ASCoVPred: a machine learning-based platform for quantitative prediction of anti-SARS-CoV-2 activity and human cell toxicity of molecules

Arun Sharma

International Centre for Genetic Engineering and Biotechnology

Neeraj Chaturvedi

International Centre for Genetic Engineering and Biotechnology

Dinesh Gupta (✉ dinesh@icgeb.res.in)

International Centre for Genetic Engineering and Biotechnology <https://orcid.org/0000-0002-7548-8835>

Article

Keywords: COVID-19, QSAR models, ASCoVPred

Posted Date: October 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-967196/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

There is an urgent need to accelerate the discovery of effective drugs for COVID-19. We have developed machine learning models for rapid discovery of molecules potentially inhibitory to SARS-CoV-2 and negligible or no human cell toxicity. The machine learning (ML) QSAR models were trained and optimized with features (descriptors and fingerprints) of the experimentally validated SARS-CoV-2 inhibitory compounds. Several molecular descriptors and fingerprints were calculated to select the decisive ones for the training and evaluation of thousands of ML models. The best-optimized models are deployed as ASCoVPred webserver and standalone software, that provides easy and free access to the models. The feature selection for selecting the best descriptors for ML models training helped identify a set of decisive descriptors and fingerprints that correlate positively or negatively with the anti-SARS-CoV-2 activity and toxicity of the compounds. Systematic prediction and optimization of compounds with the help of ASCoVPred can facilitate the discovery of novel anti-SARS-CoV-2 compounds. The ASCoVPred web server and standalone software are freely available at <http://14.139.62.220/ascovpred/>.

Introduction

According to the World Health Organization (WHO) online dashboard¹, as of October 13, 2021, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected more than 238.22 million people and is responsible for >4.85 million deaths globally. Despite having high morbidity and mortality rates, the world is limited to a few medicines to treat Coronavirus disease or COVID-19, an infectious disease caused by SARS-CoV-2. Although several vaccines have been developed to reduce the disease burden, effective antivirals are still required to treat infected and hospitalized patients. We are also not sure of another impending wave of viral infections. An antiviral drug Veklury (remdesivir), is the only U.S. Food and Drug Administration (FDA) approved available for adult and certain pediatric COVID-19 patients. Also, several drugs are in different stages of trials as the quest for an effective anti-SARS-CoV-2 continues (e.g., molnupiravir)². FDA has also issued emergency use authorization (EUA) for other types of treatments such as monoclonal antibody-based treatments³. However, treatment options, such as the development of antivirals, immunomodulators, neutralizing antibody therapies, cell therapy, etc., are ongoing and yet to pass through different clinical trials⁴. Thus, efforts with multiple approaches are continuing to discover novel anti-SARS-CoV-2 drugs as quickly as possible. However, *in vitro* discovery of novel inhibitors is tedious, labor-intensive, time-consuming, apart from being costly exercise. However, computational predictions facilitate *in vitro* discovery by shortlisting the most effective chemical entities, saving time and cost.

Recently, several drug targets have been explored to design novel drug molecules against SARS-CoV-2. Few of the identified drug targets are essential for the viral entry into target cells and survival into host cells, while others play an indispensable role in viral growth. For example, the viral surface Spike protein (S-protein) is essential for the attachment of the virus to human angiotensin-converting enzyme 2 (ACE-2), present on the target human cells^{5,6}. Thus, S-protein plays a vital role in the SARS-CoV-2 entry into human target cells is an attractive drug target. Furthermore, a human serine protease named transmembrane serine protease 2 (TMPRSS2) also plays an essential role in the entry of SARS-CoV-2 into human target cells through S-protein priming. Hence, inhibiting the TMPRSS2 activity may lower the risk of disease progression⁶⁻⁸. Thus, as S-protein and TMPRSS2 play a vital role in SARS-CoV-2 entry into human target cells, many efforts are going on to develop drugs against these viral targets⁹. There are reports of computational approaches, including molecular docking and machine learning (ML)-based classification algorithm development, to identify suitable anti-SARS-CoV2 inhibitors¹⁰⁻¹³. Moreover, drug repurposing approaches have been explored to quickly identify the approved drug molecules which may bind to SARS-CoV-2 drug targets¹⁴⁻¹⁶. Most of such computational studies have used docking or molecular dynamics, and combinations of the two, to explore FDA-approved drugs as potential SARS-CoV-2 inhibitors^{10,14,17-21}. However, systematic attempts to develop ML-based models through quantitative structure-activity relationship (QSAR) approaches are lacking, which motivated us to develop ML-based QSAR models that could rapidly screen large chemical libraries to identify anti-SARS-CoV-2 compounds.

In the present study, we have explored the publicly available National Centre for Advancing Translational Sciences (NCATS) chemical assays data^{9,22} to develop ML-based QSAR models to predict anti-SARS-CoV-2 activity (primarily related to preventing viral entry into the target human cells) and toxicity of compounds (against HEK293 and Vero E6 cells). The best models are deployed on an online webserver and standalone software "ASCoVPred" that predicts the activities of user-submitted compounds in an automated way. In the future, we plan to update the webserver and standalone software with additional ML-based prediction models with improved prediction performance.

Results

Data mining

The data from the nine assays of high-throughput screening (HTS) were passed through rigorous pre-processing steps. As mentioned in the materials and methods section, only the compounds with SMILES structures and maximum response values were used in the study, while others were discarded. Furthermore, after the descriptors and fingerprints (FPs) calculations, only the compounds with the calculated descriptor values were selected for dataset preparation.

The duplicate records were removed from the training datasets to eliminate the overfitting of models during training steps. The duplicate records correspond to the compounds with identical descriptors/FPs values and maximum response values. For the details of records after the filtering steps, see Supplementary Tables 1-9. The compounds listed in the supplementary files were used to train and evaluate the machine learning (ML)-based QSAR models. Moreover, model building and training of models was performed using only the datasets with the ratio of the number of compounds to the number of descriptors at least five, as suggested for best practices for QSAR studies²³. The number of molecules used for the training and evaluation of best prediction models and ML-based models' performances is described in the subsequent sections.

Molecules used for training and evaluation of best models

The assay-wise summary of the molecules used for models' training and evaluation is shown in Table 1. For further details of the training dataset molecules, refer to Supplementary files 1-9.

Table 1
The number of molecules used for training and evaluation of best prediction models.

Assay ID	Assay name	Number of unique molecules	Number of training dataset molecules	Number of external validation dataset molecules
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	3370	2696	674
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	3235	2588	647
6	ACE2 enzymatic activity	3376	2701	675
8	TMPRSS2 enzymatic activity	5144	4115	1029
9	3CL enzymatic activity	11007	8806	2201
14	SARS-CoV-2 cytopathic effect (CPE)	9909	7927	1982
15	SARS-CoV-2 cytopathic effect (host tox counterscreen)	9080	7264	1816
20	HEK293 cell line toxicity	9694	7755	1939
21	Human fibroblast toxicity	4491	3593	898

Performances of best prediction models

The performance measures of the best machine learning (ML) models, trained and evaluated with nine different assays' experimental data, are given in Table 2. The regression techniques (along with parameter optimization) available in the Weka (v3.8.2)²⁴ package were used to train and evaluate multiple ML models (for details, please refer to methods). However, the best results are achieved with RandomForest only, either as a main "Classifier" or a hyper-parameter for another "Classifier." Thus, five out of the best nine ML models (based on assay IDs 1, 2, 8, 15 and 20) performed best. Hence, we deployed them at the ASCoVPred webserver and standalone software. The five best ML models achieved Pearson Correlation Coefficient value (R; between actual and predicted maximum response values), greater than 0.50 for training (in five-fold cross-validation), as well as external validation dataset molecules. The ML models which did not perform well are based on the four assays, namely the "ACE2 enzymatic activity" assay (Assay ID-6), the "3CL enzymatic activity" assay (Assay ID-9), the "SARS-CoV-2 cytopathic effect (CPE)" assay (Assay ID-14) and the "Human fibroblast toxicity" assay (Assay ID-21). Due to poor performance of ML models with these four assays, the models were rejected.

Table 2
Results of the evaluation for best prediction models.

Performance evaluation of best models with training dataset compounds (five-fold cross-validation)								Performance evaluation of best models with external validation dataset compounds		
Assay ID	Assay name	Descriptors or fingerprints type	Number of input features	R	MAE	RMSE	WEKA technique used for training the model with training dataset of compounds	R	MAE	RMSE
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	SubFPC*	307	0.67	14.96	21.22	weka.classifiers.meta.AdditiveRegression (RandomForest)	0.66	17.41	24.12
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	1D & 2D	50	0.72	13.51	18.83	weka.classifiers.meta.AdditiveRegression (RandomForest)	0.74	14.44	20.12
6	ACE2 enzymatic activity	ExtendedFingerprinter#	28	0.32	14.31	37.09	weka.classifiers.trees.RandomForest	0.57	16.68	44.12
8	TMPRSS2 enzymatic activity	SubFPC*	307	0.52	11.12	36.47	weka.classifiers.trees.RandomForest	0.73	14.99	44.12
9	3CL enzymatic activity	SubFPC*#	307	0.40	5.90	11.31	weka.classifiers.meta.RandomCommittee (RandomForest)	0.45	5.11	9.12
14	SARS-CoV-2 cytopathic effect (CPE)	1D & 2D#	1444	0.50	7.43	14.59	weka.classifiers.meta.RandomSubSpace (RandomForest)	0.43	8.23	14.12
15	SARS-CoV-2 cytopathic effect (host tox counterscreen)	1D & 2D	1444	0.66	13.12	21.13	weka.classifiers.meta.AdditiveRegression (RandomForest)	0.65	14.11	21.12
20	HEK293 cell line toxicity	SubFPC*	307	0.66	26.68	34.16	weka.classifiers.meta.RandomCommittee (RandomForest)	0.68	26.41	35.12
21	Human fibroblast toxicity	1D & 2D#	45	0.43	12.61	19.69	weka.classifiers.meta.RandomCommittee (RandomForest)	0.51	11.99	18.12

*SubstructureFingerprintCount; #Not deployed on ASCoVPred webserver

Profile for the desired multi-target molecule

For any molecule to be an ideal multi-target hit (against SARS-CoV-2 and human TMPRSS2), it must be predicted as highly active by AlphaLISA assay (Assay ID-1), least active by TruHit counterscreen assay (Assay ID-2), highly active by TMPRSS2 enzymatic activity assay (Assay ID-8), least active by "SARS-CoV-2 cytopathic effect (host tox counterscreen)" assay (Assay ID-15) and HEK293 cell-line assay (Assay ID-20). The desired activity prediction profile for an ideal multi-target molecule is shown in Table 3. In addition, the scales used to assign "Activity Class", are given on the ASCoVPred webserver's prediction results page.

Table 3
The desired activity prediction profile for an ideal multi-target hit molecule.

Assay ID	Assay Name	Target Category	Activity Class
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	Viral entry	High
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Counterscreen	Low
8	TMPRSS2 enzymatic activity	Viral entry	High
15	SARS-CoV-2 cytopathic effect (host tox counterscreen)	Counterscreen	Low
20	HEK293 cell line toxicity	Counterscreen	Low

Descriptors or fingerprints associated with activities of molecules

For a QSAR study, it is vital to identify the putative descriptors or fingerprints associated with the biological activity of the molecules. Therefore, in the present study, Pearson Correlation Coefficient (R) based-analysis of molecules was performed to understand the types of descriptors or fingerprints present in highly active (most potent) versus least active molecules. Supplementary files 10-19 contain the results of these analyses.

For AlphaLISA assay (Assay ID-1), all the top 10 negatively correlated FPs achieved an R-value of greater than -0.1 (in absolute terms) with the highest value of -0.37 for FP "SubFPC307" (Supplementary file10.xlsx, Table 1). However, all the top 10 positively correlated FPs achieved R-value less than 0.1 (Supplementary file11.xlsx, Table 1). Furthermore, the analysis with the top 10 negatively correlated FPs showed higher average values for all 10 FPs in highly active compounds than least active compounds (Supplementary file10.xlsx, Table 2-4 and Figure 1). Thus, the compounds with strong potential (with maximum response value approaching -100) to disrupt Spike-ACE2 interaction possess higher average values for fingerprints "SubFPC307", "SubFPC274", "SubFPC287", "SubFPC137" and "SubFPC169" than the compounds which possess weak potential (Supplementary file10.xlsx, Figure 1).

In the case of TruHit Counterscreen assay (Assay ID-2), the top 10 negatively correlated FPs achieved R values within the range of -0.16 to -0.48 (Supplementary file12.xlsx, Table 1). However, the values ranged from 0.07 to 0.23 for the top 10 positively correlated FPs (Supplementary file13.xlsx, Table 1). The negatively correlated, top 6 FPs, namely "C2SP2", "C3SP2", "CrippenLogP", "SP-7", "MLFER_E" and "nF10Ring", showed significant differences in average values amongst highly active and least active compounds (Supplementary file12.xlsx, Tables 2-4, Figure 1). Whereas, in the case of positively correlated FPs, significant differences of average values were observed for FPs, namely "GATS1p", "MAT2Sc", "nAcid" and "MAT2Se" (Supplementary file13.xlsx, Tables 2-4, Figure 1).

For the TMPRSS2 enzymatic activity assay (Assay ID-8), the negatively correlated FPs achieved lower R-value (-0.03 to -0.05) between FP values and compounds maximum response values than positively correlated FPs (0.03 to 0.16) (Supplementary file14.xlsx, Table 1 and Supplementary file15.xlsx, Table 1). Further analysis of positively correlated FPs data showed significant differences between the average values amongst highly active and least active compounds, for some FPs, namely "SubFPC28", "SubFPC181", "SubFPC184" and "SubFPC275" (Supplementary file15.xlsx, Tables2-4, Figure 1).

For the SARS-CoV-2 cytopathic effect (host tox counterscreen) (Assay ID-15), the negatively correlated FPs achieved higher absolute R-value (-0.29 to -0.26) between FP values and compounds maximum response values than positively correlated FPs (0.14 to 0.22) (Supplementary file16.xlsx, Table 1 and Supplementary file17.xlsx, Table 1). Further analysis of negatively correlated FPs data showed significant differences between the average values amongst highly active and least active compounds, for some FPs, namely "nAromBond", "nAromAtom", "nBondsM" and "SwHBA" (Supplementary file16.xlsx, Tables2-4, Figure 1).

The analysis with human cell-line toxicity assay named "HEK293 cell line toxicity" (Assay ID-20) provided top 10 negatively correlated and top 10 positively correlated FPs with R values within the ranges -0.11 to -0.36 (Supplementary file18.xlsx, Table 1) and 0.14 to 0.23 (Supplementary file19.xlsx, Table 1), respectively. Furthermore, all the top 10 negatively (Supplementary file18.xlsx, Tables2-4, Figure 1) and positively (Supplementary file19.xlsx, Tables2-4, Figure 1) correlated FPs showed considerable differences in their average FP values amongst highly active and least active compounds.

Utility of ASCoVPred

The *in silico* designing of inhibitors is an essential task for any drug discovery strategy. Also, the real-life utility of a computational study lies in its practical day-to-day usage. Thus, to enhance the utility of the present study and best use the developed ML-based prediction models, a webserver named "ASCoVPred" has been developed. Moreover, a standalone version of the ASCoVPred has also been developed for the high-throughput screening (HTS) of compounds. Currently, the standalone version of ASCoVPred has been tested for the Linux (Ubuntu 18.04 LTS) and Mac operating system (OS) by supplying thousands of compounds as input (at a time) in SMILES format. Thus, the standalone software can be run in parallel mode to screen multiple chemical libraries simultaneously. Whereas the webserver version of ASCoVPred provides a user-friendly web interface to the best prediction models. Users interested in anti-SARS-CoV-2 drug-discovery projects can use the webserver to rapidly discover or filter a single compound within a minute. Hence, the webserver possesses the potential to speed up the SARS-CoV-2 inhibitor discovery. The current version of the ASCoVPred webserver is freely accessible from <http://14.139.62.220/ascovpred/>. The server and standalone software provide various modules to perform specialized tasks such as virtual screening of large chemical libraries, designing potent inhibitors, improving the biological activity of molecules, etc. Figure 1 provides the screenshots to display the functioning of ASCoVPred webserver and standalone software. For further details, refer to the online help on ASCoVPred webserver.

Discussion

The discovery of novel hits is a tedious process in drug discovery that requires several years of massive investments of time, manpower, resources, and money. However, using computational drug discovery resources can minimize the usage of valuable resources, improve efficiency to accelerate the drug discovery process. To facilitate drug discovery for COVID-19, the present study focuses on developing a computational resource or webserver "ASCoVPred" and standalone software that helps design efficacious and safe to use inhibitors against SARS-CoV-2. The valuable experimental activity data of compounds from HTS assays (downloaded from the NCATS portal) were utilized for training and evaluating ML-based prediction models. Finally, the best performing models are deployed as a publicly available prediction webserver, ASCoVPred and standalone software (for Linux and Mac OS).

The analysis performed with the compounds (used in the training of best models) has provided some interesting descriptors and FPs with potential roles in the determination of activities of compounds. For example, in AlphaLISA assay (Assay ID-1), the top 10 negatively correlated FPs (with the maximum response value of compounds) includes a count of chiral center specified (SubFPC307), aromatic functional groups (SubFPC274), conjugated double bonds (SubFPC287), vinylogous esters (SubFPC137), phenols (SubFPC169), arylchlorides (SubFPC171), vinylogous acids (SubFPC136), tertiary mixed amines (SubFPC33), etc. Furthermore, the highly active compounds possess higher average values for these FPs than the least active compounds (Supplementary file10.xlsx, Figure 1). Therefore, a compound with higher values corresponding to these FPs might act as a suitable inhibitor (highly active) to disrupt the Spike-ACE2 interaction, thus preventing viral entry into the human host cells.

The TruHit counterscreen assay (Assay ID-2) ensures the identification of false positives from the AlphaLISA assay, thus being helpful in the identification of non-specific inhibitors. For any compound to be a true positive, it must be detected as inactive or least active in the TruHit counterscreen assay^{9,13}. From the analysis of training dataset compounds, the top 10 negatively correlated 1D and 2D descriptors (with the maximum response value of compounds) possess

higher average values for descriptors, namely, doubly bound carbon bound to two other carbons (C2SP2), doubly bound carbon bound to three other carbons (C3SP2), Crippen's LogP (CrippenLogP), simple path, order 7 (SP-7), excessive molar refraction (MLFER_E), number of 10-membered fused rings (nF10Ring), the coefficient sum of the last eigenvector from topological distance matrix (VE1_D), maximum atom-type E-State: =C< (maxdssC), maximum atom-type E-State: =CH- (maxdssCH) and the coefficient sum of the last eigenvector from Barysz matrix / weighted by first ionization potential (VE1_Dzi), in case of highly active compounds than least active compounds (Supplementary file12.xlsx, Figure 1). Therefore, for any compound to be least active or inactive in the TruHit counterscreen assay, the ideal values for these 1D and 2D descriptors must be low. Whereas, in the case of top positively correlated 1D and 2D descriptors (with the maximum response value of compounds), a few descriptor types such as MATS2c, nAcid and MATS2e, showed significant differences of average values amongst highly active and least active molecules (Supplementary file13.xlsx, Figure 1). Therefore, for any compound to be least active or inactive in TruHit counterscreen assay, the compound may possess higher values for the descriptors, namely, Moran autocorrelation - lag 2 / weighted by charges (MATS2c), number of acidic groups (nAcid) and Moran autocorrelation - lag 2 / weighted by Sanderson electronegativities (MATS2e), than the highly active compounds.

The TMPRSS2 enzymatic activity (Assay ID-8) is specific to identification of compounds with the potential to inhibit the activity of the human TMPRSS2 enzyme, thus preventing the viral entry into the human cells⁹. Although the trained models with this assay data didn't show high performance in five-fold cross-validation (R = 0.52), better performance is observed with the external validation dataset compounds (R = 0.73, Table 2). These kinds of results point towards the reliability and good performance of a model in real-life practices. Further analysis with training dataset compounds provided top 10 negatively correlated FPs having very poor correlation (R-value) between FP values and the maximum response value of the compounds (Supplementary file14.xlsx, Table 1). However, a few of the top 10 positively correlated FPs, namely, primary arom amine (SubFPC28), hetero N nonbasic (SubFPC181), heteroaromatic (SubFPC184) and heterocyclic (SubFPC275), achieved slightly higher values of R (Supplementary file15.xlsx, Table 1). Furthermore, these FPs possess lower average values for the highly active compounds than the least active compounds (Supplementary file15.xlsx, Figure 1). Therefore, for any compound to be highly active in TMPRSS2 enzymatic assay, the values for these FPs should be ideally low.

For any compound to act as a drug molecule, it must be non-toxic for normal cells in the human body. Therefore, the toxicity prediction of compounds against normal human cells is the most important task in a drug discovery pipeline. The SARS-CoV-2 cytopathic effect assay (host tox counterscreen, Assay ID-15) can help in the estimation of toxicity of compounds against Vero E6 cells. From the FPs level analysis of training dataset compounds, few negatively correlated FPs with the maximum response values are identified. For example, FPs, namely "nAromBond", "naAromAtom", "nBondsM" and "SwHBa", showed significant differences in their average maximum response values amongst the highly active and least active compounds (Supplementary file16.xlsx, Figure 1). Thus, for a compound to act as a drug, it should ideally have lower predicted maximum response values for this assay. Moreover, the compound might possess the least number of FPs, namely, aromatic bonds (nAromBond), aromatic atoms (naAromAtom), bonds that have bond order greater than one (nBondsM); and least value for the sum of E-States for weak hydrogen bond acceptors (SwHBa). Furthermore, the HEK293 cell-line toxicity assay (Assay ID-20) also fulfills the goal of testing the compounds' activity or toxicity against human HEK293 cell lines. For any compound to behave like a drug molecule, it must be least active or inactive in this assay. Furthermore, the analysis with training dataset molecules (used in the training of best prediction model) provided the FPs that are negatively and positively correlated with the maximum response value of compounds (Supplementary files18-19.xlsx, Table 1). Interestingly, some of the top 10 negatively correlated and all top 10 positively correlated FPs (with the maximum response value of compounds) show differences amongst the highly active and least active compounds (Supplementary files18-19.xlsx, Figure 1). The negatively correlated FPs, namely, aromatic groups (SubFPC274), hetero N nonbasic groups (SubFPC181), heteroaromatic groups (SubFPC184), heterocyclic groups (SubFPC275), chiral center specified (SubFPC307), amine groups (SubFPC23), tertiary aliphatic amine groups (SubFPC26) and hetero N basic no H groups (SubFPC180), show lower average values for least active molecules w.r.t. highly active molecules (Supplementary file18.xlsx, Figure 1). Whereas the top 10 positively correlated FPs show higher average values of FPs, namely, carboxylic acid groups (SubFPC84), 1,2-Diol groups (SubFPC41), alcoholic groups (SubFPC12), secondary alcohol groups (SubFPC14), sugar pattern 2 beta (SubFPC286), sugar pattern 2 alpha (SubFPC285), sugar pattern 2 (SubFPC282), sugar pattern 1 (SubFPC281), sugar pattern combi (SubFPC283) and primary alcohol groups (SubFPC13), for the least active compounds w.r.t. highly active compounds (Supplementary file19.xlsx, Figure 1). Therefore, designing the compounds with higher counts of these FPs values would help predict non-toxic compounds (against normal HEK293 cells).

The descriptors and FPs determined through the chemicals data analysis can be very useful in designing highly (against SARS-CoV-2) or least active molecules (against normal human cells) in the future. Thus, essentially deterministic features have been identified from experimental data. However, manual optimization of multiple descriptors and FPs is a challenging task that must be performed to design the compounds with desired properties.

Although, the current version of ASCoVPred predicts the activity of input compounds with the calculated efficiency, it may not predict efficiently for targets/assays it is not trained for. This limitation can be addressed by including additional models trained with compound activity datasets representing other targets that are not included in the current version of ASCoVPred. We will continue to update ASCoVPred with additional models, apart from improving the efficiency of existing models.

Conclusion

The designing of strategies for the rapid discovery of anti-SARS-CoV-2 compounds is an urgent need of the hour. Machine learning-based approaches in drug discovery and design are time-saving and cost-effective. The present study is based on computational designing of anti-SARS-CoV-2 compounds and estimates their toxicity against normal human cells. ASCoVPred webserver and standalone software are very useful in rapidly discovering inhibitors against SARS-CoV-2 and preventing viral entry into the human host cells. Also, the toxicity of molecules against normal human cells (HEK293 and Vero E6 cell-line) can be estimated with the help of toxicity prediction models deployed on the ASCoVPred platform.

Furthermore, the functional groups associated with the anti-SARS-CoV-2 activity of the molecules may provide better insights while designing the better hit molecules. In the future, the development of more ML models (trained and evaluated with more NCATS assays data) could enhance the utility of the ASCoVPred platform. We will also continue to improve the performance of the deployed models and update those on the ASCoVPred platform.

Methods

Data source

In the present study, a total of nine high-throughput screening (HTS) assays data were downloaded from National Centre for Advancing Translational Sciences (NCATS) website and used for the machine learning (ML)-based models training and evaluation (Table 4).

Table 4
List of assays used in the present study with the data source.

Assay ID	Assay name	Target category	Download link
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	Viral entry	https://opendata.ncats.nih.gov/covid19/assay?aid=1
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Counterscreen	https://opendata.ncats.nih.gov/covid19/assay?aid=2
6	ACE2 enzymatic activity	Viral entry	https://opendata.ncats.nih.gov/covid19/assay?aid=6
8	TMPRSS2 enzymatic activity	Viral entry	https://opendata.ncats.nih.gov/covid19/assay?aid=8
9	3CL enzymatic activity	Viral replication	https://opendata.ncats.nih.gov/covid19/assay?aid=9
14	SARS-CoV-2 cytopathic effect (CPE)	Live virus infectivity	https://opendata.ncats.nih.gov/covid19/assay?aid=14
15	SARS-CoV-2 cytopathic effect (host tox counterscreen)	Counterscreen	https://opendata.ncats.nih.gov/covid19/assay?aid=15
20	HEK293 cell line toxicity	Counterscreen	https://opendata.ncats.nih.gov/covid19/assay?aid=20
21	Human fibroblast toxicity	Counterscreen	https://opendata.ncats.nih.gov/covid19/assay?aid=21

Initially, nine tab-separated files were downloaded from the NCATS website. Of the nine files, each contained information such as “Sample ID”, “Sample Name”, “PubChem SID”, “Primary MOA”, “Assay Name”, “CAS ID”, “AC50 value”, “Efficacy”, “Maximum Response”, “Drug Name”, “SMILES”, etc. Of the given columns, two columns, namely “SMILES” and “MAX_RESPONSE”, were extracted from each tab-separated file for further processing and machine learning-based models training and validation. According to the description given on the NCATS data browser²⁵, a molecule with a MAX_RESPONSE value approaching -100 can be considered as highly active while the one with MAX_RESPONSE value approaching zero or towards positive values can be considered as least active or inactive.

Description of HTS assays

The nine HTS assays (Table 4) used to test compounds’ bio-activities by NCATS can be broadly categorized into four different types viz. assay used to determine the molecules that (i) prevent viral entry into host cells (ii) prevent viral replication into host cells (iii) reverse the cytopathic effect of host cells (caused by SARS-CoV-2 virion) and (iv) show toxic effects against normal human/host cells.

AlphaLISA assay (Table 4, Assay ID-1) contains the therapeutic molecules that can potentially disrupt the interaction between SARS-CoV-2 Spike protein and the human host ACE2 receptor. The interaction between SARS-CoV-2 Spike protein and host ACE2 receptor plays a vital role during the viral entry into the human host⁶. Thus, this assay aims to identify the molecules with the potential to prevent SARS-CoV-2 entry into the human host. The second assay, i.e., TruHit counterscreen (Table 4, Assay ID-2), helps identify false-positive compounds that interfere with the AlphaLISA readout in a non-specific manner.

ACE2 enzymatic activity assay (Table 4, Assay ID-6) measures the ACE2 inhibitory potential of compounds to prevent the disruption of endogenous enzyme function. The TMPRSS2 enzymatic activity assay (Table 4, Assay ID-8) is specific to measure the TMPRSS2 inhibitory potential of compounds. Recent studies have highlighted the importance of human TMPRSS2 as an antiviral drug target and made attempts to screen the molecules with inhibitory potential against TMPRSS2^{19,26}.

The 3CL enzymatic activity (Table 4, Assay ID-9) measures the inhibitory potential of molecules against viral 3-chymotrypsin like protease (3CLpro), the main protease of SARS-CoV-2. The enzyme is pivotal for the viral replication within human host cells^{17,27}. Thus, it is an important anti-SARS-CoV-2 drug target in COVID-19 drug discovery.

CPE assay (Table 4, Assay ID-14) determines the ability of a compound to reverse the cytopathic effect caused by SARS-CoV-2 in Vero E6 host cells. This assay can be used for the high-throughput screening of compounds with antiviral activity. In contrast, there are chances that a compound itself may show cytotoxicity; therefore, a counterscreen (Table 4, Assay ID-15) to measure the toxicity of compounds against host (Vero E6) cells is used for the detection of such compounds.

The general toxicity of compounds against HEK293 and Hh-Wt cell lines is measured with the help of HEK293 cell line toxicity (Table 4, Assay ID-20) and Human fibroblast toxicity assay (Table 4, Assay ID-21), respectively. Further details about the assays used in the present study can be retrieved from the

Descriptors calculation

The SMILES structures of the molecules are used to calculate 1444 1D, 2D descriptors and 12 different types of fingerprints (FPs) calculation using open-source software PaDEL (v2.21)²⁸.

Data pre-processing

The pre-processing of data is an important task for any machine learning study. Therefore, a systematic approach is applied to process the molecular structure data while molecular descriptors and fingerprints calculation (Figure 2). The parameters opted on PaDEL software (before actually starting the descriptors / FPs calculation) are "Remove salt", "Detect aromaticity", "Standardize nitro groups", "Max. threads -1", "Max. waiting jobs -1", "Max. Running time per molecule: 12,00,000 milliseconds", and "Retain molecules order". Furthermore, only those molecules for which all the descriptor/fingerprint values are calculated that have been used for ML-based models training, external validation and further analysis.

Preparation of datasets for models training and validation

Data pre-processing and filtering are followed by redundancy removal to retrieve the dataset of unique molecules. Therefore, the molecules possessing identical descriptor or FPs values and maximum response values are included only once. The unique dataset of molecules was further split into a training dataset (80% molecules) and an external validation dataset (20% molecules). The training datasets are used for training and internal validation (through the five-fold cross-validation technique) of the ML-based models, while external validation datasets are kept separate for the final and external validation of the developed models.

Descriptors or feature selection

In the past, it has been observed that all the features or descriptors are not equally important to predict the biological activity of the molecules²⁹⁻³¹. Therefore, a feature selection technique in WEKA (v3.8.2) software²⁴ is applied to determine the most relevant descriptors and fingerprints associated with the biological activity of the molecules. Thus, "CfsSubsetEval" (with default parameter values) as "Attribute Evaluator" with "BestFirst" as "Search Method" (with default parameter values) is used as feature selection techniques for the present study. This technique has been extensively used for features selection in various QSAR and classification-related studies³²⁻³⁶.

Tools used for model building

An open-source data mining and ML tool, WEKA (v3.8.2), has been used in the present study to train and validate the prediction models. In the past, WEKA has been used to solve several classification and regression-based, important biological problems^{32-35, 37}.

Cross-validation technique used

The prioritization of best models (with the highest efficacy) is essential in any ML-based study. In the present study, thousands of models (the results for best models from each assay are given online at <http://14.139.62.220/ascovpred/supple.php>) were trained with all regression techniques available in WEKA, however, selection of the best models are made through the five-fold cross-validation technique. This technique has been widely used in previous studies to prioritize the best prediction/classification models^{34,35,38-42}. In the five-fold cross-validation technique, the whole training dataset is divided into five almost equal-sized parts and each part is used in the training and testing of the prediction models.

Formulae used to evaluate the models' performance

The in-built functions available with WEKA (v3.8.2), such as Pearson Correlation Coefficient (R; equation (1)), mean absolute error (MAE; equation (2)) and root mean squared error (RMSE; equation (3)), have been used to evaluate the models' performance through five-fold cross-validation technique. In both internal and external validation, the models with the highest R-value and lowest MAE and RMSE values are selected as the best prediction models. The latter five models are hosted on the ASCovPred webserver and standalone software for public use. The equations for the used formulae are as given below:

$$R = \frac{\left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N} \right)}{\sqrt{\left(\sum X_i^2 - \frac{(\sum X_i)^2}{N} \right) \left(\sum Y_i^2 - \frac{(\sum Y_i)^2}{N} \right)}}$$

1

$$\text{MAE} = \frac{\sum_{i=1}^N |Y_i - X_i|}{N}$$

2

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - X_i)^2}$$

3

For i^{th} compound, Y_i and X_i represent predicted and actual maximum response values, respectively. N is the total number of compounds. The value of R is used to measure the quality of a model. The value of R varies from -1 to $+1$. The negative value of R shows the negative correlation with a particular property or feature. Thus, higher the value of R , the better the model's quality in terms of the predicted maximum response value of the compounds.

Identification of descriptors of FPs associated with the maximum response

To understand the relative contribution of different descriptors and fingerprints (FPs) in determining the molecules' maximum response values, assay-wise, R values were calculated between these. The molecules with top 10 positively and negatively correlated descriptors or FPs values (Supplementary files 10-19, Table 1) were shortlisted from the training dataset molecules along with their maximum response values and sorted in the ascending order of maximum response values. Further, the top 100 records (highly active compounds) (Supplementary files 10-19, Table 2) and the lowest 100 records (least active compounds) (Supplementary files 10-19, Table 3) were selected for both positively and negatively correlated descriptors and FPs. Finally, the average values were calculated for each descriptor and FP (Supplementary files 10-19, Table 4). Grouped bar charts were plotted to compare the average values of descriptors / FPs amongst highly active and least active compounds (Supplementary files 10-19, Figure 1). Thus, the grouped bar charts depict the differences of descriptors / FPs average values amongst highly active and least active compounds.

Declarations

Acknowledgments

We acknowledge ICGB for providing the necessary infrastructure and facilities for the research.

Data availability

Publicly available data were analyzed in this study that may be obtained from:

<https://opendata.ncats.nih.gov/covid19/assays>

Code availability

The ASCoVPred standalone software is freely available from <http://14.139.62.220/ascovpred/download.php>

Contributions

DG and AS conceptualized the study, analyzed the data, and prepared the manuscript. AS and NC carried out the machine learning studies. All authors reviewed and approved the final version.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. WHO. WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int/> (2021).
2. Willyard, C. How antiviral pill molnupiravir shot ahead in the COVID drug hunt. *Nature* (2021) doi:10.1038/d41586-021-02783-1.
3. FDA. Know Your Treatment Options for COVID-19. <https://www.fda.gov/consumers/consumer-updates/know-your-treatment-options-covid-19> (2021).
4. FDA. Coronavirus Treatment Acceleration Program (CTAP). <https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap> (2021).
5. Huang, Y., Yang, C., Xu, X., Xu, W. & Liu, S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
6. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
7. Monticelli, M. *et al.* Protective Role of a TMPRSS2 Variant on Severe COVID-19 Outcome in Young Males and Elderly Women. *Genes* vol. 12 (2021).
8. Gunst, J. D. *et al.* Efficacy of the TMPRSS2 inhibitor camostat mesilate in patients hospitalized with Covid-19-a double-blind randomized controlled trial. *EClinicalMedicine* **35**, (2021).
9. NCATS. SARS-CoV-2 Assays. <https://opendata.ncats.nih.gov/covid19/assays>.
10. Hu, X. *et al.* Discovery of TMPRSS2 Inhibitors from Virtual Screening as a Potential Treatment of COVID-19. *ACS Pharmacol. Transl. Sci.* **4**, 1124–1135 (2021).
11. Huang, X., Pearce, R., Omenn, G. S. & Zhang, Y. Identification of 13 Guanidinobenzoyl- or Aminidinobenzoyl-Containing Drugs to Potentially Inhibit TMPRSS2 for COVID-19 Treatment. *International Journal of Molecular Sciences* vol. 22 (2021).
12. Sgrignani, J. & Cavalli, A. Computational Identification of a Putative Allosteric Binding Pocket in TMPRSS2. *Frontiers in Molecular Biosciences* vol. 8 279 (2021).
13. KC, G. B. *et al.* A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat. Mach. Intell.* **3**, 527–535 (2021).
14. Muratov, E. N. *et al.* A critical overview of computational approaches employed for COVID-19 drug discovery. *Chem. Soc. Rev.* **50**, 9121–9151 (2021).

15. Urbina, F., Puhl, A. C. & Ekins, S. Recent advances in drug repurposing using machine learning. *Curr. Opin. Chem. Biol.* **65**, 74–84 (2021).
16. Chen, C. Z. *et al.* Drug Repurposing Screen for Compounds Inhibiting the Cytopathic Effect of SARS-CoV-2. *Frontiers in Pharmacology* vol. 11 2005 (2021).
17. Mody, V. *et al.* Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents. *Commun. Biol.* **4**, 93 (2021).
18. Rajpoot, S., Alagumuthu, M. & Baig, M. S. Dual targeting of 3CLpro and PLpro of SARS-CoV-2: A novel structure-based design approach to treat COVID-19. *Curr. Res. Struct. Biol.* **3**, 9–18 (2021).
19. Peiffer, A. L. *et al.* TMPRSS2 inhibitor discovery facilitated through an in silico and biochemical screening platform. *bioRxiv Prepr. Serv. Biol.* 2021.03.22.436465 (2021) doi:10.1101/2021.03.22.436465.
20. Mslati, H., Gentile, F., Perez, C. & Cherkasov, A. Comprehensive Consensus Analysis of SARS-CoV-2 Drug Repurposing Campaigns. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c00384.
21. Elekofehinti, O. O., Iwaloye, O., Molehin, O. R. & Famusiwa, C. D. Identification of lead compounds from large natural product library targeting 3C-like protease of SARS-CoV-2 using E-pharmacophore modelling, QSAR and molecular dynamics simulation. *Silico Pharmacol.* **9**, 49 (2021).
22. Brimacombe, K. R. *et al.* An OpenData portal to share COVID-19 drug repurposing data in real time. *bioRxiv* 2020.06.04.135046 (2020) doi:10.1101/2020.06.04.135046.
23. Sun, G. *et al.* In Silico Prediction of O6-Methylguanine-DNA Methyltransferase Inhibitory Potency of Base Analogs with QSAR and Machine Learning Methods. *Molecules* vol. 23 (2018).
24. Eibe Frank, Mark A. Hall, and I. H. W. The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques'. Fourth Edition https://waikato.github.io/weka-wiki/citing_weka/ (2016).
25. National Center for Advancing Translational Sciences. NCATS Open Data Browser. <https://opendata.ncats.nih.gov/covid19/databrowser> (2021).
26. Shrimp, J. H. *et al.* An Enzymatic TMPRSS2 Assay for Assessment of Clinical Candidates and Discovery of Inhibitors as Potential Treatment of COVID-19. *ACS Pharmacol. Transl. Sci.* **3**, 997–1007 (2020).
27. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642–646 (2021).
28. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32**, 1466–1474 (2011).
29. Goodarzi, M., Dejaegher, B. & Vander Heyden, Y. Feature selection methods in QSAR studies. *J. AOAC Int.* **95**, 636–651 (2012).
30. Chauhan, J. S., Dhanda, S. K., Singla, D., Agarwal, S. M. & Raghava, G. P. S. QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. *PLoS One* **9**, e101079 (2014).
31. Dhall, A., Patiyal, S., Sharma, N., Devi, N. L. & Raghava, G. P. S. Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. *Comput. Biol. Med.* **137**, 104780 (2021).
32. Qureshi, A., Kaur, G. & Kumar, M. AVCpred: an integrated web server for prediction and design of antiviral compounds. *Chem. Biol. Drug Des.* **89**, 74–83 (2017).
33. Qureshi, A., Rajput, A., Kaur, G. & Kumar, M. HIVprotl: an integrated web based platform for prediction and design of HIV proteins inhibitors. *J. Cheminform.* **10**, 12 (2018).
34. Sharma, A., Gupta, P., Kumar, R. & Bhardwaj, A. dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides. *Sci. Rep.* **6**, 21839 (2016).
35. Sharma, A., Singla, D., Rashid, M. & Raghava, G. P. Designing of peptides with desired half-life in intestine-like environment. *BMC Bioinformatics* **15**, 282 (2014).
36. Sardar, R., Sharma, A. & Gupta, D. Machine Learning Assisted Prediction of Prognostic Biomarkers Associated With COVID-19, Using Clinical and Proteomics Data. *Front. Genet.* **12**, 636441 (2021).
37. Mathur, D., Singh, S., Mehta, A., Agrawal, P. & Raghava, G. P. S. In silico approaches for predicting the half-life of natural and modified peptides in blood. *PLoS One* **13**, e0196829 (2018).
38. Baltruschat, M. & Czodrowski, P. Machine learning meets pK (a). *F1000Research* **9**, (2020).
39. Vignaux, P. A., Minerali, E., Foil, D. H., Puhl, A. C. & Ekins, S. Machine Learning for Discovery of GSK3 β Inhibitors. *ACS omega* **5**, 26551–26561 (2020).
40. Zhang, L. *et al.* CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci. Rep.* **7**, 2118 (2017).
41. Liu, M. *et al.* Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. *Toxicol. Lett.* **332**, 88–96 (2020).
42. Wei, Y., Li, W., Du, T., Hong, Z. & Lin, J. Targeting HIV/HCV Coinfection Using a Machine Learning-Based Multiple Quantitative Structure-Activity Relationships (Multiple QSAR) Method. *Int. J. Mol. Sci.* **20**, (2019).

Figures

A

B

Figure 1

Screenshots of [A] ASCoVPred webserver and [B] standalone software usage.

Figure 2

(A) ASCoVPred development pipeline. (B) ASCoVPred prediction pipeline. Figure Created with BioRender.com

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile1.xlsx](#)
- [Supplementaryfile2.xlsx](#)
- [Supplementaryfile3.xlsx](#)
- [Supplementaryfile4.xlsx](#)
- [Supplementaryfile5.xlsx](#)
- [Supplementaryfile6.xlsx](#)
- [Supplementaryfile7.xlsx](#)
- [Supplementaryfile8.xlsx](#)
- [Supplementaryfile9.xlsx](#)
- [Supplementaryfile10.xlsx](#)
- [Supplementaryfile11.xlsx](#)
- [Supplementaryfile12.xlsx](#)
- [Supplementaryfile13.xlsx](#)
- [Supplementaryfile14.xlsx](#)
- [Supplementaryfile15.xlsx](#)
- [Supplementaryfile16.xlsx](#)
- [Supplementaryfile17.xlsx](#)

- [Supplementaryfile18.xlsx](#)
- [Supplementaryfile19.xlsx](#)