Journal of Medicinal Chemistry



pubs.acs.org/jmc

Article

# Efficient Identification of Anti-SARS-CoV-2 Compounds Using Chemical Structure- and Biological Activity-Based Modeling

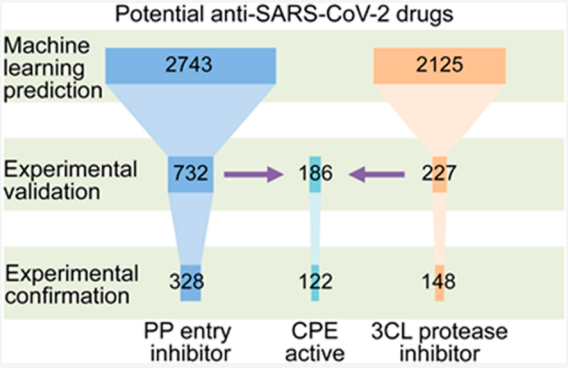
Tuan Xu, Miao Xu, Wei Zhu, Catherine Z. Chen, Qi Zhang, Wei Zheng, and Ruili Huang*

Cite This: J. Med. Chem. 2022, 65, 4590−4599

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Identification of anti-SARS-CoV-2 compounds through traditional high-throughput screening (HTS) assays is limited by high costs and low hit rates. To address these challenges, we developed machine learning models to identify compounds acting via inhibition of the entry of SARS-CoV-2 into human host cells or the SARS-CoV-2 3-chymotrypsin-like (3CL) protease. The optimal classification models achieved good performance with area under the receiver operating characteristic curve (AUC-ROC) values of >0.78. Experimental validation showed that the best performing models increased the assay hit rate by 2.1-fold for viral entry inhibitors and 10.4-fold for 3CL protease inhibitors compared to those of the original drug repurposing screens. Twenty-two compounds showed potent (<5 $\mu$M) antiviral activities in a SARS-CoV-2 live virus assay. In conclusion, machine learning models can be developed and used as a complementary approach to HTS to expand compound screening capacities and improve the speed and efficiency of anti-SARS-CoV-2 drug discovery.




## INTRODUCTION

The current global pandemic of coronavirus disease 19 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a highly infectious enveloped RNA virus.[1,2] The transmission of SARS-CoV-2 occurs primarily through inhalation of respiratory droplets of infected individuals and/or contact with virally contaminated objects.[3] The initial phase of SARS-CoV-2 infection is the entry of the virus into the host cells through the interaction of the receptor-binding domain (RBD) of the viral Spike glycoprotein with the angiotensin converting enzyme 2 (ACE2) on the cell surface.[4,5] Subsequently, SARS-CoV-2 initiates RNA replication and eventually assembles new virions that are released to infect other cells in the host.[6] Each stage of the SARS-CoV-2 life cycle (e.g., viral entry and viral replication) could be targeted for the development of specific antiviral drug candidates for COVID-19 treatment.[7,8]

Drug repurposing has been widely used to identify new clinical indications from existing drugs for the treatment of many diseases, including viral infections. During the COVID-19 outbreak, several high-throughput drug repurposing assays were developed and applied to identify potential inhibitors of SARS-CoV-2 entry and replication at the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH).[9−11] For example, the pseudotyped particle (PP) entry assay is a cell-based assay with a luminescence readout, which can facilitate the identification of viral cell entry inhibitors using pseudotyped viral particles containing coronavirus Spike glycoprotein without the viral genome.[9] Another potential target for the development of antiviral therapies is the SARS-CoV-2 3-chymotrypsin-like (3CL) protease, which plays a vital role in viral replication by cleaving the viral polyprotein to form the RNA replicase−transcriptase complex.[8] The 3CL protease assay is a fluorescence-based biochemical assay that measures the inhibitory effect of compounds on the activity of SARS-CoV-2 3CL protease.[11] However, due to limitations in resources, it is impractical to efficiently screen large chemical libraries using these assays.[12,13] In addition, there are millions of commercial compounds that are not practical to include for high-throughput screening (HTS) as they are not part of our internal compound collections.

As an alternative to physical HTS, computational models such as machine learning classifiers can make predictions on new unseen data based on previous experiences and known data properties, such that they have been widely used to virtually screen millions of compounds for potential biological activities.[14−16] Among the machine learning models, quantitative structure−activity relationship (QSAR) approaches enable the prediction of biological activities for compounds of interest

Received: August 19, 2021
Published: March 11, 2022





https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

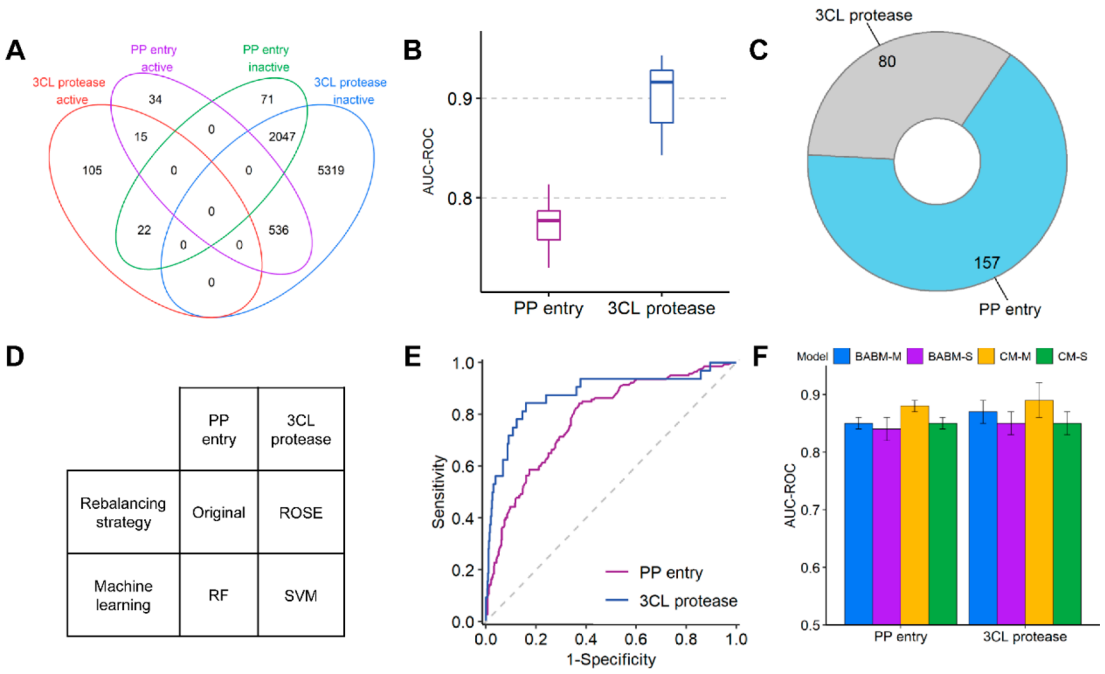



**Figure 1.** Construction and evaluation of the optimal machine learning classification models. (A) Activity distribution of compounds in the original drug repurposing screens, including the PP entry assay and the 3CL protease assay. Optimal QSAR models built on the training data set, including (B) model performance, (C) the number of selected features, and (D) data rebalancing strategies and machine learning algorithms. (E) Example ROC curves of the optimal QSAR models on the external validation data set. (F) BABM model performance measured by AUC-ROC. Abbreviations: ROSE, random oversampling examples; SVM, support vector machine; RF, random forest; QSAR, quantitative structure (ECFP4)−activity relationships; SBM, structure (ToxPrint)-based model; BABM-M, activity-based model (MLS); BABM-S, activity-based model (Sytravon); CM-M, combined model (SBM + BABM-M); CM-S, combined model (SBM + BABM-S).

as a function of similarity in chemical structure (i.e., molecular descriptors).[17,18] Unlike QSAR, biological activity-based modeling (BABM) approaches build on the hypothesis that compounds showing similar biological activity patterns tend to share similar biological targets or mechanisms of action.[16,19] The combined use of these two methods exhibits complementary advantages, such that their application domains are not limited to small molecules with well-defined structures (e.g., QSAR) or substances with available biological profiles (e.g., BABM). The PP entry and 3CL protease assays have been applied to screen several thousands of known bioactive compounds, including the NCATS Pharmaceutical Collection (NPC) of approved and investigational drugs.[9−11,20] The data generated from these assays can be used to build computational models to predict new PP entry or 3CL protease inhibitors.

Given the fast-growing number of COVID-19 patients and the current lack of effective drug treatments, there is an urgent need to accelerate efforts in exploring new potential drugs to treat COVID-19. In this study, we applied machine learning methods including both structure (QSAR)- and activity (BABM)-based approaches to build models for prediction of potential inhibitors of SARS-CoV-2 entry and 3CL protease. The selected hits predicted by these models were first tested in the SARS-CoV-2 entry assay or 3CL protease assay. The antiviral activities of compounds were then confirmed in a cell-based live SARS-CoV-2 cytopathic effect reduction (CPE) assay.[21,22]

## RESULTS

**Construction and Evaluation of the Prediction Models.** A total of 8149 unique compounds were screened against either the PP entry assay (2725 compounds) or the 3CL protease assay (8044 compounds) (Figure 1A and Table S1). These compounds fell into different activity categories (Figure 1A). For example, 570 compounds were active only in the PP entry assay, 142 compounds were active only in the 3CL protease assay, and 15 compounds were active against both PP entry and 3CL protease. On the basis of these data, we built two types of models, QSAR and BABM, for the prediction of PP entry or 3CL protease inhibitors (Figure 1B−F). We tested various parameter combinations on the training data sets to find the optimal model for each assay target (Figure 1B−D). For example, the combination of RF (machine learning algorithm), Original (rebalancing strategy), and 157 ECFP4 features (Fisher's exact test with a $P$ value of 0.02) produced the best classification performance (AUC-ROC = 0.77 ± 0.02) for predicting PP entry inhibitors, and the corresponding parameters for the best performing 3CL protease model (AUC-ROC = 0.90 ± 0.03) were SVM, ROSE, and 80 ECFP4 features (Fisher's exact test with a $P$ value of 0.05). The optimal QSAR models also showed good prediction performance on the external validation data set for PP entry (AUC-ROC = 0.78) and 3CL protease (AUC-ROC = 0.88) (Figure 1E). The feature sets that produced the optimal QSAR models for the prediction of PP entry or 3CL protease inhibitors are listed in Table S2, and the AUC-ROC values obtained from all QSAR models are listed in Table S3. We observed large differences among the models built with different methods and feature sets. The AUC-ROC values for predicting PP entry inhibitors ranged from 0.64 to 0.78 with an average of 0.73, while the AUC-ROC values for predicting 3CL protease inhibitors ranged from 0.64 to 0.90 with an average of 0.81. For the BABM models, the AUC-ROC values on the test sets ranged from 0.84 to 0.88 for the PP entry inhibitor models and from 0.85 to 0.89 for the 3CL protease


https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

inhibitor models, with the combined model [CM-M, structure (ToxPrint)-based model + BABM-M] yielding the best performance (for PP entry, AUC-ROC = 0.88 ± 0.01; for 3CL protease, AUC-ROC = 0.89 ± 0.03) (Figure 1F).

**Virtual Screening and Experimental Validation of Model-Predicted Compounds.** In an attempt to identify novel inhibitors of SARS-CoV-2 entry and 3CL protease, the optimal models were applied to virtually screen a large compound collection of ~360K compounds. For the QSAR models, a total of 4868 compounds with the highest predicted probabilities (>0.5) were collected, including 2743 predicted PP entry inhibitors and 2125 predicted 3CL protease inhibitors. The density distributions of the probabilities for both sets of predictions exhibited a single peak at relatively small probability values (peak value of <0.56) for most of the compounds (Tables S4 and S5 and Figure 2A). For the BABM models, a total of 847 compounds were collected, including 485 predicted PP entry inhibitors and 364 predicted 3CL protease inhibitors (Tables S6 and S7). After combining these compounds and filtering by structure (see Experimental Section for details), we selected a total of 1972 predicted PP entry inhibitors and 1493 predicted 3CL protease inhibitors for experimental validation (Tables S4–S7). For the QSAR models, when a default probability cutoff of 0.5 was used, the models increased the assay hit rate by 1.6-fold (from 21.5% to 34.9%) for the PP entry inhibitors and 8.4-fold (from 1.76% to 14.8%) for the 3CL protease inhibitors (Figure 2B). As the probability cutoff increased, the hit rates of the QSAR model predictions gradually increased, as well (Figure 2C). When the probability cutoff was set to 0.9, the hit rates of the QSAR model predictions reached 66% for the PP entry assay and 22% for the 3CL protease assay (Figure 2C). Moreover, the BABM models increased the assay hit rate by 2.1-fold for the PP entry (from 21.5% to 45.6%) and 10.4-fold for the 3CL protease (from 1.76% to 18.4%) (Figure 2B).

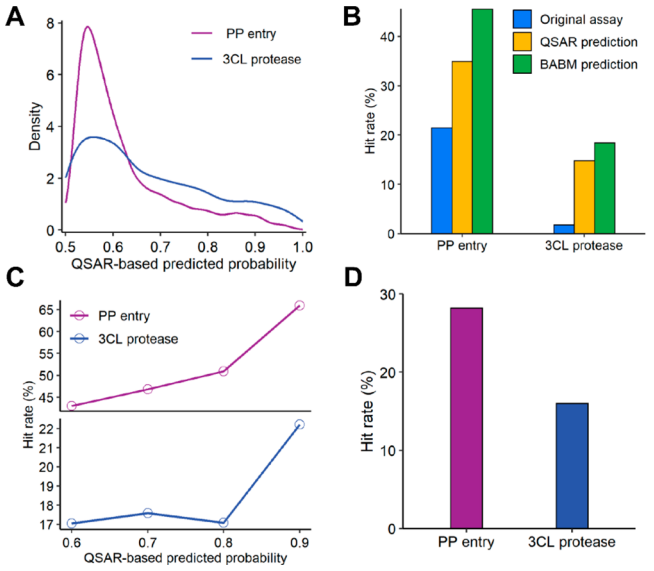



**Figure 2.** Virtual screening of a large compound library based on the optimal models. (A) Distributions of the predicted probabilities of QSAR model-identified compounds. (B) Comparison of hit rates between the original drug repurposing screens and model predictions. (C) Hit rates of the QSAR model predictions based on different prediction probability cutoffs. (D) Hit rates of the potential PP entry inhibitors and 3CL protease inhibitors in the CPE assay.

**Secondary Experimental Confirmation of Model-Predicted PP Entry and 3CL Protease Inhibitors.** To further confirm the experimentally validated predictions, a total of 672 compounds, including 446 PP entry inhibitors and 226 3CL protease inhibitors, were retested at 11 concentrations (Tables S8 and S9). For the PP entry assay, 328 of the 446 compounds remained active, yielding a confirmation rate of 74%. Of the confirmed PP entry inhibitors, 149 were known drugs or bioactive compounds while the other 179 inhibitors were diverse compounds without any well-annotated biological activity (Table S8). For the 3CL protease assay, 148 of the 226 compounds remained active, yielding an assay confirmation rate of 65%. Of the confirmed 3CL protease inhibitors, 62 were known drugs or bioactive compounds while the other 86 inhibitors were diverse compounds without any well-annotated biological activity (Table S9). The most potent PP entry inhibitor was NCGC00390584 (Exatecan; $IC_{50}$ = 3.1 nM), and the most potent 3CL protease inhibitor was NCGC00390337 (Z-DQMD-FMK; $IC_{50}$ = 0.92 $\mu$M) (Tables S8 and S9). Several representative compounds with potent inhibitory effect against SARS-CoV-2 PP entry or 3CL protease are shown in Figures 3 and 4, such as NCGC00599688 (AQ-13; $IC_{50}$ = 14.23 ± 11.66 $\mu$M) for the PP entry inhibitor and NCGC00371011 (Fluorobexarotene; $IC_{50}$ = 28.95 ± 2.35 $\mu$M) for the 3CL protease inhibitor.

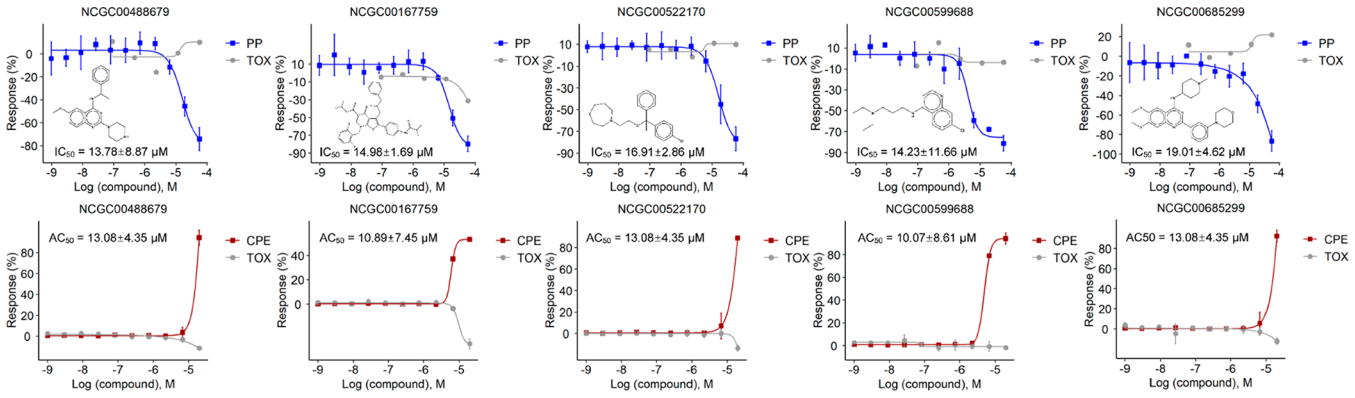



**Figure 3.** Concentration–response curves of representative PP entry inhibitors in the anti-SARS-CoV-2 PP entry assay and the CPE assay. Abbreviations: PP, PP entry assay; TOX, cytotoxicity assay; CPE, CPE assay; $IC_{50}$, half-maximal inhibitory concentration; $AC_{50}$, half-maximal activity concentration. Results are presented as mean ± standard deviation (SD), and the error bars represent the SD of two independent experiments.



https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590–4599

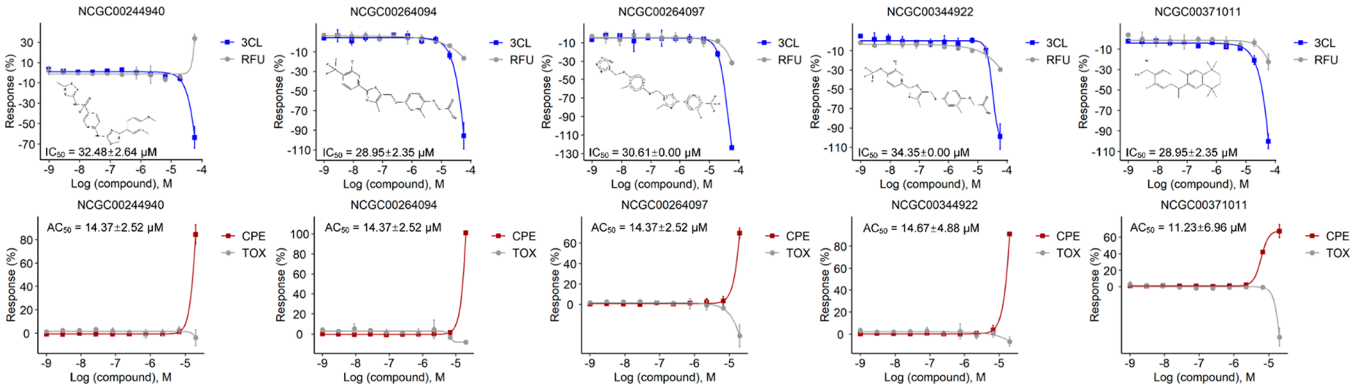



**Figure 4.** Concentration−response curves of representative 3CL protease inhibitors in the anti-SARS-CoV-2 3CL protease assay and the CPE assay. Abbreviations: 3CL, 3CL protease assay; RFU, relative fluorescence unit; TOX, cytotoxicity assay; CPE, CPE assay; $IC_{50}$, half-maximal inhibitory concentration; $AC_{50}$, half-maximal activity concentration. Results are presented as mean ± SD, and the error bars represent the SD of two independent experiments.

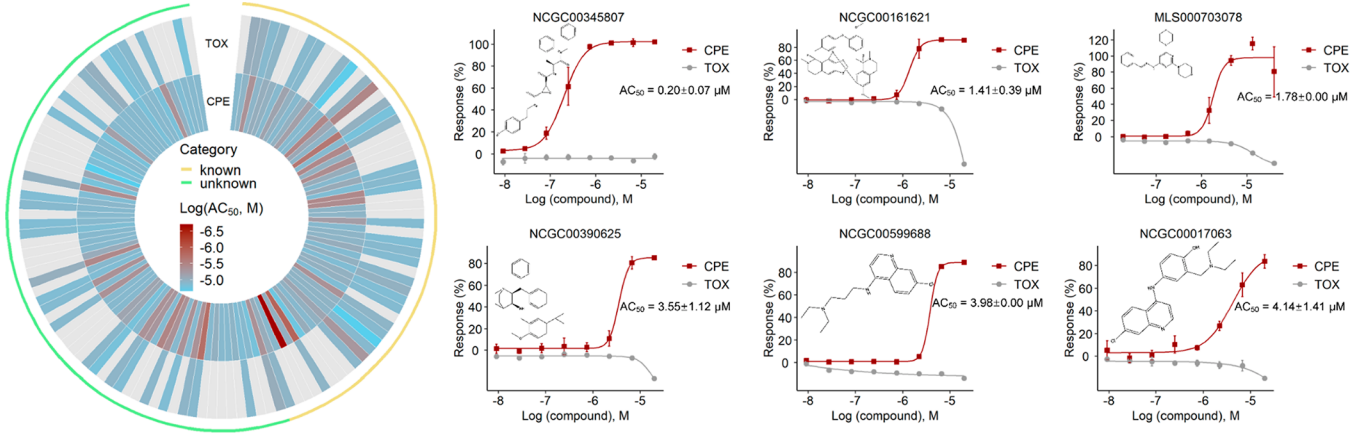



**Figure 5.** Secondary experimental confirmation of the potential CPE actives and concentration−response curves of representative compounds. The heat map shows the overall potencies of compounds confirmed in the CPE assay (inner ring) and the cytotoxicity counter screen (outer ring). The heat map is colored by compound potency, such that darker shades of red indicate more potent compounds and lighter shades of blue indicate less potent compounds. Gray indicates missing values. The outer yellow line represents known bioactive compounds, and the outer green line represents compounds with no previously reported bioactivity. Concentration−response curves are shown for representative compounds with potent anti-SARS-CoV-2 activity (<5 $\mu$M). Abbreviations: CPE, CPE assay; TOX, cytotoxicity assay; $AC_{50}$, half-maximal activity concentration. Results are presented as mean ± SD, and the error bars represent the SD of two independent experiments.

**Assessment of Compound Antiviral Activity in the SARS-CoV-2 CPE Assay.** A total of 578 PP entry inhibitors and 150 3CL protease inhibitors were further tested in a SARS-CoV-2 live virus assay, the CPE assay. The results showed that 28.2% of the PP entry inhibitors and 15.3% of the 3CL protease inhibitors were active in the CPE assay (Tables S4−S7 and S10 and Figure 2D). For secondary confirmation, the 127 compounds that showed activity in the CPE assay were retested at eight or more concentrations (instead of four or five concentrations in the first screen) to further confirm their activity and obtain more accurate potency measures. Of the 127 compounds, 122 remained active, yielding a confirmation rate of 96% for the SARS-CoV-2 CPE assay (Table S11 and Figure 5). Of the 122 confirmed CPE actives, the potencies ranged from 0.20 $\mu$M (NCGC00345807, CAA-0225) to 22 $\mu$M (NCGC00417833) with an average potency of 10.4 $\mu$M (Table S11). Moreover, 22 compounds showed potencies of <5 $\mu$M, accounting for 18% of the total CPE actives (Table S11). In addition, 55 CPE actives were known drugs or bioactive compounds, while the other 67 compounds were compounds without any well-annotated biological activity (Table S11). Six representative compounds with potent anti-SARS-CoV-2 activity (<5 $\mu$M) in the CPE assay are shown in Figure 5, including NCGC00345807 (CAA-0225; $AC_{50}$ = 0.20 ± 0.07 $\mu$M), NCGC00161621 (Cepharanthine; $AC_{50}$ = 1.41 ± 0.39 $\mu$M), MLS000703078 ($AC_{50}$ = 1.78 ± 0.00 $\mu$M), NCGC00390625 (Maropitant; $AC_{50}$ = 3.55 ± 1.12 $\mu$M), NCGC00599688 (AQ-13; $AC_{50}$ = 3.98 ± 0.00 $\mu$M), and NCGC00017063 (Amodiaquine; $AC_{50}$ = 4.14 ± 1.41 $\mu$M).

## DISCUSSION AND CONCLUSIONS

Identification of inhibitors of the entry of SARS-CoV-2 into host cells and 3CL protease from compound screening can result in lead compounds that may potentially be developed into anti-COVID-19 therapeutics. In this study, we built machine learning classification models based on the data generated from the high-throughput anti-SARS-CoV-2 drug repurposing assays and applied the optimal models to virtually screen a large collection of diverse compounds. One hundred twenty-two compounds were validated experimentally for their anti-SARS-CoV-2 activities in the live SARS-CoV-2 CPE assay.



https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

The circular topological descriptor ECFP4, which has been widely used in drug discovery,[23] was used to build the QSAR models in this study. Because the ECFP4 fingerprint has 1024 bits, feature selection was performed to avoid overfitting and possibly improve the prediction performance (Figure 1C). Because of the imbalanced nature of the original assay outcomes, i.e., the large prevalence of inactive compounds compared to active compounds in a data set, five rebalancing strategies were applied prior to modeling. The best rebalancing strategy found for the 3CL protease data set is consistent with previous research findings that ROSE worked well in improving the predictive power of the models.[24] However, the original data without applying any rebalancing achieved the best performance for the PP entry data set (Figure 1D), indicating that a balanced data set does not necessarily result in better performance. Consistent with previous reports on constructing and optimizing classification models,[25−27] optimal model performance is often the result of a combination of the best feature set, rebalancing strategy, and machine learning algorithm (Figure 1B−D). The optimal QSAR models achieved good performance on the external validation data set with AUC-ROC values of 0.78 for predicting PP entry inhibitors and 0.88 for 3CL protease inhibitors (Figure 1E). Our QSAR model for predicting 3CL protease inhibitors outperformed previously reported models (AUC-ROC = 0.71), which were constructed on the same data set as the one used in this study, but without applying any rebalancing strategy.[28] In addition, the BABM models combining activity and structure data showed better performances (AUC-ROC > 0.84) than the QSAR models (Figure 1), further confirming the value of biological activity profiles in improving prediction performance.[16] These findings are consistent with other studies demonstrating that combining in vitro bioactivity with chemical structure descriptors could improve the predictive performance of machine learning models.[27,29] To further improve our model, 99 3CL protease inhibitors with $IC_{50}$ values of <50 $\mu$M were retrieved from the literature (Table S12).[30−37] For each compound in a compound set, we identified its closest structural neighbor by calculating the Tanimoto coefficient (TC). The TC between a compound and its closest structural neighbor is defined as maxTC. The average maxTC value (AmaxTC) of the 99 3CL protease inhibitors was 0.74, which was much larger than that of the 3CL protease inhibitors in the original model training set (AmaxTC = 0.32) and the AmaxTC between the literature 3CL inhibitor set and the original model training set (AmaxTC = 0.28) (Figure S1). These findings indicate that the newly added compounds are structurally similar to each other but distinct from the original model training set; thus, adding the literature compounds increased the structural diversity of the training set. After adding the 99 3CL protease inhibitors to the model training set, we found that the RF model produced better performance (AUC-ROC = 0.86 ± 0.04) based on the 3-fold cross-validation results. Using the best threshold (0.07) defined by the Youden index, the RF model was able to correctly identify 94 (95%) of the 99 literature 3CL protease inhibitors (Table S12). The addition of the 3CL protease inhibitors from the literature helped to expand the applicability domain of our model and further demonstrated the reliability and robustness of our model.

The optimal models (including QSAR and BABM) can be used to rapidly screen large compound libraries to identify potential anti-SARS-CoV-2 compounds and prioritize them for experimental validation (Figure 2). Compared with the original HTS assays, the optimal machine learning models increased the assay hit rate by 2.1-fold for viral entry inhibitors and 10.4-fold for 3CL protease inhibitors, resulting in a hit rate of 45.6% for the PP entry assay and 18.4% for the 3CL protease assay from a large diverse compound library (Figure 2B). The model hit rates were calculated on the basis of the experimental validation results. For experimental validation, the model-predicted active compounds were tested in the same assay as the one used for the original drug repurposing screen.[9−11] The hit rate of the experimental validation screen is termed the model hit rate, which is equivalent to the positive predictive value [PPV = TP/(TP + FP)] (Figure 2B,C). As the probability cutoff increased, the model hit rates (i.e., PPVs) and specificities [TN/(TN + FP)] of the PP entry and 3CL protease models increased, while their sensitivities [TP/(TP + FN)] decreased. These hit rates are significantly higher than that of a typical HTS of a diverse compound library, which is <1.5%, demonstrating that our models could significantly improve the efficiency of anti-COVID-19 drug lead identification. Multitarget drugs, i.e., compounds that interact with multiple targets in the biological network simultaneously, have been widely used in the field of drug discovery, especially for the treatment of complex diseases.[38,39] COVID-19 has emerged as a complex disease that presents variable clinical symptoms and disease progression (e.g., asymptomatic, acute respiratory distress syndrome, and multiorgan failure).[40] In this study, the PP entry assay is a phenotypic assay that contains multiple targets in the viral entry process, while the 3CL protease assay is a single-target assay. When evaluating the antiviral activities of the hits identified from these two assays, we found that the hit rate of the PP entry inhibitors (28.2%) was higher than that of the 3CL protease inhibitors (15.3%) in the CPE assay (Figure 2D), further confirming that multitarget compounds could be more effective antivirals than single-target compounds.[41] In addition, these results also suggest that multitarget assays may be an important direction in new assay development for more efficient anti-SARS-CoV-2 drug discovery, whereas the single-target assays are more suitable for investigating compound mechanisms of action.[42] Although the addition of DTT in the 3CL protease assay could reverse the inhibitory effect of the compounds with electrophilic warheads, this condition did not show any significant impact on our results. For example, the $IC_{50}$ of NCGC00390337 (Z-DQMD-FMK, the most potent 3CL protease inhibitor in our results) was 0.92 $\mu$M, which was not significantly different from the finding of Sun et al. ($IC_{50}$ = 0.94 $\mu$M).[43] In this study, our models identified 122 compounds that were active in the SARS-CoV-2 CPE assay (Table S11), and these compounds could be further developed as potential anti-COVID-19 treatments. For example, among the 22 potent compounds with $AC_{50}$ values of <5 $\mu$M, 10 were known drugs or bioactive compounds, while the others were diverse compounds without any well-annotated biological activities. For example, four compounds (i.e., NCGC00345807, NCGC00161621, NCGC00386665, and NCGC00017063) have been reported to have anti-SARS-CoV-2 activities. The most potent compound is NCGC00345807 (CAA-0225, a cathepsin L-specific inhibitor; $AC_{50}$ = 0.20 $\mu$M) (Figure 5 and Table S11), which was also reported as a lead anti-SARS-CoV-2 compound in a previous study.[22] On the basis of our results, the anti-SARS-CoV-2 mechanism of CAA-0225 may be attributed partly to the inhibition of the entry of SARS-CoV-2 into host cells ($IC_{50}$ = 0.60 $\mu$M; efficacy = 89%) (Table S8). NCGC00161621 (Cepharanthine), an approved drug with anti-inflammatory activities, has been reported to rescue the CPE of SARS-CoV-2


https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

to full efficacy probably due to the inhibition of Spike-mediated cell entry or SARS-CoV-2 replication.[9,44,45] Consistent with these previous studies, we also found that Cepharanthine showed potency against the SARS-CoV-2 CPE effect with an $AC_{50}$ of 1.41 $\mu$M, and its potential mechanism of action is to inhibit the entry of the virus into host cells ($IC_{50}$ = 6.88 $\mu$M; efficacy = 102%) (Figure 5 and Tables S8 and S11). NCGC00386665 (Bemcentinib), a selective small-molecule inhibitor of AXL kinase, has been reported as a dual inhibitor of SARS-CoV-2 papain-like protease and 3CL protease on the basis of molecular docking. This compound was also reported to potentially reduce viral infection and block the SARS-CoV-2 Spike protein according to a multicenter, seamless, phase 2 adaptive randomization platform study.[46] Consistent with these findings, in our study Bemcentinib showed potent activity against the SARS-CoV-2 CPE with an $AC_{50}$ of 3.36 $\mu$M, potentially via inhibition of the entry of the virus into host cells ($IC_{50}$ = 19 $\mu$M; efficacy = 112%) (Tables S8 and S11). NCGC00017063 (Amodiaquine), an antimalarial and anti-inflammatory agent, was reported to show activity in the anti-SARS-CoV-2 CPE assay.[45,47,48] Consistent with the previous report, Amodiaquine showed potent activity against the SARS-CoV-2 CPE in our study with an $AC_{50}$ of 3.98 $\mu$M, potentially by inhibiting the entry of the virus into host cells ($IC_{50}$ = 19 $\mu$M; efficacy = 103%) (Figure 5 and Tables S8 and S11). In addition, the other 18 potent compounds have no previous reports on their antiviral activity; therefore, in-depth investigations are needed to explore these compounds as potential drug candidates for the treatment of COVID-19. Furthermore, 6 of the 22 potent compounds (i.e., NCGC00599688, NCGC00590975, NCGC00390625, NCGC00263128, NCGC00482724, and NCGC00484976) have been reported to have biological activities, but not antiviral activities. For example, NCGC00599688 (AQ-13), an analogue of chloroquine (CQ), is an investigational antimalarial drug.[49] NCGC00590975 (JQEZ5) is a potent pharmacologic enhancer of zeste homologue 2 (EZH2) inhibitor that exhibits significant in vivo antitumor activity in EZH2 mutant cancer models.[50] NCGC00390625 (Maropitant) is a selective neurokinin 1 receptor antagonist that is clinically used as a new antiemetic drug for dogs.[51] NCGC00263128 (PD 0220245) shows inhibitory effects on both IL-8 receptor binding and IL-8-mediated neutrophil chemotaxis.[52] NCGC00482724 (Vacquinol-1), a quinolone derivative, has been reported to display potent antitumor effects by inducing rapid cell death in glioblastomas.[53] NCGC00484976 (Z36), a novel B-cell lymphoma-extra large (Bcl-xL) protein inhibitor, could efficiently induce autophagic cell death by blocking the interaction between Bcl-xL/Bcl-2 and Beclin-1.[54] The remaining 12 compounds have no reports on their biological activities. Because the development of a new drug takes a long time, an in-depth investigation is needed to explore these 22 compounds as potential drug candidates for the treatment of COVID-19.

In summary, we applied machine learning classification models, including QSAR and BABM models, to identify inhibitors of the entry of SARS-CoV-2 into host cells and the SARS-CoV-2 3CL protease from a large diverse compound library. The optimal models significantly increased the hit rates of the anti-SARS-CoV-2 HTS assays by several-fold. Twenty-two compounds showed potent (<5 $\mu$M) anti-SARS-CoV-2 activities in a live virus assay; for 18 of them, anti-SARS-CoV-2 activities have not been reported previously. These compounds have the potential to be developed as novel anti-COVID-19 drug leads. Therefore, machine learning classification models can be used as a complementary approach to HTS to improve the speed and efficiency of anti-SARS-CoV-2 drug discovery.

## EXPERIMENTAL SECTION

**HTS Assays.** All assays were performed according to protocols described previously.[16] Briefly, the SARS-CoV-2 PP entry assay was performed in a human ACE2 (HEK293-ACE2) cell line under biosafety level 2 (BSL-2) containment. After incubation for 48 h for entry of PP into cells and luciferase reporter expression, luciferase activity was measured using Bright-Glo Luciferase Assay reagents (Promega). The 3CL protease assay was performed in black, medium-binding microplates (Greiner Bio-One) with the enzyme (50 nM) in a reaction buffer and a substrate. After incubation for 1 h for the enzyme reaction, the fluorescence intensity was measured at excitation and emission wavelengths of 340 and 460 nm, respectively. Compounds were tested as five-point 1:5 titrations (experimental validation) or 11-point 1:3 titrations (experimental confirmation), starting from 57.5 $\mu$M, for both the PP entry assay and the 3CL enzyme assay. The live SARS-CoV-2 CPE assay was performed at a contractor BSL-3 facility at the Southern Research Institute (Birmingham, AL). In addition, the cytotoxicity of these compounds was evaluated in a cell viability assay that measures intracellular ATP content using a PHERAstar FSX plate reader (BMG LABTECH). All compound libraries used in this study were assembled within NCATS. The purity of lead compounds was determined to be >95% by HPLC, and copies of the HPLC traces are provided in Figure S2.

**Data Collection for Modeling.** NCATS in-house collections of bioactive compounds, including approved and investigational drugs, were screened against the PP entry assay and the 3CL protease assay in quantitative HTS (qHTS) format.[9,11] Detailed descriptions of these high-throughput drug repurposing assays and all of the screening data are publicly available through the NCATS/NIH open science data portal (OpenData, https://opendata.ncats.nih.gov/covid19/). The qHTS data were analyzed using custom software developed in house at NCATS. Briefly, concentration−response curves were fit to a four-parameter Hill equation yielding concentrations of half-maximal activity ($AC_{50}$) and maximal response (efficacy) values.[55,56] Compounds were further designated as class 1−4 on the basis of the shape of the concentration−response curve.[57] Compounds that exhibited activation effects were assigned class 1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4, and 3 curves. Compounds that exhibited inhibitory effects were assigned class −1.1, −1.2, −1.3, −1.4, −2.1, −2.2, −2.3, −2.4, and −3 curves. Compounds that showed no significant concentration response were considered inactive and assigned class 4. For modeling purposes, we assigned each compound one of three outcomes: active (1), inactive (0) or inconclusive (exclude). For the PP entry assay, compounds that showed inhibition with >50% efficacy and were inactive or at least 6-fold less potent in the cell viability counter assay were considered active (1); compounds with a positive curve class were considered inactive (0), and all other compounds were considered inclusive and excluded from modeling. For the 3CL protease assay, compounds that showed inhibition with >50% efficacy were considered active (1), compounds with a positive curve class were considered inactive (0), and all other compounds were considered inclusive and excluded from modeling. The activity assignments for all of the compounds included in constructing the PP entry assay and 3CL protease assay models are listed in Table S1, along with their corresponding structures.

**QSAR Modeling.** The two-dimensional structures of all compounds encoded in SMILES strings were converted to the Extended Connectivity Fingerprints radius 4 (ECFP4) using the Chemistry Development Kit (CDK) package[61] in the Konstanz Information Miner open-source software (KNIME, version 4.0.2, https://www.knime.org/). ECFP4 encodes circular topological fragments into a fixed length binary fingerprint ($n$ = 1024) where the presence or absence of the feature was recorded in a binary system as 1 or 0, respectively.

Five classification machine learning models were built and tested using R version 3.4.2, including the "e1071" package for naïve Bayes (NB) and support vector machine (SVM) classifiers, the "Random


https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

Forest" package for the random forest (RF) classifier, the "nnet" package for the neural networks (NNET) classifier, and the "xgboost" package for the eXtreme gradient boosting (XGboost) classifier.[58] The NB classifier is implemented with a Laplace smoothing setup, while the SVM classifier uses a Gaussian radial basis function kernel. In addition, the other parameters of the SVM, RF, and NNET classifiers were set to default values. The parameters of the XGboost classifier were set as follows: maximum depth of a tree, 3; control the learning rate, 0.01; and subsample ratio of columns when constructing each tree, 0.5.

To optimize model performance, feature selection prior to machine learning modeling was performed using four different methods: Fisher's exact test with a *P* value, area under the receiver operating characteristic curve (AUC-ROC) value, Gini score from the RF algorithm, and Gain score from the XGboost algorithm. For the Fisher's exact test method, features were selected at five different P value cutoffs, which were in the range of 0.01−0.05 with an interval of 0.01. For the AUC-ROC method, features were selected at four different cutoffs, which were in the range of 0.52−0.58 with an interval of 0.02 using "pROC" packages.[59] For the RF and XGboost methods, features were selected using the "Random Forest" and "xgboost"[58] packages, respectively. Features ranked with Gini or Gain scores were picked at 10 intervals from the top 10 to top 50. Different feature sets generated in the feature selection process were used to build machine learning models, and their performances were evaluated.

To evaluate model performance, the data set was randomly divided into two parts: 70% for training and cross-validation and 30% for external validation. The training data set was used to tune the model parameters to yield the maximum model performance, and the external validation data set was used to evaluate the model's capacity to extrapolate to new data. Each model was evaluated by an internal 3-fold cross validation on the training data set. To ensure the robustness of our results, the cross-validation process was repeated 20 times with different random data partitions. Because the class distributions of the assay outcomes were imbalanced, the training data set was rebalanced using four different subsampling methods, including up sampling, down sampling, random oversampling examples (ROSE), and synthetic minority oversampling technique (SMOTE) via the "ROSE" and "DMwR" packages in R.[60,61] Model performance was evaluated by the AUC-ROC value, which ranged from 0.5 (a random classifier) to 1 (a perfect classifier). The combinations of feature sets, rebalancing strategies, and machine learning algorithms yielded models with different performances, and the model with the optimal performance (i.e., maximum AUC-ROC value) was used for further virtual screening.

**BABM Modeling.** Two types of bioactivity-based models (BABM-M and BABM-S) and two types of structure−activity combined modes (CM-M and CM-S) were built using compound activity profiles from two sets of qHTS assays (MLS, 225 readouts; Sytravon, 130 readouts).[16] The activity-based models were trained and tested using the activity profiles from the NCATS Pharmaceutical Collection (NPC)[20] and the Library of Pharmacologically Active Compounds (LOPAC). The ChemoTyper was used to generate structure fingerprints for all compounds for the structure−activity combined models. In the combined models, the structure fingerprint and the activity profile were concatenated to form a new fingerprint for each compound. For modeling purposes, each compound was represented as a bit vector of 1s and 0s. In a structure fingerprint, the bit value was set to 1 if the compound contains a particular structural feature and 0 if the compound does not have that feature. For activity profile data, each assay readout was treated as a feature and the feature value was set to 1 for "active" compounds and 0 for inactive compounds. The detailed modeling process was described previously.[16] Briefly, the weighted feature significance (WFS) method previously developed at NCATS was used to build the models. For each model, compounds were randomly divided into two groups of approximately equal sizes (i.e., one used for training and the other for testing). To evaluate the robustness of the models, the randomization was performed 10 times, generating 10 different training and test sets. The performance of the BABM models was evaluated by calculating the AUC-ROC value. The random data split and model training and testing were repeated 10 times, and the average AUC-ROC values were calculated for each model.

**Virtual Screening of Large Compound Libraries.** The optimal models were applied to predict the activity of the NCATS in-house collection of ∼360K compounds against viral entry, as well as 3CL protease. For the QSAR models, each compound was assigned a predicted probability to a specific outcome based on the optimal model, and compounds with the highest probabilities (>0.5) were collected. To identify novel, structurally diverse compounds, the collected compounds were subjected to a k-means clustering analysis using the Hartigan−Wong algorithm implemented in R, resulting in 100 clusters based on ECFP4 fingerprints. The compounds with top probabilities in each cluster were selected for experimental validation. For BABM models, WFS score cutoff values for model-predicted actives were determined using the ROC curves where both sensitivity and specificity were optimized. Only compounds that scored higher than the cutoff values were considered candidates for follow-up selection. For experimental validation, the selection was driven mostly by availability of assay resources and physical samples. The top-ranked compounds that met the WFS score cutoff from a model with physical samples available for cherry picking were selected to fit into one or two 1536-well plates for testing.

**Statistical Analysis.** Example curve plots were fitted using a four-parameter logistic regression with the percent assay activity as the response and log 10 compound concentration as the independent variable using the "drc" statistical package in R. Plots were generated using the "ggplot2" package in R. Representative chemical structures were drawn using ChemDraw Professional (version 17.1).

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.1c01372.

Original drug repurposing assay data, features for the optimal models, model performances, experimental validation, experimental confirmation, and HPLC traces of lead compounds (PDF)

SMILES molecular formula strings (CSV)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Ruili Huang** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States;* orcid.org/0000-0001-8886-8311; Email: huangru@mail.nih.gov

### Authors

**Tuan Xu** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States*

**Miao Xu** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States*

**Wei Zhu** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States*

**Catherine Z. Chen** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States*

**Qi Zhang** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National*




https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

*Institutes of Health (NIH), Rockville, Maryland 20850, United States*

**Wei Zheng** − *Division of Pre-clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, Maryland 20850, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.1c01372

### Author Contributions

R.H. designed the study. M.X., W. Zhu, and Q.Z. conducted the experiments. C.Z.C. and W. Zheng directed the generation of experimental data. T.X. and R.H. built the models and performed the virtual screening. T.X. and R.H. wrote the manuscript. R.H. directed the project. All authors reviewed the manuscript.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the Intramural Research Programs of the National Center for Advancing Translational Sciences, National Institutes of Health. The authors thank the NCATS OpenData Portal team for making the screening data publicly available and Katlin Recabo, Danielle Bougie, and Paul Shinn for assistance with compound management and quality control (QC).

## ■ ABBREVIATIONS USED

3CL, 3-chymotrypsin-like; $AC_{50}$, half-maximal activity concentration; ACE2, angiotensin converting enzyme 2; AUC-ROC, area under the receiver operating characteristic curve; BABM, biological activity-based modeling; BABM-M, activity-based model (MLS); BABM-S, activity-based model (Sytravon); CM-M, combined model (SBM + BABM-M); CM-S, combined model (SBM + BABM-S); COVID-19, coronavirus disease 19; FN, false negative; FP, false positive; HTS, high-throughput screening; NB, naïve Bayes; NNET, neural networks; NPC, NCATS Pharmaceutical Collection; PP, pseudotyped particle; PPV, positive predictive value; QSAR, quantitative structure−activity relationship; RBD, receptor-binding domain; RF, random forest; RFU, relative fluorescence unit; ROSE, random oversampling examples; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SBM, structure (ToxPrint)-based model; SMOTE, synthetic minority oversampling technique; SVM, support vector machine; TN, true negative; TOX, cytotoxicity assay; TP, true positive; WFS, weighted feature significance; XGboost, eXtreme gradient boosting

## ■ REFERENCES

(1) Yang, Y.; Peng, F.; Wang, R.; Guan, K.; Jiang, T.; Xu, G.; Sun, J.; Chang, C. The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. *J. Autoimmun.* **2020**, *109*, 102434.

(2) Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; Cheng, Z.; Yu, T.; Xia, J.; Wei, Y.; Wu, W.; Xie, X.; Yin, W.; Li, H.; Liu, M.; Xiao, Y.; Gao, H.; Guo, L.; Xie, J.; Wang, G.; Jiang, R.; Gao, Z.; Jin, Q.; Wang, J.; Cao, B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)* **2020**, *395*, 497−506.

(3) Rothan, H. A.; Byrareddy, S. N. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* **2020**, *109*, 102433.

(4) Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **2020**, *581*, 215−220.

(5) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **2020**, *367*, 1444−1448.

(6) V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **2021**, *19*, 155−170.

(7) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; et al. 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.

(8) 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.

(9) Chen, C. Z.; Xu, M.; Pradhan, M.; Gorshkov, K.; Petersen, J. D.; Straus, M. R.; Zhu, W.; Shinn, P.; Guo, H.; Shen, M.; Klumpp-Thomas, C.; Michael, S. G.; Zimmerberg, J.; Zheng, W.; Whittaker, G. R. Identifying SARS-CoV-2 Entry Inhibitors through Drug Repurposing Screens of SARS-S and MERS-S Pseudotyped Particles. *ACS Pharmacol. Transl. Sci.* **2020**, *3* (6), 1165−1175.

(10) Brimacombe, K. R.; Zhao, T.; Eastman, R. T.; Hu, X.; Wang, K.; Backus, M.; Baljinnyam, B.; Chen, C. Z.; Chen, L.; Eicher, T. An OpenData portal to share COVID-19 drug repurposing data in real time. *bioRxiv* **2020**, DOI: 10.1101/2020.06.04.135046.

(11) Zhu, W.; Xu, M.; Chen, C. Z.; Guo, H.; Shen, M.; Hu, X.; Shinn, P.; Klumpp-Thomas, C.; Michael, S. G.; Zheng, W. Identification of SARS-CoV-2 3CL Protease Inhibitors by a Quantitative High-throughput Screening. *ACS Pharmacol. Transl. Sci.* **2020**, *3* (5), 1008−1016.

(12) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(13) Gloriam, D. E. Bigger is better in virtual drug screens. *Nature* **2019**, *566*, 193−194.

(14) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **2019**, *18*, 435−441.

(15) Adeshina, Y. O.; Deeds, E. J.; Karanicolas, J. Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci. U S A* **2020**, *117*, 18477−18488.

(16) Huang, R.; Xu, M.; Zhu, H.; Chen, C. Z.; Zhu, W.; Lee, E. M.; He, S.; Zhang, L.; Zhao, J.; Shamim, K.; Bougie, D.; Huang, W.; Xia, M.; Hall, M. D.; Lo, D.; Simeonov, A.; Austin, C. P.; Qiu, X.; Tang, H.; Zheng, W. Biological activity-based modeling identifies antiviral leads against SARS-CoV-2. *Nat. Biotechnol.* **2021**, *39*, 747.

(17) Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(18) Zhang, R.; Li, X.; Zhang, X.; Qin, H.; Xiao, W. Machine learning approaches for elucidating the biological effects of natural products. *Nat. Prod. Rep.* **2021**, *38* (2), 346−361.

(19) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat. Commun.* **2016**, *7*, 10425.

(20) Huang, R.; Southall, N.; Wang, Y.; Yasgar, A.; Shinn, P.; Jadhav, A.; Nguyen, D. T.; Austin, C. P. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* **2011**, *3*, 80ps16.

(21) Zhang, Z. R.; Zhang, Y. N.; Li, X. D.; Zhang, H. Q.; Xiao, S. Q.; Deng, F.; Yuan, Z. M.; Ye, H. Q.; Zhang, B. A cell-based large-scale screening of natural compounds for inhibitors of SARS-CoV-2. *Signal Transduction Targeted Ther.* **2020**, *5*, 1−3.

(22) Chen, C. Z.; Shinn, P.; Itkin, Z.; Eastman, R. T.; Bostwick, R.; Rasmussen, L.; Huang, R.; Shen, M.; Hu, X.; Wilson, K. M.; Brooks, B. M.; Guo, H.; Zhao, T.; Klump-Thomas, C.; Simeonov, A.; Michael, S. G.; Lo, D. C.; Hall, M. D.; Zheng, W. Drug Repurposing Screen for Compounds Inhibiting the Cytopathic Effect of SARS-CoV-2. *Front. Pharmacol.* **2021**, *11*, 592737−592737.
(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(24) Anderson, G. B.; Oleson, K. W.; Jones, B.; Peng, R. D. Classifying heatwaves: Developing health-based models to predict high-mortality versus moderate United States heatwaves. *Clim. Change* **2018**, *146*, 439−453.
(25) Lv, Z.; Jin, S.; Ding, H.; Zou, Q. A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. *Front. Bioeng. Biotechnol.* **2019**, *7*, 215.
(26) Gong, J.; Liu, J.; Hao, W.; Nie, S.; Wang, S.; Peng, W. Computer-aided diagnosis of ground-glass opacity pulmonary nodules using radiomic features analysis. *Phys. Med. Biol.* **2019**, *64*, 135015.
(27) Xu, T.; Ngan, D. K.; Ye, L.; Xia, M.; Xie, H. Q.; Zhao, B.; Simeonov, A.; Huang, R. Predictive Models for Human Organ Toxicity Based on In Vitro Bioactivity Data and Chemical Structure. *Chem. Res. Toxicol.* **2020**, *33*, 731−741.
(28) Kc, G. B.; Bocci, G.; Verma, S.; Hassan, M. M.; Holmes, J.; Yang, J. J.; Sirimulla, S.; Oprea, T. I. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat. Mach. Intell.* **2021**, *3*, 527−535.
(29) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chem. Res. Toxicol.* **2015**, *28*, 738−751.
(30) Vankadara, S.; Wong, Y. X.; Liu, B.; See, Y. Y.; Tan, L. H.; Tan, Q. W.; Wang, G.; Karuna, R.; Guo, X.; Tan, S. T.; Fong, J. Y.; Joy, J.; Chia, C. S. B. A head-to-head comparison of the inhibitory activities of 15 peptidomimetic SARS-CoV-2 3CLpro inhibitors. *Bioorg. Med. Chem. Lett.* **2021**, *48*, 128263.
(31) Dampalla, C. S.; Kim, Y.; Bickmeier, N.; Rathnayake, A. D.; Nguyen, H. N.; Zheng, J.; Kashipathy, M. M.; Baird, M. A.; Battaile, K. P.; Lovell, S.; Perlman, S.; Chang, K. O.; Groutas, W. C. Structure-Guided Design of Conformationally Constrained Cyclohexane Inhibitors of Severe Acute Respiratory Syndrome Coronavirus-2 3CL Protease. *J. Med. Chem.* **2021**, *64*, 10047−10058.
(32) Bai, B.; Belovodskiy, A.; Hena, M.; Kandadai, A. S.; Joyce, M. A.; Saffran, H. A.; Shields, J. A.; Khan, M. B.; Arutyunova, E.; Lu, J.; Bajwa, S. K.; Hockman, D.; Fischer, C.; Lamer, T.; Vuong, W.; van Belkum, M. J.; Gu, Z.; Lin, F.; Du, Y.; Xu, J.; Rahim, M.; Young, H. S.; Vederas, J. C.; Tyrrell, D. L.; Lemieux, M. J.; Nieman, J. A. Peptidomimetic $\alpha$-Acyloxymethylketone Warheads with Six-Membered Lactam P1 Glutamine Mimic: SARS-CoV-2 3CL Protease Inhibition, Coronavirus Antiviral Activity, and in Vitro Biological Stability. *J. Med. Chem.* **2022**, *65*, 2905.
(33) Zhang, C. H.; Stone, E. A.; Deshmukh, M.; Ippolito, J. A.; Ghahremanpour, M. M.; Tirado-Rives, J.; Spasov, K. A.; Zhang, S.; Takeo, Y.; Kudalkar, S. N.; Liang, Z.; Isaacs, F.; Lindenbach, B.; Miller, S. J.; Anderson, K. S.; Jorgensen, W. L. Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent. Sci.* **2021**, *7*, 467−475.
(34) Ghosh, A. K.; Raghavaiah, J.; Shahabi, D.; Yadav, M.; Anson, B. J.; Lendy, E. K.; Hattori, S. I.; Higashi-Kuwata, N.; Mitsuya, H.; Mesecar, A. D. Indole Chloropyridinyl Ester-Derived SARS-CoV-2 3CLpro Inhibitors: Enzyme Inhibition, Antiviral Efficacy, Structure-Activity Relationship, and X-ray Structural Studies. *J. Med. Chem.* **2021**, *64*, 14702−14714.
(35) Xia, Z.; Sacco, M.; Hu, Y.; Ma, C.; Meng, X.; Zhang, F.; Szeto, T.; Xiang, Y.; Chen, Y.; Wang, J. Rational Design of Hybrid SARS-CoV-2 Main Protease Inhibitors Guided by the Superimposed Cocrystal Structures with the Peptidomimetic Inhibitors GC-376, Telaprevir, and Boceprevir. *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 1408−1421.
(36) Boras, B.; Jones, R. M.; Anson, B. J.; Arenson, D.; Aschenbrenner, L.; Bakowski, M. A.; Beutler, N.; Binder, J.; Chen, E.; Eng, H.; Hammond, H.; Hammond, J.; Haupt, R. E.; Hoffman, R.; Kadar, E. P.; Kania, R.; Kimoto, E.; Kirkpatrick, M. G.; Lanyon, L.; Lendy, E. K.; Lillis, J. R.; Logue, J.; Luthra, S. A.; Ma, C.; Mason, S. W.; McGrath, M. E.; Noell, S.; Obach, R. S.; O'Brien, M. N.; O'Connor, R.; Ogilvie, K.; Owen, D.; Pettersson, M.; Reese, M. R.; Rogers, T. F.; Rosales, R.; Rossulek, M. I.; Sathish, J. G.; Shirai, N.; Steppan, C.; Ticehurst, M.; Updyke, L. W.; Weston, S.; Zhu, Y.; White, K. M.; García-Sastre, A.; Wang, J.; Chatterjee, A. K.; Mesecar, A. D.; Frieman, M. B.; Anderson, A. S.; Allerton, C. Preclinical characterization of an intravenous coronavirus 3CL protease inhibitor for the potential treatment of COVID19. *Nat. Commun.* **2021**, *12*, 6055.
(37) Qiao, J.; Li, Y. S.; Zeng, R.; Liu, F. L.; Luo, R. H.; Huang, C.; Wang, Y. F.; Zhang, J.; Quan, B.; Shen, C.; Mao, X.; Liu, X.; Sun, W.; Yang, W.; Ni, X.; Wang, K.; Xu, L.; Duan, Z. L.; Zou, Q. C.; Zhang, H. L.; Qu, W.; Long, Y. H.; Li, M. H.; Yang, R. C.; Liu, X.; You, J.; Zhou, Y.; Yao, R.; Li, W. P.; Liu, J. M.; Chen, P.; Liu, Y.; Lin, G. F.; Yang, X.; Zou, J.; Li, L.; Hu, Y.; Lu, G. W.; Li, W. M.; Wei, Y. Q.; Zheng, Y. T.; Lei, J.; Yang, S. SARS-CoV-2 M(pro) inhibitors with antiviral activity in a transgenic mouse model. *Science* **2021**, *371*, 1374−1378.
(38) Csermely, P.; Agoston, V.; Pongor, S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* **2005**, *26*, 178−182.
(39) Ramsay, R. R.; Popovic-Nikolic, M. R.; Nikolic, K.; Uliassi, E.; Bolognesi, M. L. A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* **2018**, *7*, 3.
(40) Chen, N.; Zhou, M.; Dong, X.; Qu, J.; Gong, F.; Han, Y.; Qiu, Y.; Wang, J.; Liu, Y.; Wei, Y.; Xia, J.; Yu, T.; Zhang, X.; Zhang, L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet (London, England)* **2020**, *395*, 507−513.
(41) Shyr, Z. A.; Gorshkov, K.; Chen, C. Z.; Zheng, W. Drug discovery strategies for SARS-CoV-2. *J. Pharmacol. Exp. Ther.* **2020**, *375*, 127.
(42) Xu, T.; Zheng, W.; Huang, R. High-throughput screening assays for SARS-CoV-2 drug development: current status and future directions. *Drug Discovery Today* **2021**, *26*, 2439−2444.
(43) Sun, Q.; Ye, F.; Liang, H.; Liu, H.; Li, C.; Lu, R.; Huang, B.; Zhao, L.; Tan, W.; Lai, L. Bardoxolone and bardoxolone methyl, two Nrf2 activators in clinical trials, inhibit SARS-CoV-2 replication and its 3C-like protease. *Signal Transduction Targeted Ther.* **2021**, *6*, 212.
(44) Ohashi, H.; Watashi, K.; Saso, W.; Shionoya, K.; Iwanami, S.; Hirokawa, T.; Shirai, T.; Kanaya, S.; Ito, Y.; Kim, K. S. Multidrug treatment with nelfinavir and cepharanthine against COVID-19. *bioRxiv* **2020**, DOI: 10.1101/2020.04.14.039925.
(45) Jeon, S.; Ko, M.; Lee, J.; Choi, I.; Byun, S. Y.; Park, S.; Shum, D.; Kim, S. Identification of Antiviral Drug Candidates against SARS-CoV-2 from FDA-Approved Drugs. *Antimicrob. Agents Chemother.* **2020**, *64*, No. e00819-20.
(46) Wilkinson, T.; Dixon, R.; Page, C.; Carroll, M.; Griffiths, G.; Ho, L.-P.; De Soyza, A.; Felton, T.; Lewis, K. E.; Phekoo, K.; Chalmers, J. D.; Gordon, A.; McGarvey, L.; Doherty, J.; Read, R. C.; Shankar-Hari, M.; Martinez-Alier, N.; O'Kelly, M.; Duncan, G.; Walles, R.; Sykes, J.; Summers, C.; Singh, D. on behalf of the, A. C. ACCORD: A Multicentre, Seamless, Phase 2 Adaptive Randomisation Platform Study to Assess the Efficacy and Safety of Multiple Candidate Agents for the Treatment of COVID-19 in Hospitalised Patients: A structured summary of a study protocol for a randomised controlled trial. *Trials* **2020**, *21*, 691.
(47) Bocci, G.; Bradfute, S. B.; Ye, C.; Garcia, M. J.; Parvathareddy, J.; Reichard, W.; Surendranathan, S.; Bansal, S.; Bologa, C. G.; Perkins, D. J.; et al. Virtual and In Vitro Antiviral Screening Revive Therapeutic Drugs for COVID-19. *ACS Pharmacol. Transl. Sci.* **2020**, *3*, 1278−1292.
(48) Weston, S.; Coleman, C. M.; Haupt, R.; Logue, J.; Matthews, K.; Li, Y.; Reyes, H. M.; Weiss, S. R.; Frieman, M. B. Broad anti-coronavirus activity of Food and Drug Administration-approved drugs against SARS-CoV-2 in vitro and SARS-CoV in vivo. *J. Virol.* **2020**, *94*, No. e01218-20.
(49) Mengue, J. B.; Held, J.; Kreidenweiss, A. AQ-13-an investigational antimalarial drug. *Expert. Opin. Investig. Drugs* **2019**, *28*, 217−222.


https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599

(50) Souroullas, G. P.; Jeck, W. R.; Parker, J. S.; Simon, J. M.; Liu, J.-Y.; Paulk, J.; Xiong, J.; Clark, K. S.; Fedoriw, Y.; Qi, J.; Burd, C. E.; Bradner, J. E.; Sharpless, N. E. An oncogenic Ezh2 mutation induces tumors through global redistribution of histone 3 lysine 27 trimethylation. *Nat. Med.* **2016**, *22*, 632−640.
(51) Benchaoui, H. A.; Cox, S. R.; Schneider, R. P.; Boucher, J. F.; Clemence, R. G. The pharmacokinetics of maropitant, a novel neurokinin type-1 receptor antagonist, in dogs. *J. Vet. Pharmacol. Ther.* **2007**, *30*, 336−344.
(52) Li, J. J.; Carson, K. G.; Trivedi, B. K.; Yue, W. S.; Ye, Q.; Glynn, R. A.; Miller, S. R.; Connor, D. T.; Roth, B. D.; Luly, J. R.; Low, J. E.; Heilig, D. J.; Yang, W.; Qin, S.; Hunt, S. Synthesis and structure-activity relationship of 2-amino-3-heteroaryl-quinoxalines as non-peptide, small-molecule antagonists for interleukin-8 receptor. *Bioorg. Med. Chem.* **2003**, *11*, 3777−3790.
(53) Sander, P.; Mostafa, H.; Soboh, A.; Schneider, J. M.; Pala, A.; Baron, A.-K.; Moepps, B.; Wirtz, C. R.; Georgieff, M.; Schneider, M. Vacquinol-1 inducible cell death in glioblastoma multiforme is counter regulated by TRPM7 activity induced by exogenous ATP. *Oncotarget* **2017**, *8*, 35124−35137.
(54) Lin, J.; Zheng, Z.; Li, Y.; Yu, W.; Zhong, W.; Tian, S.; Zhao, F.; Ren, X.; Xiao, J.; Wang, N.; Liu, S.; Wang, L.; Sheng, F.; Chen, Y.; Jin, C.; Li, S.; Xia, B. A novel Bcl-XL inhibitor Z36 that induces autophagic cell death in Hela cells. *Autophagy* **2009**, *5*, 314−320.
(55) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U S A* **2006**, *103*, 11473−11478.
(56) Wang, Y.; Jadhav, A.; Southal, N.; Huang, R.; Nguyen, D. T. A grid algorithm for high throughput fitting of dose-response curve data. *Curr. Chem. Genomics* **2010**, *4*, 57−66.
(57) Huang, R. A Quantitative High-Throughput Screening Data Analysis Pipeline for Activity Profiling. In *High-Throughput Screening Assays in Toxicology*, 1st ed.; Zhu, H., Xia, M., Eds.; Humana Press: Totowa, NJ, 2016; Vol. *1473*.
(58) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y. Xgboost: extreme gradient boosting. *R Package*, ver. 0.4-2; 2015; pp 1−4.
(59) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **2011**, *12*, 77.
(60) Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *The R Journal* **2014**, *6*, 79.
(61) Torgo, L.; Torgo, M. L. Package 'DMwR'. *Comprehensive R Archive Network*, 2013.


https://doi.org/10.1021/acs.jmedchem.1c01372
J. Med. Chem. 2022, 65, 4590−4599