JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

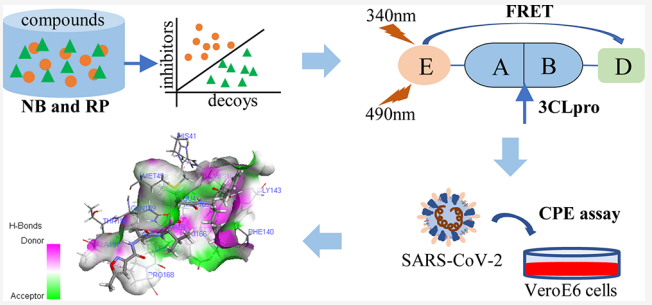
Exploration of SARS-CoV-2 3CL^{pro} Inhibitors by Virtual Screening Methods, FRET Detection, and CPE Assay

Jun Zhao,[⊥] Qinhai Ma,[⊥] Baoyue Zhang, Pengfei Guo, Zhe Wang, Yi Liu, Minsi Meng, Ailin Liu,* Zifeng Yang,* and Guanhua Du*

Cite This: https://doi.org/10.1021/acs.jcim.1c01089 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: COVID-19 caused by a novel coronavirus (SARS-CoV-2) has been spreading all over the world since the end of 2019, and no specific drug has been developed yet. 3C-like protease (3CL^{pro}) acts as an important part of the replication of novel coronavirus and is a promising target for the development of anticoronavirus drugs. In this paper, eight machine learning models were constructed using naive Bayesian (NB) and recursive partitioning (RP) algorithms for 3CL^{pro} on the basis of optimized two-dimensional (2D) molecular descriptors (MDs) combined with ECFP_4, ECFP_6, and MACCS molecular fingerprints. The optimal models were selected according to the results of 5-fold cross verification, test set verification, and external test set verification. A total of 5766 natural compounds from the internal natural product database were predicted, among which 369 chemical components were predicted to be active compounds by the optimal models and the EstPGood values were more than 0.6, as predicted by the NB (MD + ECFP_6) model. Through ADMET analysis, 31 compounds were selected for further biological activity determination by the fluorescence resonance energy transfer (FRET) method and cytopathic effect (CPE) detection. The results indicated that (+)-shikonin, shikonin, scutellarein, and 5,3′,4′-trihydroxyflavone showed certain activity in inhibiting SARS-CoV-2 3CL^{pro} with the half-maximal inhibitory concentration (IC_{50}) values ranging from 4.38 to 87.76 μM. In the CPE assay, 5,3′,4′-trihydroxyflavone showed a certain antiviral effect with an IC_{50} value of 8.22 μM. The binding mechanism of 5,3′,4′-trihydroxyflavone with SARS-CoV-2 3CL^{pro} was further revealed through CDOCKER analysis. In this study, 3CL^{pro} prediction models were constructed based on machine learning algorithms for the prediction of active compounds, and the activity of potential inhibitors was determined by the FRET method and CPE assay, which provide important information for further discovery and development of antinovel coronavirus drugs.



1. INTRODUCTION

At present, coronavirus disease 19 (COVID-19) caused by a novel coronavirus (SARS-CoV-2) is still circulating worldwide and highly contagious mutant strains have emerged. As known from the World Health Organization (WHO), the number of confirmed cases of COVID-19 worldwide had exceeded 247 million as of November 4, 2021, and the cumulative death toll had exceeded 5.0 million.[1] The rapid spread of the virus and rising infectivity have driven the global acceleration of interventions. Currently, related vaccines have been introduced into the market, and people in many countries have been vaccinated,[2] but the related adverse reactions and effective duration still need further clinical confirmation. Although there has been rapid progress in the research and development of vaccines, no specific therapeutic drug has been developed against this virus. The main strategies of drug treatment include drug repositioning, broad-spectrum screening of antiviral drugs, and discovery of new targeted drugs. However, drugs that showed certain activity in the initial stage, such as chloroquine and remdesivir, could not significantly reduce the clinical mortality in COVID-19 patients with the progression of clinical trials.[3,4] Therefore, screening all potential and available drugs aimed at the effective targets of SARS-CoV-2 is still necessary to control and alleviate the epidemic.

After entering the host cell, novel coronavirus replicates and synthesizes a large amount of genetic material and related proteins in the cell, and then, the mature virus particles are assembled in the cytoplasm and released outside the cell.[5] 3C-like protease (3CL^{pro}), also known as M^{pro}, is an essential enzyme for the replication of coronavirus, which exerts a crucial part in cutting polymers and may interfere with the host's innate antiviral immune response. The replication and proliferation of coronavirus can be effectively interfered with

Received: September 8, 2021

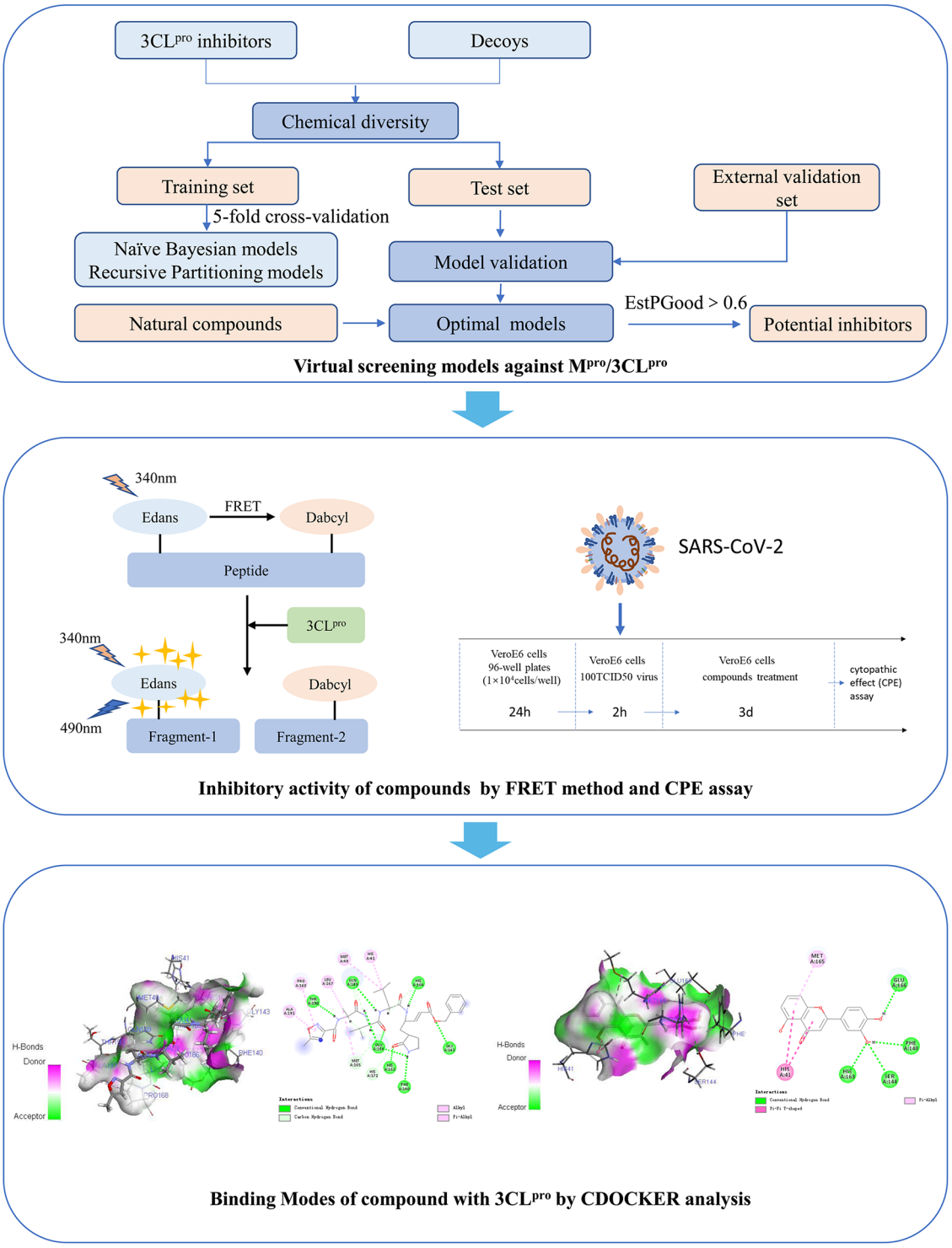


Figure 1. Specific scheme for the establishment of 3CL^{pro} inhibitor prediction models, FRET activity detection, the CPE assay, and CDOCKER analysis.

by inhibiting the activity of 3CL^{pro}.[6] 3CL^{pro} is highly conserved in different coronaviruses, so drugs targeting 3CL^{pro} can significantly reduce mutation-mediated drug resistance and show broad-spectrum antiviral activity.[7] Finding or designing 3CL^{pro} inhibitors is a potential therapeutic strategy for COVID-19.

In recent years, as a method for computer-aided drug design and high-throughput screening, computer virtual screening technology has played an important role in drug discovery and development. The most common methods are molecular docking, pharmacophore modeling, and machine learning. Compared with the traditional screening process, the machine learning approach is simple, easy, and low cost, which can greatly reduce the research time. At present, there have been compelling studies focusing on potential drugs against COVID-19 through computer virtual screening technology based on the 3CL^{pro} structure. Early in the outbreak of COVID-19, the structural sequence of 3CL^{pro} was analyzed to

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

construct its 3D homologous model, which was used to screen a database of medicinal plants containing 32297 potential antiviral plant chemical constituents, and nine potential anti-SARS-CoV-2 compounds were found.[8] Gyebi et al. detected four potential nontoxic, drug-usable plant-derived $3CL^{pro}$ inhibitors by screening 62 African plant-derived alkaloids and 100 terpenoids using molecular docking technique.[9] There have been 168 virtual screening studies for $3CL^{pro}$, but the accuracy of the screening models is limited, and most of the prediction results from models have not been verified by experiments.[10] In this paper, machine learning models were established first by naïve Bayesian (NB) and recursive partitioning (RP) algorithms for $3CL^{pro}$ to predict 5766 natural chemical components in the natural molecular database established by our laboratory. The predicted compounds were further screened by ADMET analysis, and then, the activity of screened drugs was determined by the fluorescence resonance energy transfer (FRET) method and the cytopathic effect (CPE) assay. Finally, the action mechanism of potential inhibitors was analyzed by molecular docking. The overall process is shown in Figure 1. In summary, this paper provides important information for further discovery and development of antinovel coronavirus drugs.

2. MATERIALS AND METHODS

2.1. Data Aggregation and Processing. The active ligands against $3CL^{pro}$ were collected in the BindingDB database (http://www.bindingdb.org). After removing the repetitive structures, a total of 149 active compounds (inhibitors) were obtained, and then, these active ligands were used to generate inactive compounds (decoys) in the DUD-E database (http://dude.docking.org). Based on the proportion of 3:1, inactive compounds and active compounds were stochastically grouped into a training set including 112 active compounds and 337 inactive compounds and a test set including 37 active compounds and 113 inactive compounds in DS 2018 (Discovery Studio version 2018, San Diego, CA). $3CL^{pro}$ inhibitors reported from the related literature were collected to form an external test set containing 40 active compounds and 120 inactive compounds. The symbols 1 and −1 were used to mark the activity of the active compounds and inactive compounds, respectively, in all data sets. Hydrogenation, deprotonation, and energy optimization were performed for all compounds before the molecular descriptors (MDs) were calculated.

2.2. Calculation and Optimization of Molecular Descriptors. Molecular descriptors (MDs) are employed to measure the molecular weight, atomic number, lipid−water partition coefficient, molecular polarity surface area, and other parameters. In this study, 348 molecular descriptors of the compounds in the training set were calculated by DS 2018 software, comprising 8 AlogP molecular descriptors, 35 molecular property descriptors, 43 topological molecular descriptors, 7 surface area and volume descriptors, 92 molecular property number descriptors, and 163 estate keys. The Pearson correlation coefficients were calculated to quantify the degree of correlation between 348 molecular descriptors and the activity of compounds. First, the molecular descriptor was removed when the frequency of the descriptor value was more than 50%. Then, the molecular descriptor was excluded if its Pearson correlation coefficient[11] with activity was less than 0.1. Meanwhile, of the two molecular descriptors with a correlation coefficient of more than 0.9, the one with a lower correlation coefficient with activity was discarded. Eventually, the reserved molecular descriptors were carried out by stepwise linear regression, in which the molecular descriptors were screened to construct the classification models.

2.3. Molecular Fingerprints. Molecular fingerprints characterize the molecular structure of compounds by a series of molecular fragments. In the present study, the SciTegic extended connection fingerprint ECFP was used, and to ensure that the molecular fragment size described by the molecular fingerprint was kept in the appropriate range, we used the molecular fingerprint with a diameter of 4 or 6, that is, ECFP_4 and ECFP_6, which were calculated in DS 2018 software.[12] Another MACCS molecular fingerprint using the MDL structure library containing 166 seed structures was calculated with PaDEL Descriptor software.[13]

2.4. Spatial Distribution Prediction of Compounds. The spatial distribution diversity of compounds in the training set and test set greatly affects the predictive ability of the machine classification learning model. In general, when compounds in the training set have a wider chemical spatial distribution, the established classification model will also have higher prediction precision and stronger generalization. Conversely, when the spatial distribution in the training set is narrow, the model application will be limited to a great extent. In this study, principal component analysis (PCA) and Tanimoto analysis[14] were used to investigate the chemical spatial distribution characteristics of compounds in all data sets.

2.5. Naïve Bayesian Classification Model and Recursive Partitioning Model. The NB algorithm and RP algorithm were adopted to establish classification models by learning the mapping relationship between molecular descriptors and their activity, which can predict the activity of uncertain active compounds. The NB algorithm is a probability-based algorithm developed by British mathematician Bayes.[15] The NB model was established in DS 2018 software to study how to separate inhibitors from decoys based on the compound information in the training set. The RP algorithm can classify analytical samples layer by layer according to a series of rules by simulating the human learning process.[16] The outcome of the RP model can be directly shown by the graph of a bifurcated decision tree, so the RP model is also called the decision tree model, which was also built in DS 2018 software. The minimum number of samples per node, the maximum number of nodes for each descriptor, and the maximum depth of the decision tree were respectively set to 10, 20, and 20. Each model was established using the training set, and 5-fold cross verification in the training set was carried out in the process of building each model.

2.6. Evaluation of Model. The prediction ability of each model was assessed by 5-fold cross verification in the training set and validation in the test set and external test set. The specific assessment indicators of the predictive ability included sensitivity (SE), specific (SP), overall accuracy (Q), and Matthews correlation coefficient (MCC).[17] They were calculated by Formulas 1−4.

$$SE = \frac{TP}{TP + FN} \quad (1)$$

$$SP = \frac{TN}{TN + FP} \quad (2)$$

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (4)$$

True positive (TP) indicates the number of active compounds predicted to be positive. True negative (TN) indicates the number of inactive compounds predicted to be negative. False positive (FP) indicates the number of inactive compounds predicted to be positive. False negative (FN) indicates the number of active compounds predicted to be negative. The value of MCC is between 1 and 1, where a high MCC value represents the good prediction ability of the model.

The receiver operating characteristic (ROC) curve is a curve drawn with SE as the ordinate and the false positive rate (1-SP) as the abscissa. The area under curve (AUC) is also an important evaluation index of predictive ability.[18] A higher AUC value indicates better prediction ability of the model.

2.7. FRET Detection of SARS-CoV-2 3CL^{pro} Activity. *2.7.1. Reagents.* Thirty-one predicted compounds were obtained from the Institute of Materia Medica of the Chinese Academy of Medical Sciences (Beijing, China). Dimethyl sulfoxide (DMSO), acquired from Sigma-Aldrich Company (St. Louis), was used to dissolve all compounds, and all prepared solutions were stored at −20 °C. The 2019-nCoV M^{pro}/3CL^{pro} inhibitor screening kit was purchased from Beyotime Institute of Biotechnology (Shanghai, China).

2.7.2. FRET Detection of SARS-CoV-2 3CL^{pro} Activity. The amino acid sequence of 3CL^{pro} in the 2019-nCoV M^{pro}/3CL^{pro} inhibitor screening kit is the same as that of natural novel coronavirus 3CL^{pro}. The FRET method[19] was used to detect the activity of 3CL^{pro} in this kit (Figure 2). The fluorescent donor (Edans) and fluorescent receptor (Dabcyl) were connected to both ends of the natural substrate of 2019-nCoV 3CL^{pro}, and the fluorescence of Edans could be detected when the two groups were separated by cutting substrate. The reaction was carried out in a 96-well black plate. First, 93 μL of 3CL^{pro} assay reagent and 5 μL of compounds were added successively to each sample well, and DMSO was used to replace the compound in the model well, and 93 μL of assay buffer and 5 μL of DMSO were added to the control well. The 96-well plate was oscillated for 1 min to fully mix the reaction solution, and then, 2 μL of substrate was quickly added to each well and fully mixed. The 96-well plate was incubated at 37 °C in black for 15−20 min. The fluorescence was determined by a multifunction enzyme labeling reader (SpectraMaxM5, Molecular Devices) with a 340 nm excitation wavelength and 490 nm emission wavelength. The inhibition rate of the detected compounds was calculated by formula (5). IC_{50} values (n = 3) were calculated by a nonlinear regression model (log-[inhibitor] vs normalized response-variable slope) in GraphPad Prism 7 (GraphPad Software, San Diego, CA).

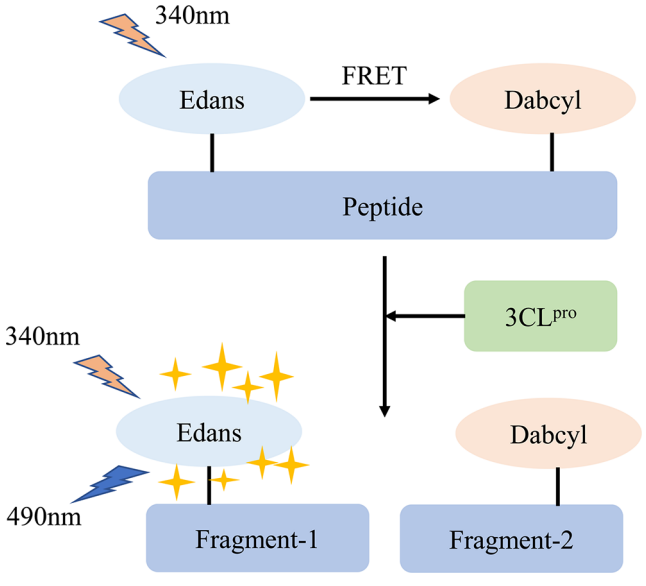


Figure 2. Detection principle of FRET for SARS-CoV-2 3CL^{pro}.

$$\text{inhibition rate (\%)} = (RFU_{enzyme} - RFU_{sample})/(RFU_{enzyme} - RFU_{control}) \times 100\% \quad (5)$$

2.8. CPE Inhibition Assay. VeroE6 cells were provided by the virus room of the State Key Laboratory of Respiratory Diseases (SKLRD), Guangzhou Institute of Respiratory Health. SARS-CoV-2 (GenBank accession no. MT123290.1, TCID50 = $10^{-6.5}$/100 μL) was obtained from the BSL-3 Laboratory of Guangzhou Customs Technology Center (Laboratory of Highly Pathogenic Microbiology of SKLRD). VeroE6 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) mixed with 10% fetal bovine serum (FBS), 100 μg/mL penicillin, and 100 μg/mL streptomycin.

VeroE6 cells were incubated in 96-well plates (1×10^4 cells/well) and cultured at 37 °C in a humidified incubator supplied with 5% CO_2. Control groups of the cell and solvent, virus group, and drug administration group were set up. After 24 h, cells were exposed to SARS-CoV-2 (100 50% tissue culture infective doses [TCID50]) for 2 h, washed, and cultured in different concentrations of compounds or fresh culture medium for 3 days. CPE was observed under a light microscope. IC_{50} values (n = 3) were calculated by the Reed−Muench method and GraphPad Prism 7. All of the above experiments were carried out in a BSL-3 laboratory.

2.9. Molecular Docking. In general, molecular docking is often used in structure-based virtual screening models to study the possible binding modes between ligands and proteins in protein complexes. Based on the CHARMm molecular force field, CDOCKER[20] in DS 2018 first randomly searches the conformations of small molecules using the molecular dynamics method and then optimizes each structure in the active site region of the receptor by simulated annealing to produce more accurate docking results. To ensure the reliability of molecular docking, we selected the crystal structure of the protein−ligand complex with a resolution of less than 2.5 Å to establish a molecular docking model. The crystal complex structure of SARS-CoV-2 3CL^{pro} and its active ligand N3 with a resolution of 2.16 Å was downloaded from the Protein Data Bank (PDB ID: 6LU7). The SARS-CoV-2 3CL^{pro} crystal complex structure was pretreated in DS 2018. The active pocket of the protein−ligand docking was defined, and then, the ligand in the SARS-CoV-2 3CL^{pro} structure was cut out and docked back to the intended active site. After docking, the molecular conformations generated by docking were compared with the original molecular conformation of

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

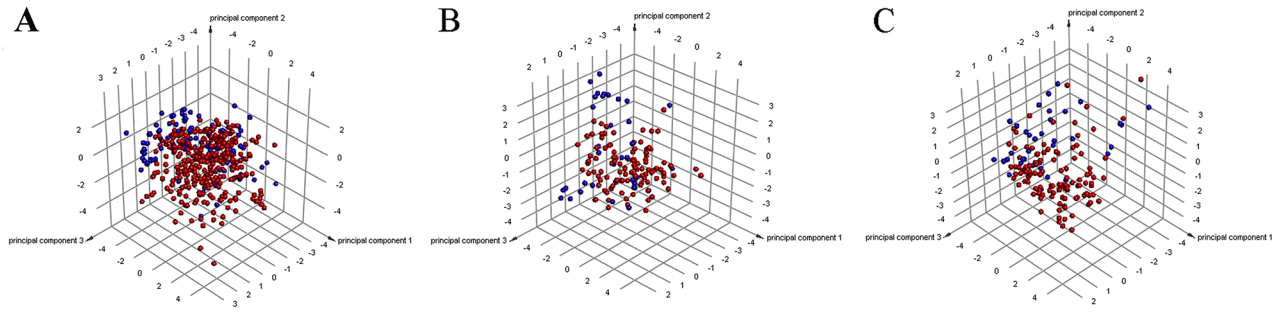


Figure 3. Chemical space analysis of the training set (A), test set (B), and external test set (C) by principal component analysis (PCA).

the ligand in the protein crystal structure, and the related root-mean-square deviations (RMSDs) were calculated. There were ten conformations, and more than half of the RMSDs were less than 2. The docking method was considered suitable for the studied system. On this basis, the compounds with potential anti-$3CL^{pro}$ activity were analyzed and verified.

3. RESULTS

3.1. Optimization of Molecular Descriptors. Through the calculation and optimization of the molecular descriptors of the training set, 12 Discovery Studio two-dimensional molecular descriptors (DS_2D_MD) including AlogP98, Es_Count_aasC, Es_Sum_ssCH2, Es_Sum_dO, logD, QED_ALOGP, QED_PSA, SAscore, Num_Rings, Num_Rings6, Num_SingleBonds, and Molecular_FractionalPolarSASA were selected for the establishment of the classification models.

Table 1. Detailed Statistical Description of the Entire Data Set

data set	inhibitors (active)	decoys (inactive)	total	Tanimoto coefficient
training set	112	337	449	0.105
test set	37	113	150	0.111
external test set	40	120	160	0.101

3.2. Chemical Spatial Diversity Analysis. PCAs of the compounds in the data sets were carried out according to the reserved 12 molecular descriptors, and the results are presented in Figure 3. The PC1 values of the compounds in the training set, test set, and external test set ranged from −6 to 6, the PC2 values ranged from −6 to 4, and the PC3 values were between −5 and 4, indicating that the chemical spatial distributions of the compounds in the three data sets were wide enough and could overlap well.

Table 3. Performance of the Five Models for the External Test Set Using Different Combinations of Molecular Descriptors

model	descriptors	number of descriptors	Q	MCC	AUC
NB-2	MD + ECFP_4	13	0.906	0.745	0.985
NB-3	MD + ECFP_6	13	0.906	0.745	0.984
RP-2	MD + ECFP_4	13	0.944	0.847	0.969
RP-3	MD + ECFP_6	13	0.944	0.847	0.969
RP-4	MD + MACCS	13	0.944	0.984	0.996

Tanimoto similarity analysis is another method commonly used to evaluate the spatial distribution of compounds in data sets. The smaller the Tanimoto similarity coefficient, the greater the diversity of compounds. We calculated the Tanimoto similarity coefficients of the chemical compounds in the training set, test set, and external test set based on the molecular fingerprint ECFP-6. As shown in Table 1, the Tanimoto similarity coefficients of the compounds in the three data sets were 0.105, 0.111, and 0.101, respectively, indicating that the compounds in the three data sets had good chemical structure diversity.

3.3. Validation of Classification Models. Based on the NB and RP algorithms, eight classification models (NB-1−NB-4 and RP-1−RP-4) were constructed using optimized 2D molecular descriptors combined with ECFP_4, ECFP_6, and MACCS molecular fingerprints. Table 2 shows the results for 5-fold cross verification and test set verification. The NB-1 and RP-1 models established only by 12 kinds of DS_2D_MD performed poorly. In the internal 5-fold cross verification of the two models, the values of MCC were 0.595 and 0.758, respectively, and in the test set verification, the values of MCC were 0.507 and 0.760, respectively. The classification models established by the combination of different molecular fingerprints and DS_2D_MD (NB-2−NB-4, RP-2−RP-4) were

Table 2. Performance of the Eight Models for the Training Set and Test Set Using Different Combinations of Molecular Descriptors

			training set			test set		
model	descriptors	number of descriptors	Q	MCC	AUC	Q	MCC	AUC
NB-1	DS_2D_MD	12	0.835	0.595	0.859	0.753	0.507	0.854
NB-2	MD + ECFP_4	13	0.982	0.953	0.991	0.980	0.946	0.999
NB-3	MD + ECFP_6	13	0.996	0.988	0.992	0.980	0.946	0.999
NB-4	MD + MACCS	13	0.871	0.718	0.940	0.893	0.756	0.969
RP-1	DS_2D_MD	12	0.891	0.758	0.968	0.900	0.760	0.926
RP-2	MD + ECFP_4	13	0.924	0.826	0.981	0.987	0.964	0.997
RP-3	MD + ECFP_6	13	0.924	0.826	0.981	0.987	0.964	0.997
RP-4	MD + MACCS	13	0.927	0.830	0.978	0.953	0.873	0.995

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

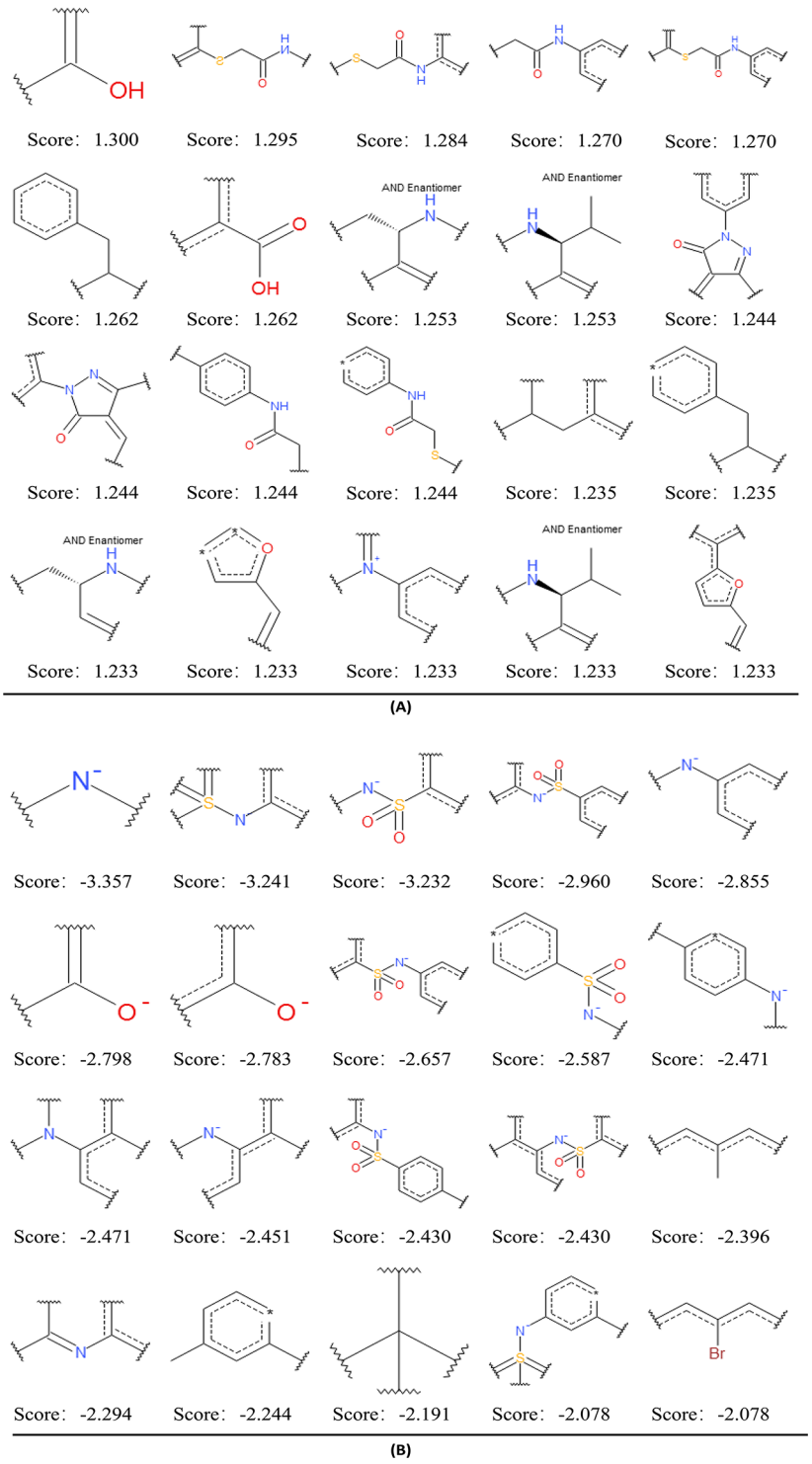


Figure 4. Examples of the top 15 good (A) and bad (B) fragments for $3CL^{pro}$ inhibition as estimated by the NB-3 model. The Bayesian score (Score) is given for each fragment.

significantly improved in terms of Q values and MCC values compared with the models established by molecular descriptors alone (NB-1 and RP-1), that is, the introduction of molecular fingerprints improved the prediction ability of the classification models to a great extent.

The NB models with molecular fingerprints ECFP_4 (NB-2) and ECFP_6 (NB-3) performed better. The MCC values of the two models in the internal 5-fold cross verification were 0.953 and 0.988, respectively, and the MCC values in the test set verification were both 0.946. The performance of the RP model with MACCS molecular fingerprint (RP-4) was better than that of the RP models with molecular fingerprints ECFP_4 (RP-2) and ECFP_6 (RP-3) in internal 5-fold cross verification, but the MCC value of RP-4 in test set verification was slightly lower than that of RP-2 and RP-3.

In addition, to further investigate the predictive ability of the models, 40 compounds with potential $3CL^{pro}$ inhibitory activity were collected from the recently published literature

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

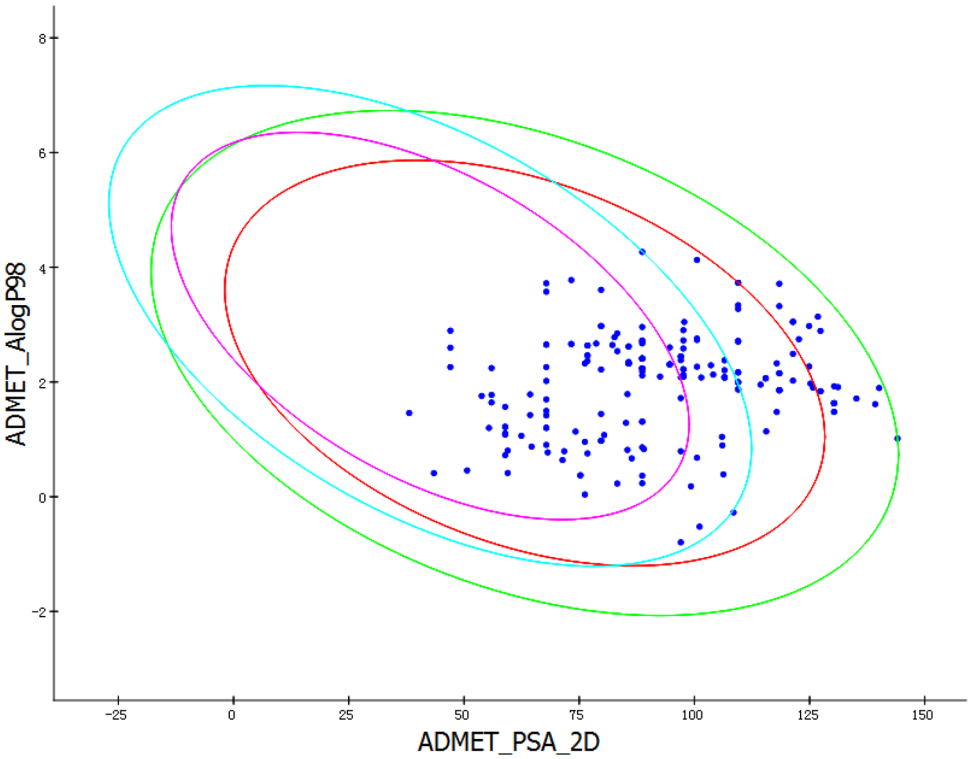


Figure 5. Distribution of ADME parameters of about 202 compounds.

Table 4. Inhibitory Activity of (+)-Shikonin, Shikonin, Scutellarein, 5,3′,4′-Trihydroxyflavone, and Ebselen against SARS-CoV-2 3CL^{pro} and 5,3′,4′-Trihydroxyflavone against SARS-CoV-2 in VeroE6 Cells[a]

	3CL^{pro} FRET assay		SARS-CoV-2 VeroE6 cells	
compounds	IC_{50} (μM)	IC_{50} (μM)	TC_{50} (μM)	SI
(+)-shikonin	4.38	>100	ND	
shikonin	4.50	>100	ND	
scutellarein	19.92	>100	>175	
5,3′,4′-trihydroxyflavone	87.76	8.217	131.66	16
ebselen	0.76	ND	ND	

[a]ND, not determined.

and combined with 120 decoys to form an external test set. The performances of NB-2, NB-3, RP-2, RP-3, and RP-4 were better in internal 5-fold cross verification and test set verification, so an external test set was carried out to further validate the above models, and the results are shown in Table 3. The NB models in external test set verification were higher in Q values and AUC values but lower in MCC values. The RP model with the MACCS molecular fingerprint had a higher Q value, MCC value, and AUC value in the external test set verification. Considering the results of internal 5-fold cross verification, test set verification, and external test set verification, five models, including NB-2, NB-3, RP-2, RP-3, and RP-4, were used to comprehensively predict the natural product molecular database of our laboratory.

3.4. Dominant and Inferior Structural Fragments Analysis. The introduction of fingerprints into the NB model provides information on the dominant and inferior structural fragments that play a crucial part in active compounds. Fifteen dominant fragments and fifteen inferior fragments were obtained by analyzing the Bayesian scores of structural fragments from the NB-3 (MD + ECFP_6) model, which provided a reference for the rational design of 3CL^{pro} inhibitors. As shown in Figure 4, most of the 15 dominant fragments contained amide bonds, and most of the 15 inferior fragments contained sulfonyl and nitrogen negative ions, which suggested that the existence of amide bonds was beneficial to inhibiting the activity of 3CL^{pro}, while the existence of sulfonyl and nitrogen negative ions was not conducive to the inhibition of 3CL^{pro} activity.

3.5. Prediction Results for Compounds. A total of 5766 natural chemical components in the database of our laboratory were predicted, among which 347 compounds were identified as active compounds by five models, and the EstPGood values of 347 compounds were more than 0.6 in the NB-3 (MD + ECFP_6) model. Further ADME analysis was carried out to remove the chemical compounds that fit any of the listed conditions: (1) the solubility was no more than 8, (2) CYP2D6 enzyme inhibition activity was true, (3) the absorption availability was greater than or equal to 2. There were 202 compounds left. The distribution of ADME parameters is given in Figure 5. After that, toxicity prediction analysis was carried out to eliminate the compounds with toxicity possibilities greater than 0.7. Finally, 139 compounds were retained, and 31 compounds (Supporting Information Table S1) were selected for further *in vitro* activity detection.

3.6. FRET Detection of SARS-CoV-2 3CL^{pro}. Taking ebselen as a reference compound, the inhibitory activity of 31 compounds on SARS-CoV-2 3CL^{pro} was detected using the FRET technique. As shown in Table 4 and Figure 6, the IC_{50} value of ebselen detected was 0.76 μM (Figure 6A), which was similar to that previously reported (IC_{50} = 0.67 μM). Among the 31 compounds, (+)-shikonin and shikonin had strong activity against SARS-CoV-2 3CL^{pro}, and the IC_{50} values were 4.38 μM and 4.50 μM, respectively. The IC_{50} value of

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

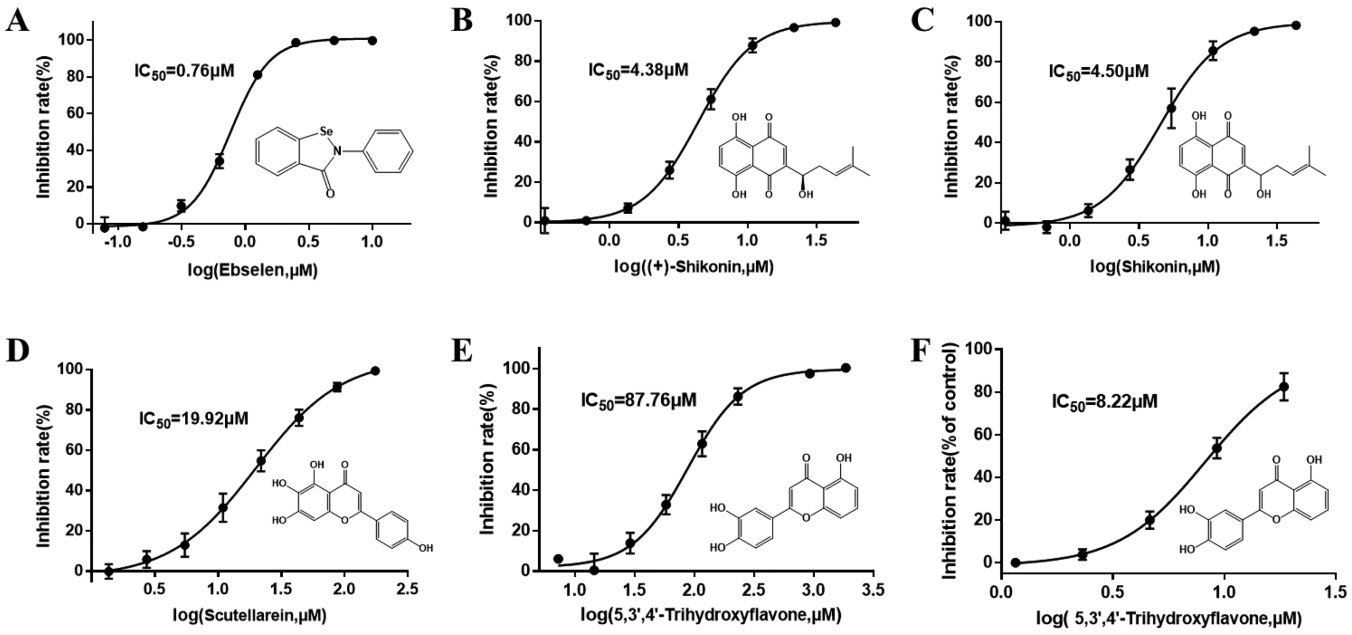


Figure 6. Inhibitory curves and IC_{50} values for the reference compound ebselen (A) and (+)-shikonin (B), shikonin (C), scutellarein (D), and 5,3′,4′-trihydroxyflavone (E) against SARS-CoV-2 $3CL^{pro}$ and 5,3′,4′-trihydroxyflavone against SARS-CoV-2 in VeroE6 cells (F).

scutellarein was 19.92 μM, and the IC_{50} value of 5,3′,4′-trihydroxyflavone was 87.76 μM. Overall, the four compounds showed definite SARS-CoV-2 $3CL^{pro}$ inhibitory activity (Figure 6B−E).

3.7. Cytopathic Effect Assay in VeroE6 Cells. On the basis of $3CL^{pro}$ inhibitory activity detection of the compounds, the active compounds were further tested for cellular-level activity inhibiting SARS-CoV-2, which was estimated through the CPE of VeroE6 cells under viral infection. It was reported that (+)-shikonin and shikonin could not inhibit the replication of SARS-CoV-2,[21] and then, the antiviral effect of scutellarein and 5,3′,4′-trihydroxyflavone was evaluated against SARS-CoV-2 in VeroE6 cells. According to the results of the CPE assay, 5,3′,4′-trihydroxyflavone showed certain antiviral effects (Figure 6F, IC_{50} = 8.22 μM). The median toxic concentration (TC_{50}) value of 5,3′,4′-trihydroxyflavone in the absence of viral infection was 131.66 μM, and the selection index (SI) was 16 (Table 4). The antiviral activity and cytotoxicity of 5,3′,4′-trihydroxyflavone showed a good tendency to separate, suggesting that 5,3′,4′-trihydroxyflavone may be a promising candidate for further research to help develop more potent $3CL^{pro}$ inhibitors against SARS-CoV-2.

3.8. Verification of Molecular Docking. Furthermore, the binding modes of 5,3′,4′-trihydroxyflavone, scutellarein, and shikonin with SARS-CoV-2 $3CL^{pro}$ were revealed by CDOCKER (Figure 7). The original ligand N3 of SARS-CoV-2 $3CL^{pro}$ could form seven hydrogen bonds with amino acid residues of Glu166, His163, Gly143, Thr190, Gln189, His164, and Phe140 and carbon−hydrogen bonds with amino acid residues of Gln189, His164, Glu166, Met165, and His172. What is more, the potential interactions also included pi−alkyl interactions with Ala191 and Pro168 and alkyl interactions with Leu167, Met49, His41, and Met165. 5,3′,4′-Trihydroxyflavone could form hydrogen bonds similar to N3 with His163, Phe140, and Glu166. In addition, 5,3′,4′-Trihydroxyflavone could form another hydrogen bond with Ser144, pi−alkyl interaction with Met165, and pi−pi T-shaped interaction with His41. Scutellarein could form hydrogen bonds, pi−alkyl bonds, and pi−pi T-shaped interaction similar to 5,3′,4′-trihydroxyflavone with His163, Phe140, Glu166, Met165, and His41. However, scutellarein also could form carbon−hydrogen bonds and pi−sulfur interaction with Arg188 and Cys145, respectively. Shikonin could interact with His163 to form hydrogen bonds similar to N3, with Gln189 to form carbon−hydrogen bonda; with Met49, His41, and Met165 to form alkyl interactions; and with Cys145 to form pi−sulfur interactions.

4. DISCUSSION AND CONCLUSIONS

To date, the spread of the novel coronavirus has disrupted the normal life order of many countries around the world and has laid a heaven burden on the country's economic development. At present, related vaccines against the virus have been introduced into the market, and people in many countries have been vaccinated, but the adverse reactions and effective duration after vaccination still need further clinical confirmation. Although relevant drugs are also under urgent development, there are still no specific drugs in the market, so screening and identifying all potential and available drugs are still important for controlling and alleviating the epidemic. $3CL^{pro}$ is an enzyme necessary for coronavirus replication that can cleave polymers to produce nonstructural proteins and may also interfere with the host's innate antiviral immune response. $3CL^{pro}$ is highly conserved in different coronaviruses and has no homologous protein in humans. Inhibiting the activity of this enzyme can effectively interfere with virus replication and proliferation and reduce mutation-mediated drug resistance.

In this study, NB and RP algorithms were used to establish classification models for $3CL^{pro}$. First, active compounds and inactive compounds of $3CL^{pro}$ were collected, and molecular descriptors were optimized by correlation evaluation and stepwise linear regression. Then, eight classification models were established based on the optimized molecular descriptors combined with ECFP_4, ECFP_6, and MACCS molecular fingerprints. According to the results of 5-fold cross verification, test set verification and external test set

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

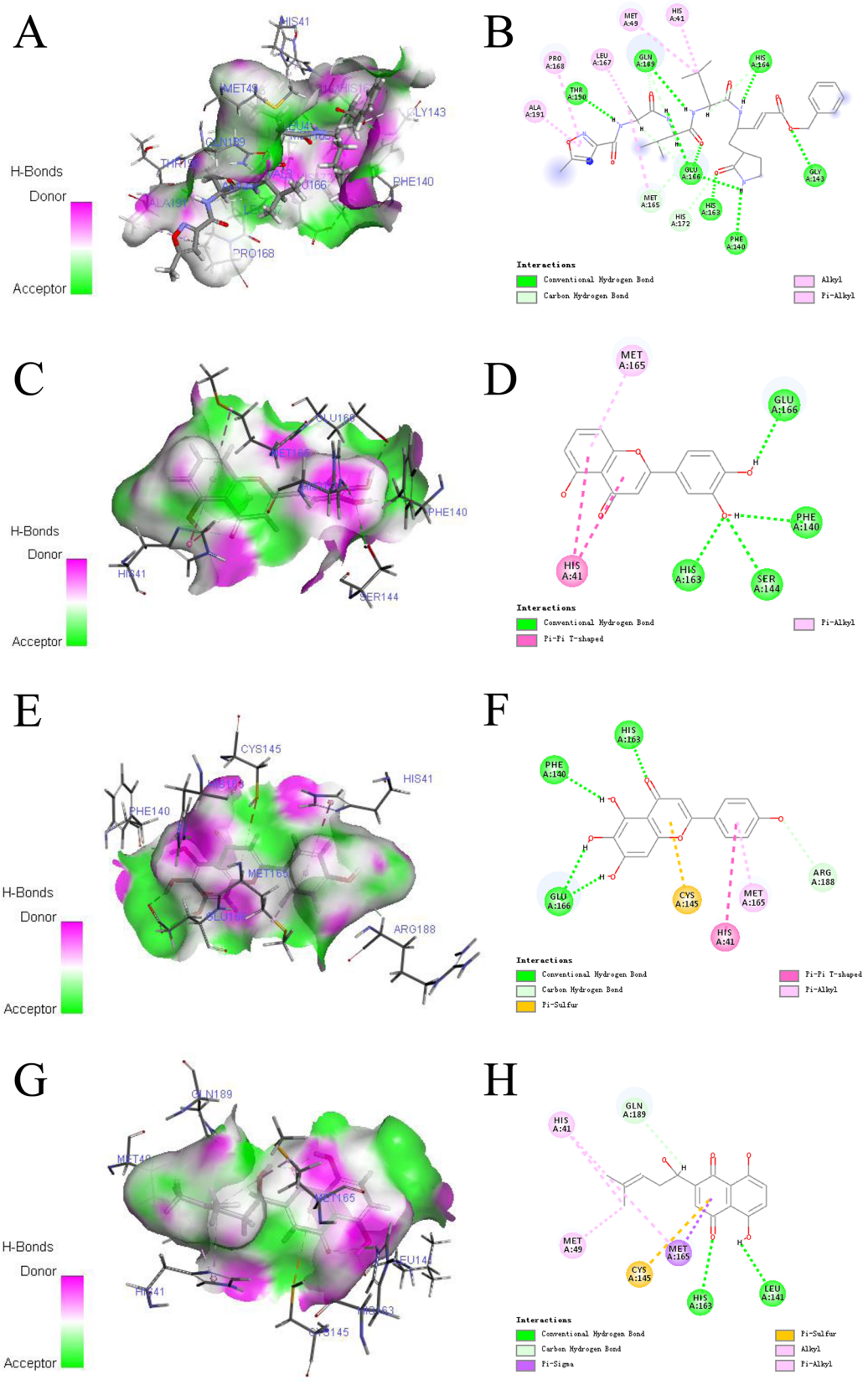


Figure 7. Receptor−ligand interactions of N3 (A, B), 5,3′,4′-trihydroxyflavone (C, D), scutellarein (E, F), and shikonin (G, H) with the active site of SARS-CoV-2 $3CL^{pro}$.

verification, the optimal models were selected. Through the prediction of the natural product molecular database collected and integrated by our previous work, 139 chemical components were predicted to be positive and had good ADMET parameters. Thirty-one compounds were further tested *in vitro* by the FRET method, among which

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(+)-shikonin, shikonin, scutellarein, and 5,3′,4′-trihydroxyflavone showed certain activity inhibiting SARS-CoV-2 3CL^{pro}. In the CPE assay, 5,3′,4′-trihydroxyflavone showed an antiviral effect. Also, the possible binding modes of 5,3′,4′-trihydroxyflavone, scutellarein, and shikonin with SARS-CoV-2 3CL^{pro} were analyzed through CDOCKER in DS 2018.

Shikonin, a purple-red tea quinone natural pigment extracted from the root of the natural plant Zongfu, possesses anticancer, anti-inflammatory, and antibacterial functions and is mainly used in the treatment of acute icteric or nonicteric hepatitis and chronic hepatitis. It has been reported that shikonin can effectively inhibit the activity of SARS-CoV-2 3CL^{pro} in FRET analysis,[22] which is consistent with the result of our study. Scutellarein, a flavonoid mainly existing in Erigeron karvinskianus, owns anti-inflammation functions, relieves pain, dispels wind and dampness, and so on. Studies have shown that it has certain inhibitory activity against coronavirus.[23] We further verified its activity in inhibiting SARS-CoV-2 3CL^{pro} by virtual screening and FRET analysis. However, shikonin and scutellarein did not show the activity of inhibiting SARS-CoV-2 in the CPE assay. There was no related report on 5,3′,4′-trihydroxyflavone having inhibitory activity on SARS-CoV-2 3CL^{pro} and SARS-CoV-2. We first found that 5,3′,4′-trihydroxyflavone had certain inhibitory effects on SARS-CoV-2 3CL^{pro} with FRET detection and SARS-CoV-2 in the CPE assay.

Based on the above analysis, NB and RP virtual screening models were established for the first time to predict the active natural products against 3CL^{pro}. The inhibitory activity of 5,3′,4′-trihydroxyflavone on SARS-CoV-2 3CL^{pro} in FRET detection and SARS-CoV-2 in the CPE assay was reported first. The binding modes of 5,3′,4′-trihydroxyflavone with SARS-CoV-2 3CL^{pro} were explained and verified by molecular docking. This study lays a foundation for further *in vivo* and clinical research and speeds up the discovery of new drugs against novel coronavirus.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01089.

Information of 31 compounds selected by optimal models and ADMET (Table S1) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ailin Liu − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*; Email: liuailin@imm.ac.cn

Zifeng Yang − *State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong 510000, China*; Email: jeffyah@163.com

Guanhua Du − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*; Email: dugh@imm.ac.cn

Authors

Jun Zhao − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*

Qinhai Ma − *State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong 510000, China*

Baoyue Zhang − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*

Pengfei Guo − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*

Zhe Wang − *Beijing Key Lab of Drug Target Identification and Drug Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*

Yi Liu − *College of Chemical Engineering, Sichuan University of Science & Engineering, Zigong, Sichuan 643000, China*

Minsi Meng − *College of Chemical Engineering, Sichuan University of Science & Engineering, Zigong, Sichuan 643000, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01089

Author Contributions

⊥Jun Zhao and Qinhai Ma are co-first authors.

Notes

The authors declare no competing financial interest.

A total of 5766 predicted compounds are derived from the natural product database of the screening Center Laboratory of Institute of Medicine, Chinese Academy of Medical Sciences, and are not open to the public. Other databases can be predicted by our model. The process of data collection and model prediction can be found in the method section of this paper, and the database involved are the BindingDB database (http://www.bindingdb.org) and the DUD-E database (http://dude.docking.org). Discovery Studio version 2018, which comes from BIOVIA, is paid software. PaDEL-Descriptor software can be downloaded at http://padel.nus.edu.sg/software/padeldescriptor.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81673480), the Drug Innovation Major Project (Nos. 2018ZX09711001-003-002, and 2018ZX09711001-012), and the CAMS Major collaborative innovation fund for major frontier research (2020-I2M-1-003).

■ REFERENCES

(1) World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard[EB/OL]. https://covid19.who.int/ (accessed March 3, 2021).

(2) World Health Organization. Draft Landscape of COVID 19 Candidate Vaccines[EB/OL]. https://www.who.int/who-documents-detail/draft-landscape-of-covid-19-candidate-vaccines (accessed March 3, 2021).

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(3) Wang, Y.; Zhang, D.; Du, G.; Du, R.; Zhao, J.; Jin, Y.; Fu, S.; Gao, L.; Cheng, Z.; Lu, Q.; Hu, Y.; Luo, G.; Wang, K.; Lu, Y.; Li, H.; Wang, S.; Ruan, S.; Yang, C.; Mei, C.; Wang, Y.; Ding, D.; Wu, F.; Tang, X.; Ye, X.; Ye, Y.; Liu, B.; Yang, J.; Yin, W.; Wang, A.; Fan, G.; Zhou, F.; Liu, Z.; Gu, X.; Xu, J.; Shang, L.; Zhang, Y.; Cao, L.; Guo, T.; Wan, Y.; Qin, H.; Jiang, Y.; Jaki, T.; Hayden, F. G.; Horby, P. W.; Cao, B.; Wang, C. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* **2020**, *395*, 1569−1578.

(4) World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—25 May 2020[EB/OL], 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---25-may-2020.

(5) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Kruger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N. H.; Nitsche, A.; Muller, M. A.; Drosten, C.; Pohlmann, S. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271−280.e8.

(6) Helmy, Y. A.; Fawzy, M.; Elaswad, A.; Sobieh, A.; Kenney, S. P.; Shehata, A. A. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. *J. Clin. Med.* **2020**, *9*, No. 1225.

(7) Yang, H.; Yang, M.; Ding, Y.; Liu, Y.; Lou, Z.; Zhou, Z.; Sun, L.; Mo, L.; Ye, S.; Pang, H.; Gao, G. F.; Anand, K.; Bartlam, M.; Hilgenfeld, R.; Rao, Z. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13190−13195.

(8) Tahir Ul Qamar, M.; Alqahtani, S. M.; Alamri, M. A.; Chen, L. L. Structural basis of SARS-CoV-2 3CL(pro) and anti-COVID-19 drug discovery from medicinal plants. *J. Pharm. Anal.* **2020**, *10*, 313−319.

(9) Gyebi, G. A.; Ogunro, O. B.; Adegunloye, A. P.; Ogunyemi, O. M.; Afolabi, S. O. Potential inhibitors of coronavirus 3-chymotrypsin-like protease (3CL(pro)): an in silico screening of alkaloids and terpenoids from African medicinal plants. *J. Biomol. Struct. Dyn.* **2020**, 1−13.

(10) Llanos, M. A.; Gantner, M. E.; Rodriguez, S.; Alberca, L. N.; Bellera, C. L.; Talevi, A.; Gavernet, L. Strengths and Weaknesses of Docking Simulations in the SARS-CoV-2 Era: the Main Protease (Mpro) Case Study. *J. Chem. Inf. Model.* **2021**, *61*, 3758−3770.

(11) Ly, A.; Marsman, M.; Wagenmakers, E. J. Analytic posteriors for Pearson's correlation coefficient. *Stat. Neerl.* **2018**, *72*, 4−13.

(12) Rabal, O.; Amr, F. I.; Oyarzabal, J. Novel Scaffold FingerPrint (SFP): applications in scaffold hopping and scaffold-based selection of diverse compounds. *J. Chem. Inf. Model.* **2015**, *55*, 1−18.

(13) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem,* **2011**, *32*, 1466−1474.

(14) Vogt, M.; Bajorath, J. Modeling Tanimoto Similarity Value Distributions and Predicting Search Results. *Mol. Inf.* **2017**, *36*, No. 1600131.

(15) Bender, A. Bayesian methods in virtual screening and chemical biology. *Methods Mol. Biol.* **2011**, *672*, 175−96.

(16) Stegmann, G.; Jacobucci, R.; Serang, S.; Grimm, K. J. Recursive Partitioning with Nonlinear Models of Change. *Multivar. Behav. Res.* **2018**, *53*, 559−570.

(17) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

(18) Janssens, A. C.; Martens, F. K. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int. J. Epidemiol.* **2020**, *49*, 1397−1403.

(19) Zauner, T.; Berger-Hoffmann, R.; Muller, K.; Hoffmann, R.; Zuchner, T. Highly adaptable and sensitive protease assay based on fluorescence resonance energy transfer. *Anal. Chem.* **2011**, *83*, 7356−7363.

(20) Gagnon, J. K.; Law, S. M.; Brooks, C. L., 3rd. Flexible CDOCKER: Development and application of a pseudo-explicit structure-based docking method within CHARMM. *J. Comput. Chem.* **2016**, *37*, 753−762.

(21) Gurard-Levin, Z. A.; Liu, C.; Jekle, A.; Jaisinghani, R.; Ren, S.; Vandyck, K.; Jochmans, D.; Leyssen, P.; Neyts, J.; Blatt, L. M.; Beigelman, L.; Symons, J. A.; Raboisson, P.; Scholle, M. D.; Deval, J. Evaluation of SARS-CoV-2 3C-like protease inhibitors using self-assembled monolayer desorption ionization mass spectrometry. *Antiviral Res.* **2020**, *182*, No. 104924.

(22) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289−293.

(23) Russo, M.; Moccia, S.; Spagnuolo, C.; Tedesco, I.; Russo, G. L. Roles of flavonoids against coronavirus infection. *Chem.-Biol. Interact.* **2020**, *328*, No. 109211.

https://doi.org/10.1021/acs.jcim.1c01089
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX