Original Research

Developing a SARS-CoV-2 main protease binding prediction random forest model for drug repurposing for COVID-19 treatment

Jie Liu*, Liang Xu*, Wenjing Guo, Zoe Li, Md Kamrul Hasan Khan, Weigong Ge, Tucker A Patterson and Huixiao Hong

National Center for Toxicological Research, U.S. Food & Drug Administration, Jefferson, AR 72079, USA
*These authors contributed equally to this paper.
Corresponding author: Huixiao Hong. Email: huixiao.hong@fda.hhs.gov

Impact statement

Although the World Health Organization (WHO) declared "with great hope" an end to COVID-19 as a global health emergency, SARS-CoV-2 infection continues around worldwide as evidenced by more than 287,000 confirmed cases reported to WHO on 14 August 2023. Therefore, effective drugs are needed for treating COVID-19 patients. New drug development is not only costly but also time-consuming; therefore, it is not ideal for emerging diseases, such as COVID-19. Repurposing approved drugs for treating COVID-19 is important to combat the COVID-19 pandemic. The most used repurposing strategy is based on human expertise. This is the first machine learning model that was developed for repurposing FDA-approved drugs for COVID-19 treatment through targeting the SARS-CoV-2 main protease. The discovered FDA-approved drugs in this study provide potential repurposing candidates for clinical consideration, thus the findings of this study are expected to advance the development of COVID-19 treatment.

Abstract

The coronavirus disease 2019 (COVID-19) global pandemic resulted in millions of people becoming infected with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus and close to seven million deaths worldwide. It is essential to further explore and design effective COVID-19 treatment drugs that target the main protease of SARS-CoV-2, a major target for COVID-19 drugs. In this study, machine learning was applied for predicting the SARS-CoV-2 main protease binding of Food and Drug Administration (FDA)-approved drugs to assist in the identification of potential repurposing candidates for COVID-19 treatment. Ligands bound to the SARS-CoV-2 main protease in the Protein Data Bank and compounds experimentally tested in SARS-CoV-2 main protease binding assays in the literature were curated. These chemicals were divided into training (516 chemicals) and testing (360 chemicals) data sets. To identify SARS-CoV-2 main protease binders as potential candidates for repurposing to treat COVID-19, 1188 FDA-approved drugs from the Liver Toxicity Knowledge Base were obtained. A random forest algorithm was used for constructing predictive models based on molecular descriptors calculated using Mold2 software. Model performance was evaluated using 100 iterations of fivefold cross-validations which resulted in 78.8% balanced accuracy. The random forest model that was constructed from the whole training dataset was used to predict SARS-CoV-2 main protease binding on the testing set and the FDA-approved drugs. Model applicability domain and prediction confidence on drugs predicted as the main protease binders discovered 10 FDA-approved drugs as potential candidates for repurposing to treat COVID-19. Our results demonstrate that machine learning is an efficient method for drug repurposing and, thus, may accelerate drug development targeting SARS-CoV-2.

Keywords: SARS-CoV-2, main protease, drug repurposing, random forest, binding activity, predictive model

Experimental Biology and Medicine **2023; 248: 1927–1936. DOI: 10.1177/15353702231209413**

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused the coronavirus disease 2019 (COVID-19) pandemic and continues to pose a threat to global public health.[1] According to the World Health Organization (WHO), as of August 14, 2023, there have been over 769 million confirmed cases of COVID-19 worldwide, including more than 287,000 on 14 August 2023, and more than 6.9 million deaths (https://covid19.who.int). Although COVID-19 vaccines have played a crucial role in reducing virus transmission and providing protection, the emergence of new variants, such as alpha, beta, gamma, delta, epsilon, omicron, arcturus, and eris, necessitates the development of updated vaccines to address vaccine-escape mutations.[2] There is a need for the scientific community and pharmaceutical companies to develop new drugs against SARS-CoV-2, especially for

ISSN 1535-3702

infected patients who do not respond to the currently available COVID-19 drugs.

To fight against COVID-19, several target proteins of SARS-CoV-2 have been identified, including the main protease, the papain-like protein, the spike protein, and the RNA-dependent RNA polymerase.[3–5] The main protease has been shown to be essential for viral replication and thus has been recognized as a potent drug target.[6,7] In particular, Paxlovid (nirmatrelvir and ritonavir) is the first oral antiviral pill approved by the Food and Drug Administration (FDA) to treat COVID-19 in adults.[8] In Paxlovid, nirmatrelvir acts as the main protease inhibitor and ritonavir inhibits the P450 3A-mediated metabolism of nirmatrelvir.[8] Currently, more than 400 three-dimensional (3D) structures of the ligand-bound main protease have been determined and deposited in the Protein Data Bank (PDB) (http://www.rcsb.org/). These structures provide valuable sources for insights into the binding mechanisms of diverse inhibitors. The relationships between the structures of main protease inhibitors and their inhibition activity have been widely discussed in many reviews.[9–13] However, combining this structural information with other approaches for the development of new main protease inhibitors has not been reported.

Although there are now significant efforts aimed at developing drugs that can inhibit the main protease, drug repurposing offers a cost-effective and time-efficient approach to address unmet medical needs.[14–16] Different from *de novo* drug design, drug repurposing uses existing drugs that were originally developed for different indications to screen drugs for new diseases. Because of the knowledge and understanding of an existing drug's safety, mechanisms of action, and pharmacokinetics, drug repurposing reduces the risks associated with drug development, and has shown promising results for the discovery of the main protease inhibitors,[17–19] especially when combined with the power of artificial intelligence approaches.[20]

Machine learning is the approach to learn from the provided data, identify the patterns within the provided data, and make predictions on new data using what it has learned. Machine learning methods have been widely used for toxicity prediction, novel drug discovery, and drug repurposing.[21–29] With the large amount of data from existing studies on the main protease binding activity, machine learning is a valuable way to identify the potential main protease binders from FDA-approved drugs.

The purpose of this study was to apply a machine learning approach for repurposing FDA-approved drugs that could bind the main protease of SARS-CoV-2 as potential candidates for the treatment of COVID-19. To achieve this goal, we curated the main protease binding activity data from public databases and the literature, constructed and validated the random forest model using cross-validations and external validation, and predicted the potential main protease binding activity of FDA-approved drugs. Our results demonstrate that machine learning can be used to assist in drug repurposing, and thus accelerate drug development targeting the SARS-CoV-2 main protease.

Materials and methods

Study design

The study design of this research is illustrated in Figure 1. The chemicals and their SARS-CoV-2 binding data curated from the PDB and scientific literature were divided into training and testing sets. To evaluate the modeling process, fivefold cross-validations were conducted on the training set. More specifically, the training set was randomly split into fivefolds. Fourfolds were then used to develop a random forest model and the remaining fold was used to test the developed model. Because the dataset is very imbalanced, downsampling the majority (binders) was used to make a balanced dataset to train a random forest model, so that the imbalance impact on the performance of the constructed model is alleviated. This process was iterated five times until each of the fivefolds was used only once as the test set. All predictions were used to calculate the performance metrics. The fivefold cross-validations were repeated 100 times to reach a robust statistical estimation of model performance. Then, the whole training set was used to construct a random forest model with the same downsampling strategy. The model was evaluated using the testing set and was applied to predict the potential binding activity of FDA-approved drugs curated from the Liver Toxicity Knowledge Base (LTKB, https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-severity-and-toxicity-dilist-dataset).[30]

Data curation

We first searched the PDB database (https://www.rcsb.org/) using the keyword "SARS-CoV-2 main protease." Then, we manually checked the hits and downloaded the structure files. Thereafter, we separated the ligands from the complexes and removed redundant ligands. Due to enzyme reactions, compounds in the complex structures in the PDB may be different from their parent compounds. In this study, parent compound structures of the bound ligand were used. Molecular weights of these compounds have a wide distribution, varying from 95 to 710 daltons (Da), with two normal distribution curves that have a mean value of 250 and 500 Da, respectively. In addition, the SARS-CoV-2 main protease binding activity data of compounds not reported in the PDB were curated from the scientific literature. The curated main protease binding activity data were experimentally tested in a transgenic mouse model,[31] identified by virtual screening of ultra-large databases and tested in binding and enzymatic assays,[32] tested using the fluorescence resonance energy transfer assay,[32] and obtained from a simplified cell-based assay.[33] We used the ligands in the PDB as binders for the training set and the binders determined by binding assays in the literature for the testing set. Finally, we proportionately divided the 245 non-binders into training and testing sets, resulting in the same ratio of binders/non-binders in the training (372 binders and 144 non-binders, Supplementary Table S1) and testing (259 binders and 101 non-binders, Supplementary Table S2) sets. A list of 1442

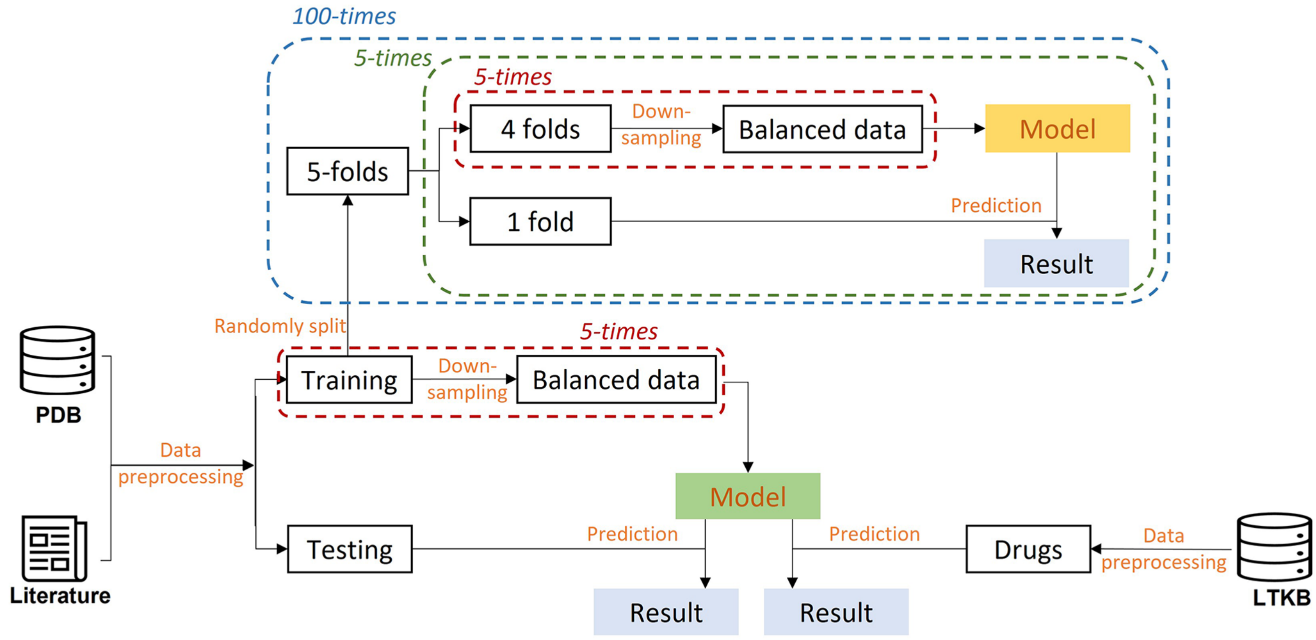

Figure 1. Study design. Data curated from the PDB and scientific literature were separated into training and testing sets. For cross-validation, the training set was randomly split into fivefolds. Downsampling binders were applied on fourfolds to get a balanced dataset for training a random forest model. The downsampling process was repeated five times using different random seeds. Next, the five random forest models were used to test the remaining fold and the predictions were combined to generate consensus predictions. The whole process was iterated five times until each of the fivefolds was used only once as a test for complete fivefold cross-validations. The fivefold cross-validations were repeated 100 times. Then, the whole training dataset was downsampled to get a balanced dataset (repeated five times) to construct random forest models. The models were evaluated using the testing set and applied to predict the potential binding activity of FDA-approved drugs curated from the LTKB.

Table 1. Datasets for model development and prediction.

Datasets	Main protease binder	Main protease non-binder	Total
Training set	372	144	516
Testing set	259	101	360
FDA-approved drugs			1188

FDA-approved drugs was acquired from the LTKB. After removing mixtures, inorganics, and duplicates, 1188 FDA-approved drugs remained for the main protease binding activity prediction (Supplementary Table S3). The curated data sets are summarized in Table 1.

Data preprocessing

Simplified molecular-input line-entry system (SMILES) codes of the compounds were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov/) and were used to generate two-dimensional (2D) chemical structures using the Online SMILES Translator and Structure File Generator (https://cactus.nci.nih.gov/translate/). The 777 molecular descriptors were calculated by Mold2[34,35] for each chemical from its 2D structure. The molecular descriptors with little information or low variance were discarded. First, molecular descriptors with zeros for more than 90% of the compounds were removed. Next, Shannon entropy analysis[34,36,37] was conducted to identify molecular descriptors with high variance. For each molecular descriptor, compounds in the training set were put into 20 groups with even bins of descriptor values. A Shannon entropy value was then calculated for the molecular descriptor according to equation (1)

$$H_n\left(p_1,p_2,\cdots,p_n\right)=-\sum_{i=1}^{n}p_i\,log_2p_i \qquad (1)$$

where p_i is the probability of chemicals in group i. The 297 molecular descriptors with Shannon entropy values greater than 3.0 were kept for model development. The 297 molecular descriptors are listed in Supplementary Table S4. Finally, the values of each molecular descriptor in all datasets (training set, testing set, and FDA-approved drugs) were normalized using equation (2)

$$V=\frac{V_o-Min_{train}}{Max_{train}-Min_{train}} \qquad (2)$$

where V is the normalized value, V_o is the original value, Min_{train} is the minimum value of the descriptor in the training set, and Max_{train} is the maximum value in the training set.

Model development and validation

The machine learning model was built using a random forest algorithm which is an ensemble learning algorithm combining all predictions from individual decision trees.[38] Due to the imbalance in classes in the training set (72.1% binders and 27.9% non-binders), we conducted downsampling of the binders to get balanced training data for model development. The downsampling process was performed by random selection of the binders (majority class) to balance with the non-binders. Five models were generated from the same imbalanced training set by downsampling the binders. Predictions from the five models for a compound were combined using majority voting strategy to make the final prediction for the compound.

Cross-validation was applied to the training set to evaluate the performance of the random forest model. In fivefold cross-validations, the training set was first randomly split into five even folds. Fourfolds were used to train the model and the remaining fold was used to test the model. This process was iterated five times until each of the fivefolds was used only once as the test set. The fivefold cross-validations were repeated 100 times to get a statistically robust estimation of model performance. The fivefold cross-validations and the results analysis were implemented using Python (3.8.5) scikit-learn packages (0.23.2), and the default parameters were applied for random forest model development.

To assess generalization of the constructed random forest model, the model that was built using the whole training set was applied to predict the potential main protease binding activity of compounds in the test set. First, the binders in the whole training set were randomly downsampled to get a balanced dataset for constructing a random forest model. The random downsampling was repeated five times, so that five models were constructed from the same training dataset. Finally, the majority voting was conducted to make final predictions using the five models.

Model performance measurement

Accuracy, sensitivity, specificity, balanced accuracy, and Matthew's correlation coefficient (MCC) were calculated by comparing the predictions from the testing or from each iteration of fivefold cross-validations with the actual binding activity data to measure the performance of the random forest models. These metrics were calculated using equations (3)–(7)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{BA} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{6}$$

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{7}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Prediction confidence analysis

A prediction from the random forest model is a probability that indicates how likely a compound is a SARS-CoV-2 main protease binder. This probability is used to not only classify the compound as binder or non-binder but also measure the confidence of the prediction. In addition to overall model performance, the confidence of the prediction results from the random forest model can be analyzed to inform better utilization of predictions in applications. The prediction confidence value of prediction was calculated from its prediction probability from the random forest model using equation (8) as in our previous studies[35,37,39]

$$\text{Prediction confidence} = \frac{|\text{prob} - 0.5|}{0.5} \tag{8}$$

where *prob* is the probability, the compound predicted is a main protease binder from the random forest model. The prediction confidence values range between 0 and 1. In our prediction confidence analysis, all predictions were first grouped into 10 sets according to their prediction confidence values, with 10 even bins of confidence values from 0 to 0.1 with an equal interval of 0.1. Next, the performance metrics were calculated for the 10 sets of predictions separately. Finally, the relationship between the prediction performance and prediction confidence level was analyzed.

Informative descriptors identification

Different molecular descriptors convey different structural information and contribute differently to the random forest models. The descriptors that were frequently used in the random forest models are informative descriptors to the model and should be important for a compound to interact with the main protease. To identify such informative descriptors, we first calculated an importance value for each descriptor using sum of the importance values of the descriptor output from the 500 random forest models constructed in the 100 times fivefold cross-validations. Molecular descriptors were then ranked by their importance values. The top-ranked descriptors were identified as the informative descriptors.

Applicability domain analysis

Applicability domain (AD) of a model is the structural space of chemicals that are used to train the model. Chemicals within the AD of a model are structurally similar as the training chemicals of the model and should be more accurately predicted. Therefore, AD analysis is important for assessing predictions of predictive models.[40–42] To conduct AD analysis on the developed random forest model, we first ranked descriptors using their importance values in the random forest model. Then, the top-ranked descriptors that count for 90% of the total importance values were used to define AD of the model, that is the hyperbox with boundary from the minimum to maximum for each of the selected descriptors. When all descriptors of a chemical are within the boundary, the chemical is inside the model's AD. However, if one or more of these descriptors of a chemical are out of the boundary, the chemical is outside the model's AD and the distance of the chemical to the AD is calculated. We calculated the distance of a compound to the AD of the random forest model using equation (9)

$$Distance = \sqrt{{d_1}^2 + {d_2}^2 + \ldots + {d_n}^2} \tag{9}$$

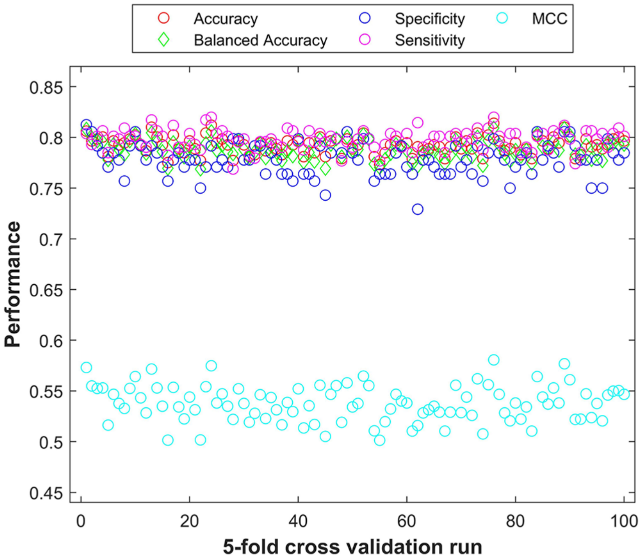

Figure 2. Performance of the fivefold cross-validations. The x-axis indicates 100 iterations of the fivefold cross-validations. The y-axis gives performance metrics values. Accuracy, balanced accuracy, specificity sensitivity, and MCC values are plotted as red circles, green diamonds, blue circles, magenta circles, and cyan circles, respectively.

where d_i ($i = 1, 2, \ldots, n$) is the distance of molecular descriptor i to the ith dimension of the AD. The distance was set to zero when a chemical is on the boundary of or inside the AD.

After defining AD for a random forest model, we first calculated distances to the AD of the random forest for all chemicals predicted by the model. We then examined the performance difference between the predictions inside and outside the AD. For predictions outside the AD, we further examined the relationship between the prediction performance and distances to the AD.

Results

Model performance evaluation

For the 100 iterations of fivefold cross-validations, the performance metrics were calculated and plotted (Figure 2). The average accuracy, sensitivity, specificity, balanced accuracy, and MCC of the random forest models in the 100 iterations of fivefold cross-validations were 0.783 ± 0.011, 0.785 ± 0.013, 0.778 ± 0.021, 0.782 ± 0.012, and 0.523 ± 0.023, respectively. Figure 2 illustrates that the constructed models show good performance and all performance metrics have small standard deviations among the 100 iterations of cross-validations, indicating the random forest models were not impacted much by the random divisions of the dataset into fivefolds.

The random forest was further evaluated using the testing set. The model was built using the whole training set. Comparing the actual SARS-CoV-2 main protease binding activity with the predictions from the random forest model on the testing set showed accuracy, sensitivity, specificity, balanced accuracy, and MCC of 0.514, 0.432, 0.723, 0.578, and 0.143, respectively, indicating that the developed model has good predictive power for predicting the main protease binding on unseen compounds.

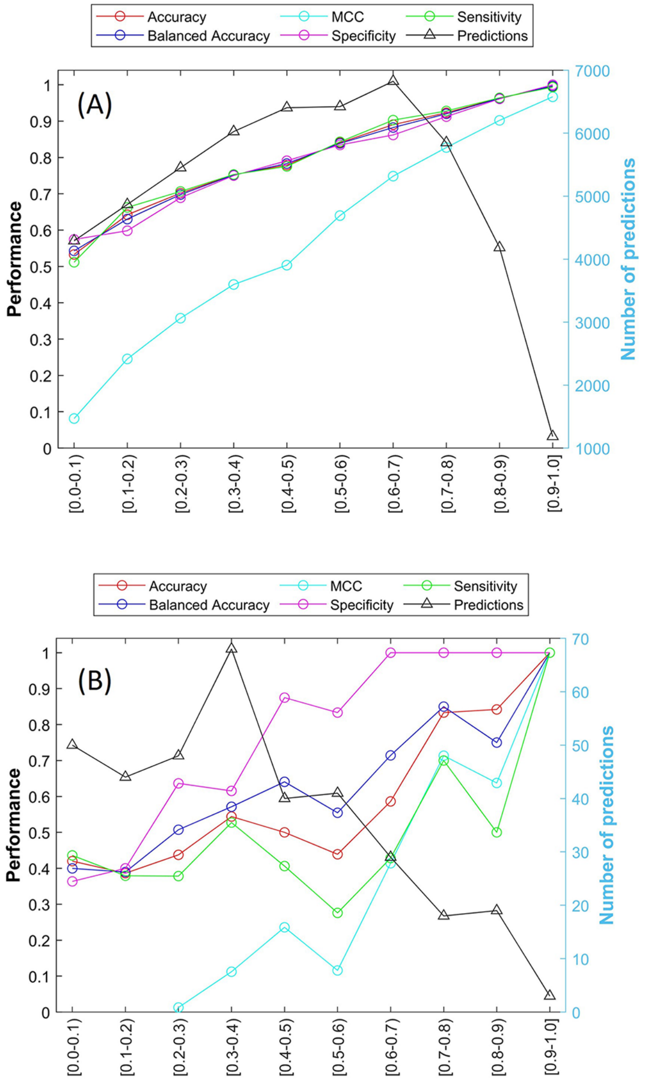

Figure 3. Prediction confidence analysis for the predictions from the 100 iterations of fivefold cross-validations (A) and for the predictions on the testing set (B). The x-axes show the 10 prediction confidence levels. The left y-axes give performance metrics values. The right y-axes depict the numbers of predictions. Accuracy, balance accuracy, MCC, specificity, and sensitivity are plotted as the red, blue, cyan, magenta, and green circles, respectively. The numbers of predictions are represented by black triangles.

Prediction confidence analysis

The prediction confidence analysis was conducted on the predictions from 100 iterations of fivefold cross-validations and predictions on the testing set. The performance metrics (accuracy, sensitivity, specificity, balanced accuracy, and MCC) and the number of predictions at different prediction confidence levels are plotted in Figure 3(A) for the fivefold cross-validations and in Figure 3(B) for predictions on the testing set. Clear trends were observed in the prediction confidence analysis results: the performance of predictions (accuracy, sensitivity, specificity, balanced accuracy, and MCC) improved when the prediction confidence level increased. Interestingly, more predictions are at higher confidence levels, but much less predictions are at very high confidence levels. The prediction confidence analysis results

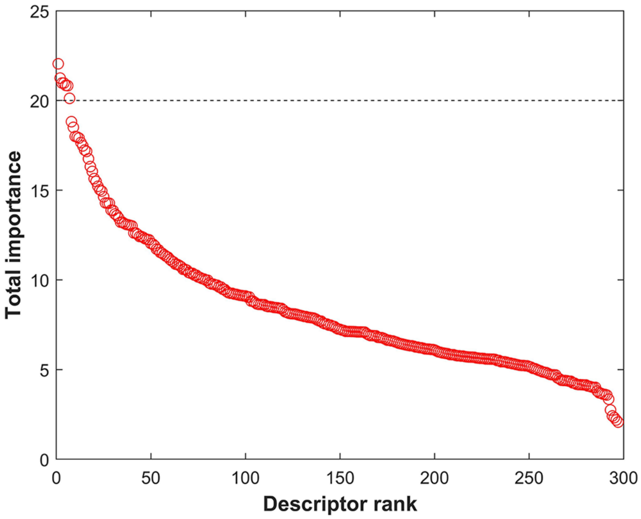

Figure 4. Importance of molecular descriptors. The x-axis shows the rank of the 297 molecular descriptors and the y-axis depicts importance value. Each red circle represents a descriptor. The dashed line indicates importance value 20.

demonstrated that the prediction confidence derived from the predictions from the random forest model provides an additional metric to help more appropriately use predictions from the developed random forest model in practical applications.

Informative descriptors identification

Molecular descriptors represented the chemical features and are important for the interpretation of the relationship between structures of chemicals and their potential main protease binding activity. To identify the informative molecular descriptors used in the random forest models, we first added up the importance values from the 500 models in the 100 times fivefold cross-validations for each molecular descriptor as the importance value of the descriptor. Then, the 297 molecular descriptors were ranked according to their importance values as plotted in Figure 4. Seven molecular descriptors with importance values higher than 20 were identified as the informative molecular descriptors to the random forest models and convey key structural information for SARS-CoV-2 main protease binding. These seven descriptors are D285 (structural information content order-3 index), D289 (complementary information content order-2 index), D290 (complementary information content order-3 index), D417 (topological structure autocorrelation length-3 weighted by atomic masses), D418 (topological structure autocorrelation length-4 weighted by atomic masses), D430 (topological structure autocorrelation length-8 weighted by atomic van der Waals volumes), and D470 (Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities).[34] Complementary information content indexes measure molecular shape–related properties, such as symmetry. Topological structure autocorrelations weighted by physical–chemical properties measure distributions of the properties along the topological structures. These identified informative molecular descriptors indicate that the physical–chemical properties and molecular shape and size of a compound are important for the compound to bind the main protease.

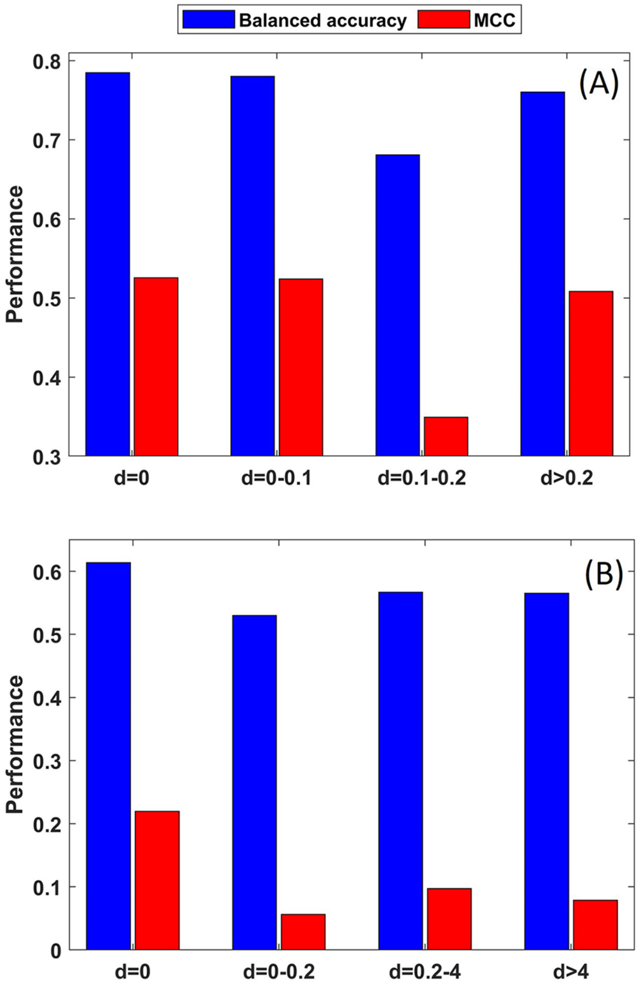

Figure 5. Prediction performance on compounds with different distances to the AD of the models in the cross-validations (A) and testing set prediction (B). The x-axes indicate distance to the AD. The y-axes give performance metrics values. Balanced accuracy and MCC are plotted as blue and red bars, respectively.

AD analysis

The AD of the constructed random forest represents the chemical space of the compounds that were used to train the model. Therefore, compounds outside the AD are less similar to the training chemicals and should be less accurately predicted by the model compared to the compounds inside the AD. The distances to the AD for the compounds predicted by the constructed random forest models in the fivefold cross-validations and testing set predictions were calculated. Prediction performance on compounds inside the AD and at different distance ranges from the AD was calculated and summarized in Figure 5. For both the cross-validations (Figure 5[A]) and the testing set predictions (Figure 5[B]), the compounds inside the AD had better prediction performance (greater balanced accuracy and MCC) than the compounds outside the AD, demonstrating AD analysis is useful to assess prediction reliability of the developed random forest model. Surprisingly, compounds with different distances to the AD did not show much difference in prediction performance. This observation may be explained by the algorithmic characteristic of random forest which is a decision tree–based machine learning algorithm. In a decision tree, a node is split using a cut-off value on a selected

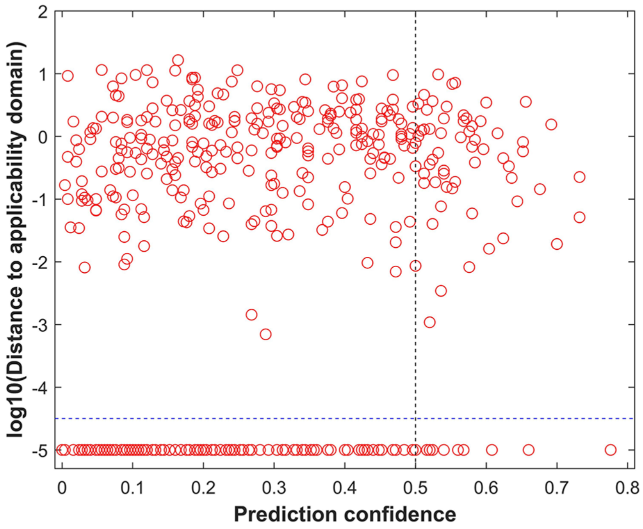

Figure 6. AD and prediction confidence analysis for FDA-approved drugs predicted as potential SARS-CoV-2 main protease binders. The x-axis shows the prediction confidence value. The y-axis indicates the log10 values of the distance to the AD. The drugs are represented by red circles. The value −5 is used to represent drugs inside the AD and the horizontal dashed line indicates the separation of drugs inside and outside the AD. The vertical dashed line represents the prediction confidence value at 0.5.

molecular descriptor and the magnitude of values of the select descriptor does not impact the split.

FDA-approved drugs

To identify potential candidates targeting the SARS-CoV-2 main protease for repurposing for COVID-19 treatment from FDA-approved drugs, the random forest model constructed using the whole training set was applied to predict the potential main protease binding activity of the 1188 FDA-approved drugs. Of the 1188 drugs, 472 were predicted as potential main protease binders by the random forest model. To identify repurposing candidates, we first calculated the prediction confidence values using the prediction probability values from the model and the distances to the AD of the model using the important descriptors determined in the model for the 472 drugs. The prediction confidence values and distances to the AD were plotted in Figure 6. As it can be seen in Figure 6, most of the 472 drugs are outside the AD of the model or have low prediction confidence (less than 0.5). Overall, 10 drugs were predicted to be SARS-CoV-2 binders with high prediction confidence (greater or equal to 0.5) and are inside the AD of the random forest model that were identified as potential candidates to repurpose for COVID-19 treatment targeting the main protease. Table 2 lists these 10 FDA-approved drugs as potential candidates for repurposing to COVID-19 treatment through binding the SARS-CoV-2 main protease.

Discussion

During the COVID-19 pandemic, many people were infected by SARS-CoV-2. Although several drugs have been approved by the FDA to treat COVID-19 patients, more effective drugs are needed. However, new drug development is a long and expensive process.[43–45] Hence, repurposing an existing drug to a new usage is a speedy and efficient approach. Previous studies have identified several targets on SARS-CoV-2 and the main protease has been extensively studied because there is no homology in humans and because of its conservative structure.[46–49] In this study, we explored the utilization of machine learning for identifying SARS-CoV-2 main protease binders as potential candidates for repurposing FDA-approved drugs for COVID-19 treatment. The compounds with SARS-CoV-2 main protease binding activity data were curated from the PDB and scientific literature for developing and evaluating machine learning models using a random forest algorithm for predicting SARS-CoV-2 main protease binding activity of compounds. Using the developed random forest model, we identified 10 drugs from 1188 FDA-approved drugs that were retrieved from the LTKB as potential candidates to repurpose for COVID-19 treatment.

The curated data have much more binders than non-binders, making both the training and testing data sets imbalanced. Imbalance in training data sets is a long-standing issue in machine learning as models trained on imbalanced data sets have bias to the majority class of samples when predicting new samples. Several approaches have been explored to solve the imbalance issue, but it remains a challenge in machine learning. In this study, a downsampling majority approach was applied to alleviate the impact of the imbalanced training data on the constructed random forest models. The random forest models constructed with downsampling the majority (binders) in the 100 iterations of fivefold cross-validations had very similar sensitivity and specificity (magenta and blue circles in Figure 2) with average values 0.785 and 0.778, respectively. Our results suggest that downsampling majority is a useful method for reducing impact of imbalanced training sets in machine learning.

A good machine learning model not only predicts a sample belonging to a class but also quantifies how likely the sample should belong to the predicted class, commonly termed as prediction confidence. The random forest models developed in this study predicted compounds as SARS-CoV-2 binders or non-binders using the probability value that measures the likelihood of the predicted compound to be a binder. To assess the usefulness of the quantified predictions from the random forest models, prediction confidence analysis was performed on the results from both fivefold cross-validations (Figure 3[A]) and testing set prediction (Figure 3[B]). Our results showed that prediction performance of predictions at a high confidence level is better than predictions at a low confidence level. Our findings suggest that prediction confidence from the developed random forest model provides a complementary metric to the main protease binding activity prediction in using the model in applications.

Different molecular descriptors made different contributions to the developed random forest model. The descriptors that contributed more to the model are informative to the model and should play important roles for a compound to bind the main protease. To find the possible relation of structural features of a compound and its main protease binding activity, we identified seven informative molecular descriptors and further found the molecular physical–chemical properties and shape and size of a compound play vital roles in the main protease binding of chemicals. These structural features may be useful in designing more potent SARS-CoV-2 main protease binders.

Table 2. Candidates for drug repurposing to treat COVID-19.

Drug	ATC code	DrugBank ID	Use
Ciclopirox	D01AE14	DB01188	Treat dermal infections
Procarbazine	L01XB01	DB01168	Treat Hodgkin's disease
Polystyrene sulfonate	V03AE01	DB01344	Treat hyperkalemia
Metyrapone	V04CD01	DB01011	Test hypothalamic-pituitary adrenocorticotropic hormone function
Ketamine	N01AX03	DB01221	Anesthetic agent
Benzylpenicilloyl polylysine	J01CR50	DB00895	Detect immunoglobulin E antibodies
Isocarboxazid	N06AF01	DB01247	Treat symptoms of depression
Edrophonium	V04CX07	DB01010	Test myasthenia gravis
Avanafil	G04BE10	DB06237	Treat erectile dysfunction
Afatinib	L01XE13	DB08916	Treat non–small cell lung cancer

ATC: Anatomical Therapeutic Chemical.

AD is important to evaluate the extrapolation capability of a machine learning model to samples that are not similar to the training samples. It is expected that samples similar to the training samples should be more accurately predicted by the machine learning model than samples that are not similar to the training samples. There are many methods to define AD for a machine learning model. Considering the characteristic of a random forest algorithm, the dissimilarity values (distances to the AD) of compounds outside the AD can be quantified. However, the impact of such quantification may not linearly relate to prediction performance. Our AD analysis results from the fivefold cross-validations (Figure 5[A]) and testing set predictions (Figure 5[B]) revealed that predictions inside the AD are more accurate than predictions outside the AD, but no large difference in performance was found for predictions at different distances to the AD. Our findings confirmed that the distances to the AD calculated in this study linearly correlate with prediction performance. A suitable distance calculation method to generate quantifications linearly related to prediction performance is needed and deserves for further investigation.

The goal of this study was to identify FDA-approved drugs that are SARS-CoV-2 main protease binders as candidates for drug repurposing. We identified 11 drugs that were predicted to be main protease binders with high-prediction confidence values ($\geqslant 0.5$) using the developed random forest model and are inside the AD of the model. Of these 11 approved drugs, suprofen was discontinued in the USA and the rest 10 drugs (Table 2) were suggested as potential candidates for repurposing to treat COVID-19. Ciclopirox is an antifungal drug that has antibacterial and anti-inflammatory properties. Procarbazine is an antineoplastic agent. Polystyrene sulfonate is a potassium-binding resin and an effective topical microbicide and spermicide. Metyrapone is an inhibitor of endogenous adrenal corticosteroid synthesis. Ketamine is a rapid-acting general anesthetic. Benzylpenicilloyl polylysine is penicilloyl bound to polylysine and the major determinant of penicillin metabolism. Isocarboxazid is a monoamine oxidase inhibitor. Suprofen is a non-steroidal anti-inflammatory analgesic and antipyretic and is no longer approved for use in the USA. Edrophonium is a cholinesterase inhibitor. Avanafil is a phosphodiesterase-5 inhibitor. Afatinib is a 4-anilinoquinazoline tyrosine kinase inhibitor and an antineoplastic agent. Binding free energy calculations showed that afatinib can act as a potential inhibitor of the main protease.[50] Additional experimental testing is required to confirm the potential main protease binding activity of these identified FDA-approved drugs.

Multiple machine learning models for SARS-CoV-2 main protease binding activity prediction have been developed and applied in drug design and discovery for COVID-19 treatment. Combining with other approaches, such as molecular docking and molecular dynamics simulation, regression models have been developed using machine learning algorithms for improving potency of candidate compounds.[51–53] Regression models are not able to distinguish between active and inactive compounds, and thus are not suitable for assisting in drug repurposing. Gomes *et al.*[54] used a pipeline consisting of molecular docking, metadynamics, and machine learning models for screening SARS-CoV-2 main protease inhibitors from compounds in DrugBank. Similar to our approach, they selected 74 SARS-CoV-2 main protease structures from PDB and generated 50 decoys as negatives for each positive. Compared to our model, their model trained on much fewer positives. Moreover, they did not report model performance but only reported the enrichment effect of their whole pipeline. Classification models for predicting SARS-CoV-2 main protease inhibitors have also been developed using a variety of machine learning algorithms, such as random forest, k-nearest neighbors, support vector machine, and Naïve Bayes.[55–58] Compared to our models, these models were developed based on datasets with much fewer SARS-CoV-2 main protease inhibitors. Moreover, annotations of positives and negatives used to generate the training sets are not consistent. For example, after collecting compounds with IC_{50} values from literature, Ferdous *et al.*[55] discarded compounds with IC_{50} of 1–10 µM and assigned compounds with $IC_{50} < 0.5\,\mu M$ as positives and compounds with $IC_{50} > 10\,\mu M$ as negatives, while Mekni *et al.*[56] excluded compounds with $IC_{50} > 98\,\mu M$ as negatives. It is worth noting that the SARS-CoV-2 main protease has multiple binding sites that can bind compounds with distinct structural features.[59] The positives in our training set are curated from PDB. They bind to Site 1[59] and are "seen" binding to the protease. Therefore, the quality of the training set is high. Another advantage of our model is that it provides prediction confidence and AD that are not available from previously published models, but are especially useful for identifying highly reliable candidates from FDA-approved drugs for repurposing to treat COVID-19.

In conclusion, we developed a random forest model for SARS-CoV-2 main protease binding activity prediction. The model was evaluated using fivefold cross-validations and a testing set. The model achieved good prediction performance. We conducted prediction confidence analysis and AD analysis for the model and demonstrated that prediction confidence and inside or outside the AD of a prediction are useful for assessing the quality of the prediction from the random forest model. Coupling prediction confidence and AD analysis, we identified 10 FDA-approved drugs as potential candidates for drug repurposing to COVID-19 treatment. Our results indicated that machine learning may be an efficient method for drug repurposing and thus accelerate drug development for the treatment of COVID-19 through targeting the SARS-CoV-2 main protease.

AUTHORS' CONTRIBUTIONS

JL, W Guo, KHK, and W Ge developed and evaluated the model. LX and ZL curated and processed data. JL and LX wrote the first draft of the article. HH and TAP revised the article and generated the final version.

ACKNOWLEDGEMENTS

This research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (ZL and KHK), administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the USA Department of Energy and the USA Food and Drug Administration.

DISCLAIMER

This article reflects the views of the authors and does not necessarily reflect those of the USA Food and Drug Administration.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID IDS

Jie Liu https://orcid.org/0000-0001-6988-5773
Wenjing Guo https://orcid.org/0000-0002-0814-6982
Md Kamrul Hasan Khan https://orcid.org/0009-0004-9835-5594
Huixiao Hong https://orcid.org/0000-0001-8087-3968

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Sharma A, Tiwari S, Deb MK, Marty JL. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. *Int J Antimicrob Agents* 2020;**56**:106054
2. Cohen J. Omicron sparks a vaccine strategy debate. *Science* 2021;**374**:1544–5
3. Hardenbrook NJ, Zhang P. A structural view of the SARS-CoV-2 virus and its assembly. *Curr Opin Virol* 2022;**52**:123–34
4. van de Leemput J, Han Z. Understanding individual SARS-CoV-2 proteins for targeted drug development against COVID-19. *Mol Cell Biol* 2021;**41**:e0018521
5. Lee M, Major M, Hong H. Distinct conformations of SARS-CoV-2 omicron spike protein and its interaction with ACE2 and antibody. *Int J Mol Sci* 2023;**24**:3774
6. Yang H, Rao Z. Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nat Rev Microbiol* 2021;**19**:685–700
7. Gao K, Wang R, Chen J, Tepe JJ, Huang F, Wei GW. Perspectives on SARS-CoV-2 main protease inhibitors. *J Med Chem* 2021;**64**:16922–55
8. Lamb YN. Nirmatrelvir plus ritonavir: first approval. *Drugs* 2022;**82**:585–91
9. Mengist HM, Dilnessa T, Jin T. Structural basis of potential inhibitors targeting SARS-CoV-2 main protease. *Front Chem* 2021;**9**:622898
10. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung SH. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. *J Med Chem* 2016;**59**:6595–628
11. 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 LW, 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**:289–93
12. La Monica G, Bono A, Lauria A, Martorana A. Targeting SARS-CoV-2 main protease for treatment of COVID-19: covalent inhibitors structure-activity relationship insights and evolution perspectives. *J Med Chem* 2022;**65**:12500–34
13. Hu Q, Xiong Y, Zhu GH, Zhang YN, Zhang YW, Huang P, Ge GB. The SARS-CoV-2 main protease (M^{pro}): structure, function, and emerging therapies for COVID-19. *MedComm* **2020** 2022;**3**:e151
14. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**:41–58
15. Luo H, Mattes W, Mendrick DL, Hong H. Molecular docking for identification of potential targets for drug repurposing. *Curr Top Med Chem* 2016;**16**:3636–45
16. Ye H, Wei J, Tang KL, Feuers R, Hong HX. Drug repositioning through network pharmacology. *Current Topics in Medicinal Chemistry* 2016;**16**:3646–56
17. Oerlemans R, Ruiz-Moreno AJ, Cong Y, Dinesh Kumar N, Velasco-Velazquez MA, Neochoritis CG, Smith J, Reggiori F, Groves MR, Domling A. Repurposing the HCV NS3-4A protease drug boceprevir as COVID-19 therapeutics. *RSC Med Chem* 2020;**12**:370–9
18. Pathak N, Chen YT, Hsu YC, Hsu NY, Kuo CJ, Tsai HP, Kang JJ, Huang CH, Chang SY, Chang YH, Liang PH, Yang JM. Uncovering flexible active site conformations of SARS-CoV-2 3CL proteases through protease pharmacophore clusters and COVID-19 drug repurposing. *ACS Nano* 2021;**15**:857–72
19. Ambrosio FA, Costa G, Romeo I, Esposito F, Alkhatib M, Salpini R, Svicher V, Corona A, Malune P, Tramontano E, Ceccherini-Silberstein F, Alcaro S, Artese A. Targeting SARS-CoV-2 main protease: a successful story guided by an in silico drug repurposing approach. *J Chem Inf Model* 2023;**63**:3601–13
20. Prasad K, Kumar V. Artificial intelligence-driven drug repurposing and structural biology for SARS-CoV-2. *Curr Res Pharmacol Drug Discov* 2021;**2**:100042
21. Williams AH, Zhan CG. Staying ahead of the game: how SARS-CoV-2 has accelerated the application of machine learning in pandemic management. *BioDrugs* 2023;**37**:649–74
22. Rajput A, Bhamare KT, Thakur A, Kumar M. Anti-biofilm: machine learning assisted prediction of IC(50) activity of chemicals against biofilms of microbes causing antimicrobial resistance and implications in drug repurposing. *J Mol Biol* 2023;**435**:168115
23. Feng H, Jiang J, Wei GW. Machine-learning repurposing of DrugBank compounds for opioid use disorder. *Comput Biol Med* 2023;**160**:106921
24. Azevedo P, Pecanha BRB, Flores-Junior LAP, Alves TF, Dias LRS, Muri EMF, Lima C. In silico drug repurposing by combining machine

learning classification model and molecular dynamics to identify a potential OGT inhibitor. *J Biomol Struct Dyn*. Epub ahead of print 13 April 2023. DOI: 10.1080/07391102.2023.2199868

25. Hong H, Tong W, Xie Q, Fang H, Perkins R. An in silico ensemble method for lead discovery: decision forest. *SAR QSAR Environ Res* 2005;**16**:339–47
26. Luo H, Ye H, Ng HW, Shi L, Tong W, Mendrick DL, Hong H. Machine learning methods for predicting HLA-peptide binding activity. *Bioinform Biol Insights* 2015;**9**:21–9
27. Huang Y, Li X, Xu S, Zheng H, Zhang L, Chen J, Hong H, Kusko R, Li R. Quantitative structure-activity relationship models for predicting inflammatory potential of metal oxide nanoparticles. *Environ Health Perspect* 2020;**128**:67010
28. Wang Z, Chen J, Hong H. Developing QSAR models with defined applicability domains on PPARgamma binding affinity using large data sets and machine learning algorithms. *Environ Sci Technol* 2021;**55**:6857–66
29. Ji Z, Guo W, Wood EL, Liu J, Sakkiah S, Xu X, Patterson TA, Hong H. Machine learning models for predicting cytotoxicity of nanomaterials. *Chem Res Toxicol* 2022;**35**:125–39
30. Chen M, Zhang J, Wang Y, Liu Z, Kelly R, Zhou G, Fang H, Borlak J, Tong W. The liver toxicity knowledge base: a systems approach to a complex end point. *Clin Pharmacol Ther* 2013;**93**:409–12
31. Qiao J, Li YS, Zeng R, Liu FL, Luo RH, Huang C, Wang YF, Zhang J, Quan B, Shen C, Mao X, Liu X, Sun W, Yang W, Ni X, Wang K, Xu L, Duan ZL, Zou QC, Zhang HL, Qu W, Long YH, Li MH, Yang RC, Liu X, You J, Zhou Y, Yao R, Li WP, Liu JM, Chen P, Liu Y, Lin GF, Yang X, Zou J, Li L, Hu Y, Lu GW, Li WM, Wei YQ, Zheng YT, Lei J, Yang S. SARS-CoV-2 M(pro) inhibitors with antiviral activity in a transgenic mouse model. *Science* 2021;**371**:1374–8
32. Luttens A, Gullberg H, Abdurakhmanov E, Vo DD, Akaberi D, Talibov VO, Nekhotiaeva N, Vangeel L, De Jonghe S, Jochmans D, Krambrich J, Tas A, Lundgren B, Gravenfors Y, Craig AJ, Atilaw Y, Sandstrom A, Moodie LWK, Lundkvist A, van Hemert MJ, Neyts J, Lennerstrand J, Kihlberg J, Sandberg K, Danielson UH, Carlsson J. Ultralarge Virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J Am Chem Soc* 2022;**144**: 2905–20
33. Resnick SJ, Iketani S, Hong SJ, Zask A, Liu H, Kim S, Melore S, Nair MS, Huang Y, Tay NES, Rovis T, Yang HW, Stockwell BR, Ho DD, Chavez A. A simplified cell-based assay to identify coronavirus 3CL protease inhibitors. 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7457602/
34. Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* 2008;**48**:1337–44
35. Hong H, Thakkar S, Chen M, Tong W. Development of decision forest models for prediction of drug-induced liver injury in humans using a large set of FDA-approved drugs. *Sci Rep* 2017;**7**:17311
36. Godden JW, Stahura FL, Bajorath J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J Chem Inf Comput Sci* 2000;**40**:796–800
37. Liu J, Guo W, Dong F, Aungst J, Fitzpatrick S, Patterson TA, Hong H. Machine learning models for rat multigeneration reproductive toxicity prediction. *Front Pharmacol* 2022;**13**:1018226
38. Breiman L. Random Forests. *Mach Learning* 2001;**45**:5–32
39. Hong H, Rua D, Sakkiah S, Selvaraj C, Ge W, Tong W. Consensus modeling for prediction of estrogenic activity of ingredients commonly used in sunscreen products. *Int J Environ Res Public Health* 2016;**13**:958
40. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model* 2008;**26**:1315–26
41. Klingspohn W, Mathea M, Ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. *J Cheminform* 2017;**9**:44
42. Kar S, Roy K, Leszczynski J. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. *Methods Mol Biol* 2018;**1800**:141–69
43. Simoens S, Huys I. R&D costs of new medicines: a landscape analysis. *Front Med* 2021;**8**:760762
44. Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How much does it cost to research and develop a new drug? A systematic review and assessment. *Pharmacoeconomics* 2021;**39**:1243–69
45. Kumar P, Bhardwaj T, Kumar A, Gehi BR, Kapuganti SK, Garg N, Nath G, Giri R. Reprofiling of approved drugs against SARS-CoV-2 main protease: an in-silico study. *J Biomol Struct Dyn* 2022;**40**:3170–84
46. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, Wang Q, Xu Y, Li M, Li X, Zheng M, Chen L, Li H. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* 2020;**10**:766–88
47. Ma CL, Xia ZL, Sacco MD, Hu YM, Townsend JA, Meng XZ, Choza J, Tan HZ, Jang J, Gongora MV, Zhang XJ, Zhang FS, Xiang Y, Marty MT, Chen Y, Wang J. Discovery of di- and trihaloacetamides as covalent SARS-CoV-2 main protease inhibitors with high target specificity. *J Am Chem Soc* 2021;**143**:20697–709
48. Durojaiye AB, Clarke JRD, Stamatiades GA, Wang C. Repurposing cefuroxime for treatment of COVID-19: a scoping review of in silico studies. *J Biomol Struct Dyn* 2021;**39**:4547–54
49. Huff S, Kummetha IR, Tiwari SK, Huante MB, Clark AE, Wang SB, Bray W, Smith D, Carlin AF, Endsley M, Rana TM. Discovery and mechanism of SARS-CoV-2 main protease inhibitors. *J Med Chem* 2022;**65**:2866–79
50. Arun KG, Sharanya CS, Abhithaj J, Francis D, Sadasivan C. Drug repurposing against SARS-CoV-2 using E-pharmacophore based virtual screening, molecular docking and molecular dynamics with main protease as the target. *J Biomol Struct Dyn* 2021;**39**:4647–58
51. Saar KL, McCorkindale W, Fearon D, Boby M, Barr H, Ben-Shmuel A, Consortium CM, London N, von Delft F, Chodera JD, Lee AA. Turning high-throughput structural biology into predictive inhibitor design. *Proc Natl Acad Sci U S A* 2023;**120**:e2214168120
52. Nguyen TH, Tam NM, Tuan MV, Zhan P, Vu VV, Quang DT, Ngo ST. Searching for potential inhibitors of SARS-COV-2 main protease using supervised learning and perturbation calculations. *Chem Phys* 2023;**564**:111709
53. Janairo GIB, Yu DEC, Janairo JIB. A machine learning regression model for the screening and design of potential SARS-CoV-2 protease inhibitors. *Netw Model Anal Health Inform Bioinform* 2021;**10**:51
54. Gomes IS, Santana CA, Marcolino LS, Lima LHF, Melo-Minardi RC, Dias RS, de Paula SO, Silveira SA. Computational prediction of potential inhibitors for SARS-COV-2 main protease based on machine learning, docking, MM-PBSA calculations, and metadynamics. *PLoS ONE* 2022;**17**:e0267471
55. Ferdous N, Reza MN, Hossain MU, Mahmud S, Napis S, Chowdhury K, Mohiuddin AKM. Mpropred: a machine learning (ML) driven Web-App for bioactivity prediction of SARS-CoV-2 main protease (Mpro) antagonists. *PLoS ONE* 2023;**18**:e0287179
56. Mekni N, Coronnello C, Langer T, Rosa M, Perricone U. Support vector machine as a supervised learning for the prioritization of novel potential SARS-CoV-2 main protease inhibitors. *Int J Mol Sci* 2021;**22**:7714
57. Samad A, Ajmal A, Mahmood A, Khurshid B, Li P, Jan SM, Rehman AU, He P, Abdalla AN, Umair M, Hu J, Wadood A. Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation. *Front Mol Biosci* 2023;**10**:1060076
58. Verma AK, Aggarwal R. Repurposing potential of FDA-approved and investigational drugs for COVID-19 targeting SARS-CoV-2 spike and main protease and validation by machine learning algorithm. *Chem Biol Drug Des* 2021;**97**:836–53
59. Xu L, Chen R, Liu J, Patterson TA, Hong H. Analyzing 3D structures of the SARS-CoV-2 main protease reveals structural features of ligand binding for COVID-19 drug discovery. *Drug Discov Today* 2023;**28**:103727

(Received August 24, 2023, Accepted September 26, 2023)