Identification of Natural Compounds with Anti-SARS-CoV-2 Activity using Machine Learning, Molecular Docking and Molecular Dynamics Simulation Studies

Arifuzzaman[1Ψ], Mohadese Mohammadi[2Ψ], Fatema Hashem Rupa[3], Mohammad Firoz Khan[4*], Ridwan Bin Rashid[4] and Mohammad A. Rashid[5]

[1]Governance Innovation Unit (GIU), Prime Minister's Office, Dhaka-1215, Bangladesh
[2]Department of Chemistry, Tehran University, Tehran, Iran
[3]Upazila Health Complex, Kalmakanda, Netrokona, Bangladesh
[4]Computational Chemistry and Bioinformatics Laboratory, Department of Pharmacy
State University of Bangladesh, Dhaka- 1205, Bangladesh
[5]Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Dhaka
Dhaka- 1000, Bangladesh

(Received: May 2, 2022; Accepted: June 5, 2022; Published (Web): June 22, 2022)

ABSTRACT: The coronavirus pandemic of 2019 (COVID-19) has adversely affected public health and the socioeconomic situation worldwide. Although there is no therapeutic drug to treat COVID, several treatment options are being considered to alleviate symptoms. Hence, researches on prophylactic treatment for COVID are being encouraged. Searching natural products is a rational strategy since it has served as a valuable source of lead compounds in drug discovery. In this study, three machine learning approaches, including Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM), have been used to develop the classification model. The molecular docking was performed on AutoDock vina. Further, molecular dynamics (MD) simulation of the potential inhibitors was conducted using the AmberTools package. The accuracy for SVM, RF and GBM was found to be 60.45 %, 63.43 % and 64.93 %, respectively. Further, the model has demonstrated specificity range of 41.67 % to 50.00 % and sensitivity range of 74.32 % to 79.73 %. Application of the model on the NuBBE database, a repository of natural compounds, led us to identify 322 unique natural compounds, likely possessing anti-SARS-CoV-2activity. Further, molecular docking study has yielded three flavonoids and one lignoid compounds with comparable binding affinities to the standard compound. In addition, MD showed that these compounds form stable complexes with different magnitude of binding energy. The *in silico* investigations suggest that these four compounds likely demonstrate their anti-SARS-CoV-2activity by inhibiting the main protease enzyme. Our developed and validated *in silico* high-throughput investigations may assist in identifying and developing antiviral drug-like compounds from natural sources.

Key words: SARS-CoV-2, COVID-19, main protease, natural products, high throughput screening, machine learning, molecular docking, molecular dynamics simulation.

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been identified as the causative agent for the outbreak of coronavirus disease 2019 (COVID-19), the symptoms of which ranges from common cold, including fever, dry cough, fatigue, chest discomfort and in severe cases dyspnea and respiratory failure.[1-3] COVID-19 has been declared as a pandemic by the World Health Organization (WHO). Being highly transmissible, this disease has spread rapidly all over the world[4,5] and has infected at least 180 million individuals with a death toll of over 3.9 million as of June 2021. Usually, the children and young adults remain asymptomatic whereas older people or people with co-morbidities are prone to develop severe disease, respiratory failure and death.[6] The SARS-CoV-2is a (+)ss-RNA virus. The RNA encodes 4 structural

Correspondence to: Mohammad Firoz Khan
E-mail:firoz@sub.edu.bd
[Ψ]These authors contributed equally

Dhaka Univ. J. Pharm. Sci. **21**(1): 1-13 2022 (June)
DOI: https://doi.org/10.3329/dujps.v21i1.60391

[spike (S), envelop (E), membrane (M) and nucleocapsid (N)], 16non-structural and 9 accessory proteins (Figure 1).

SARS-CoV-2 spreads very rapidly and hence, demands an urgent need for treatment options until a vaccine, effective against all variants can be introduced. Manufacturing a vaccine that provides long term protection against this virus would be very challenging since it is RNA virus and has high mutation rate (Forni and Mantovani 2021). Besides, partially effective vaccine may not be efficient in controlling infectious diseases.[7,8] Therefore, the need for effective antiviral drugs against SARS-CoV-2 is urgent and should be prioritized. However, none of the synthetic or semisynthetic antiviral drugs are not devoid of side effects.[9] On the contrary, drugs from natural sources have little or no side effects, thus searching for new antiviral lead compounds from natural sources is still a rational strategy. Moreover, these leads can be modified and optimized to get more desirable effects.

In the current investigation, we have performed a high throughput screening using three machine learning approaches, including Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM) to develop a model using the SARS-CoV-23CL protease or main protease inhibitors available in the ChEMBL database.[10] The developed models were then subjected to identify natural compounds having anti-SARS-CoV-2 activity from the NuBBE database[11], a database of natural products of Brazilian biodiversity. The potential inhibitors suggested by the machine learning methods were then subjected for molecular docking study which suggests four promising main protease inhibitors [podocarpusflavone A (PF), 7-O-methoxyquercitrin (MQ) and proanthocyanidin (PA) and chimarrhinin (CM)] having comparable binding affinity to the standard compound (V7G). Further, the four inhibitors-enzyme complexes were subjected for molecular dynamic simulation for a period of 100 ns in explicit water.

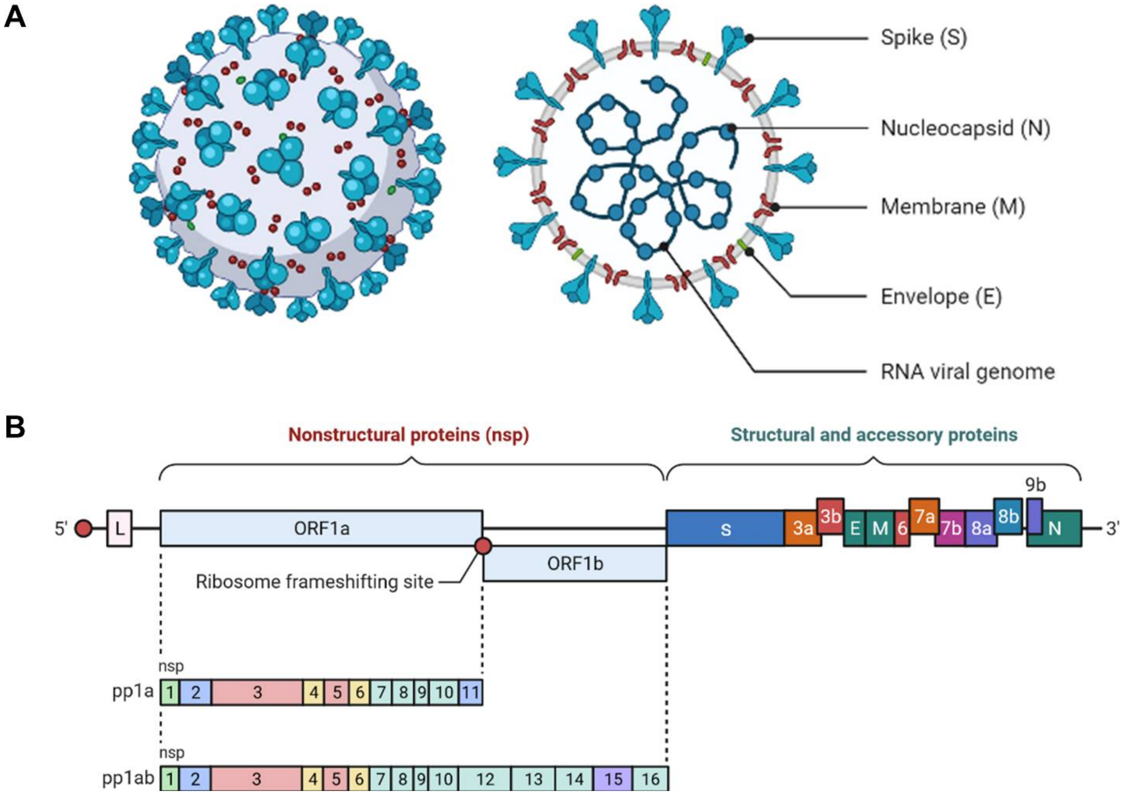


Figure 1. Structure and genome organization of SARS-CoV-2 virus.[12,13] The Figure was created in BioRender.com

MATERIALS AND METHODS

Dataset building. The SARS-CoV-23CL protease or main protease inhibitors reported in the ChEMBL database[10] were curated in March 2021. After removing duplicate data, a total of 6424 inhibitors were obtained. Compounds that do not bind to the receptor is required to test false positives. Negative results are seldom reported and an accurate choice of decoys is still a subject of intense research effort.[14] Therefore, we have divided the dataset into ‘active’ if its IC_{50} or *% inhibition* is ≤ 1 μM or > 50 % and represented as ‘1’ and ‘inactive’ if its IC_{50} or *% inhibition* is > 1 μM or < 1 % and represented as ‘0’.[15] In this way, the dataset of 535 compounds was divided into295 (~ 55 %) active and 240 (~ 45 %) inactive ligands. The dataset was then split into training and test sets containing 401 (75 %) and 134 (25 %) ligands, respectively.

Descriptors calculation. The 2048 bits Morgan fingerprints descriptors was calculated for each of the inhibitor using the Python toolkit RDKit.[16] The fingerprints represent a sub-structural feature and can differentiate two dissimilar chemical structures. For a particular structure, the presence and absence of a substructure are represented by ‘1’ and ‘0’, respectively.

Machine learning methods. Three machine learning methods were used to build the classification models which were Support Vector Machine (SVM) is a powerful machine learning method (suitable for classification of nonlinear problems)[17], Random Forest (RF) method[18] (a widely used method due to its prediction accuracies, ease of use and robustness to adjustable parameters)[19] and Gradient Boosting Machine (GBM) method (a competitive and highly robust method for classification problems).[20]

Model validation. The accuracy, sensitivity, specificity, Matthews correlation coefficient and Cohen’s kappa statistic of the test set were used to evaluate the performances of the models using the following equations.[21, 22]

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \qquad \text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Matthews correlation coefficient} - \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}}$$

$$\text{Cohen's kappa statistic} = \frac{p_o - p_e}{1 - p_e}$$

Here,
TP = True positive indicating the number of actives predicted correctly,
TN = True negative indicating the number of inactives predicted correctly
FP = False positive indicating the number of inactives mispredicted as actives,
FN = False negative indicating the number of actives mispredicted as inactives.
p_o = Observed agreement
p_e = Expected agreement

Application of the developed model to identify new inhibitors from natural products. NuBBE database is a library of natural products of Brazilian biodiversity[11] containing a large variety of chemical classes and structural types of secondary metabolites obtained from plants, fungi, insects, marine organisms and bacteria. A total of 2111 compounds of different chemical classes, including alkaloids, amino acids and peptides, aromatic derivatives, chalcones, flavonoids, lignoids, lipids, carbohydrates, phenylpropanoids, polyketides, tannins and terpenes, were downloaded from the database[23]. These compounds were then subjected to the developed model to identify natural products exhibiting anti SARS-Cov-2 activity by inhibiting the main protease enzyme.

Molecular docking. Preparation of target protein. The main protease of SARS-CoV-2 is critical for the assembly of viral replication-transcription complex and the release of viral proteins.[24] Thus, it plays an essential role in the replication and pathogenicity of the virus.[24,25] Therefore, the main protease serves as an attractive drug target to combat viral replication and pathogenesis.[24,26]

The ligand-bound conformation of protein structure is the prerequisite to perform molecular docking study since during docking, software searches complementary binding site/(s) for ligand within the search space of the target protein.[27] Moreover, the structure should be solved with a reasonable accuracy, which is reflected in the statistics for data processing and refinement of the X-ray crystal structure. The refinement statistics such as R_{work}/R_{free}, RMS deviation from ideality (bonds and angles) and validation parameters including Ramachandran outliers, rotamer outliers, bad bond or bad angle count indicate the quality of the built model structure. In our current investigation, we have selected the crystal structure of SARS-CoV-2 main protease complexed with V7G (GRL-024-20) (PDB ID: 6XR3)[28] since the structure was solved at 1.4 Å resolution with R_{work}/R_{free} of 0.143/0.187 and the validation parameters indicate good quality of the build model structure. Water molecules and ligands of the model structure were removed and polar hydrogen atoms were added to the protein using PyMOL.[29] Energy minimization was performed by YASARA force field level of theory in the YASARA Energy Minimization Server.[30] After energy minimization, the macromolecule was prepared for docking using MGLTools.[31]

Preparation of ligands. The molecular geometry of the identified compounds was optimized with the Universal Force Field (UFF) level of theory. Then, appropriate partial charges were assigned to the structures using Open Babel.[32]

Validation of docking protocol. The docking of the target protein with the ligand was conducted using AutoDock vina.[33] The docking method was first validated by re-docking V7G into the binding pocket of main protease. Different volumes of boxes with a center of the binding site were used to get the optimal box volume. The (20.00 × 20.00 × 20.00) $Å^3$ box centered at the binding site resulted in the most optimal docking performance (the root mean square deviation (RMSE) was <2 for all non-hydrogen atoms) (Figure2). Thus, the docking method has reasonable accuracy and reproducibility and can be used for further docking experiments.

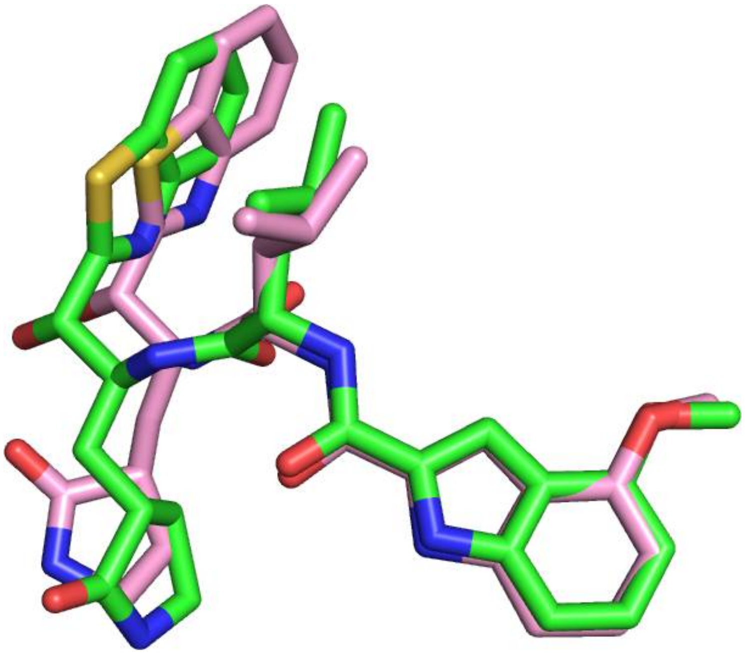


Figure 2. Validation of docking protocol. The experimental and docked V7G are represented in green and pink color, respectively.

Protein-ligand docking. The docking of the identified ligands was conducted using the optimal box volume and center. Throughout the docking study, the flexible ligand searched its complementary site(s) within the search space of rigid macromolecule. Ligands with the lowest binding affinity and promising binding pose were chosen as the best conformation. The interactions of ligands with main protease were analyzed by PyMOL.[29]

Molecular dynamics simulation. The best docking pose of the compounds was selected to examine the interactions of ligand with the active site of enzyme. The molecular dynamic simulation was performed using the selected compounds for 100 ns in explicit water.The AmberTools package was applied asfull Amber topology/coordinate files for all ligands.[34] The VDW and the bonded parameters of different ligands were computed using the antechamber program of AmberTools with the general amber force field (GAFF) [35],while the protein was modelled by theAMBERff14SB force field.[36] The RESP charge model was used to obtain the partial atomic charges.[37] The TIP3P water model was used to solvate the compounds and ions added to the box to neutralize the system. The periodic boundary condition (PBC) was applied in three dimensions. All MD simulations were performed by a parallel version of SANDER in AmberTools 19 software package.[38] Before performing the MD simulation of protein-ligand complexes, the steepest descent algorithm was examined to minimize their energy and a leap-frog algorithm was considered to integrate their motion.[39] In this process, the effect of long-range electrostatic interactions of molecules was observed using the Particle mesh Ewald (PME) method.[40] To simulate the limitations of H-bonds in this process, the LINKS algorithm was studied in both equilibration and production runs.[41]The cutoff for nonbonded interactions was set to 12.0 nm. After the energy minimization, the system was simulated for 20 ps in the canonical ensemble (NVT) and with a 1 ns time-step in the NPT ensemble. The Langevin dynamics [42] and Parrinello-Rahman [43] models were studied to couple the temperature and pressure of the system using coupling constants of 0.1 and 0.5 ps, respectively.

RESULTS AND DISCUSSION

Support vector machine (SVM) model. Three parameters such as a kernel function, a kernel coefficient and a penalty parameter are required to build an SVM model. The radial based function (rbf) was used as a kernel function in the SVM model. Optimized values for other hyperparameters such as the penalty parameter C and the kernel coefficient gamma were obtained by a 5-fold cross-validated grid search process. The best values or C and gamma were found 10 and0.0001, respectively.

Random forest (RF) model. To build the RF model two parameters are required:the n_estimators and the max_features.These parameters were also obtained by a grid search process following the technique described above. The optimal values of n_estimators and max_features were found 100 and 204, respectively.

Gradient boosting machine (GBM) model. In the GBM, only the n_estimators were optimized by a grid search process following the technique described earlier. The other parameters such as subsample and max_features, were set to 0.5. The optimum value of n_estimators was found to be 100.

Performance metrics of the models. The accuracy of the training set for SVM, RF and GBM were 60.45 %, 63.43 % and 64.93 %, respectively, indicating excellent performance of the developed methods. The models were then subjected to validate the test set. The performance metrics such as accuracy, sensitivity, specificity, Matthews correlation coefficient and Cohen’s kappa statistics of the test set are presented in Table 1. The performance metrics suggest that all the three methods have good specificity, accuracy and sensitivity.

Identification of potential main protease inhibitors from natural products using machine learning. The SVM, RF and GBM suggested 89, 735 and 802 natural compounds, respectively, would have anti-SARS-CoV-2 activity with a probability of at least 0.5. To make a robust prediction, molecules with at least 70 % probability of being active were selected. The application of this filter yielded 322 unique compounds (10, 270 and 315 compounds from SMV, RF and GBM, respectively). These compounds were then further screened using molecular docking study.

Table 1. Parameters and Performances of Support Vector Machine, Random Forest and Gradient Boosting.

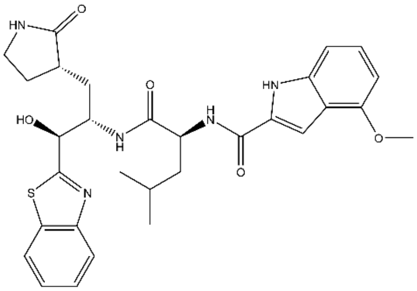
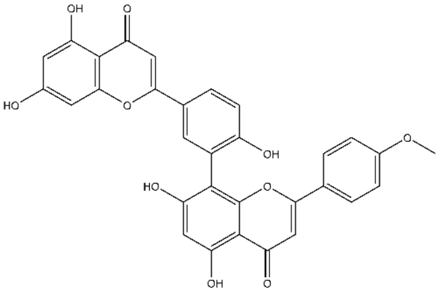
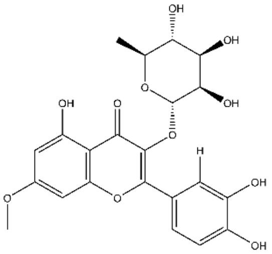
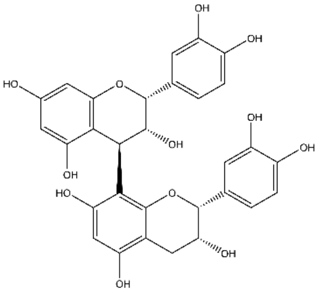
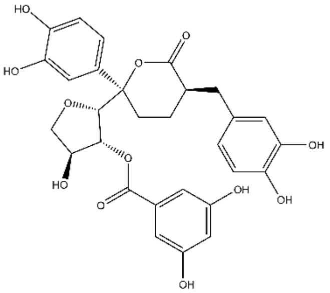
Models	Test set				
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Matthews correlation coefficient	Cohen's kappa statistic
Support vector machine	60.45	75.68	41.67	0.18	0.18
Random forest	63.43	74.32	50.00	0.25	0.25
Gradient boosting	64.93	79.73	46.67	0.28	0.27

Identification of potential main protease inhibitors from natural products using molecular docking. The molecular docking study revealed that the standard compound V7G interacted with the main protease with a binding affinity of -8.9 Kcal/mol. Therefore, we set a threshold of -8.8 Kcal/mol binding affinity to short list potential inhibitors from the docking study. Application of this threshold yielded three flavonoids [NuBBE_199: Podocarpusflavone A (PF), NuBBE_321: 7-O-Methoxyquercitrin (MQ) and NuBBE_359: Proanthocyanidin (PA)] and one lignoid [NuBBE_510: Chimarrhinin (CM)] compounds. Table 2 represents binding affinity of standard compound and the identified potential natural inhibitors of SARS-CoV-2 main protease. Molecular docking study demonstrated that the identified lignoid showed higher binding affinity (-9.0 Kcal/mol) than the standard compound.

Molecular mechanism of main protease inhibition. The molecular mechanism of main protease inhibition was also explored using molecular docking study. The docking study revealed that all the inhibitors occupy the same position to that of GRL-024-20 in the main protease/GRL-024-20 complex structure [28]. The PFshowed identical binding affinity (-8.9 Kcal/mol) to that of standard compound GRL-024-20and forms hydrogen bonding with Thr25, Thr26, Gly143, His163, His164 andGlu166 and approaching within 4.0 Å from the side chain of Leu27, Phe140, Leu141, Asn142, Ser144, Cys145 and Met165(Figure 3A). The MQ interacts with the enzyme with a binding affinity of -8.8 Kcal/mol and makes hydrogen bonding withTyr54, Leu141, Asn142, Gly143, Ser144, His163, Glu166 and Arg188, and is positioned withing 4.0 Å from the side chain of His41, Cys44, Met49, Pro52, Leu141, Cys145, His164 and Met165 (Figure 3B). PA displayed similar binding affinity to that of MQ and forms hydrogen bonding with Thr25, His41, Leu141, Ser144, Cys145, His163 and Gln189, and located within 4.0 Å from Thr26, Leu27, Met49, Tyr54, Asn142, Gly143, His164, Met165, Glu166, Asp187 and Arg188 (Figure 3C). The lignoid CM showed the highest binding affinity (-9.0 Kcal/mol) and makes hydrogen bonding with His41, Tyr54, Leu141, Asn142, Gly143, Ser144, Cys145, Glu166 and Arg188. This complex is further stabilized by van der Waals interactions with Leu27, Met49, His163, Met165, Asp187 and Gln189 (Figure 3D).

Evaluation of flexibility and conformational differences between apo and bounded main protease. MD simulation was performed on the four complexes of ligands and the main protease. The root mean square deviation (RMSD) of Cα atoms of the protein backbone was monitored throughout the 100 ns simulation to test the stability of trajectories derived from MD simulation and it is shown in Figure 4.MD simulation of the inhibitor protease system indicates that the protease structure remains very similar to that of the x-ray structure with an RMSD of 0.174 nm (Figure 4).Analysis of the

Table 2. Compound code, common name, chemical structure and binding affinity with main protease or main protease of SARS-CoV-2.

Compound code	Common name	Chemical structure	Binding affinity (Kcal/mol)
V7G	GRL-024-20		-8.9
NuBBE_199	Podocarpusflavone A		-8.9
NuBBE_321	7-O-methoxyquercitrin		-8.8
NuBBE_359	Proanthocyanidin		-8.8
NuBBE_510	Chimarrhinin		-9.0

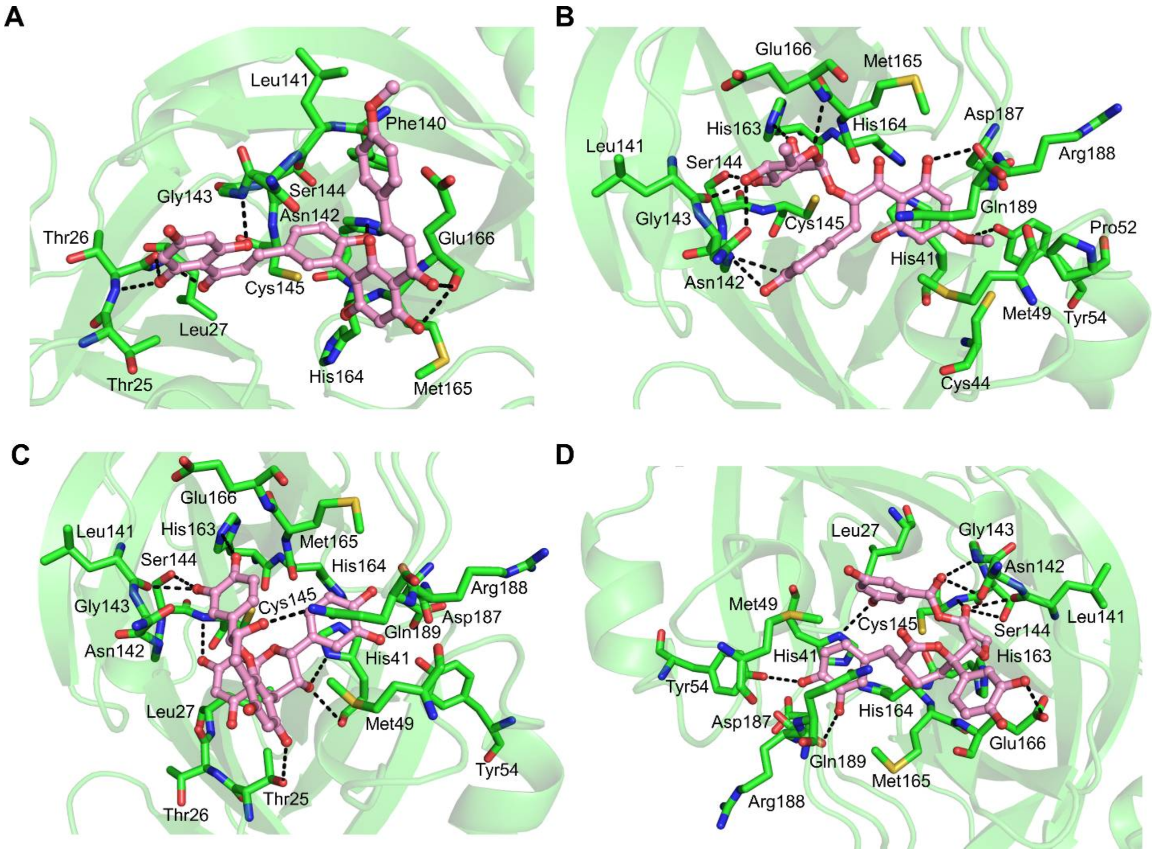


Figure 3. Interactions of (A) Podocarpusflavone A (NuBBE_199), (B) 7-O-Methoxyquercitrin (NuBBE_321), (C) Proanthocyanidin (NuBBE_359) and (D) Chimarrhinin (NuBBE_510) at the binding site of main protease enzyme. The ligand and interacting residues of protein are presented in pink and green color, respectively. Black dash indicates hydrogen bonding.

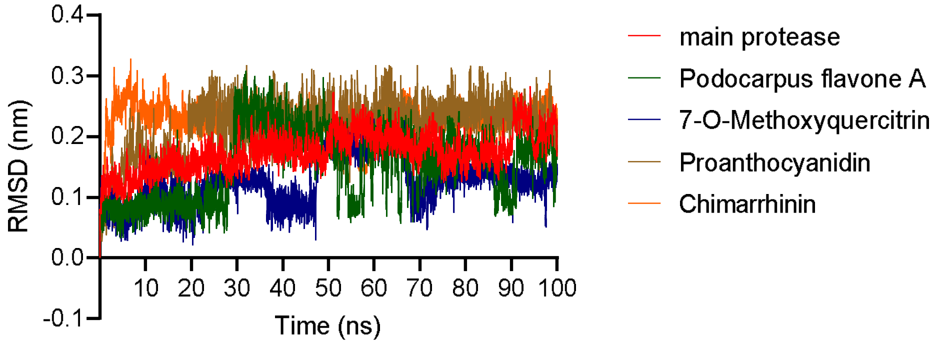


Figure 4. RMSD of alpha carbon atoms of main protease and its four complexes as a function of simulation time.

RMSD showed that the overall RMSD of the Cα atoms of PF, MQ,PA and CM were about 0.154 nm, 0.126 nm, 0.216 nm and 0.226 nm, respectively suggesting the freedom of protein movement due to the former two (PF and MQ) complex formation is less than the latter two complexes (PA and CM).

The root mean square fluctuation (RMSF) indicates the fluctuation of every single residue from its reference position and is an important parameter for assessing amino acid motion restrictions due to interaction with ligands. The results of the RMSF analysis indicate that the main fluctuations correspond to residues located in the ligand-binding cavityformed by the domains I and II (six-stranded antiparallel β barrels) of the enzyme (Figures 3 and5)[44].In domain I,the mobility is highest for CM followed by MQ whereas the highest fluctuations were observed for PA followed by CM in domain II. The fluctuations in domain III, which is responsible for regulating the dimerization, do not significantly affect the substrate-binding site of the enzyme. Therefore, they have not been considered.

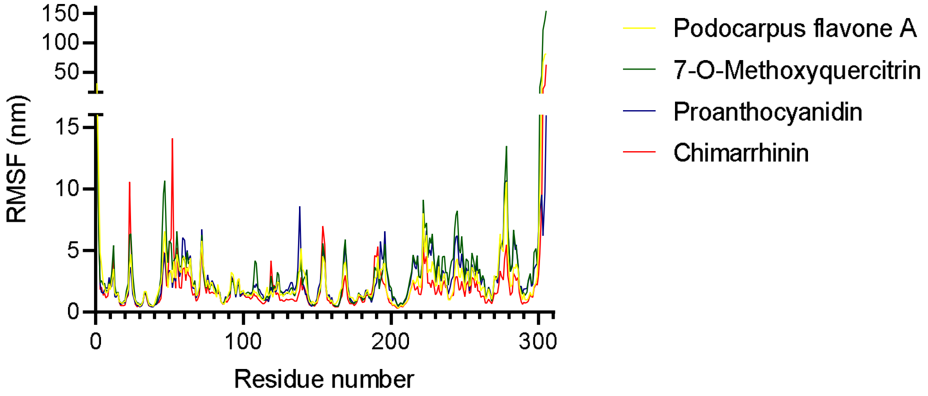


Figure 5. RMSF of residues of main protease\Podocarpus flavone A, \7-O-methoxyquercitrin, \Proanthocyanidin and \Chimarrhinincomplexes.

Table 3. Hydrogen bond formation analysis between inhibitors and main protease (M_{pro}).

Inhibitors	Residue	Donor atom	Acceptor atom	Occupancy (%)
Podocarpusflavone A (PF)	Thr26	H (M_{pro})	O35(PF)	5.80
	Asn119	H (M_{pro})	O38(PF)	26.10
	His164	H10 (PF)	O (M_{pro})	24.70
7-O-methoxyquercitrin (MQ)	Asn142	H21 (MQ)	OD1 (M_{pro})	10.40
	Gly143	H (M_{pro})	O32 (MQ)	2.10
	Glu166	H (M_{pro})	O31 (MQ)	10.80
	His164	H23 (MQ)	O (M_{pro})	16.76
	Thr26	H18 (MQ)	O (M_{pro})	15.22
Proanthocyanidin (PA)	Gln192	HE22 (M_{pro})	O39 (PA)	44.22
	Ser46	H4 (PA)	O (M_{pro})	10.32
	His164	H10 (PA)	O (M_{pro})	10.32
Chimarrhinin (CM)	His41	H39 (CM)	O (M_{pro})	7.74
	Leu141	H23 (CM)	O (M_{pro})	2.52
	Asn142	HD21 (M_{pro})	O35 (CM)	6.80
	Gly143	H (M_{pro})	O34 (CM)	2.05
	Ser144	H (M_{pro})	O33 (CM)	11.96
	Glu166	H23 (CM)	OE1 (M_{pro})	10.62
	Arg188	H27 (CM)	O (M_{pro})	8.21
	Glu166	H22 (CM)	OE2 (M_{pro})	35.55

Hydrogen bond analysis. MD simulation of the protein-ligand complex was performed over the 100 ns time to determine the hydrogen bonding stability of PF, MQ, PA and CM with the main protease. The hydrogen bond analysis was conducted using Amber Tools. The applied threshold of hydrogen bond formation was 3 Å with an angle of 150□.The analysis of the H-bonding is shown in Table3. Table 3 shows that PF forms three stable H-bonding with Thr26, Asn119and His164.The stability of H-bonding is more substantial with Asn119 (26.1 %) and His164 (24.7 %) as indicated by the occupancy. Further, MQ forms H-bonding with Gly143, Glu166, Asn142, His164 and Thr26. Among these, ligand forms stable interactions with Thr26 and His164 with 15.22 % and 16.76 %occupancy, respectively. The PA makes H-bonding with Gln192, Ser46 and His164. The occupancy of these H-bonding suggests that the ligand forms stable interaction with Gln192 (44.22 %). In addition, CM forms H-bonding with Asn142, Gly143, Ser144, His41, Leu141, Arg188 and Glu166. The H-bonding between the ligand and Glu166 was most stable, accounting for 35.55% of the simulation time.

Free energy of binding. The free energy of binding between ligand-receptor interactions was evaluated using the MM-GBSA method. The results of free energy of binding are presented in Table 4. From Table 4 it is clear that the van der Waals energy (ΔEvdW) term is the major contributing factor for ligand binding energy in PF, MQ and PA complexes. The van der Waals energy (ΔEvdW) and the electrostatic energy (ΔEEEL) terms contribute to the CM-main protease binding energy. The total binding energy of all the complexes declines mainly due to the polar solvation free energy (ΔE_{GB}). The binding free energies of PF, MQ, PA and CM with the main protease were found to be -17.59 Kcal/mol, -21.27 Kcal/mol, -32.86 Kcal/mol and -31.54 Kcal/mol, respectively.

Table 4. MM-GBSA binding energies (kcal/mol).

Energy	Podocarpus flavone A	7-O-Methoxyquercitrin	Proanthocyanidin	Chimarrhinin
Van der Waal energy (ΔEvdW)	-29.07	-34.34	-46.37	-44.73
Electrostatic energy (ΔEEEL)	-15.29	-14.95	-20.88	-47.09
Polar solvation energy (ΔE_{GB})	30.81	32.75	39.84	67.01
SASA energy (ΔE_{SURF})	-4.03	-4.73	-5.44	-6.73
Gas-phase energy (ΔG_{GAS})	-44.37	-49.29	-67.25	-91.82
ΔG_{SOL}	26.78	28.02	34.40	60.28
$\Delta G_{Binding\ energy}$	-17.59	-21.27	-32.86	-31.54

The outbreak of COVID-19 has led to a global crisis with increasing morbidity. The scale and rapid contagious nature of this disease demands an urgent need for treatment options before an effective vaccine can be produced. There is a long series of infectious diseases in which vaccines are only partially effective [7] and hence, a series of vaccine defeats.[8] Besides, the ongoing research on COVID-19 in laboratories worldwide are adding new data at a tremendous pace, making it difficult to predict what kind of vaccine can be truly effective.[7] Further, the SARS-CoV-2 is an RNA virus and generally have a high mutation rate which also represent a challenge to develop effective vaccines against these viruses.[7]

Therefore, the need for effective antiviral drugs against SARS-CoV-2 is urgent and should be prioritized. However, none of the synthetic or semisynthetic antiviral drugs are not devoid of side effects.[9] On the contrary, drugs from natural sources have little or no side effects. Therefore, to identify anti-SARS-CoV-2 drugs from natural sources we performed a high throughput screening using machine learning models followed by molecular docking. In the current study, a natural compound database (NuBBE) was subjected to *in silico* screening to identify natural compounds with anti-SARS-CoV-2 activity. Before that, we trained and validated our model using the main protease inhibitors curated from the ChEMBL database.[10] Our classification models (SVM, RF and GBM) predicted that 322 unique compounds would exhibit anti-SARS-CoV-2 property by inhibiting the main protease.

Further, our molecular docking study revealed that three flavonoids (PF, MQ and PA) and one lignoid (CM)compounds possess comparable binding affinity (-8.8 Kcal/mol to -9.0 Kcal/mol) to the standard compound, GRL-024-20 (-8.9 Kcal/mol) and occupy the same binding position to that of GRL-024-20in the main protease/GRL-024-20 complex structure.[28].In addition, MD simulation showed that all the compounds form stable complex with different magnitude of binding energy. The main protease is a cysteine protease with a catalytic dyad composed of Cys145 and His41.[24,44] Therefore, compounds interacting with these residues most likely inhibit main protease activity.[24] All the four compounds interact with the catalytic dyad and hence, likely inhibit the enzyme action. A literature survey revealed that flavonoids and lignoids inhibits the main protease and thereby, demonstrated their anti-viral activity.[45-49] Therefore, our current developed and validated machine learning methods have reliable prediction accuracy and successfully identified natural compounds with main protease inhibitory activity. Besides, our molecular docking and MD simulation studies revealed the likely mechanism of main protease inhibition by the identified compounds.

CONCLUSION

Natural products have been a useful repository of organic compounds for drug discovery. However, exploring the drugs from this resource has diminished in the past two decades, in part because of technical barriers to screening natural products in high-throughput assays against molecular targets.[50] This problem can be overcome by the application of computational techniques. These techniques involve training of models that relate molecular features to targeted activities,[51] thereby, allow potential compounds to be screened *in silico*, reducing costs and saving time.[15] In the current study, we have developed and validated a high throughput *in silico* screening method to identify potential compounds from natural sources. Our *in silico* screening yields four compounds as potential main protease inhibitors. Further, we have explored the likely mechanism of main protease inhibition by the identified compounds. Our current study will assist in identifying novel drugs from natural sources and optimize them to have more desired and less adverse effects. However, further *in vitro* biophysical and biochemical research is recommended to validate the *in silico* results.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data and material are included in the article.

Competing interests

The authors declare that there is no conflict of interest.

Funding

Not applicable.

Acknowledgement

The authors wish to thank State University of Bangladesh for providing Computational Lab facility.

References

1. Kong, W.H., Li, Y., Peng, M.W., Kong, D.G., Yang, X.B., Wang, L. and Liu, M.Q. 2020. SARS-CoV-2 detection in patients with influenza-like illness. *Nat. microbiol.* **5**, 675-678.
2. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W. and Lu, R. 2019. A novel coronavirus from patients with pneumonia in China. *N. Engl. J. Med.* **2020**.
3. Gralinski, L.E. and Menachery, V.D. 2020. Return of the Coronavirus: 2019-nCoV. *Viruses.* **12**, 135.
4. Wu, J.T., Leung, K. and Leung, G.M. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet.* **395**, 689-697.
5. Hui, D.S., Azhar, E.I., Madani, T.A., Ntoumi, F., Kock, R., Dar, O., Ippolito, G., Mchugh, T.D., Memish, Z.A. and Drosten, C. 2020. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health-The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* **91**, 264-266.
6. Hu, B., Guo, H., Zhou, P. and Shi, Z.L. 2020. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **2020**, 1-14.
7. Forni, G. and Mantovani, A. 2021. COVID-19 vaccines: where we stand and challenges ahead. *Cell Death Differ.* **28**, 626-639.

8. Forni, G., Mantovani, A., Moretta, L., Rezza G. and M, B. 2018. Vaccines. *Accademia Nazionale dei Lincei.* **2018**.

9. Koch, O., Sheehy, S., Sargent, C., Democratis, J., Abbas, S., Schiefermueller, J. and Angus, B.J. 2010. Antiviral drugs. In *side effects of drugs annual*, Elsevier. Vol. *32*, pp 529-553.

10. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L. and McGlinchey, S. 2014. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*. **42**, D1083-D1090.

11. Saldívar-González, F.I., Valli, M., Andricopulo, A.D., da Silva Bolzani, V. and Medina-Franco, J.L. 2018. Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. *J. Chem. Inf. Model.***59**, 74-85.

12. Ghaleh, H.E.G., Karimi, M.R., Rezayat, P., Bolandian, M., Nodoushan, M.M. and Farzanehpour, M. 2020. The possible role of novel coronavirus 2019 proteins in the development of drugs and vaccines. *J. Appl. Biotechnol. Rep.* **7**, 63-73.

13. Sapokta, A. 2020. Structure and genome of SARS-CoV-2 (COVID-19) with diagram. *Microbe Notes, available at: microbenotes. com/structure-and-genome-of-sars-cov-2* **2020**.

14. Lagarde, N., Zagury, J.F.O. and Montes, M. 2015. Benchmarking data sets for the evaluation of virtual ligand screening methods: review and perspectives. *J. Chem. Inf. Model.* **55**, 1297-1307.

15. Brenner, M.P. and Colwell, L.J. 2016. Predicting protein-ligand affinity with a random matrix framework. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13564-13569.

16. Landrum, G. 2006. RDKit: Open-source cheminformatics.

17. Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Mach. learn.* **20**, 273-297.

18. Breiman, L. 2001. Random forests. *Mach. learn.* **45**, 5-32.

19. Qin, Z., Xi, Y., Zhang, S., Tu, G. and Yan, A. 2019. Classification of cyclooxygenase-2 inhibitors using support vector machine and random forest methods. *J. Chem. Inf. Model.* **59**, 1988-2008.

20. Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 1189-1232.

21. Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004* **1996**.

22. Li, B.K., Cong, Y., Yang, X.G., Xue, Y. and Chen, Y.Z. 2013. *In silico* prediction of spleen tyrosine kinase inhibitors using machine learning approaches and an optimized molecular descriptor subset generated by recursive feature elimination method. *Comput. Biol. Med.* **43**, 395-404.

23. Valli, M., Dos Santos, R.N., Figueira, L.D., Nakajima, C.H., Castro-Gamboa, I., Andricopulo, A.D. and Bolzani, V.S. 2013. Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* **76**, 439-444.

24. Breidenbach, J., Lemke, C., Pillaiyar, T., Schäkel, L., Al Hamwi, G., Diett, M., Gedschold, R., Lopez, V., Mirza, S. and Namasivayam, V. 2021. Targeting the Main Protease of SARS□CoV□2: From the Establishment of High Throughput Screening to the Design of Tailored Inhibitors. *Angew. Chem. Int. Ed.* **2021**.

25. Liang, H., Zhao, L., Gong, X., Hu, M. and Wang, H. 2021. Virtual screening FDA approved drugs against multiple targets of SARS□CoV□2. *Clin. Transl. Sci.* **2021**.

26. Zhang, L., Lin, D., Kusov, Y., Nian, Y., Ma, Q., Wang, J., Von Brunn, A., Leyssen, P., Lanko, K. and Neyts, J. 2020. α-Ketoamides as broad-spectrum inhibitors of coronavirus and enterovirus replication: structure-based design, synthesis, and activity assessment. *J. Med. Chem.* **63**, 4562-4578.

27. Khan, M.F., Nahar, N., Rashid, R.B., Chowdhury, A. and Rashid, M.A. 2018. Computational investigations of physicochemical, pharmacokinetic, toxicological properties and molecular docking of betulinic acid, a constituent of *Corypha taliera* (Roxb.) with phospholipase A2 (PLA2). *BMC Complement Altern. Med.* **18**, 48.

28. Anson B., Ghosh A.K. and Mesecar, A. 2020. X-ray Structure of SARS-CoV-2 main protease bound to GRL-024-20 at 1.45 A. **2020**.

29. DeLano, W. L. 2002. The PyMOL user's manual. *DeLano Scientific, San Carlos, CA, 452*.

30. Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D. and Karplus, K. 2009. Improving physical realism, stereochemistry, and side□chain accuracy in homology modeling: four approaches that performed well in CASP8. *Proteins: Structure, Function, and Bioinformatics* **77**, 114-122.

31. Dallakyan, S. 2010. MGLTools. *Reference Source* **2010**.

32. O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G. R. 2011. Open Babel: An open chemical toolbox. *J. cheminformatics* **3**, 33.

33. Trott, O. and Olson, A.J. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455-461.

34. Case, D., Berryman, J., Betz, R., Cerutti, D., Cheatham, T., Darden, T., Duke, R., Glese, T., Gohlke, H. and Gotz, A. 2015. *Amber 2015. San Francisco: University of California*; Technical report 2015.

35. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A. 2004. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157-1174.

36. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C. 2015. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696-713.

37. Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J.C., Cieplak, P. and Dupradeau, F.Y. 2011. RED Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **39** (suppl_2), W511-W517.

38. Case, D., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham III,T., Cruzeiro, V., Darden, T., Duke, R., Ghoreishi, D. and Gilson, M. 2018. AMBER 2018: San Francisco. California.

39. Hockney, R. 1988. Computer simulation using particles, ed. RW Hockney and JW Eastwood. CRC Press.

40. Darden, T., York, D. and Pedersen, L. 1993. Particle mesh Ewald: An N□ log (N) method for Ewald sums in large systems. *J. Chem. Phys*. **98,** 10089-10092.

41. Hess, B., Bekker, H., Berendsen, H. J. and Fraaije, J. G. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463-1472.

42. Goga, N., Rzepiela, A., De Vries, A., Marrink, S. and Berendsen, H. 2012. Efficient algorithms for Langevin and DPD dynamics. *J. Chem. Theory Comput.* **8**, 3637-3649.

43. Parrinello, M. and Rahman, A. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. phys.* **52**, 7182-7190.

44. Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. and Hilgenfeld, R. 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science **368***, 409-412.

45. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. and Jung, S.H. 2016. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. *J. Med. Chem.* **59**, 6595-6628.

46. Xue, G., Gong, L., Yuan, C., Xu, M., Wang, X., Jiang, L. and Huang, M. 2017. A structural mechanism of flavonoids in inhibiting serine proteases. *Food Funct.* **8**, 2437-2443.

47. Ryu, Y.B., Jeong, H.J., Kim, J.H., Kim, Y.M., Park, J.Y., Kim, D., Naguyen, T.T. H., Park, S.J., Chang, J.S. and Park, K.H. 2010. Biflavonoids from Torreya nucifera displaying SARS-CoV 3CLpro inhibition. *Bioorg. Med. Chem.* **18**, 7940-7947.

48. Jo, S., Kim, S., Shin, D.H. and Kim, M.S. 2020. Inhibition of SARS-CoV 3CL protease by flavonoids. *J. Enzyme Inhib. Med. Chem.* **35**, 145-151.

49. Wen, C.C., Kuo, Y.H., Jan, J.T., Liang, P.H., Wang, S.Y., Liu, H.G., Lee, C.K., Chang, S.T., Kuo, C.J. and Lee, S.S. 2007. Specific plant terpenoids and lignoids possess potent antiviral activities against severe acute respiratory syndrome coronavirus. *J. Med. Chem.* **50**, 4087-4095.

50. Harvey, A.L., Edrada-Ebel, R. and Quinn, R.J. 2015. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14**, 111-129.

51. Morrone, J.A., Weber, J.K., Huynh, T., Luo, H. and Cornell, W.D. 2020. Combining docking pose rank and structure with deep learning improves protein-ligand binding mode prediction over a baseline docking approach. *J. Chem. Inf. Model.* **2020**.