Development of Machine Learning Models to Predict a Chemical's Anti-SARS-CoV-2 Activities

Beihong Ji, Yuhui Wu, Elena N Thomas, Jocelyn N Edwards, Xibing He, Junmei Wang*

Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15261, USA.

* Corresponding author, junmei.wang@pitt.edu

Abstract

The COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has put global health systems at risk, leading to an urgent need for effective treatment for infection of this coronavirus. To accelerate the identification of novel drug candidates for COVID-19 treatment in the drug discovery process, we reported a series of ML-based models to accurately predict the anti-SARS-CoV-2 activities of screening compounds. Those models were trained and evaluated using the experimental data deposited in the COVID-19 OpenData Portal which is hosted by NCATS (https://ncats.nih.gov/expertise/covid19-open-data-portal). We explored 6 popular ML algorithms in combination with 15 molecular descriptors for molecular structures from 9 screening assays. Of note, 6 screening assays of the same datasets were also adopted by KC et al. to construct prediction models which were deployed in the REDIAL-2020 model suite (Nature Machine Intelligence, **3**, 527–535, 2021). The impacts of ML algorithms and molecular descriptors on model performance were investigated. As a result, the model constructed using the k-nearest neighbors (KNN) method and the hybrid molecular descriptor, GAFF+RDKit, achieved the best performance. We evaluated the model performance on 28 drugs which have been applied in clinical trails of treating COVID-19. The overall performance of our developed models was better than REDIAL-2020. For the external CPE dataset, 79% of compounds were correctly predicted by using our model, significantly better than REDIAL-2020 (66.7%). For the external 3CL assay, the percentages of correct predictions by our predictors (38.1%) are not as high as REDIAL-2020 (61.9%). However, our models achieved more accurate predictions for the 100 druglike compounds selected as negative control. Furthermore, we reconstructed another 3CL model by utilizing the screen data from the study by Kuzikov, et al. The classification model achieved the best performance on the prediction of positive control, albeit its performance is lower than

REDIAL-2020 on the prediction for the negative control. A web server (https://clickff.org/amberweb/covid-19-cp) was developed to enable users to forecast anti-SARS-CoV2 activities of arbitrary compounds. The web portal provides users a fast and reliable way to identify potential compound candidates for COVID-19 treatment, which highly reduces the time and cost of experiments on anti-SARS-CoV activity.

Introduction

Since 2019, the COVID-19 pandemic has outbroken and put global health systems at risk[1]. So far, this novel coronavirus disease, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to more than 250 million people infected with mortality reaching over 5 million[2]. Although the mortality and spread of COVID-19 has been suppressed due to rapidly increasing vaccination rates, there is still an urgent need for effective drug treatment for the large-scaling infection of this coronavirus. To speed up the identification of novel candidates for COVID-19 treatment in the drug discovery process, machine learning (ML) has stood out as a powerful tool for its efficiency and reliability in drug screenings[3-5].

In 2020, KC, et al.[6] proposed a suite of ML models to forecast activities of small molecules for SARS-CoV-2 from molecular structures related to several SARS-CoV-2 assays. The models they developed, coined "REDIAL-2020", offer a convenient and efficient way to screen novel molecules for anti-SARS-CoV-2 activities. In this work, we further improved the performance of the prediction models by exploring different ML models and molecular descriptors regarding molecular structures. Moreover, we created prediction models for three more screening assays. In total six popular ML algorithms which include support vector machine (SVM)[7], logistic regression (LR)[8], decision tree (DT)[9], Random Forest (RF)[10], k-nearest neighbors (KNN)[11], and complement Naïve Bayes (NB)[12], were applied to construct prediction models. A variety of molecular descriptors, including fingerprint (FP2, FP3, FP4 and MACCS), atom type counts (GAFF), molecular properties (RDKit), were first applied to construct models. Their model performance was applied to guide the design of hybrid molecular descriptors. In total, 9 mixed-type molecular descriptors were exploited, including RDKit+FP2, RDKit+FP3, RDKit+FP4, RDKit+MACCS, GAFF+FP2, GAFF+FP3, GAFF+FP4, GAFF+MACCS, and GAFF+RDKit. All the 15 molecular

descriptors were applied to train experimental screening data collected in the COVID-19 OpenData Portol (https://ncats.nih.gov/expertise/covid19-open-data-portal) which was hosted by National Center for Advancing Translational Sciences (NCATS). The 9 screening assays studied in this study belong to four categories, which are (1) viral replication, (2) live virus infectability, (3) viral entry, and (4) counterscreen. Both KC et al. and we studied the first six screening assays: 3CL enzymatic activity (3CL) in Catogory 1, SARS-CoV-2 cytopathic effect CPE in Catogory 2, Spike-ACE2 protein-protein interaction AlphaLISA assay and ACE2 enzymatic activity in Category 3, SARS-CoV-2 cytopathic effect counterscreen and Spike-ACE2 protein-protein interaction TruHit conunterscreen in Category 4. Besides the above six assays, we also studied three more screening assays: TMPRSS2 enzymatic activity in Category 3, HEK 293 cell line toxicity and human fibroblast toxicity in Category 4.

For the only assay in Category 1, the papain-like proteinase 3CL cleaves SARS-CoV-2 polyprotein into individual proteins, which is a key process in the viral life cycle[13]. Inhibiting polyprotein cleavage can interrupt viral replication, making 3CL an attractive target in drug discovery and development. For assays in Category 2, the SARS-CoV-2 cytopathic effect (CPE) assay serves to measure the potential of compounds to reverse the cytopathic effect of the virus in Vero E6 host cells[14]. Thus, this assay can identify compounds with the potential to protect host cells from the CPE of the virus. Three assays belong to Category 3, measuring the ability of a compound inhibiting viral entry. The surface angiotensin-converting enzyme type 2 (ACE2) has been known as the primary host factor identified and targeted by SARS-CoV-2 virions[15,16]. The attachment of viral capsid to the host cell is facilitated by the SARS-CoV-2 Spike protein binding to the host ACE2, which trigger a multistep process of viral entry resulting in delivery of the viral genome to the cytosol, the site of replication. As a result, the disruption of the Spike-ACE2

interaction can cripple SARS-CoV-2 virions to infect host cells. The Spike-ACE2 protein-protein interaction AlphaLISA assay is used to measure the ability of therapeutics (small molecules, etc.) to disrupt the interaction between the Spike protein and ACE2. On the other hand, ACE2 plays a role in cleaving angiotensin I hormone into the vasoconstricting angiotensin II and acts as a counter-balance to ACE. Although inhibition of the Spike-ACE2 interaction could stop viral entry, off-target effects on endogenous ACE2 function may lead to disruption of critical vasodilation pathways. After the ACE2 binding, transmembrane protease serine 2 (TMPRSS2), a host protease which is essential for Spike protein priming, has been shown playing important a role in virus-host cell membrane fusion and further infection[17]. Therefore, the ACE2 and TMPRSS2 enzymatic assays can be applied to screen compounds with the ability to interrupt endogenous enzyme functions.

There are four assays in the counterscreen category. The counterscreen of the CPE assay is host cell tox counterscreen (cytotox) and is used to measure cell cytotoxicity. Another counterscreen is the Spike-ACE2 protein-protein interaction TruHit assay that serves to identify false positives. The function of this assay is to investigate whether the activity found in a AlphaLISA assay is caused by the interference with the assay system itself or not. Two extra HEK293 cell line toxicity and human fibroblast toxicity assays are used as cell viability assays that evaluate the general human cell toxicity of compounds.

In Figure 1, we illustrate the preferred response for each assay. For the assays in the first three categories, active response is preferred, while for the assays in the conterscreen category, we expect negative response which means no interference in the companion assays.

Next, we constructed a series of models by ML methods that can screen and identify compounds with anti-SARS-CoV-2 activities. The impacts of ML algorithms and molecular

descriptors on the model performance in regard to different coronavirus-related assays were explored in this study. The final satisfactory models were deployed in a web server with multiple molecular input formats, allowing a user to forecast the activities of arbitrary small molecules against viral replication, viral entry and live virus infectivity. The web server may provide users a convenient way to prioritize virtual screening drugs prior to in vitro or in vivo assays in rational drug discovery for preventing and treating COVID-19.

Results

Data Set Preparation.

As shown in Table 1, for most assays, the total compounds in the active and inactive sets are imbalanced. For the 3CL, Fibroblast, CPE, cytotox, ACE2, and TMPRSS2 assays, the total inactives are approximately five times greater than the actives. To resolve the issue of data imbalance, there are typically two resampling methods: undersampling and oversampling[18]. Undersampling refers to randomly removing some subjects from the majority class to match the counts of samples in the minority class. In the oversampling process, a sample of synthetic data for minority class was generated to match the number of samples in majority. Considering the dramatic difference of numbers of active and inactive compounds in some assays, we performed oversampling to resample the imbalanced data. We first applied class weight to balance the data, i.e., the undersampled class had larger weight and the total weight of each class was roughly same. However, the F1 scores of the models for most assays are unsatisfactory. For example, for the CPE assay, the average F1 score of all SVM models for test sets is only 0.42. Thus, we adopted another commonly used technique, Synthetic Minority Over-sampling Technique (SMOTE) implemented in a python program. SMOTE works by looking at examples that are close in the feature space for

the minority class and draws a new sample at points along the line between the examples in the feature space[19]. By applying SMOTE technique, the average F1 score for the CPE assay under the same condition was improved to 0.55. We employed SMOTE package for almost all assays except for 3CL assay. Considering the complexity of the data in 3CL assay, we adopted a simpler RandomOverSampler[20] method to balance the data set.

For the 3CL, CPE, ACE2 and TMPRSS2 screening assays, the severity of data imbalance is very high, since for each assay the number of inactive compounds is even 10-fold larger than that of actives. Thus, we randomly divided inactive compounds for each of the four assays into 2-4 subsets based on the actual number of inactives. One inactive subset was randomly selected to participate ML-based model construction, while the rest of sample sets were taken as external datasets for further model validation. Of note, the dataset consist of all actives and inactives in the selected subset of inactives is still unbalanced. The distribution of active and inactive compounds in all assays is shown in Figure 2.

In total, five metrics have been applied to evaluate the performance of a ML-based model, including the area under a receiver operating characteristic curve (ROC AUC), accuracy (ACC), F1 score, precision (PRE) and recall (REC). ROC AUC illustrates the diagnostic ability of a binary classifier; ACC and F1 measure the accuracy of a prediction model; PRE is the fraction of true positive among those classified as positive; and REC, also known as sensitivity in binary classification, is the fraction of true positive among those should have been classified as positive. All the above metrics range from 0 to 1, where 1 indicates the best scenarios and 0 indicates the worst. The defintions of those metrics are presented in the supporting information of this work.

ML-Based Model Performance

For each screening assay collected in the COVID-19 open-data-portal, we constructed binary classification models with 15 different molecular descriptors using six different ML algorithms. The model performance measured by the validation and test sets under all conditions is presented in Table 2 and Figures S1-S2. Overall, the model performance measured by the validation sets are comparable for all assays (Table 2). Specifically, the KNN model stands out as it has higher scores of AUC (0.91) and REC (0.94), as well as comparable scores of ACC (0.80), F1(0.82) and PRE (0.73) compared to other ML methods. Meanwhile, the model performance measured by the test sets is slightly lower than that of the validation sets. The overall ranking for test sets of different machine learning models is the same, with the KNN method outperforms other ML models. The performance of KNN is generally satisfactory, which achieves the highest scores of ACC (0.68), F1 (0.69) and REC (0.71), and relatively high scores of AUC (0.74) and PRE (0.67) among all ML algorithms.

We then evaluated the performance of the KNN models constructed for all screening assays. Table 3 listed the average scores of metrics for all KNN models constructed using different molecular descriptors for individual assays. Essentially, there is no dramatic difference of those measured metrics among all the screening assays, indicating KNN is a promosing ML algorithm to be applied to construct prediction models for screening data. Notably, for test sets, the average scores of some metrics, such as AUC for CPE, cytotox and TruHit assays, are higher than the best scores reported in KC, et al.'s study. The scores of metrics for model evaluation that are better than those by KC et al. were highlighted with blue and bold font in Table 3. For example, the average AUC, ACC, F1, PRE scores of KNN model for CPE assay is 0.75, 0.69, 0.71, and 0.68, respectively, which are much higher than the values in KC et al.'s model (0.651, 0.643, 0.626, 0.661, and 0.651, correspondingly).

Impact of Molecular Descriptor on Model Performance

We compared the impact of different molecular descriptors on the model performance. For the sake of comparison, we generated heatmaps which illustrate the values of a metric with colors (the more reddish a color is, the high the value, while the more bluish a color is, the lower the value).

The heatmaps illustrated in Figure 3 depict the overall performance of fifteen molecular descriptors applied for the construction of KNN models for 9 screening assays. All the five metrics should be considered to identify the best descriptor for all screening assays. Overall, the GAFF+RDKit descriptor outperformed others since there are the least amount of blue grids for it cross all the metrics and assays. GAFF, the abbreviation of General AMBER Force Field, is designed to describe subtle chemical environments using atom types.[21]. GAFF was parametrized to be consistent with AMBER biomolecular force fields for studying protein-ligand and nucleic acid-ligand interactions. It can describe a wide range of organic or pharmaceutical molecules that are constituted of H, C, N, O, S, P, F, Cl, Br and I. Utlizing the companion software tool, Antechamber, GAFF atom type-based descriptor can be automatically generated for arbitrary organic molecules that can be modelled by GAFF.[22]. Unlike fingerprint-based descriptors which only indicate the existing or not existing of a certain substructure or structural pattern, GAFF descriptor encodes the total occurances of subtle chemical environment in a molecule. On the other hand, RDKit is a popular descriptor kit collecting molecule-level properties. By combining the features of both GAFF and RDKit, the hybdrid descriptor, GAFF/RDKit, can better discriminate the actives from the inactives for all the screening assays than either of single type of descriptor. Table 4 shows the performance for KNN models of GAFF+RDKit molecular descriptor for all

assays. For the six assays studied by both KC et al. and us, our models achieved better scores than KC et al's for more than half of the performance metrics.

Evaluation of Model Predictivity Using External Test Sets

To confirm the reliability of the constructed models, we additionally evaluated the performance of the models in best scenarios (KNN algorithm and GAFF+RDKit fingerprint) on 4 different categories of external test sets compiled from different sources, which are: (1) the NCATS compounds not participating model construction, (2) the reported drugs/compounds that have been used or tested in COVID-19 treatment, (3) the reported compounds which are active in SARS-CoV2-related bioassays, and (4) the screening compounds from ZINC database[23] (https://zinc.docking.org/) serving the negative control, i.e., those compounds are assumed as inactives. The model performance can be critically evaluated by using the five metrics (AUC, ACC, F1, PRE and REC) for the four external test sets.

Test Set 1 – NCATS Screening Compounds. As described in the **Data Set Preparation** session, we have randomly divided inactive compounds for each of the four assays (3CL, CPE, ACE2 and TMPRSS2) into 2-4 subsets based on the actual number of inactives. While one inactive subset (s1) was randlomly selected to participate ML-based model construction, we used inactive compounds from other sample sets (s2, s3, s4) to conduct external prediction. For the sake of computing the five metrics, we included the actives of each assay in the test sets. The predicted results of external datasets are displayed in Figure 4. A striking feature of this figure is that the sensitivity scores of most assays are very high (> 0.90), likely due to the actives also participated in model construction. The specificity scores of those external test sets, ranged from 0.60 to 0.86, are comparable to those reported for the test sets in model construction (Table 3). The similar

specificity scores suggest that our models were not overfitted. Note that sensitivity measures the percentage of compounds predicted to be active out of the compounds which are active confirmed in bioassay, while specificity measures the percentage of the compounds predicted to be inactive out of the compounds which are inactives confirmed in bioassay.[24] Encouragingly, the specificity scores of external test sets are relatively high and comparable to the sensitivity scores, indicating that our models have the ability to rule out both the false positives and false negatives at the same time.

Test Set 2: known anti-SARS-CoV2 drugs in multiple assays. To validate the practibility of our models, we collected 28 compounds[25,26] that have been used or tested in COVID-19 treatment. 22 out of 28 compounds are approved drugs. We predicted their activities in different assays using both "REDIAL-2020" by KC et al. and our models. Table S1 lists the prediction results for each assay by utilizing REDIAL-2020 and our webtool, COVID-19-CP, as detailed below. For all 28 compounds, the screening activities reported by NCATS Covid-19 OpenData Portal served as references. For a compound, if the predicted activity, active or inactive, is the same as the measured one, the number of correct predictions increases one, otherwise zero. If the predictions for the six assays (3CL, CPE, cytotox, ACE2, AlphaLISA, TruHit) are all correct, the number of correct predictions is 6. We calculated the number of correct predictions for each compound by using REDIAL-2020 and our predictor COVID-19-CP. Overall, the predicted results of COVID-19-CP are better than REDIAL-2020 in term of the number of correct predictions. The numbers of correct predictions of 13 compounds by COVID-19-CP are higher than those by REDAIL-2020, while 8 compounds are lower, and the rest of 7 compounds are equal. When the performance of a specific assay is concerned, the percentage of correct prediction differs between REDAIL-2020 and COVID-19-CP from one assay to another. The percentages of correct

prediction are similar for cytotox (~60%) and TruHit (~40%); COVID-19-CP has larger percentages of correction prediction for 3CL (79% vs 68%), CPE (82% vs 68%) and ACE2 (61% vs 50%), while REDIAL-2020 achives a better performance for AlphaLISA (75% vs 61%).

The following are the prediction results for some interesting compounds. Although the experimental activity of Lopinavir is measured inactive, it was actually reported as 3CL protease inhibitors[27]. REDAIL-2020 predicted it inactive, while COVID-19-CP predicted the compound active. Two more interesting compounds are chloroquine and hydroxychloroquine, which were hypothesized to be ACE2 blockers, however, the ACE2 assay suggests those two compounds are inactive. As shown in Table S1, COVID-19-CP made correct prediction, in contrast, REDIAL-2020 made the opposite prediction.

A set of 10 drug molecules predicted to have potential to be repurposed to treat COVID-19 are shown in Figure 5. Those drug molecules meet the following two criteria: 1. The predicted accuracy of COVID-19-CP is higher than that of REDIAL-2020; 2. The number of correct predictions by COVID-19-CP is larger than 3, i.e., the overall prediction accuracy is higher than 50%. The assays correctly predicted by COVID-19-CP are colored in green. It is shown that our predictor can correctly predict the activities of these drugs for most of assays. In addition, the developed models can predict activities of extra assays. For example, ribavirin and nafamostat are all TMPRSS2 inhibitors[28,29], and their activities on TMPRSS2 assays were correctly predicted by our models.

Test set 3: additional active compounds in individual bioassays. We further evaluated the performance of COVID-19-CP using an external CPE dataset which was also adopted by KC et al. in their model evaluation process. Table S2 lists the names and SMILES of the 24 drugs which are actives in CPE assay. Among the 24 compounds, 19 of them were correctly predicted

as active by our model, while only 5 of them were predicted as inactive. Thus, our model achieved a prediction accuracy of 79.2% for the external data set. In contrast, the percentage of correct prediction by REDIAL-2020 was 66.7%, significantly lower than our model. The prediction results by REDIAL-2020 and COVID-19-CP for 21 3CL inhibitors collected from Kuzikov, et al[30] were also compared. As shown in Table S3, the percentages of correct predictions by COVID-19-CP (38.1%) are not as high as REDIAL-2020 (61.9%). Thus COVID-19-CP achieved a comparable performance to REDIAL-2020 for the CPE external test set rather than the 3CL test set. To improve the prediction performance of COVID-19-CP for the 3CL test set, we reconstructed the model using the sreen data reported in Kuzikov, et al.'s study. The detail for the model reconstruction is described in the next section.

Test set 4: screening compounds serving as negative control. The above three test sets mainly evaluated the models' ability to identify true actives, while this test set can be applied to assess the models' ability to reduce false positives. To this aim, we randomly collected 100 screening compounds from the ZINC database and assuming those compounds are inactive in the screenings. Table S4 lists the activities predicted by REDIAL-2020 and COVID-19-CP for four assays (3CL, CPE, AlphaLISA for REDIAL-2020 and COVID-19-CP, TMPRSS2 for COVID-19-CP) which directly measure a compound's antiviral activities. The predictions of 3CL, CPE and AlphaLISA assays by REDIAL-2020 and COVID-19-CP were compared. According to Table S4, 33 out of 100 compounds have fewer positive predictions by COVID-19-CP than REDIAL-2020, while 31 compounds have fewer positive predictions by REDIAL-2020 than COVID-19-CP. The results indicate that COVID-19-CP performs better than REDIAL-2020 not only for the known inhibitors (positive control), but also for the screening compounds (negative control). Among the 100 screening compounds, 37 and 10 were predicted to be active for AlphaLISA assay and 3CL

assay by COVID-19-CP, respectively. It is of noted that the proportion of inactives in the screening set is close to the proportion of actives in the total dataset for both assays. Specifically, the active rate is 1018/(2269+1018)=31.0% for the AlphaLISA assay, while the value is 431/(431+4716)=8.4% for the 3CL assay. The similar positive rates of both assays demonstrate the high reliability of our models.

New 3CL model construction

We constructed a second model for the 3CL protease using the screen data reported by Kuzikov, et al. to further improve the performance of COVID-19-CP. The structures in SMILES and the inihibition data of screening compounds were first collected and 7662 compounds left after data cleaning. Then the compounds were ranked based on their percent inhibition values, and those with percent inhibition values larger than 25% were allocated into the active set, while the rest of compounds were randomly allocated into two inactive subsets. In detail, there are 342 compounds in active set, 3665 in s1 inactive subset, and 3655 compounds in s2 inactive subset. As we did for the NCATS 3CL assay, s1 subset was selected to construct and test the ML-based models, while s2 served as an external dataset for further model validation.

The treatment of data imbalance and model construction were the same as we did for the NCATS 3CL assay. Again, the KNN algorithm and RDKit+GAFF molecular descriptor were employed for model construction. Table 5 shows the scores of performance metrics for the new 3CL model. According to the table, for both validation and test sets, every metric has better predicted score for the new 3CL model than that constructed using the NCATS data. The sensitivity and specificity scores of s1 are 0.62 and 0.89, which are better than the corresponding scores of NCATS 3CL s1 subset predicted by the old 3CL model (0.57 and 0.83). As for the

external sets, the sensitivity and specificity scores of s2 by the new 3CL model are 0.93 and 0.89, respectively, both higher than the corresponding values achieved by the old 3CL model for the NCATS 3CL s2 (0.89 and 0.86) and s3 (0.89 and 0.86) subsets.

In addition to the evaluation of inactive subsets, we evaluated the the new 3CL model for 3 test sets as we did for the 3CL model constructed using the NCATS 3CL assay. For the known anti-SARS-CoV2 drugs test set, the new model did not make changes on the overall performance of COVID-19-CP. There are still 13 compounds with correct predictions by COVID-19-CP higher than those by REDAIL-2020, while 8 compounds are lower, and the rest of 7 compounds are equal. The prediction results by the new model for the drug molecules in 3CL assay were summarized in Table S5. The third test set is the active compounds in single 3CL assay. Since the compounds are originally from Kuzikov, et al's study, the performance of the new model for this set is reasonably better than the old model, with 16 out of 21 compounds predicted active against only 8 compounds predicted active by the old model (Table S6). The fourth test set is the 100 screening compounds which serve as the negative control. Compared to the positive effect brought by the new 3CL model, the overall performance of COVID-19-CP on the prediction of screening compounds slightly decreased. As shown in Table S7, 30 out of 100 compounds have fewer predictions as "active" by COVID-19-CP than REDIAL-2020, which is less than the number of compounds (37) with fewer predictions as "active" by REDIAL-2020 than COVID-19-CP.

Overall, the second 3CL model can improve the performance of COVID-19-CP on the positive control but negatively affect the performance of COVID-19-CP on the negative control. Therefore, to obtain a more promising and accurate prediction results, we suggested adopt the concensus strategy which simultaneously considers the prediction results of both 3CL models constructed in this work and the REDIAL model.

Dissemination of Prediciton Models Via COVID-19-CP Web portal

To facilitate the dissemination of the prediction models, we developed a Web portal (https://clickff.org/amberweb/covid-19-cp). Users can access the web server that is integrated with the optimal KNN models and GAFF+RDKit molecular features for fast screening compounds that have potential treatment for for COVID-19. Specifically, a user can open the webpage from a web browser, then input a molecular structure via different methods, and then submit the job to obtain the predicted activities of all 9 assays. Users can not only upload a mol2 or sdf file, but also draw 2-dimentional structures of compounds with a molecular Editor. Once the web portal receives the molecular structure (mol2/sdf/smi format), it will automatically generate GAFF+RDKit descriptors and feed the input data to the trained KNN models. After processing for a short time, the built-in models will provide the predicted activities of the input compound in 9 screening assays, which are 3CL, HEK293, Fibroblast, CPE, cytotox, ACE2, AlphaLISA, TruHit, TMPRSS2. Figure 6 shows a sample submission page and the output page of the web portal. To summarize, an ideal anti-SARS-CoV-2 compound candidate meets the following criteria: (i) active in 3CL assay, (ii) inactive in HEK293 assay, (iii) inactive in Fibroblast assay, (iv) active in CPE assay, (v) inactive in cytotox assay, (vi) active in ACE2 assay, (vii) active in AlphaLISA assay, (viii) inactive in TruHit assay, (ix) active in TMPRSS2 assay.

As shown in Figure 6, in the output page, a structure similarity search section is provided immediate after the table summarizing the prediction result for users to search similar compounds in three databases, Drugbank,[31,32] ChEMBL[33] and ZINC.[23] The defaulted cutoff value for similarity search is 0.8, indicating compounds with similarity equal to or higher than 0.8 compared to the query molecule found in the database will be outputed. However, if the applied cutoff doesn't lead

to any hit, the most similar compound will be outputted. Users then can adjust the cutoff based on the Tanimoto coefficient of the most similar compound.

Conclusion

We introduce a series of predictive models to accurately forecast the anti-SARS-CoV-2 activities of screening compounds. We explored 6 different ML algorithms in combination with 15 molecular descriptors for 9 screening assays belonging to four categories. We found that the developed predictive models utilizing the KNN method using the hybrid molecular descriptor, GAFF+RDKit, achieved the best overall performance for all nine assays. Among the 4 common performance metrics (AUC, ACC, F1, PRE), our optimal prediction models achieved better predictedscores for 6 assays than those proposed in KC, et al's study. We have extensively evaluated the predictive models using four external test sets including a negative control test set consisting of 100 druglike screening compounds from ZINC database. The second 3CL model utilizing the screen data from Kuzikov, et al's study has significantly improved the performance of positive prediction, but decreased the performance of negative prediction as well, suggesting there is a trade-off on different performance metrics for a given model. As such, the consensus score of multiple models, especially those were constructed using different descriptors and machine learning algorithms, can significantly improve the prediction accuracy. We have developed a webtool, COVID-19-CP, allowing users to predict a compound's anti-SARS-CoV-2 activities using vertail input formats, and searching similar compounds from three mainstream databases. The combination use of both webportals can facilitate users to screen potential antiviral compounds targeting SARS-CoV2 with enhanced prediction accuracy.

Material and Methods

Data preparation and molecular representations

The compounds used for model training and testing for all assays were collected from the NCATS COVID-19 OpenData portal (https://opendata.ncats.nih.gov/covid19/). After removing duplicated compounds in each assay, we separated the compounds into active and inactive sets, based on whether the assay had half-maximal activity concentration (AC50) data. For some assays, the numbers of active and inactive are significantly unbalanced ($N_{inactive}/N_{active} > 10$), so we randomly put an inactive into one of the several subsets, and all subsets have similar numbers of inactives. We only used one subset of the inactive to construct models, and used the others as external test sets to evaluate the model performance. Specifically, 3CL has four subsets of inactives (s1, s2, s3, and s4), ACE2 has two, CPE has two, TMPRSS2 has three subsets of inactives. For the above assays, only the first subset (s1), was applied in the model construction.

To construct a machine learning model of an assay, we constructed a test dataset by randomly selecting 20% of molecules in active set, and the same number of molecules from the inactive set. For the rest of the compounds, we conducted stratified 10-fold cross validation by leveraging StratifiedKFold, a scikit-learn module built in python[34]. Note that numbers of actives and inactives in the training sets are unbalanced, therefore, wee applied the RandomOverSampler (for 3CL assay) and SMOTE algorithms (for all other assays)[19] to overcome the data unbalance issue. Counts of active and inactive molecules in training (mean counts), validation (mean counts) sets and test sets for nine assays were summarized in Table 1.

Generally, the collected molecules were converted into three types of descriptors: i) fingerprint-based, ii) Physicochemical, iii) force field-based. Class i) includes FP2 (1024 bits), FP3 (55 bits), FP4 (307 bits) and MACCS (166 bits). Among them, FP2 is a path-based fingerprint

that indexes small molecule fragments based on linear segments of up to 7 atoms, while FP3, FP4 and MACCS are substructure-based fingerprints based on sets of SMARTS patterns. All fingerprint-based descriptors were obtained using Open Babel program version 2.3.1 (http://openbabel.org)[35]. RDKit molecular descriptor has 208 bits of vectors, belongs to class ii), and was obtained using RDKit program[36]. GAFF is a force-field based molecular descriptor. It has 47 bits of vectors and contains parameters for a wide breadth of molecules comprised of H, C, N, O, S, P and the halogens. In addition to the above six single descriptors, we also combined them to generate nine hybrid molecular descriptors: RDKit+FP2, RDKit+FP3, RDKit+FP4, RDKit+MACCS. GAFF+FP2, GAFF+FP3, GAFF+FP4, GAFF+MACCS, and GAFF+RDKit. Specifically, before the feature matrix served as the input data for ML models, its molecular descriptors containing RDKit features were standardized into matrix with values ranging from zero to one. This conversion was implemented utilizing *MinMaxScaler* in scikit-learn module.

Model construction

Several ML classifiers were constructed for each assay using 15 molecular descriptors and 6 ML algorithms. ML algorithms applied in the study include support vector machine (SVM), logistic regression (LR), decision tree (DT), Random Forest (RF), k-nearest neighbors (KNN) and complement Naïve Bayes (NB). The description and hyperparameters of those ML algorithms are shown in supplemental information. For each assay, all models were trained and validated using partitioned data sets through the built classifiers in scikit-learn module. The data in the separated test sets for each assay is then used for further model evaluation after the training.

Model evaluation and performance metrics

The performance of constructed models were evaluated by five metrics: area under the curve (AUC) of receiver operating characteristic (ROC) curve, accuracy (ACC), F1-score, precision (PRE), and recall (REC). All of the metrics are ranged in [0,1], in which 0 indicates the worst and 1 indicates the best scenarios. Theoretically for AUC of ROC, a random model will have an AUC of 0.5. ACC measures the proportion of all correct cases among total evaluated cases. PRE is the measurement of the correct positive predictions from all predicted positive cases, while REC measures the correct positive predictions from all actual positive cases. F1-score is the harmonic mean of PRE and REC. The above five metrics were utilized to evaluate the performance of validation and test sets. Additionally, for the evaluation of external datasets, we employed sensitivity and specificity metrics to measure the model performance. Sensitivity measures the percentage of compounds which received a positive prediction on this test out of the percentage of those which actually have the condition, whereas specificity measures the fraction of compounds which had a negative result on the test out of those which actually have no condition. Formulas of all metrics in the study are described in the provided supplemental information.

Acknowledgement

This work was supported by the following funds from the National Science Foundation (NSF) and National Institutes of Health (NIH): NIH R01GM079383, and NSF 1955260. The authors also thank the computing resources provided by the Center for Research Computing (CRC) at University of Pittsburgh.

Conflict of Interest

There are no conflicts to declare.

References

1 Tanne, J. H. *et al.* Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. *BMJ (Clinical research ed.)* **368**, m1090, doi:10.1136/bmj.m1090 (2020).

2 Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. (2020).

3 Hossain, B. *et al.* Surgical Outcome Prediction in Total Knee Arthroplasty Using Machine Learning. **25**, 105--115 (2019).

4 Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* **13**, 8-17, doi:10.1016/j.csbj.2014.11.005 (2015).

5 Assaf, D. *et al.* Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and Emergency Medicine* **15**, 1435-1443, doi:10.1007/s11739-020-02475-0 (2020).

6 Kc, G. B. *et al.* A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nature Machine Intelligence* **3**, 527-535, doi:10.1038/s42256-021-00335-w (2021).

7 Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273-297, doi:10.1007/BF00994018 (1995).

8 Wright, R. E. in *Reading and understanding multivariate statistics.* 217-244 (American Psychological Association, 1995).

9 Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. An introduction to decision tree modeling. **18**, 275-285, doi:https://doi.org/10.1002/cem.873 (2004).

10 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32, doi:10.1023/A:1010933404324 (2001).

11 Peterson, L. E. {K}-nearest neighbor. *Scholarpedia* **4**, 1883, doi:10.4249/scholarpedia.1883 (2009).

12 Rish, I. in *IJCAI 2001 workshop on empirical methods in artificial intelligence.* 41-46.

13 Chen, Y. W., Yiu, C.-P. B. & Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Res* **9**, 129-129, doi:10.12688/f1000research.22457.2 (2020).

14 Xu, T., Zheng, W. & Huang, R. High-throughput screening assays for SARS-CoV-2 drug development: Current status and future directions. *Drug Discov Today* **26**, 2439-2444, doi:10.1016/j.drudis.2021.05.012 (2021).

15 Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e278, doi:10.1016/j.cell.2020.02.052 (2020).

16 Millet, J. K. & Whittaker, G. R. Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* **517**, 3-8, doi:10.1016/j.virol.2017.12.015 (2018).

17 Huang, Y., Yang, C., Xu, X.-f., Xu, W. & Liu, S.-w. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica* **41**, 1141-1149, doi:10.1038/s41401-020-0485-4 (2020).

18 Batuwita, R. & Palade, V. in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 1-8.
19 Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V. J. J. o. a. i. r. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **61**, 863-905 (2018).
20 Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* **18**, 559-563 (2017).
21 Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. **25**, 1157-1174, doi:https://doi.org/10.1002/jcc.20035 (2004).
22 Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **25**, 247-260, doi:10.1016/j.jmgm.2005.12.005 (2006).
23 Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* **52**, 1757-1768, doi:10.1021/ci3001277 (2012).
24 Lalkhen, A. G., McCluskey, A. J. C. e. i. a. c. c. & pain. Clinical tests: sensitivity and specificity. **8**, 221-223 (2008).
25 Gil, C. *et al.* COVID-19: drug targets and potential treatments. *Journal of medicinal chemistry* **63**, 12359-12386 (2020).
26 Lemaitre, F. *et al.* Potential drug–drug interactions associated with drugs currently proposed for COVID-19 treatment in patients receiving other treatments. *Fundamental & Clinical Pharmacology* **34**, 530-547 (2020).
27 Meini, S. *et al.* Role of Lopinavir/Ritonavir in the treatment of Covid-19: a review of current evidence, guideline recommendations, and perspectives. *Journal of clinical medicine* **9**, 2050 (2020).
28 Unal, M. A. *et al.* Ribavirin shows antiviral activity against SARS-CoV-2 and downregulates the activity of TMPRSS2 and the expression of ACE2 in vitro. *Canadian journal of physiology and pharmacology* **99**, 449-460 (2021).
29 Hoffmann, M. *et al.* Nafamostat mesylate blocks activation of SARS-CoV-2: new treatment option for COVID-19. *Antimicrobial agents and chemotherapy* **64**, e00754-00720 (2020).
30 Kuzikov, M. *et al.* Identification of inhibitors of SARS-CoV-2 3CL-pro enzymatic activity using a small molecule in vitro repurposing screen. *ACS pharmacology & translational science* **4**, 1096-1110 (2021).
31 Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34**, D668-D672 (2006).
32 Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091-1097, doi:10.1093/nar/gkt1068 (2014).
33 Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100-D1107, doi:gkr777 [pii] 10.1093/nar/gkr777 (2012).
34 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. **12**, 2825-2830 (2011).
35 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminf.* **3**, 33, doi:10.1186/1758-2946-3-33 (2011).
36 Landrum, G. RDKit: Open-source cheminformatics. (2006).

Tables

Table 1. Summary of counts for datasets.

Assay		Total		Training (90%) and validation (10%)		Test	
Assay Abbr.*	Category	Actives	Inactives	Actives	Inactives	Actives	Inactives
3CL	1	431	4716	345	4624	86	86
CPE	2	841	4808	648	4623	168	168
ACE2	3	203	1574	162	1533	41	41
AlphaLISA	3	1018	2269	812	2060	204	204
TMPRSS2	3	194	1597	155	1558	39	39
cytotox	4	1685	7844	1325	7494	337	337
TruHit	4	1030	2257	819	2045	206	206
HEK293	4	4307	5303	3376	4395	861	861
Fibroblast	4	590	4004	467	3868	118	118

*3CL: 3CL enzymatic activity; CPE: SARS-CoV-2 cytopathic effect CPE; ACE2: ACE2 enzymatic activity assay; AlphaLisa: Spike-ACE2 protein-protein interaction AlphaLISA assay; TMPRSS2: TMPRSS2 enzymatic activity assay; Cytotox: SARS-CoV-2 cytopathic effect counterscreen assay; TruHit: Spike-ACE2 protein-protein interaction TruHit conunterscreen assay; HEK293: HEK 293 cell line toxicity assay; Fibroblast: human fibroblast toxicity assay.

Table 2. Average scores of metrics for models with six ML algorithms of all molecular descriptors in nine assays.

Datasets	Metrics	SVM	LR	DT	RF	KNN	NB
Validation	AUC	0.88	0.88	0.81	0.87	0.91	0.77
	ACC	0.82	0.82	0.75	0.79	0.80	0.71
	F1	0.82	0.83	0.76	0.79	0.82	0.72
	PRE	0.81	0.83	0.74	0.77	0.73	0.69
	REC	0.84	0.83	0.80	0.82	0.94	0.76
Test	AUC	0.73	0.75	0.67	0.74	0.74	0.69
	ACC	0.66	0.66	0.63	0.67	0.68	0.64
	F1	0.59	0.58	0.62	0.65	0.69	0.64
	PRE	0.74	0.76	0.64	0.69	0.67	0.64
	REC	0.52	0.51	0.63	0.63	0.71	0.65

Table 3. Average scores of metrics for KNN models of all molecular descriptors for each assay. The values highlighted in blue and bold font indicating the reported values are higher than those in the study by KC, et al. Note that TMPRSS2, HEK293 and Fibroblast are assays only studied in this work.

Datasets	Metrics	3CL	CPE	ACE2	AlphaLISA	TMPRSS2	cytotox	TruHit	HEK293	Fibroblast
Validation	AUC	0.93	0.93	0.93	0.87	0.94	0.93	0.89	0.81	0.93
	ACC	0.84	0.82	0.80	0.78	0.80	0.83	0.80	0.74	0.82
	F1	0.86	0.84	0.83	0.80	0.83	0.85	0.82	0.75	0.84
	PRE	0.77	0.74	0.73	0.73	0.71	0.76	0.75	0.72	0.74
	REC	0.97	0.98	0.98	0.89	0.99	0.97	0.90	0.79	0.97
Test	AUC	0.66	**0.75**	0.67	0.76	0.71	**0.81**	**0.81**	0.78	0.70
	ACC	0.62	**0.69**	0.62	0.69	0.65	**0.74**	**0.74**	0.71	0.66
	F1	0.59	**0.71**	0.62	0.70	0.67	**0.75**	**0.74**	0.72	0.66
	PRE	0.65	**0.68**	0.63	0.68	0.64	**0.72**	0.72	0.70	0.65
	REC	0.54	0.74	0.62	0.73	0.70	0.79	0.77	0.74	0.67

Table 4. Scores of metrics for KNN models of GAFF+RDKit molecular descriptor for each assay. The values highlighted in blue and bold font indicating the reported values are higher than those in the study by KC, et al. Note that TMPRSS2, HEK293 and Fibroblast are assays only studied in this work.

Datasets	Metrics	3CL	CPE	ACE2	AlphaLISA	TMPRSS2	cytotox	TruHit	HEK293	Fibroblast
Validation	AUC	0.94	0.95	0.95	0.90	0.95	0.95	0.92	0.84	0.94
	ACC	0.86	0.84	0.83	0.81	0.82	0.85	0.83	0.76	0.83
	F1	0.88	0.86	0.85	0.82	0.85	0.87	0.84	0.77	0.85
	PRE	0.80	0.76	0.75	0.76	0.74	0.78	0.78	0.74	0.75
	REC	0.98	0.99	0.99	0.90	1.00	0.99	0.93	0.80	0.99
Test	AUC	**0.75**	**0.82**	0.75	**0.80**	0.74	**0.83**	**0.84**	0.82	0.72
	ACC	0.68	**0.76**	0.68	0.74	0.67	**0.76**	**0.76**	0.75	0.69
	F1	0.63	**0.77**	0.70	0.74	0.68	**0.77**	**0.78**	0.76	0.69
	PRE	**0.77**	**0.74**	0.67	0.75	0.66	**0.75**	0.72	0.72	0.68
	REC	0.53	0.80	0.72	0.73	0.70	0.80	0.84	0.80	0.71

Table 5. Score of metrics for KNN model of GAFF_RDKit molecualr descriptor for the second 3CL model.

Datasets	AUC	ACC	F1	PRE	REC
Validation	0.98	0.95	0.95	0.91	1.00
Test	0.77	0.76	0.72	0.85	0.62

Figures

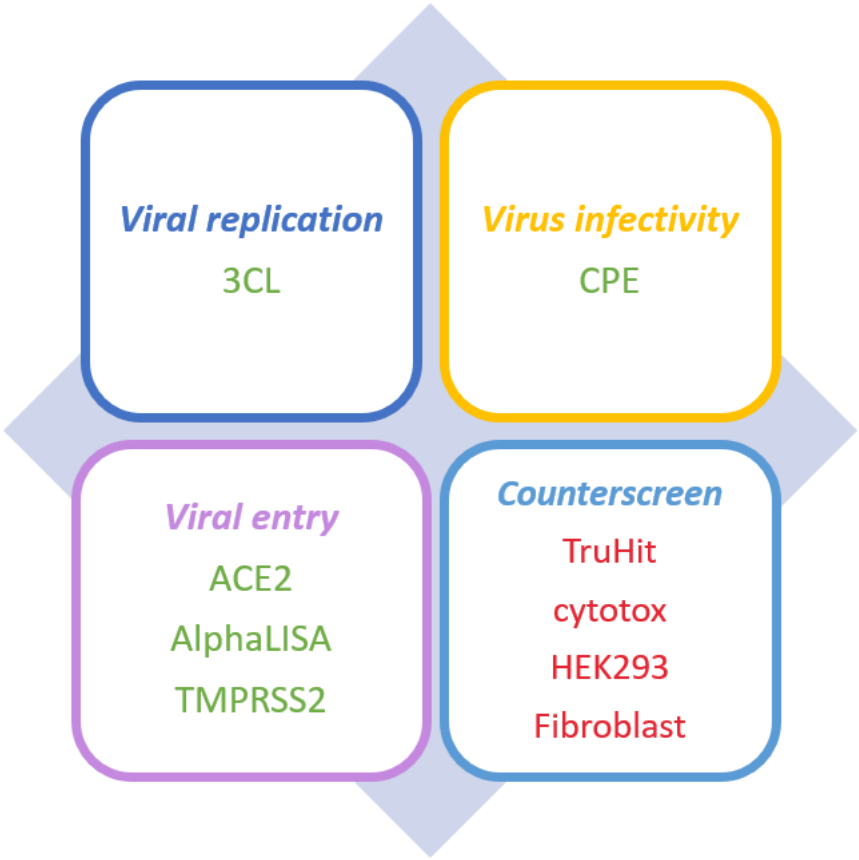

Figure 1. The desirable profile for the predicted anti-SARS-CoV-2 activities of the promising compound among 9 assays. Green color indicates “active” is preferred, while red color indicates “inactive” is preferred.

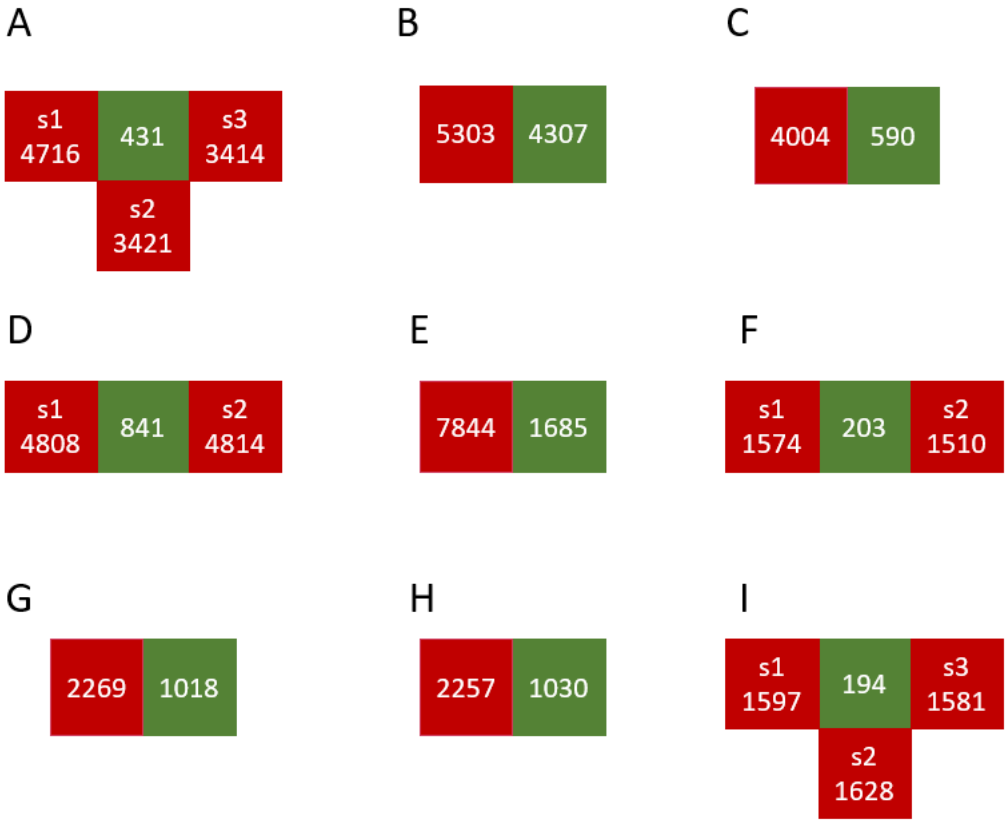

Figure 2. The counts of compounds in active sets (green) and inactive sets (red) for each assay. *s* refers to sample of inactive compounds (s1: sample 1, s2: sample 2, s3: sample 3, s4: sample 4). A-I are different assays. A. 3CL, B. HEK293, C. Fibroblast, D. CPE, E. cytotox, F. ACE2, G. AlphaLISA, H. TruHit, I.
TMPRSS2.

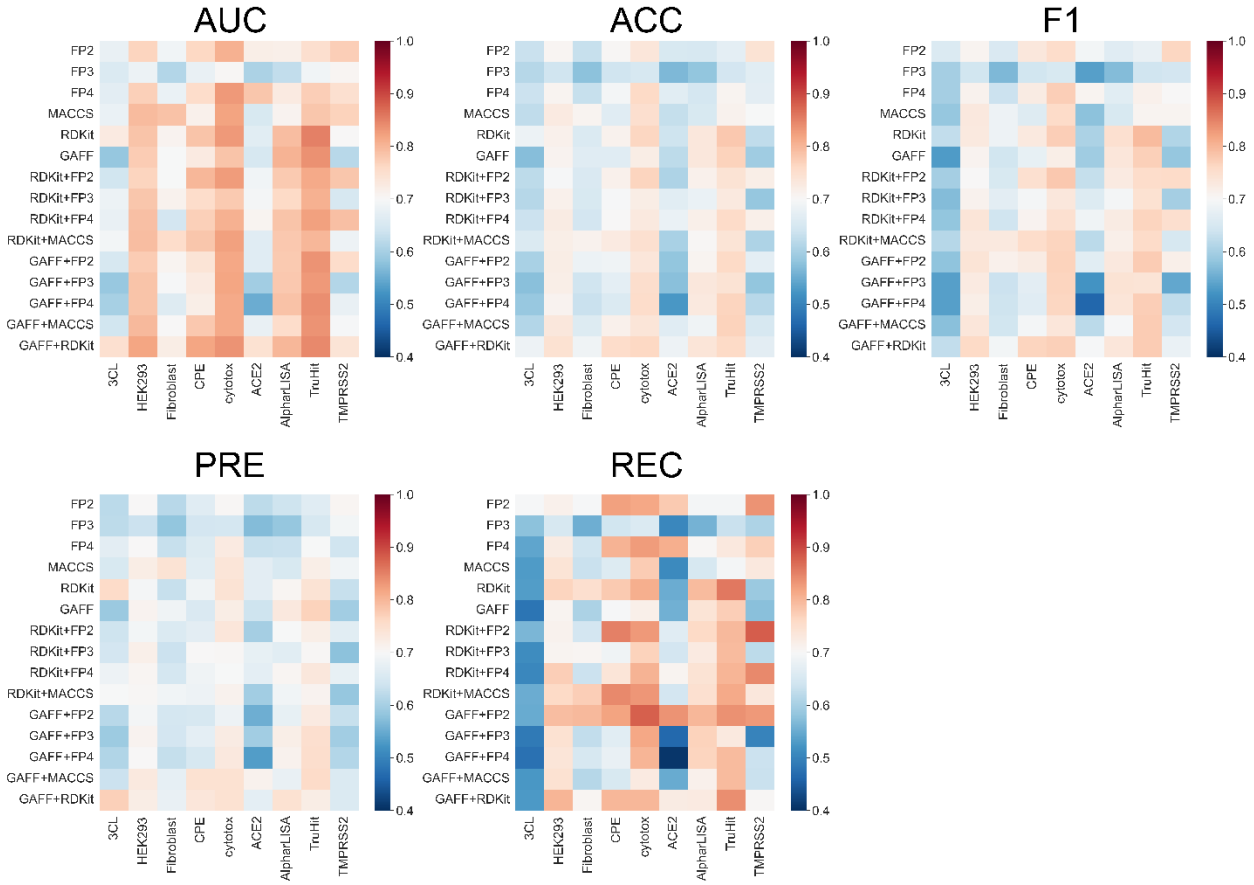

Figure 3. Heatmaps of metrics AUC, ACC, F1, PRE and REC for KNN models of different molecular descriptors.

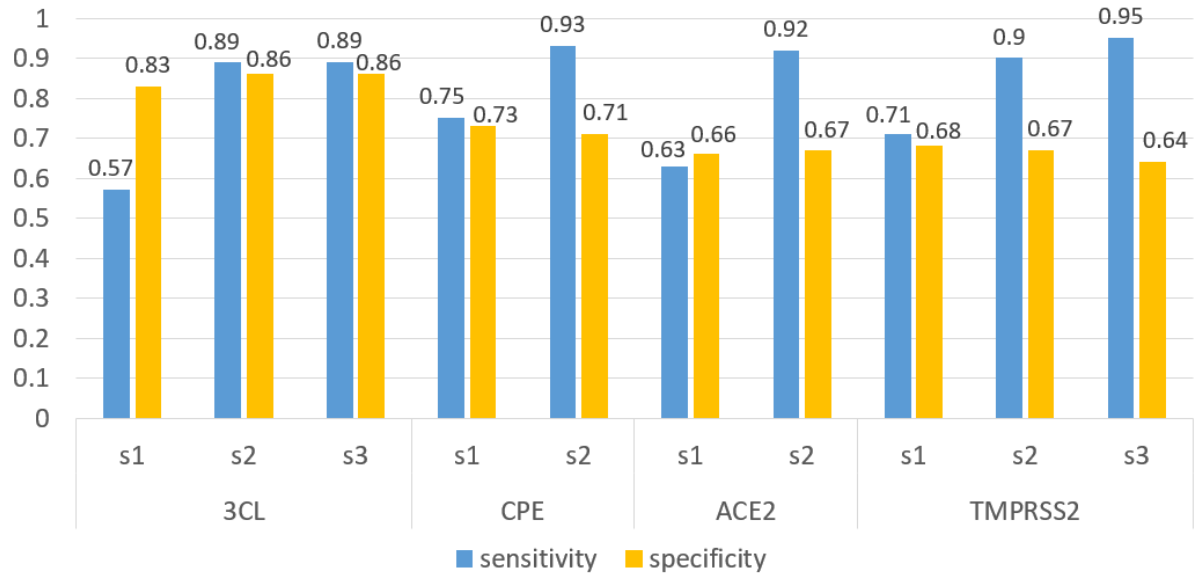

Figure 4. Sensitivity and specificity of test sets and sample sets in 3CL, CPE, ACE2, TMPRSS2 assays.

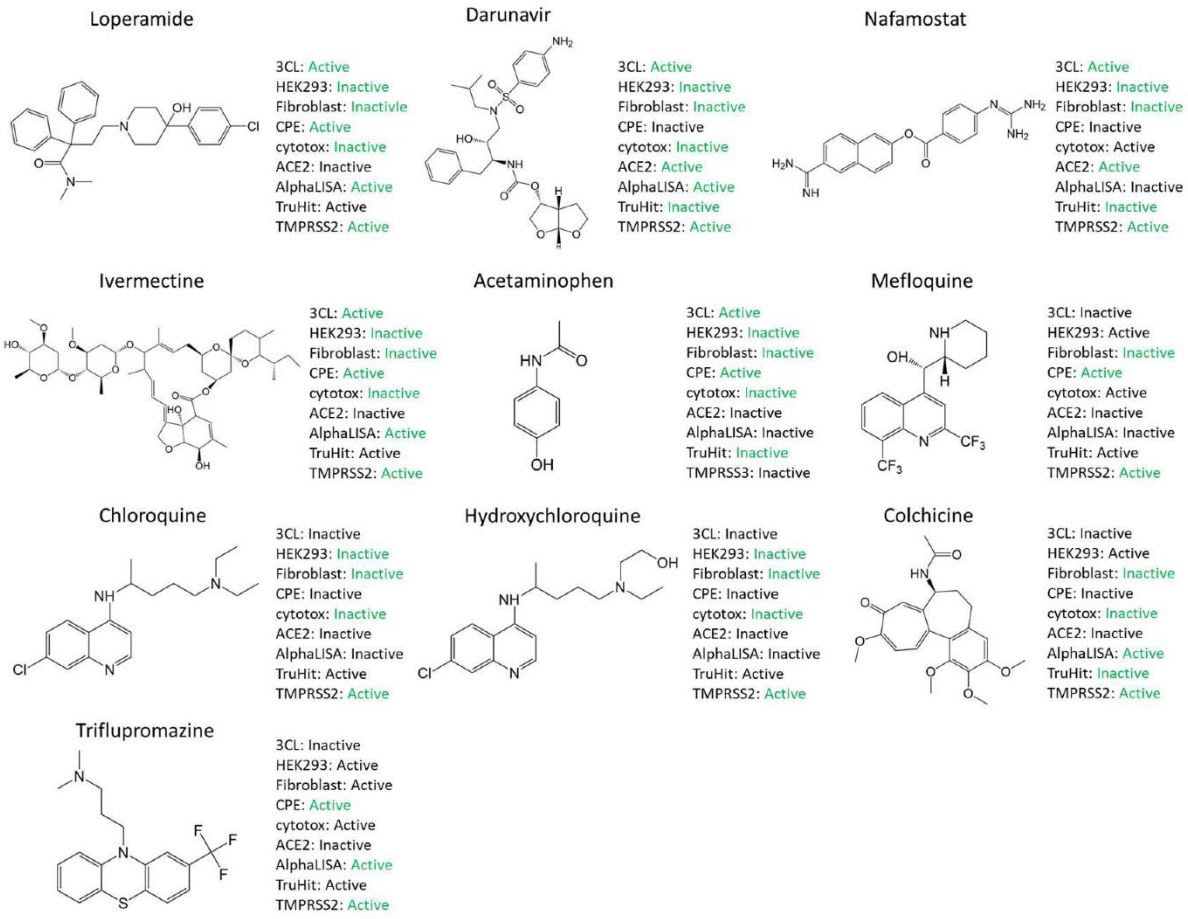

Figure 5. the structures of 9 potential candidates and the predictions of assays that they have passed.

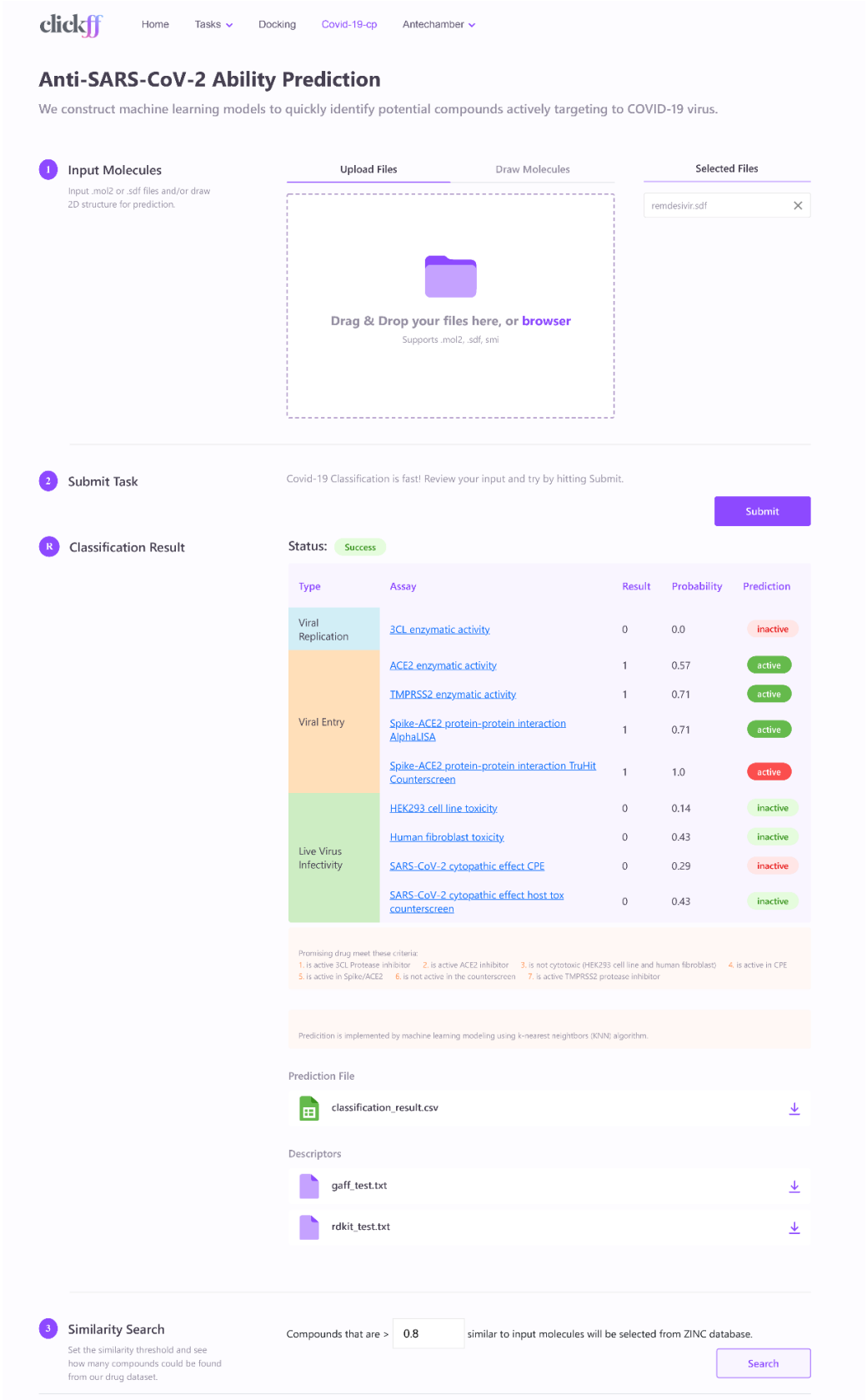

Figure 6. User interface of the web portal.