Indian Journal of Biochemistry & Biophysics
Vol. 59, September 2022, pp. 879-891
DOI: 10.56042/ijbb.v59i9.56888

Evaluation of predictive machine learning models for drug repurposing against delta variant of SARS-CoV-2 spike protein

Sudipta Dash[1]*, Ishani Ishani[2], Dibyajit Lahiri[3] & Moupriya Nag[3]

[1]Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur-721 302, West Bengal, India

[2]Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, Greater Noida-201 314, Uttar Pradesh, India

[3]Department of Biotechnology, University of Engineering and Management, Kolkata-700 156, West Bengal, India

Received 05 November 2021; *revised 26 August 2022*

Drug repurposing is a major approach used by researchers to tackle the COVID-19 pandemic which has been worsened by the current surge of delta variant in many countries. Though drugs like Remdesivir and Hydroxychloroquine have been repurposed, studies prove these drugs have insignificant effect in treatment. So, in this study, we use the already FDA approved database of 1615 drugs to apply semi-flexible and flexible molecular docking methods to calculate the docking scores and identify the best 20 potential inhibitors for our modelled delta variant spike protein RBD. Then, we calculate 2325 1-D and 2-D molecular descriptors and use machine-learning algorithms like K-Nearest Neighbor, Random Forest, Support Vector Machine and ensemble stacking method to build regression-based prediction models. We identify 15 best descriptors for the dataset all of which were found to be inversely correlated with ligand binding. With only these few descriptors, the models performed excellently with an area under curve (AUC) value of 0.952 in Regression Error Characteristic curve for ensemble stacking. Therefore, we comment that these 15 descriptors are the most important features for the binding of inhibitors to the spike protein and hence these should be studied properly in terms of drug repurposing and drug discovery.

Keywords: Drug repurposing, Machine learning, Molecular docking, Regression model, SARS-CoV-2

The recent outbreak of Coronavirus (COVID-19) is known to be caused by a recently discovered virus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). COVID-19, a novel coronavirus disease, was declared as a pandemic on March 11, 2020 by the World Health Organization (WHO). Globally, as of 19th September 2021, 221,648,869 confirmed cases of COVID-19 have been reported, including 46,97,099 deaths, notified to the WHO[1]. These numbers are rapidly rising worldwide, thereby creating an evolving emergency situation. SARS-CoV-2 virus undergoes mutation over time, like other viruses. The bulk of the changes have little to no impact on the virus's characteristics. Scientists keep track of all variants, but some may be classified as variants of interest (VOIs), concern, or high consequence because of how fast they spread, how severe their symptoms are, and how they all are cured. VOIs include the alpha (B.1.1.7), beta (B.1.351), gamma (P.1), and delta (B.1.617.2) strains[2]. According to reports, the Delta version is more than twice as infectious than previous variants.

Some evidences show that the Delta form causes more serious illness in unvaccinated people than prior variants[3].

The delta variant is defined by (G142D), 19R, 157del, 156del, R158G, T478K, L452R, D614G, D950N, P681R, mutations in the spike protein. Most of these mutations could have an effect on immune responses aimed towards the receptor binding protein's major antigenic areas (452 and 478), as well as the deletion of a portion of the N terminal domain (156 and 157). The P681R mutation alters an amino acid immediately next to the furin cleavage site, which is a crucial step in allowing the virus to enter human cells and therefore increasing viral infectivity[4,5].

The receptor-binding domain (RBD) of SARS-CoV-2's spike protein interacts with the human angiotensin-converting enzyme 2 (ACE2) receptor to initiate SARS-CoV-2 infection[6]. The medications hydroxychloroquine (HCQ) and remdesivir (RDV), which were previously used for different reasons, are being repurposed to treat COVID-19[7]. A recent study

*Correspondence:
E-mail: sudipta.sd76@gmail.com
Suppl. Data available on respective page of NOPR

established the use of drugs used against hepatitis C virus like velpatasvir, vitamin D derivatives like ergocalciferol, and drugs used to prevent asthma symptoms like zafirlukast as strong potential inhibitors of S-protein-hACE2-binding using rigid docking and molecular dynamics simulations[8]. Another study utilized the machine learning model based on the Naive Bayes algorithm for the repurposing of therapeutic agents for the treatment of COVID-19[9].

Here, we have performed docking studies and virtual screening of 1615 FDA approved drugs in order to find potential inhibitors which can be repurposed for the treatment of COVID-19. The repurposed drug compounds discovered in this study may be investigated in the lab for their effectiveness against S-protein binding to hACE2 and could lead to a quick and effective COVID-19 treatment. We have also built a predictive model using machine learning regression algorithms and determined 15 features that are capable of defining the correlation with the calculated docking scores in the best manner.

Materials and Methods

Protein 3D model generation

A partial surface glycoprotein of SARS-CoV-2 (GenBank ID: QUX03821.1) delta variant (B.1.617.2 lineage) was downloaded from the NCBI protein database[10]. This protein sequence was uploaded in six 3D protein modelling servers which include HHpred, IntFOLD, RaptorX, SPARKS-X, PHYRE2, and PSIPRED (Fig. 1).

HHpred server helps in detection of homology between protein sequences and prediction of protein structure by the implementation of pairwise comparison of profile hidden Markov models (HMMs)[11]. The FASTA sequence was first uploaded in the HHpred server[12,13] where PDB_mmCIF70_13_Jul was selected as the structural/domain database, Vir_SARS-CoV-2_31_Mar_2020 was selected as the proteome, and the job was submitted. One model was generated by selecting the top 5 hits as a template while another model was generated by selecting all the hits which had 100% identity (top 9 hits) as a template.

Next, the IntFOLD Integrated Protein Structure and Function Prediction Server (Version 5.0)[14,15] was used for 3D model generation. In this server, the single letter code of the target protein sequence was simply entered in the query box and submitted for prediction. Five 3D protein models were generated using this server.

The RaptorX[16–18] web server predicts the 3D protein structure using an ultra-deep convolutional residual neural network either from a primary sequence or a multiple sequence alignment. This server also generated five tertiary structure models for the input protein sequence.

SPARKS-X[19] web server recognizes the protein fold of the input target sequence to generate tertiary structure models. This server generated four tertiary structure models for the SARS-CoV-2 protein sequence.

Next, the Phyre2 Protein Homology/analog Y Recognition Engine V 2.0[20] was used for 3D model

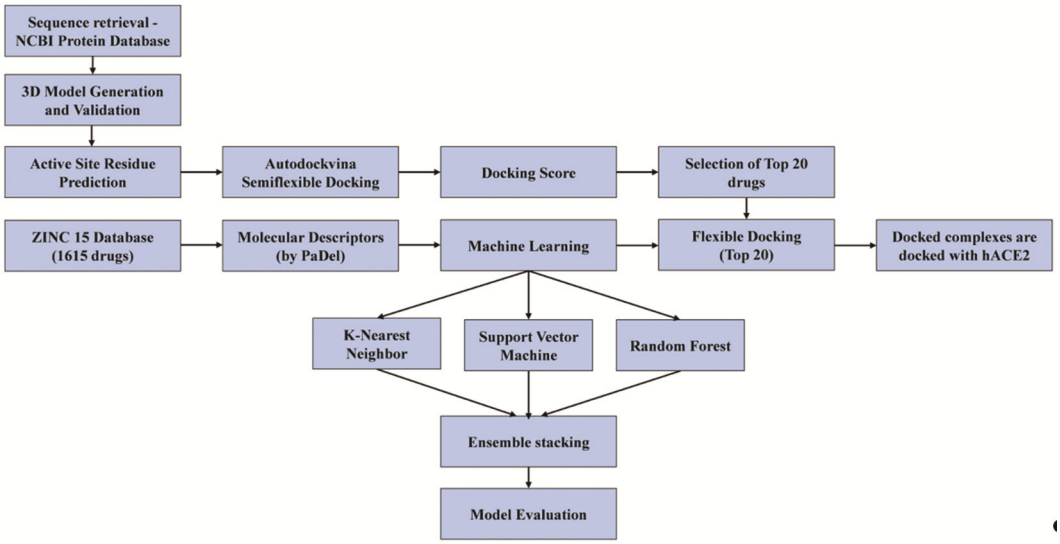


Fig. 1 — The overall scheme of work followed in this paper

generation. Two models were generated through this server. For the first model, normal modelling mode was selected while extensive modelling mode was selected for the second model.

Finally, the DMP fold 1.0 option was selected on the PSIPRED[21,22] home page to generate tertiary structure models using the sequence data. DMP fold uses deep learning to predict interatomic distances bounds, torsion angles, and the network of hydrogen bonds on the main chain, which are used to construct models in an iterative manner. This server provided two tertiary structure models for the SARS-CoV-2 protein sequence.

Protein structure validation

All the protein tertiary structure models generated from the above mentioned servers were further validated using the QMEAN4[23–25], PROCHECK[26,27], ProSA[28,29], and PROQ[30] servers.

Active site residue prediction

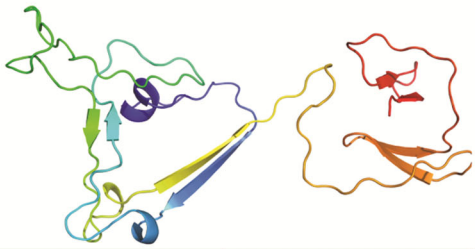
The crystal structure of SARS-CoV-2 spike receptor-binding domain bound with human Angiotensin Converting Enzyme 2 (hACE2) was retrieved from RCSB PDB (PDB ID: 6M0J). This crystal structure was composed of two chains: the chain A consisting of the hACE2 protein and chain E consisting of the SARS-CoV-2 spike protein (Fig. 2A). The chain E of 6M0J was superimposed with the HHpred5 tertiary model predicted for the spike protein using PyMOL[31].

To find the interface residues in the PDB crystal structure of 6M0J, first the file obtained from RCSB PDB was processed by removing the water molecules and other unnecessary ligands. Then, interface Residues. py python script was run in PyMOL which provided us with a list of interface residues (Fig. 2B & C) on chain A and chain E of 6M0J.

Preparation of drug molecules

1615 available FDA approved drug molecules were retrieved from the ZINC15[32] database in .sdf format. Two existing potential SARS-CoV-2 drug molecules namely, remdesivir and hydroxychloroquine were downloaded from DrugBank and included in our drug molecule set for comparison. Open babel 3.0.0[33] software was used to convert all the drugs from .sdf format to .pdb files. These .pdb files were then converted into an Autodock-specific coordinate file format, known as PDBQT format. This conversion was carried out using a python script, prepare_ligand.py, which can be found in the MGL Tools package. PDBQT is identical to PDB, except it also contains AutoDock 4 (AD4) atom types ('T') and partial charges ('Q'). The atoms of the drug molecules must be given the proper AutoDock atom types,

A



B

Interacting residues on chain A hACE2 protein:	Interacting residues on chain E spike protein:
SER-19, GLN-24, THR-27, PHE-28, ASP-30, LYS-30, HIS-34, GLU-35, GLU-37, ASP-38, TYR-41, GLN-42, LEU-45, LEU-79, MET-82, TYR-83, GLN-325, GLY-326, ASN-330, LYS-353, GLY-354, ASP-355, ARG-357, ALA-356, ARG-393	LYS-417, VAL-445, GLY-446, TYR-449, TYR-453, LEU-455, PHE-456, TYR-473, ALA-475, GLY-476, GLU-484, PHE-486 , ASN-487, TYR-489 , GLN-493, GLY-496, GLN-498, THR-500 , ASN-501, GLY-502, VAL-503, TYR-505

C

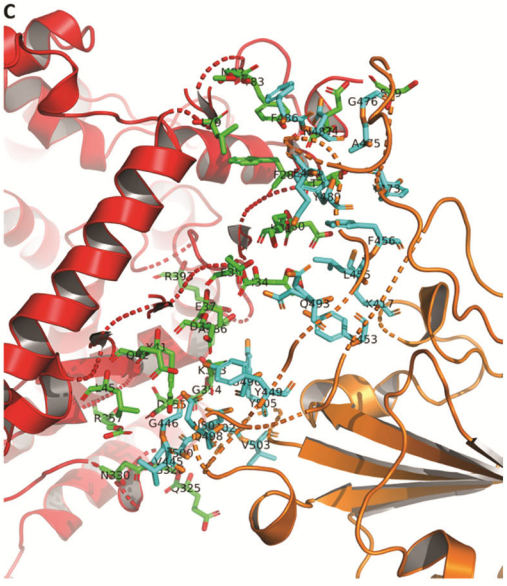


Fig 2 — (A) 3D structure of the best model (HHpred5 model) as concluded from the validation scores for the spike protein of delta variant of SARS-CoV-2; and (B) List of interacting residues on chain A (hACE2) and chain E (spike) of crystal structure (6M0J) determined using python script in PyMOL. C. Co-crystallised structure of hACE2 protein (chain A - red) and SARS-CoV-2 Spike protein (chain E - orange) (PDB ID: 6M0J). Interacting residues of the respective proteins are shown in stick representation and are colored in green and cyan in hACE2 and SARS-CoV-2 Spike protein, respectively.

Gasteiger charges must be added if required, non-polar hydrogens must be merged, aromatic carbons must be detected, and the 'torsion tree' must be set up.

Preparation of target protein molecule

To prepare the HHpred5 protein molecule, first the water molecules were removed from the tertiary structure of protein as the water molecules may interfere with the docking process. Further, polar hydrogens and 6.0 Kollman charges were added to the protein structure, such that the charge was equally distributed across the protein using AutoDock Tools-1.5.6[34,35]. To guarantee compatibility with Autodock vina while conducting molecular docking simulations, the generated protein structure was stored in .pdbqt format.

Docking grid generation

The interface residues of spike protein of 6M0J that interact with the hACE2 protein were chosen as the active binding site for drug molecules. A large grid box enclosing all the interface residues of the spike protein was created to cover all of the potential ligand binding sites in the protein structure. Grid option of AutoDock Tools was used to create a grid box of 56 × 60 × 112 with a spacing of 0.375. The binding pocket was set at x = 78.999, y = −17.378, and z = −13.748 (Suppl. Fig. 1B). The grid box defines a space which AutoDock explores to find the best possible conformation of any given ligand and determine the highest binding affinity. The values for exhaustiveness, number of modes and energy range were set at 12, 9 and 3, respectively.

Semi-flexible and flexible docking

First, semi-flexible docking was performed on all the drug molecules where the protein target was kept rigid and the drug molecules as flexible to attain a degree of freedom torsions bridged by the rotational parameter. Semi-flexible docking was performed three times for all drug molecules using AutoDock Vina[36,37] and Dmean (mean of the docking values) values were calculated. Based on the Dmean values, top 20 drug molecules with the lowest binding energy and therefore the highest binding affinity, along with hydroxychloroquine and Remdesivir were chosen for flexible docking. In flexible docking, both the ligand and the receptor molecule are treated as flexible and allowed to attain different conformations while docking. For flexible docking, the previously determined interface residues of the spike protein were chosen as flexible residues for the HHpred5 receptor molecule using AutoDock Tools. AutoDock Vina was used to perform flexible docking three times for the chosen drug molecules using docking parameters same as that used for semi-flexible docking. All the residues interacting with each of the top 20 drugs and the 2 potential drugs in the output complexes obtained from flexible docking, and the kind of interactions occurring between the ligand and the receptor molecule were determined using Discovery Studio Visualizer[38].

Protein-protein docking studies

All the complexes obtained after flexible docking were further docked with hACE2 protein (chain A of 6M0J) using the ZDOCK[39] server, one by one. A similar protein-protein docking was also performed between hACE2 (chain A of 6M0J) and HHpred5 protein in order to analyse the effect of the drug molecules on the binding efficiency of hACE2 and spike protein (Suppl. Fig. 2)[40,41].

Molecular descriptor calculation and dataset preparation

Molecular descriptors are the defining mathematical representations of the transformed chemical properties of the molecules. With the help of PaDel (version 2.21) software[42], all the 1D and 2D descriptors were calculated. PubChem fingerprinter was used to calculate the molecular fingerprints. All the calculated 2325 descriptors and fingerprints were then cleaned and processed by python (version 3.8.8) programming. A 90% cut-off was used to identify the useful descriptors for the dataset (*i.e.* the descriptors which showed zero or null values for more than 90% of the molecules were removed). Rescaling (whitening) of the dataset was done by scikit-learn object StandardScaler to standardize the distributions by following formulae:

$$z = \frac{(x - \mu)}{\sigma}$$

where, z is the standardized value of x. μ and σ are the mean and the standard deviation of the descriptor, respectively. This standardized data was fed as features to train the machine learning algorithms.

Application of machine learning algorithms

Three well established machine learning regression algorithms - K-nearest neighbour (KNN)[43], Random Forest (RF)[44] and Support Vector Machine (SVM)[45] were used to build up the predictive models. To detect the best hyperparameters that enhance the learning process, hyperparameter tuning was done for all the used algorithms by GridSearchCV[46]. For KNN,

'n_neighbors', 'weights' and 'algorithm' were the selected hyperparameters for tuning. 'N_estimators' and 'max_features' were used for Random Forest. 'C', 'gamma' and 'kernel' were tuned for SVM. Train-test-split method was used for dividing the dataset into training and test set in 70:30 ratio. All the models were generated initially with 5-fold cross-validation based on the scoring function of mean absolute error. The best detected parameters were used for training the model and making predictions on the test set.

Ensemble method

Ensemble learning (stacking) method was used to combine the three base models and generate a meta-model that increases the model-score along with reducing the bias. Scikit-learn's 'Stacking Regressor' API was used for ensemble learning. 'Voting Regressor' was used as an ensemble meta-estimator to build the stacked model based on the best hyperparameters of the three selected base models.

Feature selection and new model generation

To detect only the most important features out of the large set of descriptors (features) used for building the initial models, feature selection method of 'Select K Best' API was used. Top 15 features out of 1350 were selected by this method based on the feature-scores (F-scores) (Suppl. Fig. 3). With only these 15 features, we repeated the application of algorithms and ensemble stacking steps again to build new regression models and evaluate their performance with reduced features. However, 10-fold cross-validation was used at this stage to reduce further bias.

Model evaluation

For the assessment of the general merits of the regression models, Regression Error Characteristic (REC) curves were generated. The Area Under Curve (AUC) values calculated from the REC-curves can be used to study the overall performance of the models[47]. Briefly in our work, we used SlickML (version 0.1.3) machine learning library to plot the regression metrics from the test data predicted by the models[48].

Results and Discussions

3D model generation and validation

Due to the unavailability of a crystal structure of the delta variant of the spike protein, the tertiary protein structure was modelled using 5 web servers. In total, 20 models were generated using these servers which were further validated using 4 scoring schemes. The scores obtained from these servers for each of the 20 models are given in (Table 1).

The cut-off criteria used to evaluate a good quality model were: QMEAN4 value should be close to zero (Suppl. Table 1), more than 90% of the residues should be present in the most favourable region in the Ramachandran plot generated using PROCHECK (Suppl. Table 2), LG score should be greater than 4 and MaxSub score less than −0.8 for the results obtained from PROQ server (Supplementary Table 3), and Z-Score values obtained from ProSA should be close to zero (Suppl. Table 4).

Many models were able to qualify the cutoff criteria set for each of the validation scores, but only the results of HHpred5 and IntFOLD4 remain consistently good in all the validation servers. The structure of these two models were visualized using PyMOL and were found to be very similar. Also, the overall validation scores of HHpred5 are better than that of IntFOLD4. Therefore, based on these criteria the HHpred (5 hits) or HHpred5 model (Fig. 2A) was found to be the best and thereby, it was selected for subsequent analysis.

Further, the RMSD value was obtained by superimposing the spike protein of the crystal structure (chain A) and the modelled spike protein. The RMSD value was found to be 0.592 which indicates that our model HHpred5 is very similar to the crystal structure of spike protein in 6M0J.

Semiflexible and flexible docking

1615 FDA approved drugs from ZINC15 database along with hydroxychloroquine and remdesivir from DrugBank database were docked with the HHpred5 receptor molecule using AutoDock vina. A lower docking score represents stronger binding efficiency

Table 1 — HHpred5 model validation results obtained from QMEAN4. PROCHECK, PROQ, and ProSA servers

Model	QMEAN4	PROCHECK Ramachandran Plot analysis % of residues in				PROQ		ProSA
		Most Favoured Region	Additional Allowed Region	Generously Allowed Region	Disallowed Regions	LGscore	MaxSub	Z-Score
HHpred5	−1.62	90.7	9.3	0	0	11.177	−0.83	−1.69

of the ligand with the spike protein. To validate the docking results, all the drugs were docked three times while keeping the parameters constant for all the trials. Any deviation in the docking scores across the three trials were noted (Fig. 3A). ZINC000169289388 had the lowest Dmean score of −12.6 and no deviation was observed across the three trials. It was followed by ZINC000096006020 which had a Dmean score of −12.13333333 and a standard deviation of 0.152752523. All our potential drugs have a Dmean score less than −10.4. Whereas, hydroxychloroquine and remdesivir have a Dmean score of −4.97 and −8.1, respectively, which are relatively higher than any of our selected drugs (Table 2). This indicates that hydroxychloroquine and remdesivir have a lower binding efficiency to the spike protein when compared to the top 20 drugs mentioned in the list[37].

Next, flexible docking was performed for these top 20 drugs along with hydroxychloroquine and remdesivir using AutoDock vina. Again, to validate the results of flexible docking each drug molecule was docked three times. The deviation in the docking scores across the three trials were noted. The results of flexible docking and the ranking of the drugs based on FDmean remain consistent with that obtained from

Table 2 — Dmean (mean of rigid docking scores), Dstd (standard deviation in rigid docking scores), FDmean (mean of flexible docking scores), and FDstd (standard deviation in flexible docking scores) values obtained after performing rigid docking and flexible docking for the top 20 drugs

ZINC_ID	Dmean	Dstd	FDmean	FDstd
ZINC000169289388	−12.6	0	−12.5333	0.04714
ZINC000096006020	−12.133	0.153	−12.0667	0.329983
ZINC000085537053	−11.933	0.289	−11.6667	0.094281
ZINC000169621228	−11.567	0.058	−11.4667	0.124722
ZINC000052955754	−11.5	0	−11.4333	0.094281
ZINC000203757351	−11.5	0	−11.4	0.141421
ZINC000169621219	−11.3	0.346	−11.3667	0.169967
ZINC000169289767	−11.167	0.231	−11.3333	0.124722
ZINC000003978005	−11	0	−11.3	0.141421
ZINC000003932831	−11	0	−11.2667	0.188562
ZINC000085536932	−10.9	0.436	−11.2333	0.124722
ZINC000253630390	−10.7	2.18E-15	−11.0333	0.402768
ZINC000012503187	−10.6	0	−10.9	0.496655
ZINC000164528615	−10.6	0.436	−10.6333	0.094281
ZINC000169344691	−10.567	0.666	−10.5667	0.169967
ZINC000253387843	−10.5	1.039	−10.4667	0.04714
ZINC000242548690	−10.467	0.115	−10.4	0.08165
ZINC000169621220	−10.4	0	−10.4	0.08165
ZINC000053683151	−10.4	0	−10.3667	0.04714
ZINC000008220909	−10.4	0	−10.3	0.141421
HCQ	−4.967	0.125	−5.1	0.294392
Remdesivir	−8.1	0	−8.16667	0.169967

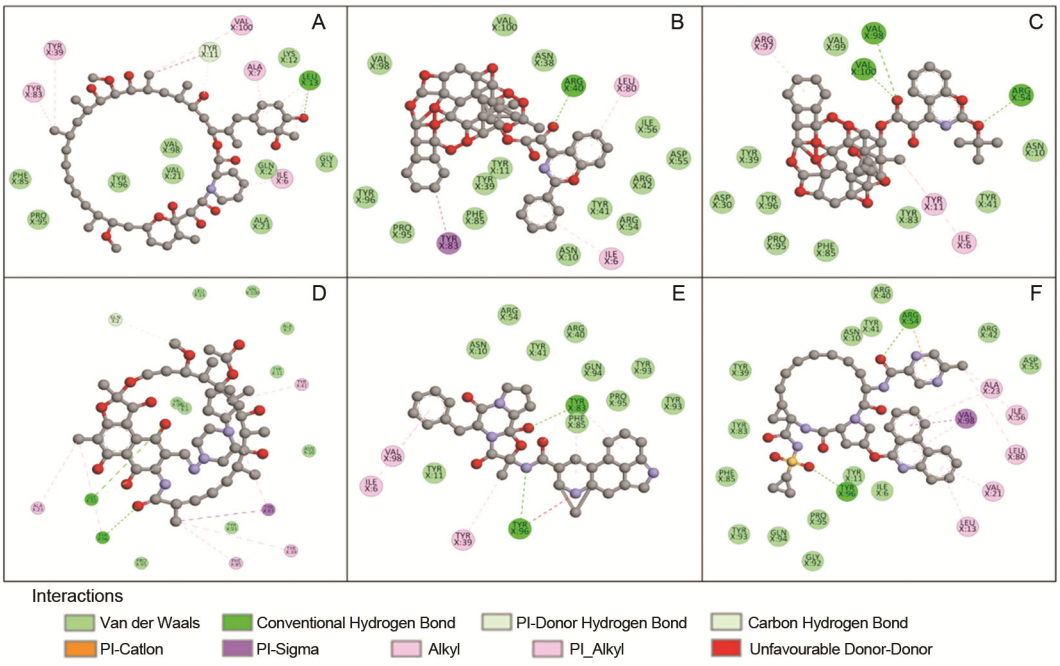


Fig. 3 — The scatter plot at the left represents percent change in the standard deviation of docking score with respect to mean score vs. mean docking score. The red dots depict the value of top 20 drugs while the blue dots represent the values of the rest of the drugs. Red dots which are closer to zero represent drugs that have minimum deviation in their docking scores across the three trials; Heatmap (at the right) indicating the types and number of non-covalent interactions between the selected drug molecules and SARS-CoV-2 Spike protein residues. As a residue might have several interactions, and only the important residues for the hACE2 protein are displayed in Fig. B, the number of interactions per drug molecule can be greater than the number of residues shown in Fig. B

Dmean scores. Here, ZINC000169289388 had the lowest FDmean score of −12.5333 with a small standard deviation of 0.04714. The second-best potential drug on our list was ZINC000096006020 which had a FDmean score of −12.0667 and a standard deviation of 0.329983. Again, hydroxychloroquine and remdesivir which have a FDmean score of −5.1 and −8.17, respectively, have a poor binding efficiency to the spike protein when compared to the top 20 drugs in our list (Table 2). This conclusion is consistent with the results of some previous research studies which indicate that the repurposed drugs, Remdesivir and hydroxychloroquine, show no significant or non-significant impact on the survival of patients infected with SARS-CoV-2[49,50].

The non-covalent bonds and interactions between the spike protein and the top 20 drugs with highest binding efficiency were visualized using Discovery Studio Visualizer (Suppl. Fig. 2A-P). All the drugs reported in the list had 1 or more hydrogen bonds and 7 or more van der waals interactions (Fig. 3B). The top two drugs, ZINC000169289388 (Fig. 4A) and ZINC000096006020 (Fig. 4B), which had the highest binding efficiency with the spike protein based on Dmean and FDmean scores were observed to form a total of 18 bonds each with the spike protein receptor molecule. ZINC000169289767 which has the 8th best docking score was observed to form the maximum number of bonds with the spike protein. ZINC000169289767 formed 27 bonds which consisted of 3 conventional hydrogen bonds with the residues GLY-1, LEU-13, and VAL-98, 1 carbon hydrogen bond with ASP-13, 1 pi-pi T shaped interaction with TYR-39, 4 alkyl interactions with ILE-6, TYR-11, and VAL-21, 3 pi-alkyl interactions with the residues VAL-21, and LEU-13, 3 salt bridges with the residues GLY-1, ARG-54, and ARG-97, and 12 van der waals interactions. The drugs hydroxychloroquine and remdesivir formed 15 and 22 bonds, respectively, with the spike protein (Suppl. Fig. 2O & P). The bonds formed by hydroxychloroquine consisted of 0 hydrogen bonds, 2 alkyl interactions with ILE-6 and PRO-95, 3 pi-alkyl interactions with the residues ILE-6, TYR-39, and PHE-85, 1 halogen bond with ASP-30, and 9 Van der Waals interactions. On the other hand, the bonds formed by remdesivir consisted of 1 conventional hydrogen bond with ARG-54, 1 carbon hydrogen bond with ARG-40, 1 pi-sigma interaction with TYR-11, 3 alkyl interactions with the residues VAL-98, and LEU-80, 3 pi-alkyl interactions with the residues TYR-11, TYR-83, PHE-85, and 13 Van der Waals interactions.

Protein-protein docking

In order to determine if our reported potential drugs are capable of inhibiting the binding or reducing the binding efficiency between the spike protein and hACE2 protein, protein-protein docking was performed using ZDOCK. ZDOCK provides a score to indicate the binding efficiency between any two proteins. A higher ZDOCK score indicates better binding efficiency[39,51]. First, a protein-protein docking was performed between the hACE2 protein (chain A of 6M0J) and the modelled spike protein (HHpred5). This was performed to determine the binding efficiency of spike protein with hACE2 protein in the absence of any drug molecule. Then, the output complexes obtained after flexible docking of the ligands with the spike protein were docked individually keeping the hACE2 protein as receptor.

Based on the results obtained after protein-protein docking (Table 3), we inferred that ZINC000169289388 which was observed to have the best binding efficiency with spike protein based on the docking scores, is in contrast playing a very minor

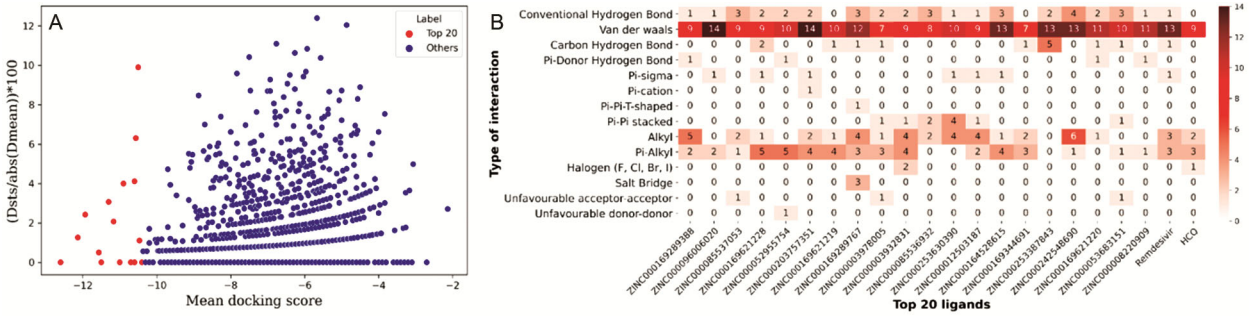


Fig 4 — 2D interaction diagram of top 6 potential drugs that can inhibit the binding of spike protein and hACE2 protein. (A) ZINC000169289388, (B) ZINC000096006020, (C) ZINC000085537053, (D) ZINC000169621228, (E) ZINC000052955754, (F) ZINC000203757351.

Table 3 — Scores obtained on Spike Protein docking with hACE2 protein in the presence and absence of the selected drugs performed using ZDOCK

Protein 1 (Receptor)	Protein 2 (Spike protein - Ligand complex)	ZDOCK Score
hACE2 (6M0J:A)	HHPRED	2003.497
hACE2 (6M0J:A)	HHPRED + ZINC000085536932	2024.133
hACE2 (6M0J:A)	HHPRED + ZINC000169289388	2001.068
hACE2 (6M0J:A)	HHPRED + ZINC000203757351	1992.046
hACE2 (6M0J:A)	HHPRED + ZINC000012503187	1982.698
hACE2 (6M0J:A)	HHPRED + ZINC000003932831	1980.457
hACE2 (6M0J:A)	HHPRED + ZINC000169621219	1927.258
hACE2 (6M0J:A)	HHPRED + ZINC000053683151	1914.12
hACE2 (6M0J:A)	HHPRED + ZINC000169289767	1906.348
hACE2 (6M0J:A)	HHPRED + ZINC000169344691	1891.584
hACE2 (6M0J:A)	HHPRED + ZINC000242548690	1820.48
hACE2 (6M0J:A)	HHPRED + ZINC000008220909	1794.348
hACE2 (6M0J:A)	HHPRED + ZINC000085537053	1776.154
hACE2 (6M0J:A)	HHPRED + ZINC000052955754	1699.153
hACE2 (6M0J:A)	HHPRED + ZINC000253387843	1659.921
hACE2 (6M0J:A)	HHPRED + ZINC000169621228	1354.333
hACE2 (6M0J:A)	HHPRED + ZINC000169621220	1236.522
hACE2 (6M0J:A)	HHPRED + ZINC000164528615	1203.584
hACE2 (6M0J:A)	HHPRED + ZINC000096006020	1163.582
hACE2 (6M0J:A)	HHPRED + ZINC000253630390	1115.343
hACE2 (6M0J:A)	HHPRED + ZINC000003978005	1092.568
hACE2 (6M0J:A)	HHPRED + HCQ	1815.96
hACE2 (6M0J:A)	HHPRED + Remdesivir	1874.678

role in inhibiting the interaction of spike protein with the hACE2 protein as it is able to reduce the binding affinity of spike protein and hACE2 protein from 2003.497 to 2001.068, which is a minute change. On the other hand, ZINC000096006020 which has the second highest binding efficiency with spike protein, is also capable of inhibiting the interaction of spike protein and hACE2 protein to a great extent.

It reduces the binding affinity of spike protein with hACE2 from 2003.497 to 1163.582. Also, ZINC000003978005 which ranks 9th on the list of drugs with highest binding efficiency with the spike protein is found to be capable of inhibiting the interaction between spike protein and hACE2 protein to the largest extent, that is, a reduction of 54.5% is observed in the binding score as it goes down from 2003.497 to 1092.568. The ZDOCK score obtained for hydroxychloroquine and remdesivir are 1815.96 and 1874.678 which indicates that these drugs are relatively less efficient in inhibiting the interaction between spike protein and hACE2 protein when compared to most of the drugs listed in the (Table 3).

Application of machine learning algorithms

Regression based prediction models were built using the selected methods (KNN, RF, SVM, Ensemble stacking) on the docking scores generated by Autodock VINA and molecular descriptors calculated by PaDel. The cross-validated mean absolute errors (MAE) were used as the parameters for comparing the performance of each algorithm on the training set. When the entire dataset was used, all the algorithms were found to be performing well with MAE ranging from −0.3 to −0.4 (Fig. 5A & B). Then the top 15 features were selected by Select K Best f-statistics for regression to evaluate the best features. Briefly, the F-score can be calculated by following formula:

$$\text{F-score} = \frac{\left(\frac{RSS_1 - RSS_2}{k_2 - k_1}\right)}{\left(\frac{RSS_2}{n - k_2}\right)}$$

where, residual sum of squares of the two compared models are represented as RSS_1 and RSS_2, respectively. k_1 and k_2 are the number of free fitting parameters the first and second model have, and n is the total number of data samples.

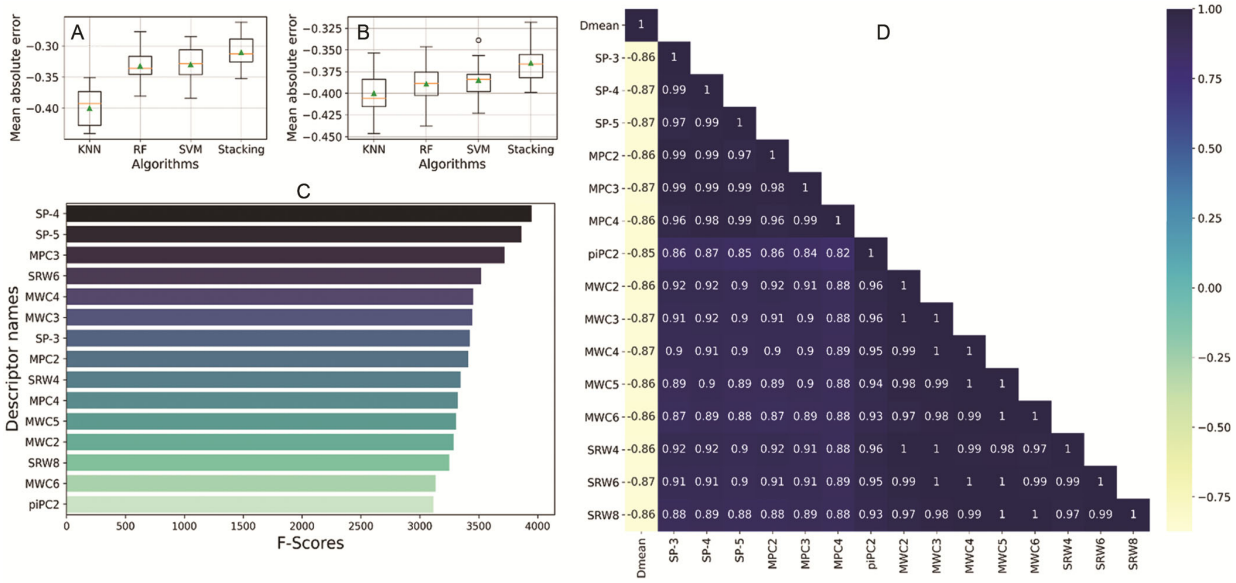


Fig. 5 — Performance analysis of the selected algorithms on the training set in presence of all the calculated descriptors (A) and in presence of only top 15 selected descriptors; (B) The bar plot in figure; (C) represents the top 15 selected features in terms of F-score. The heatmap in figure; and (D) represents the correlation between the descriptors and mean docking score (Dmean)

These 15 features are the most 'important features' for the regression analysis of our dataset. In other words, it can also be stated that these 15 features have the best-defined correlation with our calculated docking scores. The relative contribution of the top 15 descriptors is given in the bar plot (Fig. 5C). The Pearson coefficient analysis of these features shows that there is a high correlation themselves. However, all of them are consistently negatively correlated with Dmean (Fig. 5D).

From Table 4, it is clear that the KNN algorithm is least affected by feature reduction. Ensemble stacking method, as expected, performed better than the three algorithms used in the experiment having an R2 value of 0.915 with all the descriptors and 0.88 with only 15 descriptors.

Evaluation of the regression models

Evaluation of the predicted models was done by REC curves which provide statistics similar to the receiver operating characteristic (ROC) curves for classification problems. It helps in understanding the performance of the generated models across a wide range of possible errors. The cumulative distribution function of the error of the models (Fig. 6), shows that all the regressors are excellent and almost equally fit on the dataset with ensemble stacking method performing comparatively better having an area under curve (AUC) value of 0.952. The other calculated statistics and detailed evaluation results from

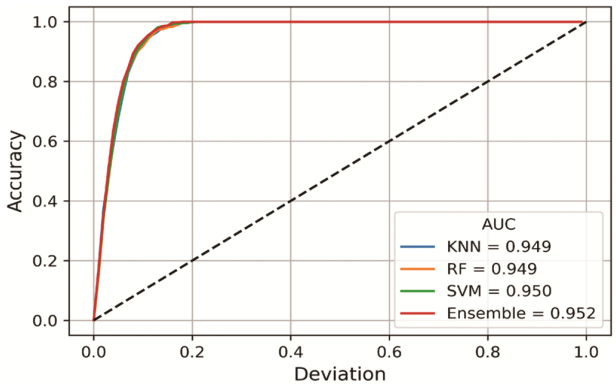


Fig. 6 — Evaluation of the models by Regression Error Characteristic (REC) Curve

Table 4 — Comparative evaluation of the algorithms between the use of all the descriptors and only top 15 descriptors

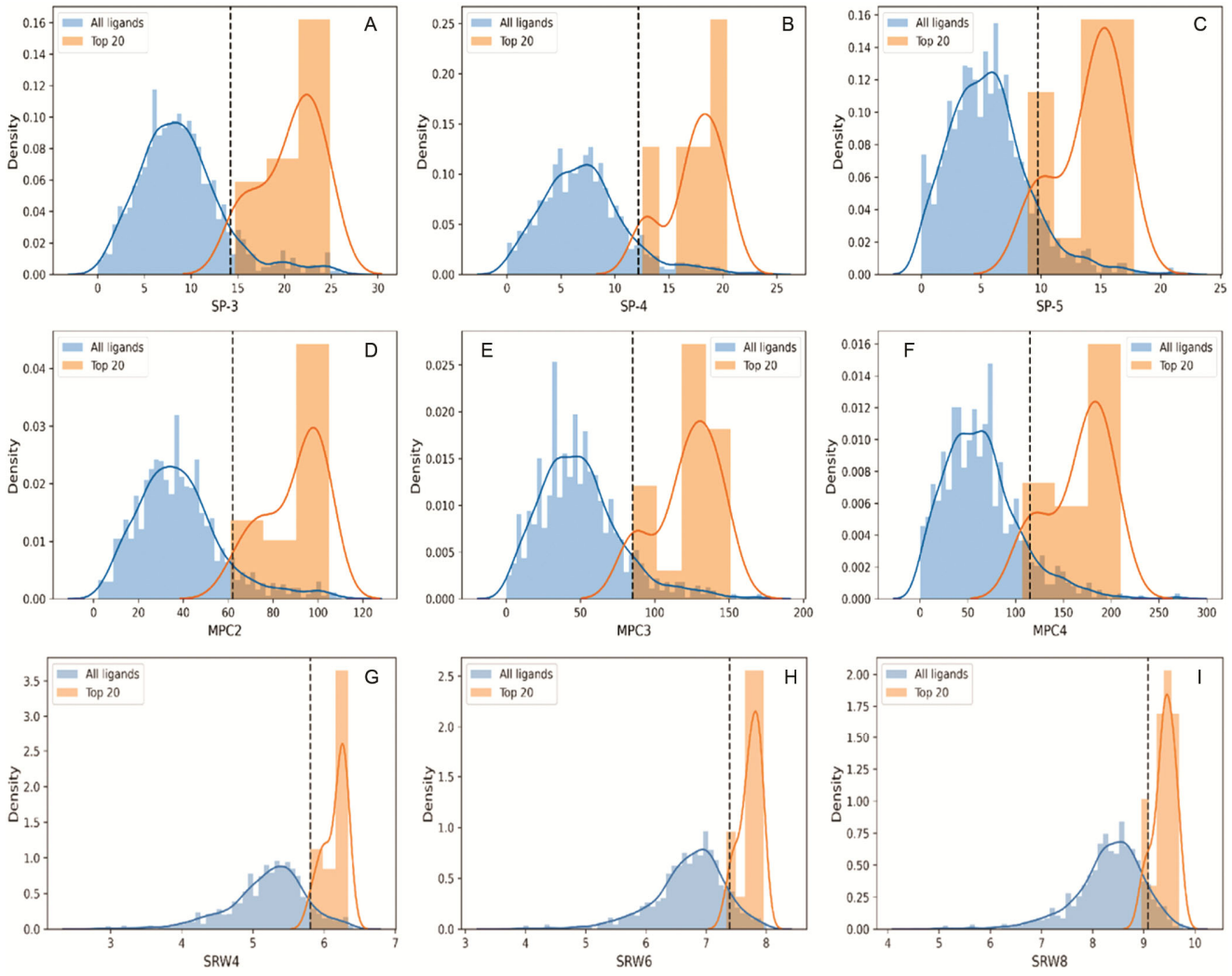
	KNN	RF	SVM	Ensemble
With all the descriptors				
Mean absolute error	0.37034	0.30945	0.30575	0.30311
Root Mean squared error	0.51281	0.42915	0.45328	0.41888
Median absolute error	0.26667	0.22325	0.20517	0.20864
Explain variance score	0.87312	0.91173	0.90174	0.91567
R2 score	0.87312	0.91114	0.90087	0.91534
With the top 15 descriptors				
Mean absolute error	0.3893	0.38347	0.38412	0.36493
Root Mean squared error	0.5353	0.53466	0.52615	0.49783
Median absolute error	0.26794	0.27463	0.27736	0.28044
Explain variance score	0.86208	0.86227	0.86661	0.88047
R2 score	0.86174	0.86207	0.86643	0.88042

SlickML are documented in the (Suppl. Table 5 and Figs 4-7).

Analysing top 15 descriptor properties

An analysis of the molecular descriptors was done for all the drug molecules used in this study. Based on the values of these important descriptors for the top 20 potential inhibitors in our list, we determined a possible range of values for which they can be associated with the inhibition property of the interaction between spike protein and hACE2 protein. Based on the boxplot (Suppl. Fig. 3), we can infer that a potential drug/inhibitor should have SP-3 value around 20, SP-4 value around 15, SP-5 value in the range of 115 to 140, MPC2, MPC3, MPC4, MWC3, MWC5, MWC6, SRW4, SRW6, SRW8 values around 10, piPC2 value in the range of 15-25, MWC2 value in the range of 75-100, and MWC4 value in the range of 140-190. Here, SP-3, SP-4, SP-5 descriptors indicate the chi path (simple and valence chi chain descriptors of the order 3,4, and 5, respectively), and piPC2 is the conventional bond order ID number of order 2 that define the molecular connectivity as stated by Kier and Hall[52]. MPC2, MP3, MPC4 descriptors represent the molecular path counts of the order 2,3, and 4, respectively, MWC2, MWC3, MWC4, MWC5, MWC6 indicate molecular walk count of the order 2,3,4,5, and 6, respectively, and SRW4, SRW6, SRW8 indicate the self-returning walk count of the order 4, 6, and 8, respectively[53].

We also analysed the distribution map for each of the important descriptors separately (Fig. 7). The best 20 inhibitors were overlaid on top of the entire drug set. Surprisingly, a distinct distribution pattern is observed. All the predicted potential inhibitors belong to the higher end in the normal distribution of all the ligands. Moreover, almost all the values for these



(*contd.*)

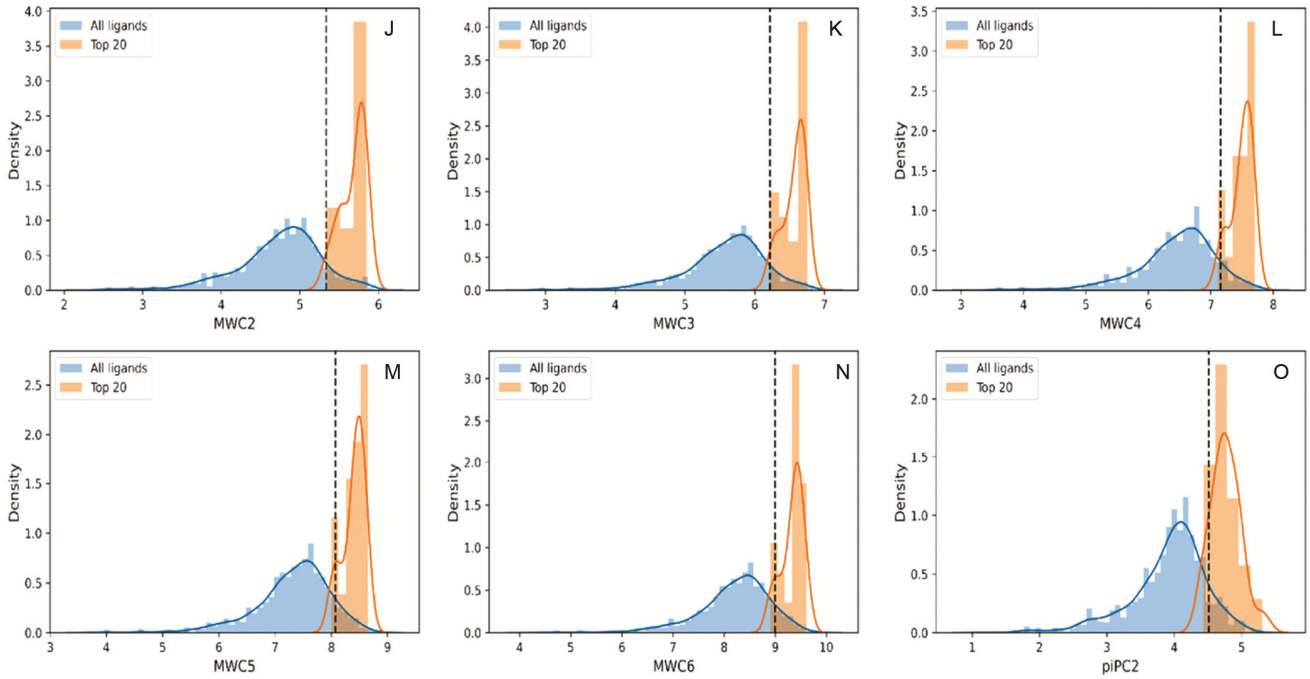


Fig. 7 — Comparative study of the distribution of values of the top 15 descriptors for 20 best ligands in the background of all the ligands. The dotted vertical line represents the 90th percentile of the distribution of each particular descriptor for the entire dataset

inhibitors were found to be in the 90th percentile range of the distribution. This apparently means that there is a precise relation between the properties of these important descriptors that contribute to the efficient binding of these molecules to the targeted delta strain spike model.

Conclusion

In conclusion, we present a promising virtual screening technique for discovering compounds that may limit and/or inhibit interactions between the human host and the SARS-CoV-2 virus. The most effective ligands, according to our hypothesis, are those that bind strongly to the spike protein at its binding region with hACE2 protein or at the interface of Spike protein-human ACE2 complex. To uncover drug compounds with such high binding affinities, we used a combination of machine learning and rigorous docking experiments in our high-throughput screening technique. Based on the Dmean and FDmean scores obtained for the top 20 drug molecules, we employ the validated machine learning model, after training with K-nearest neighbour, Random Forest and Support Vector Machine models, to search for top 15 'most important' descriptors (for *e.g.*, SP-3, MPC2) that portray a strong correlation with the docking results. The range of values determined for these descriptors can be used to determine a potential inhibitor of spike-hACE2 protein complex by searching through drug libraries with the same chemical features. We also evaluate our predictive model, created from the combination of three base models, using regression models. With our statistical analysis, we propose calculation and close observation of these descriptors, while performing virtual screening and lead generation, which could be highly beneficial especially in terms of drug repurposing considering the SARS-CoV-2 spike proteins.

Acknowledgement

We thank Anwesha Chatterjee for her help and support in molecular docking and data curation. We also acknowledge Indian Institute of Technology Kharagpur for accessibility of the tools for revision and drafting the paper.

Conflicts of interest

All authors declare no conflicts of interest.

References

1 WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data, https://COVID-19.who.int/table, (accessed 9 September 2021).

2 Tracking SARS-CoV-2 variants, https://www.who.int/emergencies/emergency-health-kits/trauma-emergency-surgery-kit-who-tesk-2019/tracking-SARS-CoV-2-variants, (accessed 9 September 2021).

3 CDC, Coronavirus Disease 2019 (COVID-19), https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html, (accessed 9 September 2021).

4 Bolze A, Cirulli ET, Luo S, White S, Cassens T, Jacobs S, Nguyen J, Ramirez JM, Sandoval E, Wang X, Wong D, Becker D, Laurent M, Lu JT, Isaksson M, Washington NL & Lee W, *Rapid displacement of SARS-CoV-2 variant B.1.1.7 by B.1.617.2 and P.1 in the United States*. *Medrxiv*, (2021).

5 Li B, Deng A, Li K, Hu Y, Li Z, Xiong Q, Liu Z, Guo Q, Zou L, Zhang H, Zhang M, Ouyang F, Su J, Su W, Xu J, Lin H, Sun J, Peng J, Jiang H, Zhou P, Hu T, Luo M, Zhang Y, Zheng H, Xiao J, Liu T, Che R, Zeng H, Zheng Z, Huang Y, Yu J, Yi L, Wu J, Chen J, Zhong H, Deng X, Kang M, Pybus OG, Hall M, Lythgoe KA, Li Y, Yuan J, He J & Lu J, Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *Nat Commun*, 13 (2021) 460.

6 XieY, Karki CB, Du D, Li H, Wang J, Sobitan A, Teng S, Tang Q & Li L, Spike Proteins of SARS-CoV and SARS-CoV-2 Utilize Different Mechanisms to Bind With Human ACE2. *Front Mol Biosci*, **7** (2020) 392.

7 Dang A, Bn V & Dang S, Hydroxychloroquine and Remdesivir in COVID-19: A critical analysis of recent events. *Indian J Med Ethics*, 5 (2020) 01.

8 Kumar NRP & Shetty NS, Machine learning approach for COVID-19 crisis using the clinical data. *Indian J Biochem Biophys*, 5 (2020) 57.

9 Mohapatra S, Nath P, Chatterjee M, Das N, Kalita D, Roy P & Satapathi S, Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking. *PLoS One*, 15 (2020) e0241543.

10 NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 44 (2016) D7.

11 Söding J, Biegert A & Lupas AN, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, 33 (2005) W244.

12 Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN & Alva V, Protein sequence analysis using the MPI bioinformatics toolkit. *Curr Protoc Bioinforma*, 72 (2020) e108.

13 Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN & Alva V, A completely reimplemented mpi bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol*, 430 (2018) 2237.

14 McGuffin LJ, Adiyaman R, Maghrabi AHA, Shuid AN, Brackenridge DA, Nealon JO & Philomina LS, IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res*, 47 (2019) W408.

15 McGuffin LJ, Shuid AN, Kempster R, Maghrabi AHA, Nealon JO, Salehe BR, Atkins JD & Roche DB, Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. *Proteins*, 86 (2018) 335.

16 Ma J, Wang S, Wang Z & Xu J, Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, 32 (2015) 3506.

17 Wang S, Sun S, Li Z, Zhang R & Xu J, Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*, 13 (2017) e1005324.

18 Wang S, Li W, Zhang R, Liu S & Xu J, CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res*, 44 (2016) W361.

19 Yang Y, Faraggi E, Zhao H & Zhou Y, Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinforma Oxf Engl*, 27 (2011) 2076.

20 Kelley LA, Mezulis S, Yates CM, Wass MN & Sternberg MJE, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10 (2015) 845.

21 Buchan DWA & Jones DT, The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res*, 47 (2019) W402.

22 Greener JG, Kandathil SM & Jones DT, Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun*, 10 (2019) 3977

23 Benkert P, Biasini M & Schwede T, Toward the estimation of the absolute quality of individual protein structure models, *Bioinformatics*, 27 (2011) 343.

24 Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J & and Schwede T, QMEAND is co-distance constraints applied on model quality estimation. *Bioinformatics*, 36 (2020) 1765.

25 Studer G, Biasini M & Schwede T, Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). *Bioinformatics*, 30 (2014) i505.

26 Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R & Thornton JM, AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 8 (1996) 477.

27 Laskowski RA, MacArthur MW, Moss DS & Thornton JM, PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*, 26 (1993) 283.

28 Sippl MJ, Recognition of errors in three-dimensional structures of proteins. *Proteins Struct Funct Genet*, 17 (1993) 355.

29 Wiederstein M & Sippl MJ, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, 35 (2007) W407.

30 Wallner B & Elofsson A, Can correct protein models be identified? *Protein Sci Publ Protein Soc*, 12 (2003) 1073.

31 Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 2.5, 2010.

32 Sterling T & Irwin JJ, ZINC 15 – Ligand Discovery for Everyone. *J Chem Inf Model*, 55 (2015) 2324.

33 O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T & Hutchison GR, Open Babel: An open chemical toolbox. *J Cheminformatics*, 3 (2011) 33.

34 Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS & Olson AJ, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30 (2009) 2785.

35 Sanner MF, Python: A programming language for software integration and development. *J Mol Graph Model*, 17 (1999) 57.

36 Eberhardt J, Santos-Martins D, Tillack AF & Forli S, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model*. 61 (2021) 3891.

37 Trott O & Olson AJ, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31 (2010) 455.

38 Discovery Studio Modeling Environment, Release 4.5. (BIOVIA, Dassault Systèmes, San Diego) 2015.

39 Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T & Weng Z, ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinforma Oxf Engl*, 30 (2014) 1771.

40 Pierce BG, Y. Hourai Y & Weng Z, Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PLoS One*, 6 (2011) e24657.

41 Pierce B, Tong W & Weng Z, M-ZDOCK: a grid-based approach for Cn symmetric multimer docking, *Bioinformatics*. 21 (2005) 1472.

42 Yap CW, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 32 (2011) 1466.

43 Mack YP, Local Properties of k-NN Regression Estimates. *J Algebr Discrete Methods*, 2 (1981) 311.

44 Breiman L, Random Forests. *Mach Learn*, 45 (2001) 5.

45 Mozer MC, Jordan MI & Petsche T, *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, (MIT Press) 1997.

46 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A & Cournapeau D, Scikit-learn: Machine Learning in Python. *J Mach Learn Res*, 12 (2011) 28250.

47 Bi J & Bennett KP, in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, (AAAI Press, Washington, DC, USA) 2003, 43.

48 *SlickML: Slick Machine Learning in Python*. SlickML, (2021).

49 Dyer O, Covid-19: Remdesivir has little or no impact on survival, WHO trial shows. *BMJ*, 371 (2020) m4057.

50 Romão VC, Cruz-Machado AR & Fonseca JE, No evidence so far on the protective effect of hydroxychloroquine to prevent COVID-19: comment by Joob and Wiwanitkit, *Ann Rheum Dis*, 80 (2021) e22.

51 Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R & Weng Z, Integrating statistical pair potentials into protein complex prediction. *Proteins Struct Funct Bioinforma*, 69 (2007) 511.

52 Kier LB & Hall LH, *Molecular connectivity in chemistry and drug research*, (Academic Press, New York) 1976.

53 Consonni V & Todeschini R, *Recent Advances in QSAR Studies: Methods and Applications*, eds. T. Puzyn, J. Leszczynski and M. T. Cronin, (Springer Netherlands, Dordrecht) 2010, 29.