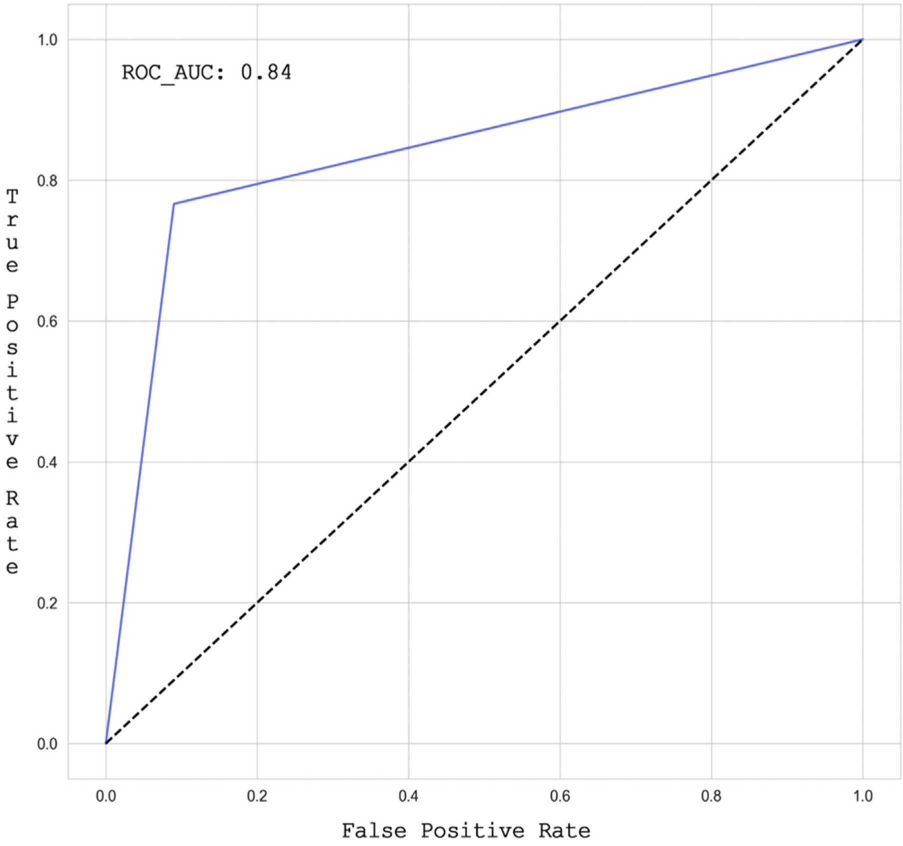


FIGURE 28.3 The ROC curve of random forest classifier (ROC_AUC score: 0.84).

are black boxes in nature and how they come about their prediction with the imputed features is not known [9]. It is therefore imperative to provide some sort of model explanation for every selected model. We used the Skater python library to provide a suitable model explanation. Skater explains machine learning models using two major plots: feature importance plot and partial dependence plot (PDP). The feature importance plot ranks features in the order of their importance to the model. It revealed SLogP_VSA4, SMR_VSA4, vsa_other, apol, and GCUT_SLOGP_0 as the top descriptors important to the model (Fig. 28.4).

The PDP investigates the marginal effect of different values of a descriptor on a predicted outcome of the machine learning model [33]. For categorical models, PDP investigates the marginal effect of the descriptor on the prediction probability of a class. For our study, the class in consideration is the active class.

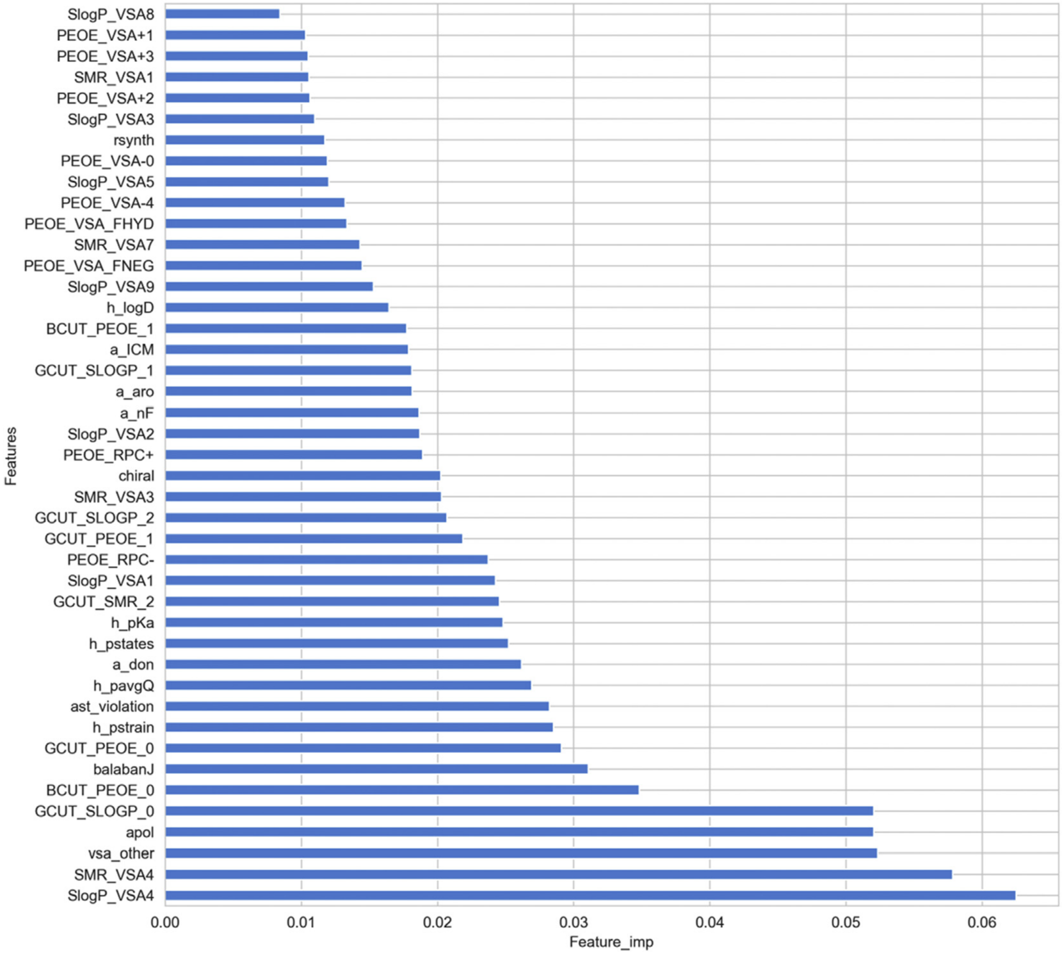


FIGURE 28.4 Feature importance of the random forest classifier model.

The PDP suggests that increasing the values of SLogP_VSA4 has the highest marginal effect on the predicted outcome (i.e., active class) of the model. SMR_VSA4 had the next highest marginal effect; however, this effect was only visible within a narrow range of 0.8–1.0. Further increase in values of SMR_VSA4 resulted in decreased effects, which were stable afterward (Fig. 28.5). VSA_other, apol, and GCUT_SLOGP_0 had an equal marginal effect on the active class predictions (Fig. 28.5).

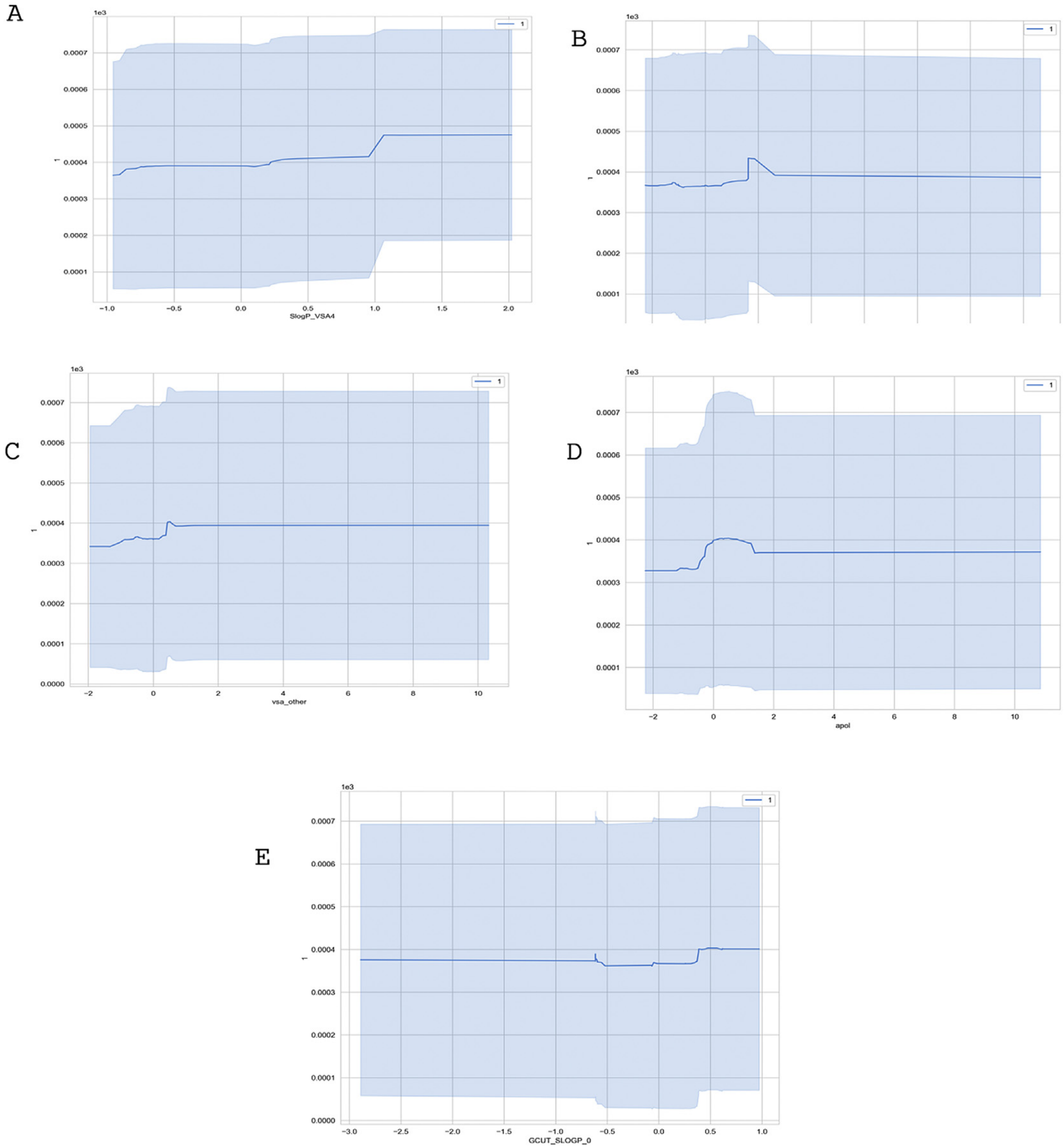


FIGURE 28.5 Partial dependence plot of random forest classifier model: (A) SLogP_VSA4, (B) SMR_VSA4, (C) vsa_other, (D) apol, and (E) GCUT_LOGP_0.

3.5.1 Neural network

In this study, NN was used to build a model on the inhibitors of TMPRSS 2. The NN model was built using the TensorFlow Python library [20]. The training set consists of 1576 compounds, test set for evaluation consisted of 675 compounds, and validation data was 10% of the training set.

As noted in non-NN, hyperparameters of NNs are also quite extensive; in order to solve this problem, python Scikit-learn Keras wrapper and GridSearchCV library were used to search for the optimum values: epoch, batch size, and the optimization function. The optimum values obtained from this exhaustive search after a 10-fold cross-validation are epoch: 20, batch size: 5, and optimizer: Adam. These values were therefore used to build the NN.

Another common problem of NN is overfitting of the model to the training data, in order to avoid this a regularization kernel (l2 regularize :0.01) was set for each layer of the neural architecture and finally an early stopping callback was set when compiling the neural architecture.

The model was trained over eight epochs (although the epoch parameter was set to 20 but was stopped early at the eighth epoch). The model training loss and training categorical accuracy, validation loss, and categorical accuracy were measured at each epoch and plotted (Fig. 28.6).

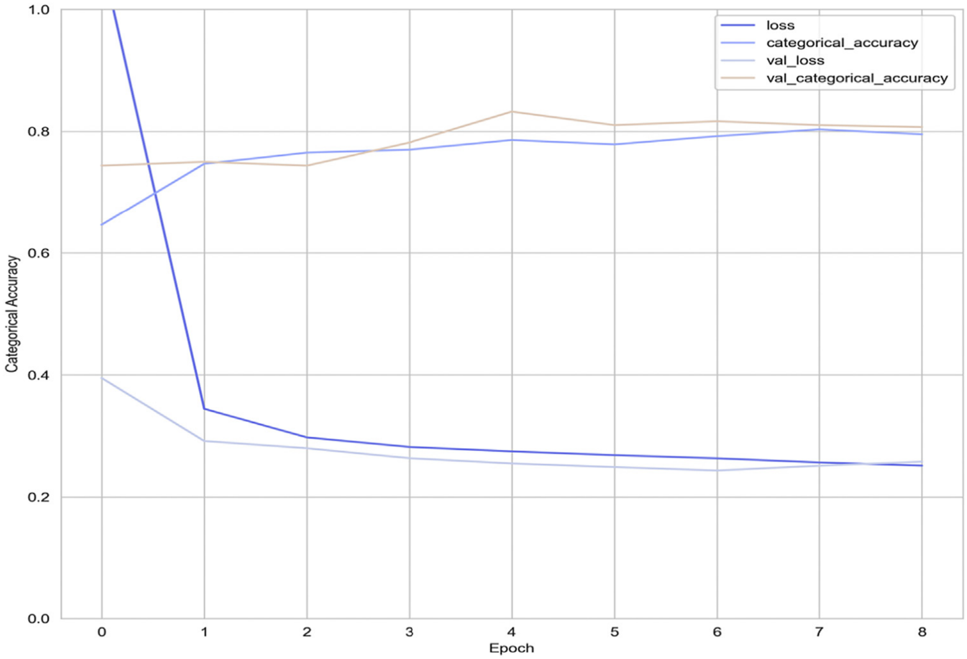


FIGURE 28.6 Model loss and accuracy, and validation loss and accuracy.

3.5.2 Neural network performance evaluation

The eighth epoch had a training loss of 0.26 and training categorical accuracy of 0.80, validation loss was 0.25, and validation categorical accuracy was 0.81 (Table 28.2). The eighth epoch was the best and saved. The model was finally evaluated with the test set with a categorical accuracy of 0.84 and a loss of 0.46 (Table 28.2).

A confusion matrix was constructed (Fig. 28.7) and categorical metrics evaluated (Table 28.2) (it should be noted, however, that evaluation is on the eighth epoch only).

The NN model had the following metric score: precision 0.79, recall (sensitivity) 0.82, F1 0.80, specificity 0.85, and balanced accuracy 0.84 (Table 28.2). The results, therefore, suggest that the NN model is reliable and can be used for extrapolations.

Table 28.2 Theeighth epoch Neural Network categorical evaluation.

Metrics	Score
Training loss	0.26
Training categorical accuracy	0.83
Validation loss	0.25
Validation categorical accuracy	0.81
Test loss	0.84
Test categorical accuracy	0.46
Recall (sensitivity)	0.82
Precision	0.79
Accuracy	0.84
Error rate	0.12
F1	0.80
Specificity	0.85
Balanced accuracy	0.84

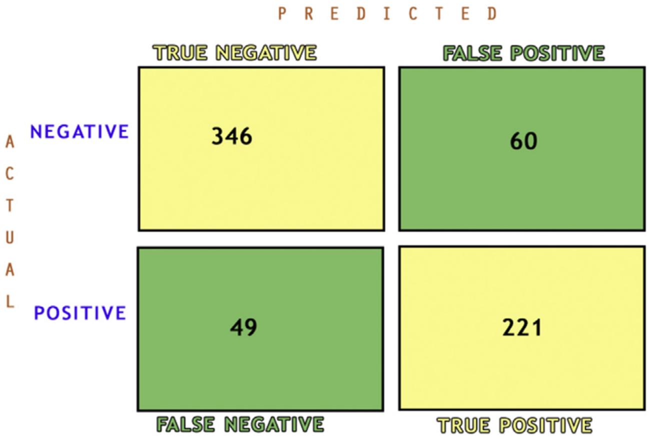


FIGURE 28.7 Theeighth epoch neural network confusion matrix.

3.5.3 Model interpretation

As stated earlier, model interpretation is important in understanding machine learning model predictions. The Skater python library was also used to provide explanations for the NN model (Feature Importance, PDP) (Figs. 28.8 and 28.9). Of the 43 descriptors used to build the NN model, SLogP_VSA4, PEOE_VSA_FHYD, a_nf, and SMR_VSA4 were the top important features (Fig. 28.8). A PDP showed the marginal effect of these descriptors on the predicted outcome (active class). Of these top descriptors, the a_nf descriptor had the highest marginal effect on the active class prediction probability, while PEOE_VSA_FHYD had the lowest contribution to the marginal effect on the active class prediction probability (Fig. 28.9). However, increasing the values of these descriptors had an increasing marginal effect on the model predictions (Fig. 28.9).

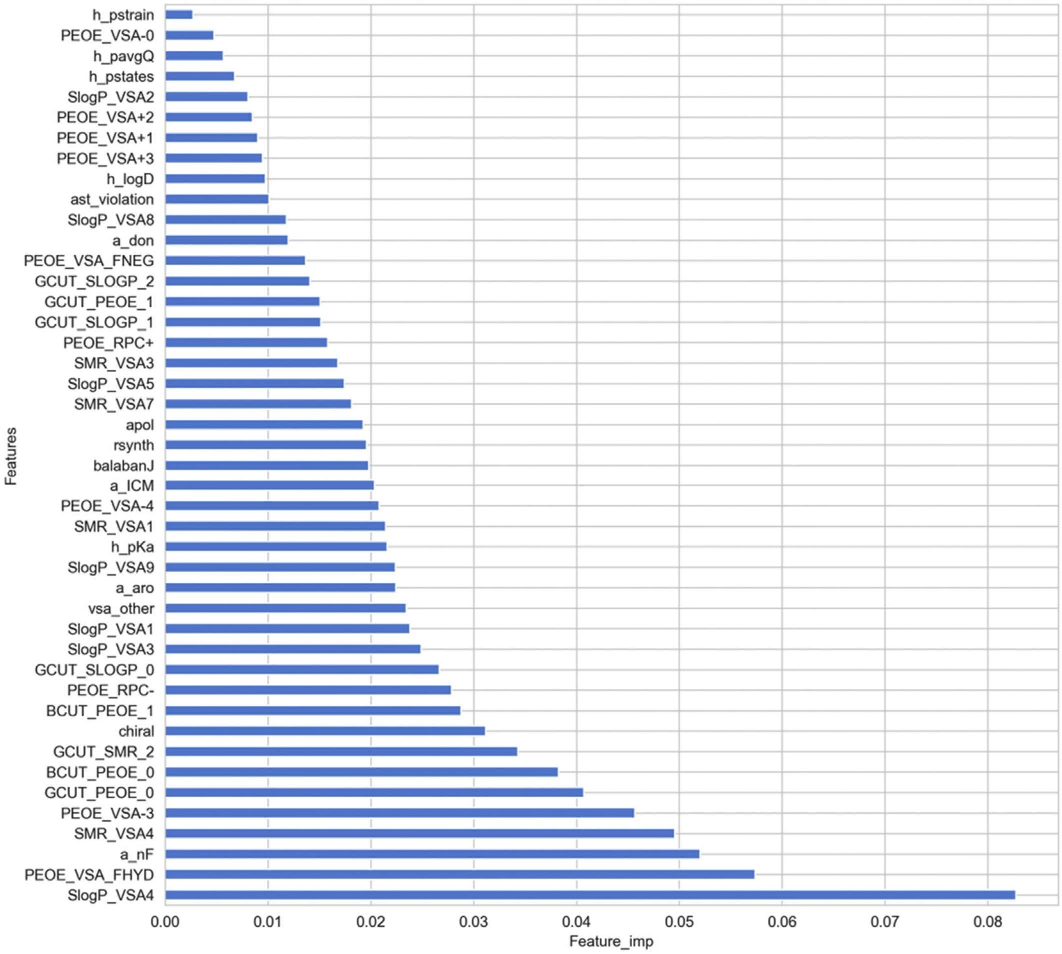


FIGURE 28.8 Feature importance of deep neural network model.

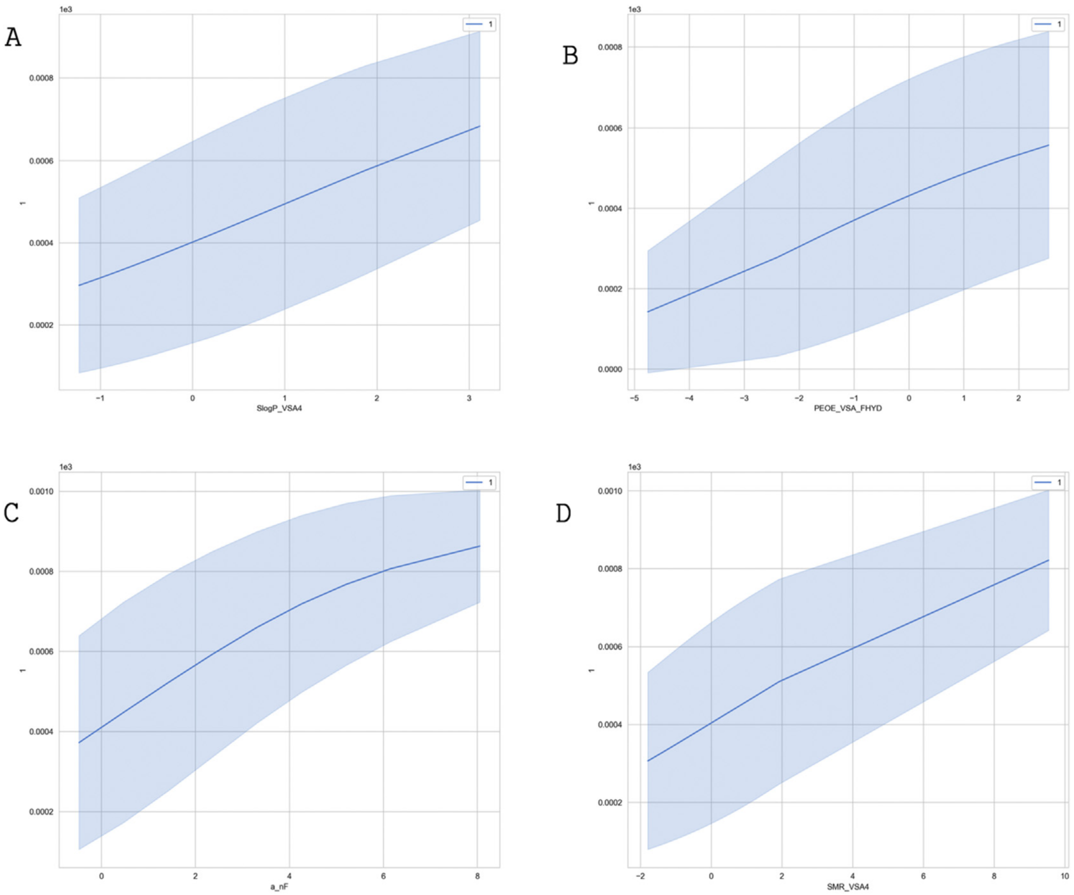


FIGURE 28.9 Partial dependence plot of deep neural network model: (A) SLogP_VSA4, (B) PEOE_VSA_FHYD, (C) a_nf, and (D) SMR_VSA4.

3.5.4 Model comparison

Although the primary aim is to compare both models, we evaluated both models on some performance metrics to see which model was the best performer.

From this comparison (Table 28.3), NN had the lowest error rate and the highest balanced accuracy, also the sensitivity of NN to true positive is about a fold higher than

Table 28.3 Machine learning model comparison.

Metric	Random forest classifier	Neural network
Sensitivity (recall)	0.77	0.82
Specificity	0.85	0.85
Error rate	0.18	0.12
Balanced accuracy	0.81	0.84

RFC. This, therefore, suggests that although both NN and RFC had a high level of accuracy and equal level of true negative identification (specificity), NN would most likely screen the database with a high level of sensitivity to compounds with inhibitory activities.

3.5.6 Database screening

Having developed and validated these two models (RFC and NN), we screened the SCUBIDOO database [20] sample of 100,000 compounds. The random forest model classified over 80,000 compounds as active and the NN model classified 3000 compounds as active (Fig. 28.10). The high number of predicted active compounds by RFC was however not surprising, which is due to the random forest low sensitivity score (Table 28.3). However, in order to filter down this number, a python script was written to filter down the base of the compound on two criteria: synthesis probability and drug-likeness. The script reduced the RFC model hits to 1600 compounds and the NN model hits to 250 compounds. Furthermore, using ADMET properties on SwissADME [34] the compounds were further screened down to 784 and 70, respectively.

3.6 Physiochemical properties of transmembrane protease serine 2 isoform 1

The TMPRSS2 isoform 1 physiochemical properties were predicted and analyzed using ProtPara webserver. The protein sequence for TMPRSS2 isoform 1 consisted of 346 amino acid residues, with Ser (8.7%) and Met and Phe (2.6%) amino acid residues having the highest and lowest composition, respectively, as indicated in Fig. 28.11. The computed pI for the amino acid residues for TMPRSS2 isoform 1, i.e., the isoelectric pH,

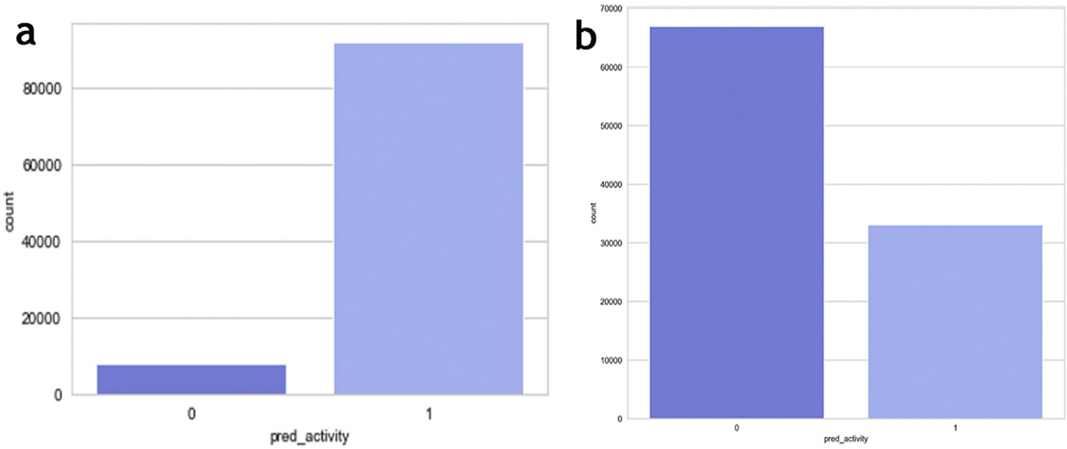


FIGURE 28.10 SCUBIDOO database screening activity prediction (1: active, 0: nonactive): (A) random forest classifier and (B) neural network.

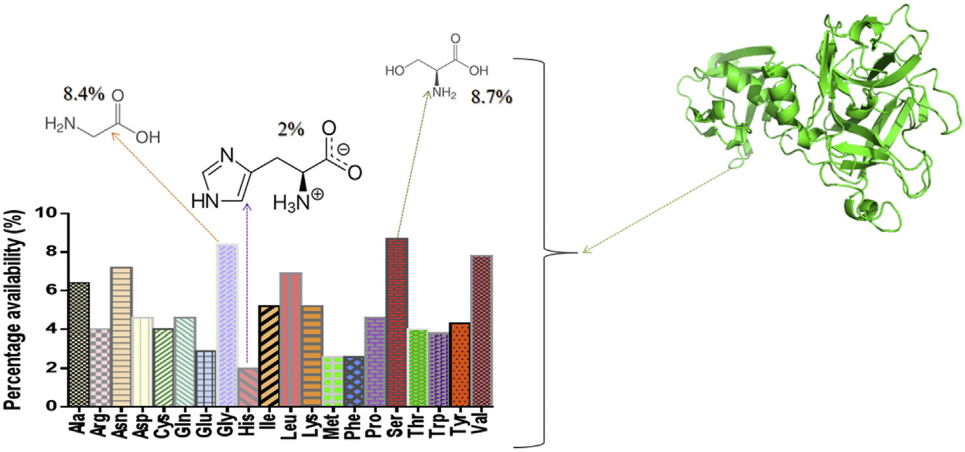


FIGURE 28.11 Amino acid residue composition of transmembrane protease serine 2 isoform 1.

was 8.58 (pI > 7); this revealed the slight alkaline nature of the amino acid residues of the model protein. In experimental studies like isoelectric focusing and 2D electrophoresis, the isoelectric pH of a protein plays a critical role.

The stability of the protein was computed by analyzing its instability index. The index for this protein was 35.18. This protein may be stable because its predicted value is within the range of 40 [35,36].

3.7 Subcellular localization of transmembrane protease serine 2 isoform 1

The Hum-mPLoc3 webserver [36] was used to predict the subcellular localization of TMPRSS2 isoform 1. The analysis revealed that the human TMPRSS2 isoform 1 is predominantly a plasma membrane protein (Fig. 28.12).

3.8 Homology modeling for predicting 3D structure of the human BAG3

Quaternary structures with complex interactions and their physiologic roles are necessary for detailed comprehension of the human system. Experimental elucidation of protein structures using either nuclear magnetic resonance spectroscopy or X-ray crystallography is more realistic but time consuming and occasionally unsuccessful in case of membrane proteins; thus these necessitate the use of homology modeling techniques. Serine protease hepsin of template 5ce1.1.A with a 33.82% sequence identity and 2.5 Å resolution was used in the homology modeling of TMPRSS2 isoform 1.

Global Model Quality Estimate helps in the identification and selection of an optimal template in the modeling process. For this study, the score was 0.48. The sequence

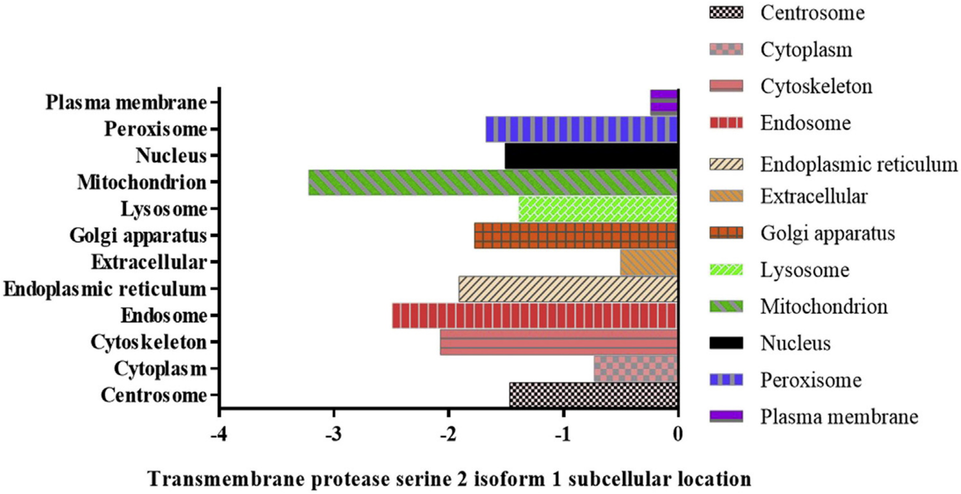


FIGURE 28.12 Subcellular localization of transmembrane protease serine 2 isoform 1.

similarity to the query sequence was 0.38, which covers the range of 183–528 of the amino acid residues. The quality of the built model was assessed by the QMEAN scoring function, which uses potentials of mean force to generate global and per residue quality estimates. The QMEAN score [37], which is the best of the model generated, was −1.39, as indicated in Fig. 28.13D. The quality of the model developed was assessed using SAVES [38], PROSESS [24], and PROSA webservers [25]. The percentage of residues in the most favored regions and the percentage of Phi/Psi pairs in the favored regions of the Ramachandran plot of the nonoptimized protein was 86% and 92% compared to the 93.08% and 97.60% expected values, respectively, as indicated in Fig. 28.13A.

The overall quality factor of the model was assessed further using ERRAT and was observed to be 92.923%. A good model of high resolution is expected to have ERRAT quality values around 95% higher. Thus the model developed has a lower resolution, as indicated in Fig. 28.13B.

One of the computational limitations of protein modeling is the variation of the experimental and predicted model from the native true structure of a protein, thus necessitating the need for refinement and optimization of protein models. In this study, an improvement in the model was made using 3Drefine [22], which led to the generation of five refined protein models from an initially input SWISS model.

3Drefine works by optimizing hydrogen bonding network in addition to atomic-level energy minimization. The best model of the five generated, as indicated in Fig. 28.14D, has the lowest potential energy according to 3Drefine force field. Quality check was

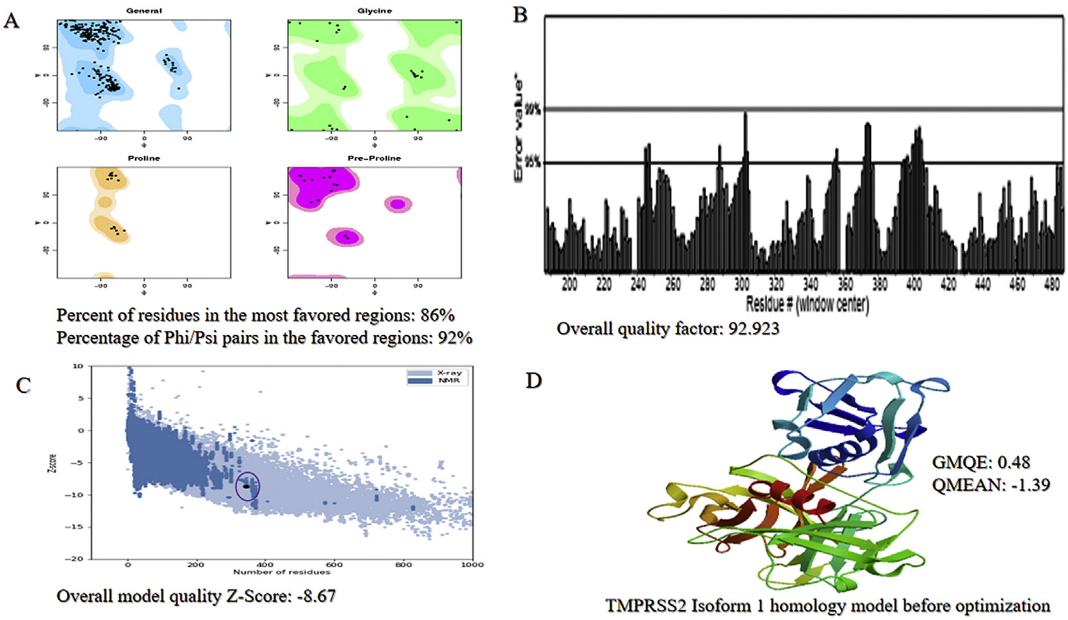


FIGURE 28.13 (A–D) Homology model protein, quality, and validation check of transmembrane protease serine 2 (TMPRSS2) isoform 1 before optimization.

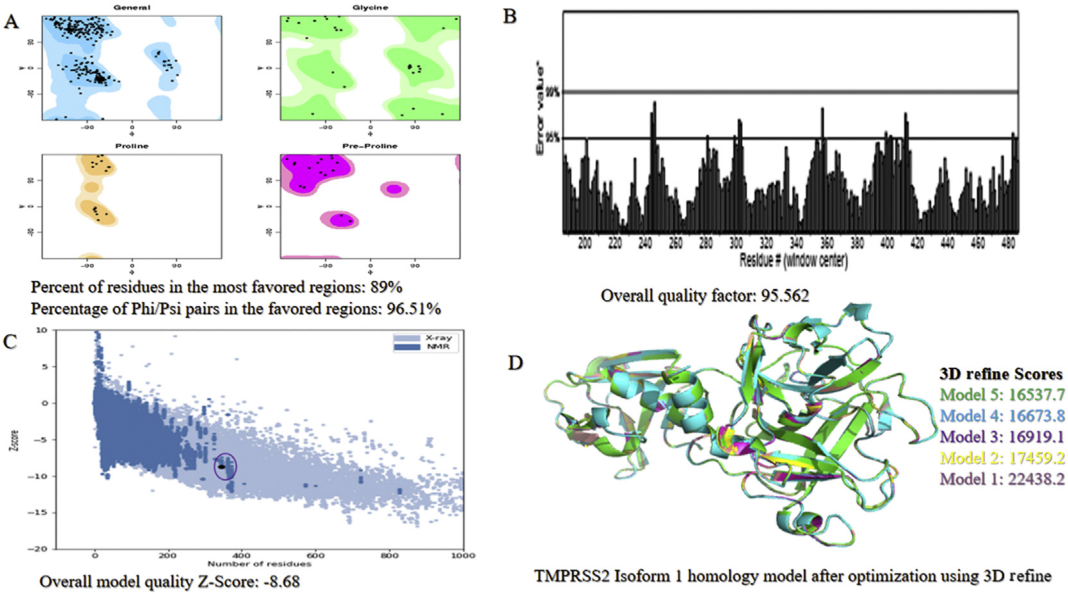
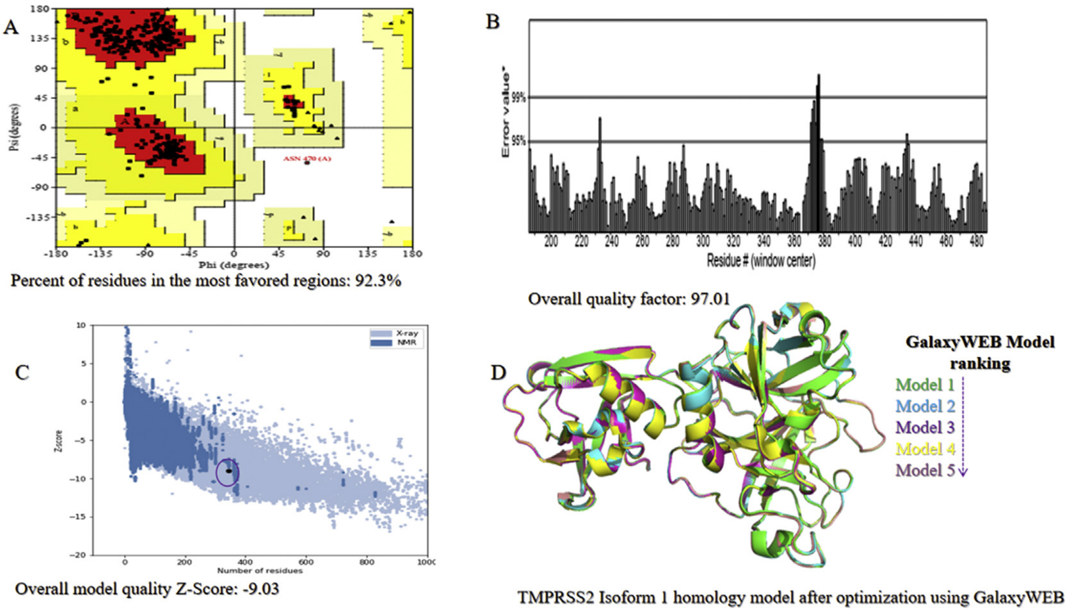


FIGURE 28.14 (A–D) Homology model protein, quality, and validation check of transmembrane protease serine 2 (TMPRSS2) isoform 1 after optimization using 3Drefine.

carried out on the 3Drefine model having the lowest potential energy. An improvement in the quality of the model was observed compared to the unrefined model, as indicated in Fig. 28.13D. The percentage of residues in the most favored regions and the percentage of Phi/Psi pairs in the favored regions of the Ramachandran plot of the 3Drefine-optimized protein was 89% and 96.51% compared to the 86% and 92% of the unrefined model, respectively. But this improvement was below the expected values 93.08% and 97.60%. The ERRAT and Z-score qualities also improved and were 95.5% and −8.68, respectively, as indicated in Figs. 28.14 and 28.15.

The quality of the 3Drefine model was further optimized using GalaxyWEB [39], which works on the basis of rebuilding the side chain and overall protein relaxation using MDS. The best of the five models generated was model one. Quality check and validation was carried out on this model also with a dramatic increase in the quality of the protein. The percentage of residues in the most favored regions of the Ramachandran plot was 92%, which exceeded the expected by 2%. The ERRAT quality factor score was also improved by 97.01%, which also exceeded the expected by 2%. The quality of the model assessed by PROSA also indicated an improvement in Z-score, as indicated in Fig. 28.15C. Thus compared to the unrefined and 3Drefine, the GalaxyWEB generated the best model as corroborated by the quality and validation scores.



Percent of residues in the most favored regions: 92.3%

Overall quality factor: 97.01

Overall model quality Z-Score: -9.03

TMPRSS2 Isoform 1 homology model after optimization using GalaxyWEB

FIGURE 28.15 (A—D) Homology model protein, quality, and validation check of transmembrane protease serine 2 (TMPRSS2) isoform 1 after optimization using GalaxyWEB.

3.9 Binding pocket prediction of transmembrane protease serine 2 isoform 1

Using a template-free machine learning algorithm, P2Rank embedded in PrankWeb [29,30] was used to predict the binding site of ligand on TMPRSS2. P2Rank works based on points situated on solvent-accessible protein surface from local chemical neighborhood ligandability. The resulting binding sites are formed via the cluster of points with high ligandability score. The predicted binding sites are indicated in Table 28.1.

3.10 Molecular docking analysis

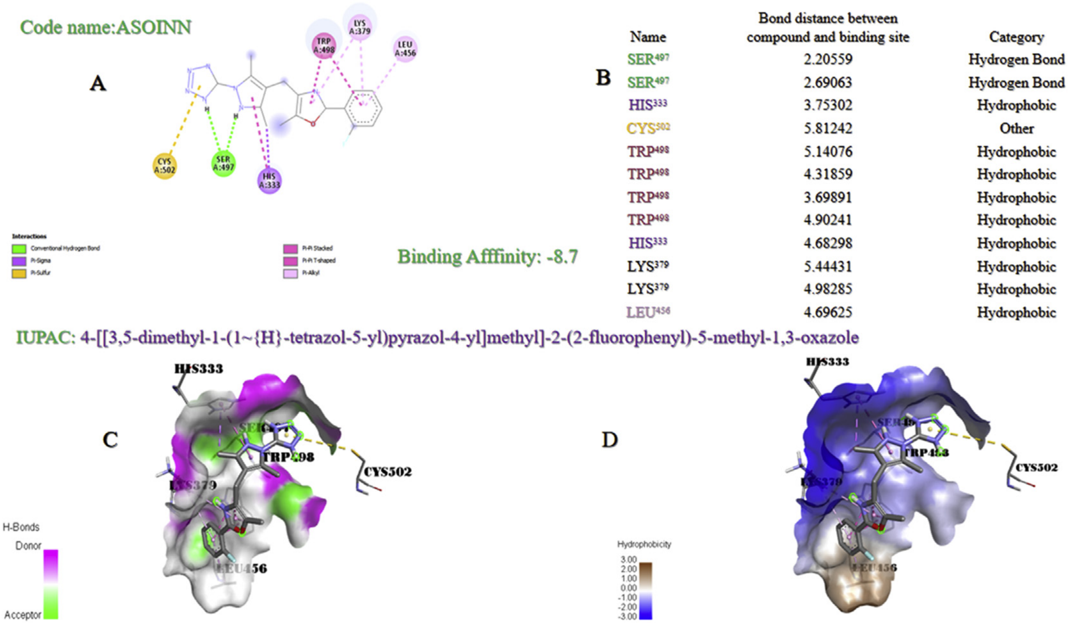
The set of computational methodology used in the analysis of large databases with the aim of identifying potential hit candidates is referred to as virtual screening [40–42]. Ligand-based and structure-based (or receptor-based) are the two main types of virtual screening methods used in potential hit identification. Both the methods were carried out. In this study, through ligand-based virtual screening, 1600 and 250 compounds obtained from RFC and NN models were screened to 70 and 784 compounds using SwissADME, respectively. In ligand-based virtual screening of the compounds obtained from RFC and NN using SwissADME the compounds were screened to 70 and 784, respectively; all the compounds passed the Lipinski's [43], Ghose's [44], Oprea's [45], Veber's [46], Varma's [47], Egan's [48], and Muegge's [49] rules filters for drug-likeness evaluation. In addition to these, no PAINS (pan assay interference compounds) [50] and Brenk alert [51] were recorded and all passed the lead-likeness rules.

These compounds were then used in receptor-based virtual screening so as to obtain potential leads compared to camostat; the established standard drug was used as an inhibitor of TMPRSS2. The leads obtained from the receptor-based virtual screening of the compounds are indicated in Figs. 28.16 and 28.17A–D. The binding affinities were −8.7, −8.4, −8.4, −8.5, and −8.5 kcal/mol, respectively, compared to −7.4 kcal/mol obtained for camostat, as indicated in Fig. 28.18. All the amino acids involved in bound interactions are within the range of predicted binding pockets, as indicated in Table 28.4.

The compound ASOINN indicated in Fig. 28.16 was the lead obtained from the structural virtual screening of the 70 compounds. The lead forms nine hydrophobic bond interactions, which included two HIS^{333}, four TRP^{498}, two LYS^{379}, and LEU^{456}. In addition to these, two SER^{497} bound interactions were formed.

Compounds ASOIRFC1–4 as indicated in Fig. 28.17A–D were obtained from TMPRSS2 isoform 1 virtual screening of 784 compounds, which were obtained from the SwissADME-screened RFC 1600 compounds.

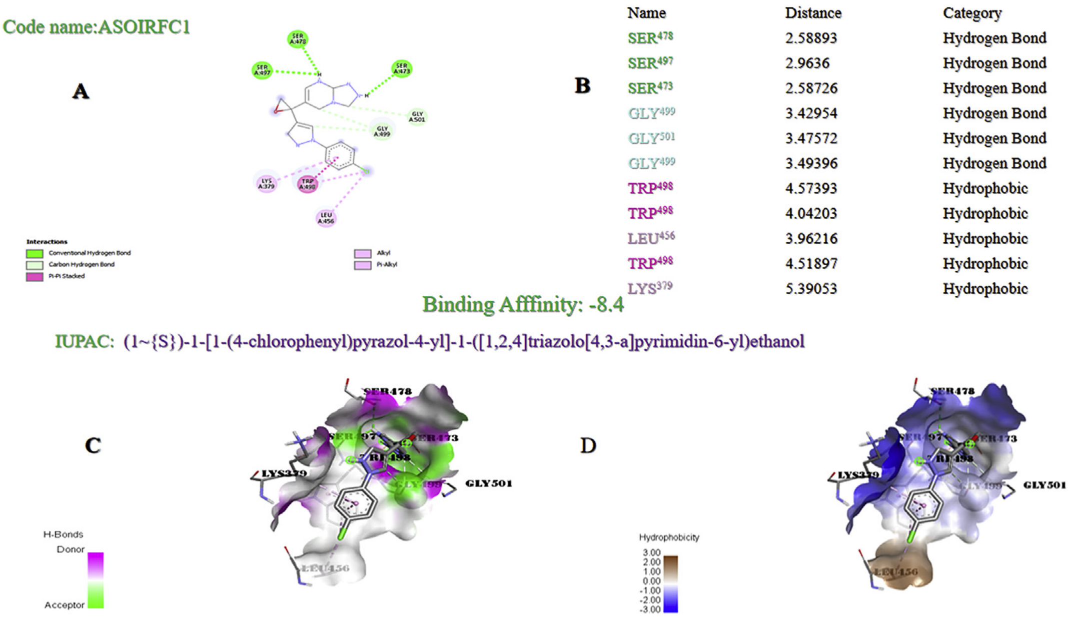
As indicated in Fig. 28.17A, ASOIRFC1 formed six hydrogen bond interactions, which included SER^{478}, SER^{497}, SER^{473}, two GLY^{499}, and GLY^{501}, as well as five hydrophobic bonds (three TRP^{498}, LEU^{456}, and LYS^{379}); these are in contrast to the nine hydrophobic interactions and two hydrogen bonds formed in ASOINN.



Name	Bond distance between compound and binding site	Category
SER^{497}	2.20559	Hydrogen Bond
SER^{497}	2.69063	Hydrogen Bond
HIS^{333}	3.75302	Hydrophobic
CYS^{502}	5.81242	Other
TRP^{498}	5.14076	Hydrophobic
TRP^{498}	4.31859	Hydrophobic
TRP^{498}	3.69891	Hydrophobic
TRP^{498}	4.90241	Hydrophobic
HIS^{333}	4.68298	Hydrophobic
LYS^{379}	5.44431	Hydrophobic
LYS^{379}	4.98285	Hydrophobic
LEU^{456}	4.69625	Hydrophobic

IUPAC: 4-[[3,5-dimethyl-1-(1~{H}-tetrazol-5-yl)pyrazol-4-yl]methyl]-2-(2-fluorophenyl)-5-methyl-1,3-oxazole

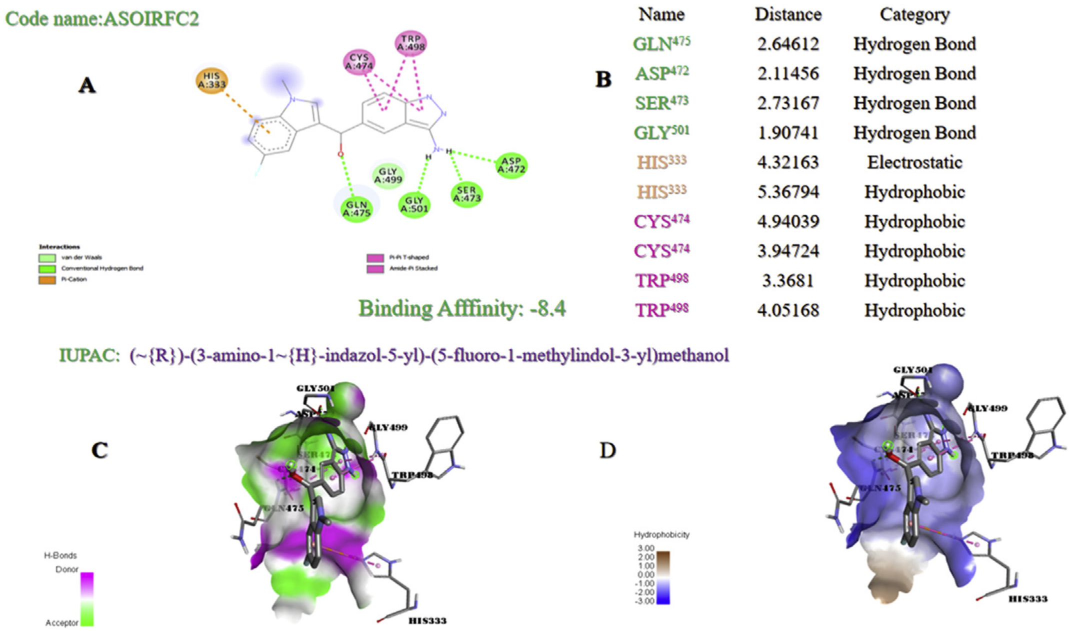
FIGURE 28.16 (A—D) ASSOINN interaction with transmembrane protease serine 2 isoform 1 at the binding site.



Name	Distance	Category
SER^{478}	2.58893	Hydrogen Bond
SER^{497}	2.9636	Hydrogen Bond
SER^{473}	2.58726	Hydrogen Bond
GLY^{499}	3.42954	Hydrogen Bond
GLY^{501}	3.47572	Hydrogen Bond
GLY^{499}	3.49396	Hydrogen Bond
TRP^{498}	4.57393	Hydrophobic
TRP^{498}	4.04203	Hydrophobic
LEU^{456}	3.96216	Hydrophobic
TRP^{498}	4.51897	Hydrophobic
LYS^{379}	5.39053	Hydrophobic

IUPAC: (1~{S})-1-[1-(4-chlorophenyl)pyrazol-4-yl]-1-([1,2,4]triazolo[4,3-a]pyrimidin-6-yl)ethanol

FIGURE 28.17A (A) ASOIRFC1 interaction with transmembrane protease serine 2 isoform 1 at the binding site.



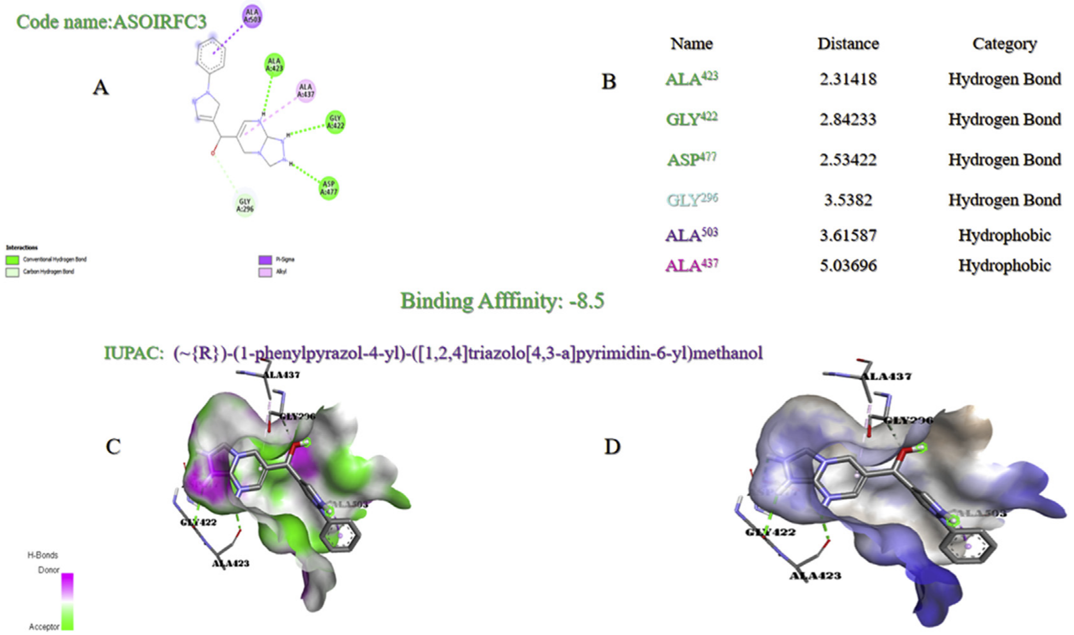
Code name:ASOIRFC2

Name	Distance	Category
GLN475	2.64612	Hydrogen Bond
ASP472	2.11456	Hydrogen Bond
SER473	2.73167	Hydrogen Bond
GLY501	1.90741	Hydrogen Bond
HIS333	4.32163	Electrostatic
HIS333	5.36794	Hydrophobic
CYS474	4.94039	Hydrophobic
CYS474	3.94724	Hydrophobic
TRP498	3.3681	Hydrophobic
TRP498	4.05168	Hydrophobic

Binding Afffinity: -8.4

IUPAC: (~{R})-(3-amino-1~{H}-indazol-5-yl)-(5-fluoro-1-methylindol-3-yl)methanol

FIGURE 28.17B (B) ASOIRFC2 interaction with transmembrane protease serine 2 isoform 1 at the binding site.



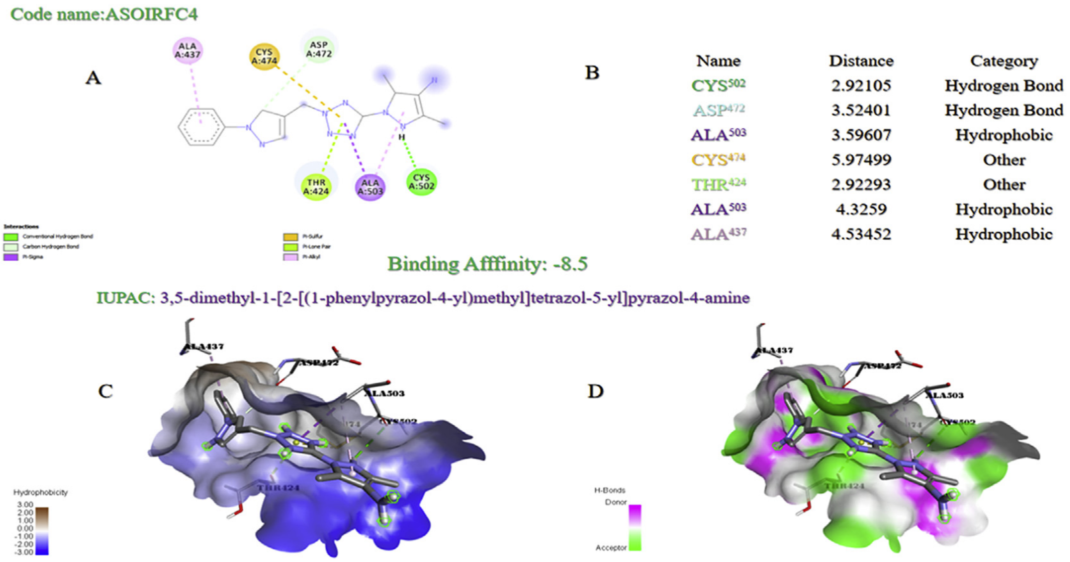
Code name:ASOIRFC3

Name	Distance	Category
ALA423	2.31418	Hydrogen Bond
GLY422	2.84233	Hydrogen Bond
ASP477	2.53422	Hydrogen Bond
GLY296	3.5382	Hydrogen Bond
ALA503	3.61587	Hydrophobic
ALA437	5.03696	Hydrophobic

Binding Afffinity: -8.5

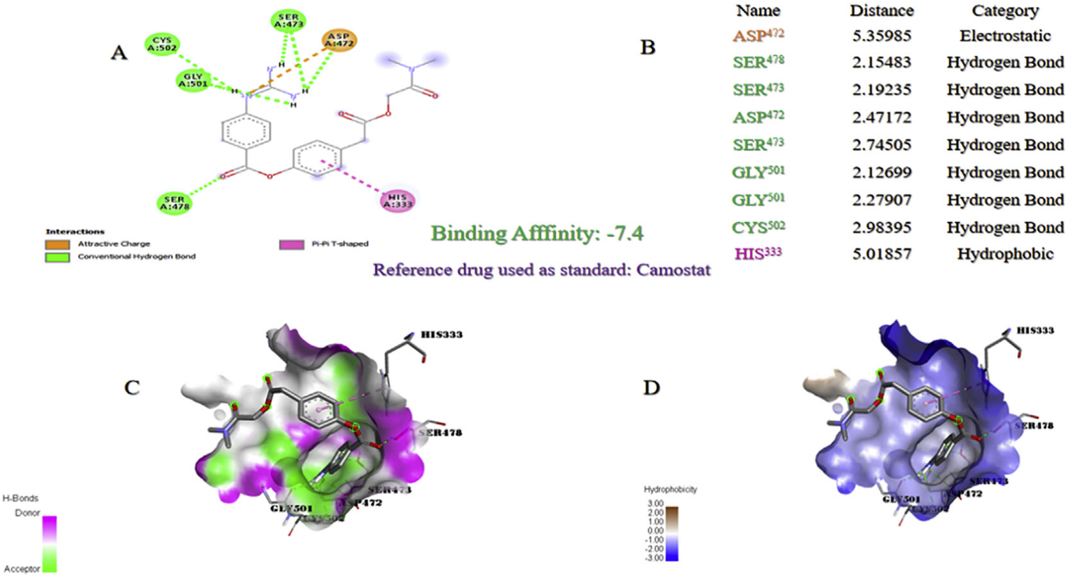
IUPAC: (~{R})-(1-phenylpyrazol-4-yl)-([1,2,4]triazolo[4,3-a]pyrimidin-6-yl)methanol

FIGURE 28.17C (C) ASOIRFC3 interaction with transmembrane protease serine 2 isoform 1 at the binding site.



Name	Distance	Category
CYS502	2.92105	Hydrogen Bond
ASP472	3.52401	Hydrogen Bond
ALA503	3.59607	Hydrophobic
CYS474	5.97499	Other
THR424	2.92293	Other
ALA503	4.3259	Hydrophobic
ALA437	4.53452	Hydrophobic

FIGURE 28.17D (D) ASOIRF4 interaction with transmembrane protease serine 2 isoform 1 at the binding site.



Name	Distance	Category
ASP472	5.35985	Electrostatic
SER478	2.15483	Hydrogen Bond
SER473	2.19235	Hydrogen Bond
ASP472	2.47172	Hydrogen Bond
SER473	2.74505	Hydrogen Bond
GLY501	2.12699	Hydrogen Bond
GLY501	2.27907	Hydrogen Bond
CYS502	2.98395	Hydrogen Bond
HIS333	5.01857	Hydrophobic

FIGURE 28.18 Camostat interaction with the binding site of transmembrane protease serine 2 isoform 1.

ASOIRFC2 indicated in Fig. 28.17B formed five hydrophobic interactions, four hydrogen interactions, and one electrostatic interaction, which are not formed in ASOINN and ASOIRFC1.

Table 28.4 Predicted binding sites and scores.

Name	Score	Predicted residues of the binding pockets
Pocket 1	7.8177	GLY^{296} ILE^{418} SER^{419} GLY^{420} GLY^{422} ALA^{423} THR^{424} ASN^{435} ALA^{436} ALA^{437} ASN^{470} VAL^{471} ASP^{472} SER^{473} ASP^{477} CYS^{502} ALA^{503}
Pocket 2	7.7105	VAL^{317} HIS^{333} CYS^{334} GLU^{336} LEU^{339} GLU^{426} ASP^{472} SER^{473} CYS^{474} GLN^{475} GLY^{476} ASP^{478} TRP^{497} GLY^{498} SER^{499} GLY^{501} CYS^{502} GLY^{509}
Pocket 3	7.2816	ARG^{184} ARG^{187} LEU^{188} GLY^{190} PHE^{193} MET^{225} TYR^{227} ARG^{277} CYS^{278} ALA^{280} CYS^{281} VAL^{283} SER^{487} ARG^{489} SER^{491}

As indicated in Fig. 28.17C, ASOIRFC3 formed four hydrogen bond interactions, which included ALA^{423}, GLY^{422}, ASP^{477}, and GLY^{296}. In addition to these, two hydrophobic interactions, including ALA^{503} and ALA^{437}, were formed. ASOIRFC3 so far formed the lowest number of interactions.

ASOIRFC4 as indicated in Fig. 28.17D formed seven bond interactions, which included two hydrogen bonds, three hydrophobic interactions, pi-lone pair, and pi-sulfur type bonds.

Camostat, the standard drug used as an inhibitor of TMPRSS2, forms seven hydrogen bonds and one electrostatic and hydrophobic bond, as indicated in Fig. 28.18.

Hydrogen bonds are the prevailing directional intermolecular interactions in biological complexes, and the predominant contribution to the specificity of molecular recognition. In drug design, hydrogen bonds are exploited to gain specificity owing to their strict distance and geometric constraints. Furthermore, previous studies indicated that, contribution of hydrogen bond to free energy is dependent on local environment: a solvent-exposed hydrogen-bond contributes significantly less to net interaction energy

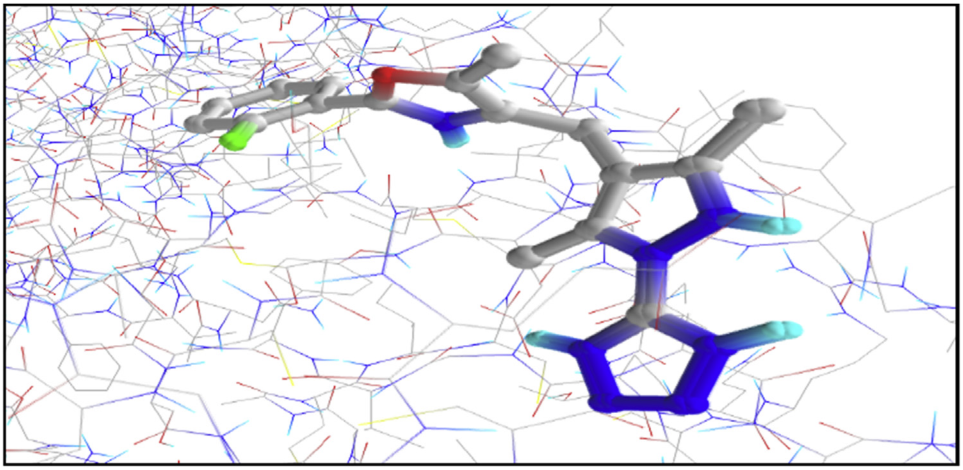


FIGURE 28.19 Validation of docking, comparability of the redocked binding mode, and the pose of ASOINN with the accompanying residues of transmembrane protease serine 2 isoform 1 binding pocket.

than the same hydrogen-bond in a buried hydrophobic pocket [51a,51b]. To validate the docking protocol, we redocked on of the lead (ASOINN) compound into the predicted binding pocket of TMPRSS2. As indicated in Fig. 28.19, the redocked pose overlapped almost completely with the previous orientation, thus indicating the reliability of the docking protocol and the scores obtained.

Hydrogen bonds are nonbonded interactions; they are specific, directional, and in a short range. They occur via a covalent interaction between hydrogen atoms and electronegative atoms, which usually involves N, S, or O. The strength of hydrogen bonds is optimal with alignment of the atoms involved in bond formation, and this occurs especially when the H donor points directly at the electron acceptor pair. However, the strength of hydrogen bonds is weaker than that of covalent or ionic bonds, contributing to the specificity of molecular recognition [52,53]. Heavy atoms in N—H···O, N—H···N, and O—H···O hydrogen bonds were all found to be separated by similar median distances of approximately 3.0 Å in previous works. In this study, the bond distances for majority of the lead complexes are within the required threshold of median distance 3.0 Å [58]. Hydrophobic contacts are by far the most common interactions in protein—ligand complexes, accounting for 66,772 contacts at a distance cut-off of 4.0 between a carbon and a carbon, halogen, or sulfur atom. The group formed by an aliphatic carbon in the receptor and an aromatic carbon in the ligand is the most populous, accounting for over 42,000 interactions [59]. This suggests that aromatic rings are common in small molecule inhibitors. In fact, one or more aromatic rings are present in 76 percent of marketed drugs, with benzene being by far the most commonly encountered ring system [60]. Not surprisingly, leucine side-chains are the most frequently involved in hydrophobic interactions, followed by valine, isoleucine, and alanine side-chains [58]. Based on the requirements of absorption and permeation of drug molecules, the number of hydrogen bonds has been shown to be limited. The Lipinski rule-of-five, for example, states that drug molecules with hydrogen bond acceptors greater than 10 and hydrogen bound donors greater than 5 are most likely to have poor absorption or permeation properties [62]. A single hydrogen bond formed either by intra- or intermolecular interaction is weaker than ionic or covalent bond but stronger than van der Waals interaction, thus unable to aid a drug-receptor interaction alone. Moreover, the formation of multiple drug-receptor hydrogen bond interactions conferred stability, which is an essential feature of drug-receptor interactions [57].

In this study, the lead compounds, ASOINN and ASORFC1—4, are within the range of accepted number of hydrogen bounds in terms of acceptors and donors; thus they all may have good absorption or permeation properties. In addition to these, although all the compounds are within the range of accepted number of hydrogen, only ASOIRF1 has the highest number of hydrogen bonds (6) compared to camostat (7); thus ASOIRF1 may be more stable than ASOINN and ASOIRF2—4.

Previous studies have shown that synergistic receptor-ligand H-bond pairings potentiate high-affinity binding, which correspond to an increase in binding affinity [54]. In addition to these, hydrogen bonding and optimized hydrophobic interactions have been shown to both stabilize the ligands at the target site and help alter binding affinity and drug efficacy [55,56]. The recorded increase in binding affinities of ligands ASOINN and ASORFC1−4 in contrast to camostat may be due to the observed increase in the hydrophobic and hydrogen bound interactions, in addition to electrostatic and π-interactions.

3.11 Conclusion

The search for the therapeutic treatment of SARS-CoV-2 infection (coronavirus disease 2019 [COVID-19]) is not only of utmost importance but also time sensitive. Hence a fast method to screen for plausible therapeutic drug against different targets in SARS-CoV-2 has been employed all over the world. In this study, we selected TMPRSS2 isoform 1 as a therapeutic target for SARS-CoV-2 and employed the power of machine learning to develop two models (random forest and NNs) based on 2251 inhibitors of serine proteases downloaded from CHEMBL. These models have been used to screen a sample of SCUBIDOO database (M: 99,977), which is a database of 21,000,000 virtual compounds. We have therefore identified five possible lead compounds having shown good ADMET properties, binding affinity, and molecular interaction with TMPRSS2.

To further improve this work, the five lead compounds can be used to search for more similar compounds in the SCUBIDOO database and finally validated experimentally.

References

[1] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K. Chu, T. Bleicker, S. Brünink, J. Schneider, M.L. Schmidt, D.G. Mulders, Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR, Euro Surveill. 25 (3) (2020) 2000045.

[2] A.R. Fehr, R. Channappanavar, S. Perlman, Middle East respiratory syndrome: emergence of a pathogenic human coronavirus, Annu. Rev. Med. 68 (2017) 387−399.

[3] World Health Organization. (2020). Coronavirus disease (COVID-19).

[4] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (10223) (2020) 497−506.

[5] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (10223) (2020) 470−473.

[6] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. (2020).

[7] P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C.L. Huang, H.D. Chen, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (7798) (2020) 270−273.

[8] I. Glowacka, S. Bertram, M.A. Müller, P. Allen, E. Soilleux, S. Pfefferle, I. Steffen, T.S. Tsegaye, Y. He, K. Gnirss, D. Niemeyer, Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response, J. Virol. 85 (9) (2011) 4122−4134.

[9] S. Matsuyama, N. Nagata, K. Shirato, M. Kawase, M. Takeda, F. Taguchi, Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2, J. Virol. 84 (24) (2010) 12658−12664.

[10] A. Shulla, T. Heald-Sargent, G. Subramanya, J. Zhao, S. Perlman, T. Gallagher, A transmembrane serine protease is linked to the severe acute respiratory syndrome coronavirus receptor and activates virus entry, J. Virol. 85 (2) (2011) 873−882.

[11] N. Iwata-Yoshikawa, T. Okamura, Y. Shimizu, H. Hasegawa, M. Takeda, N. Nagata, TMPRSS2 contributes to virus spread and immunopathology in the airways of murine models after coronavirus infection, J. Virol. 93 (6) (2019) e01815−e01818.

[12] K. Shirato, K. Kanou, M. Kawase, S. Matsuyama, Clinical isolates of human coronavirus 229E bypass the endosome for cell entry, J. Virol. 91 (1) (2017) e01387−16.

[13] K. Shirato, M. Kawase, S. Matsuyama, Wild-type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry, Virology 517 (2018) 9−15.

[14] P. Zmora, A.S. Moldenhauer, H. Hofmann-Winkler, S. Pöhlmann, TMPRSS2 isoform 1 activates respiratory viruses and is expressed in viral target cells, PLoS One 10 (9) (2015).

[15] MOE (The Molecular Operating Environment), software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7. http://www.chemcomp.com.

[16] R.S. Olson, J.H. Moore, TPOT: a tree-based pipeline optimization tool for automating machine learning, in: Automated Machine Learning, Springer, Cham, 2019, pp. 151−160.

[17] T. Chen, C. Guestrin, XGBoost: reliable large-scale tree boosting system, in: Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2015, pp. 13−17.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825−2830.

[19] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, Tensorflow: a system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265−283.

[20] F. Chevillard, P. Kolb, SCUBIDOO: a large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability, J. Chem. Inf. Model. 55 (9) (2015) 1824−1835.

[21] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T.G. Cassarino, M. Bertoni, L. Bordoli, T. Schwede, SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information, Nucleic Acids Res. 42 (W1) (2014) W252−W258.

[22] D. Bhattacharya, J. Nowotny, R. Cao, J. Cheng, 3Drefine: an interactive web server for efficient protein structure refinement, Nucleic Acids Res. 44 (W1) (2016) W406−W409.

[23] L. Heo, H. Park, C. Seok, GalaxyRefine: protein structure refinement driven by side-chain repacking, Nucleic Acids Res. 41 (W1) (2013) W384−W388.

[24] M. Berjanskii, Y. Liang, J. Zhou, P. Tang, P. Stothard, Y. Zhou, J. Cruz, C. MacDonell, G. Lin, P. Lu, D.S. Wishart, PROSESS: a protein structure evaluation suite and server, Nucleic Acids Res. 38 (Suppl. 2) (2010) W633−W640.

[25] M. Wiederstein, M.J. Sippl, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, Nucleic Acids Res. 35 (Suppl. 2) (2007) W407−W410.

[26] M. Berjanskii, J. Zhou, Y. Liang, G. Lin, D.S. Wishart, Resolution-by-proxy: a simple measure for assessing and comparing the overall quality of NMR protein structures, J. Biomol. NMR 53 (3) (2012) 167−180.

[27] Studio D. Version 3.0, Accelrys Software Inc, San Diego, 2010.

[28] V.K. Garg, H. Avashthi, A. Tiwari, P.A. Jain, P.W. Ramkete, A.M. Kayastha, V.K. Singh, MFPPI−Multi FASTA ProtParam interface, Bioinformation 12 (2) (2016) 74.

[29] L. Jendele, R. Krivak, P. Skoda, M. Novotny, D. Hoksza, PrankWeb: a web server for ligand binding site prediction and visualization, Nucleic Acids Res. 47 (W1) (2019) W345–W349.

[30] R. Krivák, D. Hoksza, P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure, J. Cheminf. 10 (1) (2018) 39.

[31] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477.

[32] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerg. Artif. Intell. Appl. Comput. Eng. 160 (2007) 3–24.

[33] Q. Zhao, T. Hastie, Causal interpretations of black-box models, J. Bus. Econ. Stat. 1–0 (2019).

[34] R. Cipollone, P. Ascenzi, P. Tomao, F. Imperi, P. Visca, Enzymatic detoxification of cyanide: clues from *Pseudomonas aeruginosa* Rhodanese, J. Mol. Microbiol. Biotechnol. 15 (2–3) (2008) 199–211.

[35] O. Carugo, K. Djinović-Carugo, Half a century of Ramachandran plots, Acta Crystallogr. Sect. D Biol. Crystallogr. 69 (8) (2013) 1333–1341.

[36] H. Zhou, Y. Yang, H.B. Shen, Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features, Bioinformatics 33 (6) (2017) 843–853.

[37] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F.T. Heer, T.A. de Beer, C. Rempfer, L. Bordoli, R. Lepore, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (W1) (2018) W296–W303.

[38] C. Colovos, T.O. Yeates, Verification of protein structures: patterns of nonbonded atomic interactions, Protein Sci. 2 (9) (1993) 1511–1519.

[39] J. Ko, H. Park, L. Heo, C. Seok, GalaxyWEB server for protein structure prediction and refinement, Nucleic Acids Res. 40 (W1) (2012) W294–W297.

[40] D.E. Clark, What has virtual screening ever done for drug discovery? Expert Opin. Drug Discov. 3 (8) (2008) 841–851.

[41] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening, Methods 71 (2015) 58–63.

[42] A. Kumar, K.Y. Zhang, Hierarchical virtual screening approaches in small molecule drug discovery, Methods 71 (2015) 26–37.

[43] C.A. Lipinski, Poor aqueous solubility—an industry wide problem in drug discovery, Am. Pharm. Rev. 5 (3) (2002) 82–85.

[44] A.K. Ghose, T. Herbertz, R.L. Hudkins, B.D. Dorsey, J.P. Mallamo, Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery, ACS Chem. Neurosci. 3 (1) (2012) 50–68.

[45] T.I. Oprea, Current trends in lead discovery: are we looking for the appropriate properties? Mol. Divers. 5 (4) (2000) 199–208.

[46] M.P. Pollastri, Overview on the rule of five, Curr. Protoc. Pharmacol. 49 (1) (2010) 9–12.

[47] M.V. Varma, S. Khandavilli, Y. Ashokraj, A. Jain, A. Dhanikula, A. Sood, N.S. Thomas, O. Pillai, P. Sharma, R. Gandhi, S. Agrawal, Biopharmaceutic classification system: a scientific framework for pharmacokinetic optimization in drug research, Curr. Drug Metabol. 5 (5) (2004) 375–388.

[48] W.J. Egan, Predicting ADME properties in drug discovery, Drug Des. Struct. Ligand-Based Approaches (2010) 165–180.

[49] I. Muegge, Pharmacophore features of potential drugs, Chem. Euro. J. 8 (9) (2002) 1976–1981.

[50] J.B. Baell, L. Ferrins, H. Falk, G. Nikolakopoulos, PAINS: relevance to tool compound discovery and fragment-based screening, Aust. J. Chem. 66 (12) (2014) 1483–1494.

[51] R. Brenk, A. Schipani, D. James, A. Krasowski, I.H. Gilbert, J. Frearson, P.G. Wyatt, Lessons learnt from assembling screening libraries for drug discovery for neglected diseases, ChemMedChem 3 (3) (2008) 435−444.

[51a] B.K. Shoichet. No free energy lunch. Nature Biotechnol. 25 (10) (2007) 1109−1110.

[51b] A. S. El-Magboub, Computational Models for Drug Design and Delivery (Doctoral dissertation, University of Southern California), 2017.

[52] L. Schaeffer, The role of functional groups in drug−receptor interactions, in: The Practice of Medicinal Chemistry, Academic Press, 2008, pp. 464−480.

[53] M. Nishio, The CH/π hydrogen bond in chemistry. Conformation, supramolecules, optical resolution and interactions involving carbohydrates, Phys. Chem. Chem. Phys. 13 (31) (2011) 13873−13900.

[54] D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S.M. Dann, T.C. Savidge, Regulation of protein-ligand binding affinity by hydrogen bond pairing, Sci. Adv. 2 (3) (2016) e1501240.

[55] R. Patil, S. Das, A. Stanley, L. Yadav, A. Sudhakar, A.K. Varma, Optimized hydrophobic interactions and hydrogen bonding at the target-ligand interface leads the pathways of drug-designing, PLoS One 5 (8) (2010).

[56] W.M. Eldehna, S.M. Abou-Seri, A.M. El Kerdawy, R.R. Ayyad, A.M. Hamdy, H.A. Ghabbour, M.M. Ali, D.A. El Ella, Increasing the binding affinity of VEGFR-2 inhibitors by extending their hydrophobic interaction with the active site: design, synthesis and biological evaluation of 1-substituted-4-(4-methoxybenzyl) phthalazine derivatives, Eur. J. Med. Chem. 113 (2016) 50−62.

[57] T.J. Maher, D.A. Johnson, Receptors and drug action, in: Foye's Principles of Medicinal Chemistry, vol. 85, 2008.

[58] R.F. de Freitas, M. Schapira, A systematic analysis of atomic protein−ligand interactions in the PDB, Med. Chem. Comm. 8 (10) (2017) 1970−1981.

[59] T.J. Ritchie, S.J. Macdonald, J. Med. Chem. 57 (17) (2014) 7206−7215.

[60] R.D. Taylor, M. MacCoss, A.D. Lawson, J. Med. Chem. 57 (14) (2014) 5845−5859.

[61] M.O. Idris, A.A. Yekeen, O.S. Alakanse, O.A. Durojaye, J. Biomol. Struct. Dyn. (2020) 1−19.

[62] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (2017) 42717.