Machine Learning Assisted Drug Discovery for SARS-CoV-2

Authors: Arun Sharma, Neeraj Chaturvedi, Dinesh Gupta*

(August 25th, 2022)

Presented by

Arun Sharma, Ph.D.

Senior Information Scientist

Translational Bioinformatics Group

DBT Apex BTIC, ICGEB, New Delhi

Email: arun@icgeb.res.in; bioinfo.arun@gmail.com

Reference: https://assets.researchsquare.com/files/rs-967196/v1/6e040cdf-a7d7-4af6-b201-00c95c013278.pdf?c=1635494778

Content

- About SARS-CoV-2
- Need for anti-SARS-CoV-2 drug discovery
- Current status of drug discovery for SARS-CoV-2
- Machine learning (ML) approaches used by our group for anti-SARS-CoV-2 drug discovery
- Performance of ML models with training and external validation datasets
- Deployment of the best ML models on ASCoVPred web-server and standalone
- Conclusion

About SARS-CoV-2

- Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a single-stranded RNA-enveloped virus that causes Coronavirus disease (COVID-19) in humans.

- Entire genome is 29881 bp in length (GenBank no. MN908947), encoding 9860 amino acids.

- Genes express structural and nonstructural proteins.

- The S, E, M, and N genes encode structural proteins.

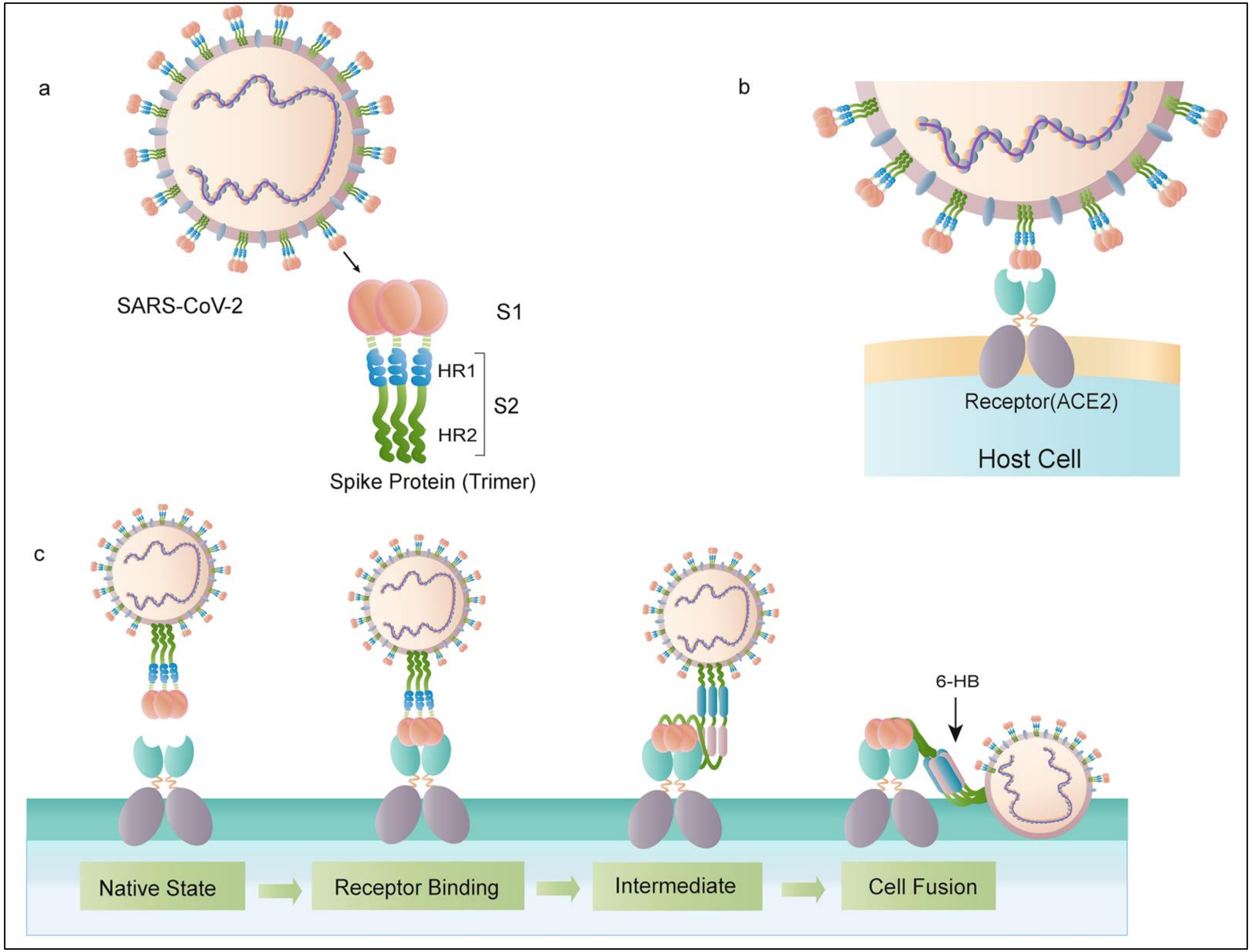
- Nonstructural proteins: 3-chymotrypsin-like protease, papain-like protease, and RNA-dependent RNA polymerase (RdRp).

- A large number of **glycosylated S proteins cover the surface of SARS-CoV-2** and **bind to the host cell receptor angiotensin-converting enzyme 2 (ACE2), mediating viral cell entry.**

- When the S protein binds to the receptor, TM protease serine 2 (TMPRSS2), a type 2 TM serine protease located on the host cell membrane, promotes virus entry into the cell by activating the S protein.

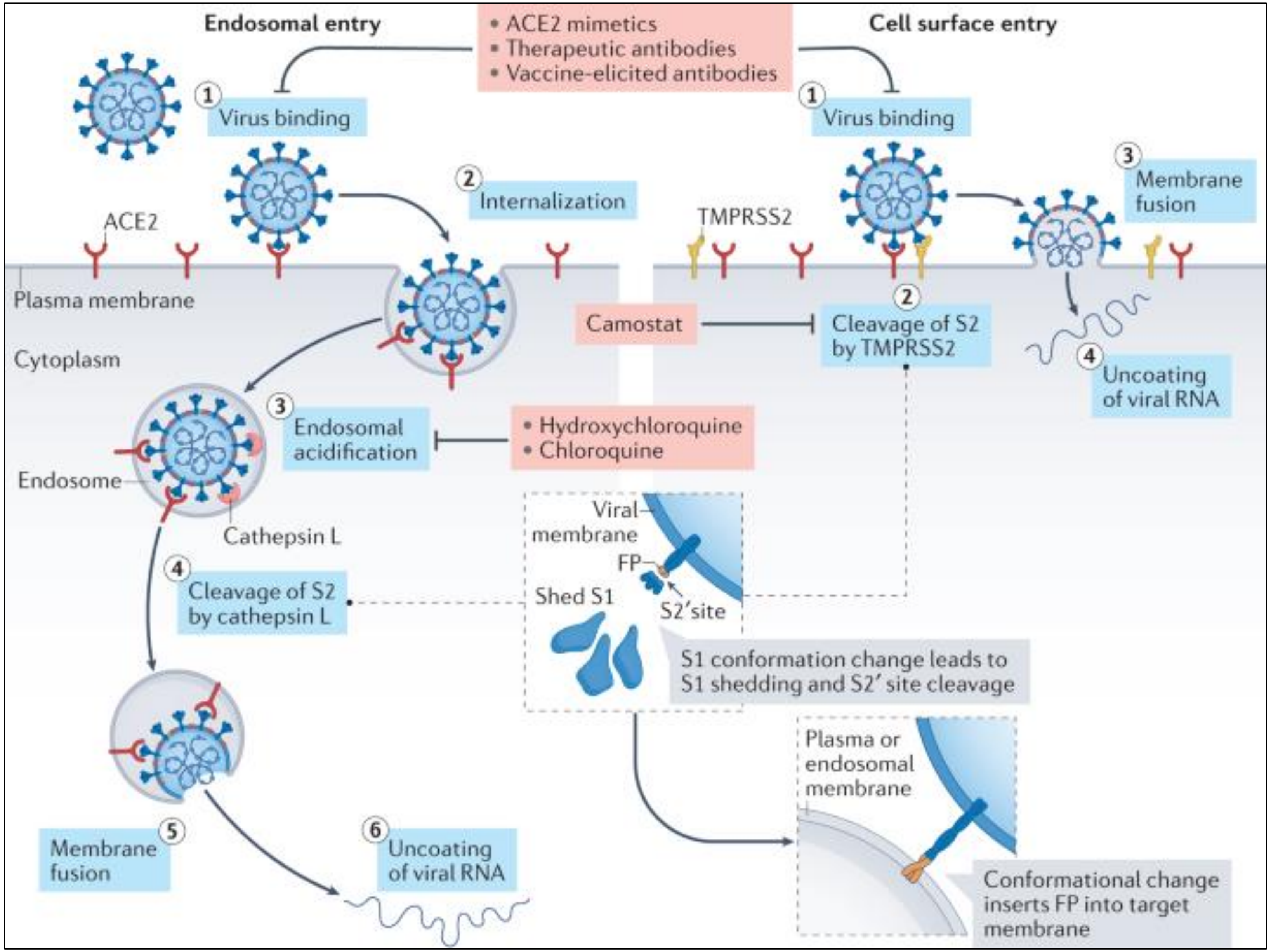
https://www.nature.com/articles/s41580-021-00418-x

Figure 1. a The schematic structure of the S protein. b The S protein binds to the receptor ACE2. c The binding and virus–cell fusion process mediated by the S protein.



https://www.nature.com/articles/s41401-020-0485-4

Figure 2. Modes of SARS-CoV-2 virus entry in the host cells. The S protein binds to the receptor, TM protease serine 2 (TMPRSS2), a type 2 TM serine protease located on the host cell membrane, promotes virus entry into the cell by activating the S protein.



https://www.nature.com/articles/s41580-021-00418-x

Need for anti-SARS-CoV-2 drug discovery

- According to the World Health Organization (WHO) online dashboard, as of August 24, 2022, the SARS-CoV-2 has infected more than 595.22 million people.
- Also, this coronavirus is responsible for >6.45 million deaths globally.
- Limited options are available to treat the COVID-19 disease.
- Although several vaccines have been developed to reduce the disease burden, effective antivirals are still required to treat infected and hospitalized patients.

https://covid19.who.int/
https://www.fda.gov/drugs/emergency-preparedness-drugs/coronavirus-covid-19-drugs

Current status of drug discovery for SARS-CoV-2

- **Veklury** (remdesivir), is the U.S. Food and Drug Administration (FDA) approved drug available for adult and certain pediatric COVID-19 patients.

- **Olumiant** (baricitinib) is approved for the treatment of COVID-19 in hospitalized adults requiring supplemental oxygen, non-invasive or invasive mechanical ventilation, or extracorporeal membrane oxygenation (ECMO).

- Several drugs are in different stages of trials as the quest for an effective anti-SARS-CoV-2 continues (e.g., molnupiravir).

- FDA has also issued emergency use authorization (EUA) for other types of treatments, such as monoclonal antibody-based treatments.

- Treatment options, such as the development of antivirals, immunomodulators, neutralizing antibody therapies, cell therapy, etc., are **ongoing and yet to pass through different clinical trials**.

- However, ***in vitro*** **discovery of novel inhibitors** is **tedious, labor-intensive, time-consuming and costly exercise**.

- Computational predictions facilitate *in vitro* discovery by shortlisting the most effective chemical entities, saving time and cost.

https://www.fda.gov/drugs/emergency-preparedness-drugs/coronavirus-covid-19-drugs

- 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.

- S-protein plays a vital role in the SARS-CoV-2 entry into human target cells that also makes it an attractive drug target.

- 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.

- **S-protein** and **TMPRSS2** play a **vital role in SARS-CoV-2 entry into human target cells**.

- Computational approaches, including molecular docking and machine learning (ML)-based classification algorithm development, have been used to identify suitable anti-SARS-CoV-2 inhibitors.

- Moreover, drug repurposing approaches have also been explored to quickly identify the approved drug molecules which may bind to SARS-CoV-2 drug targets.

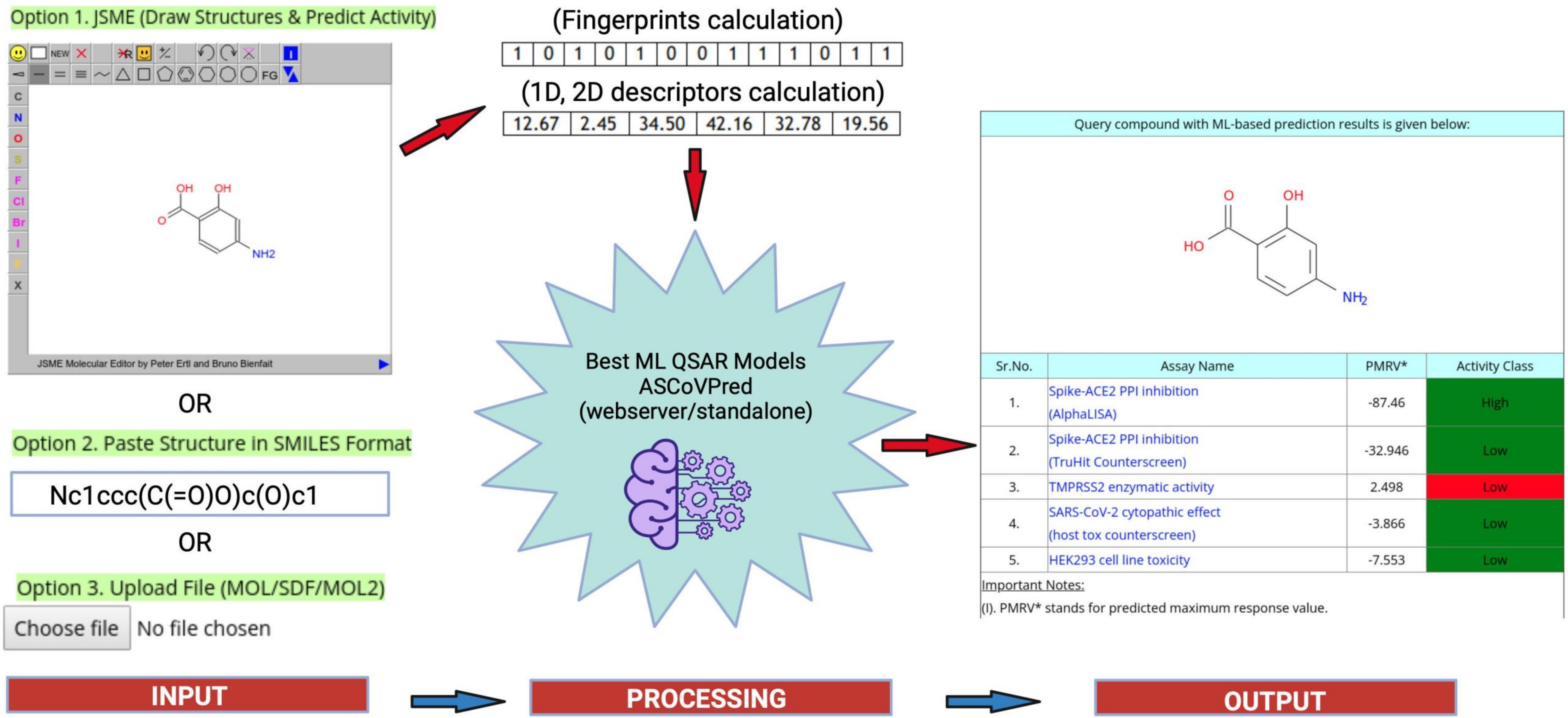
- **Most of such computational studies** have used **docking or molecular dynamics** and combinations of the two to explore FDA-approved drugs as potential SARS-CoV-2 inhibitors.

- 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**.

ML approaches used by our group for anti-SARS-CoV-2 drug discovery

(https://assets.researchsquare.com/files/rs-967196/v1/6e040cdf-a7d7-4af6-b201-00c95c013278.pdf?c=1635494778)

Figure 3. Anti-SARS-CoV-2 activity and human cell toxicity prediction of molecules through ASCoVPred webserver.



ASCoVPred webserver & standalone URL: http://192.168.5.81/ascovpred/

- **Data source:** 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 (http://192.168.5.81/ascovpred/supple.php).

- The nine HTS assays used to test compounds' bio-activities by NCATS can be **broadly categorized into four different types:**

(i) Prevent viral entry into host cells.

(ii) Prevent viral replication into host cells.

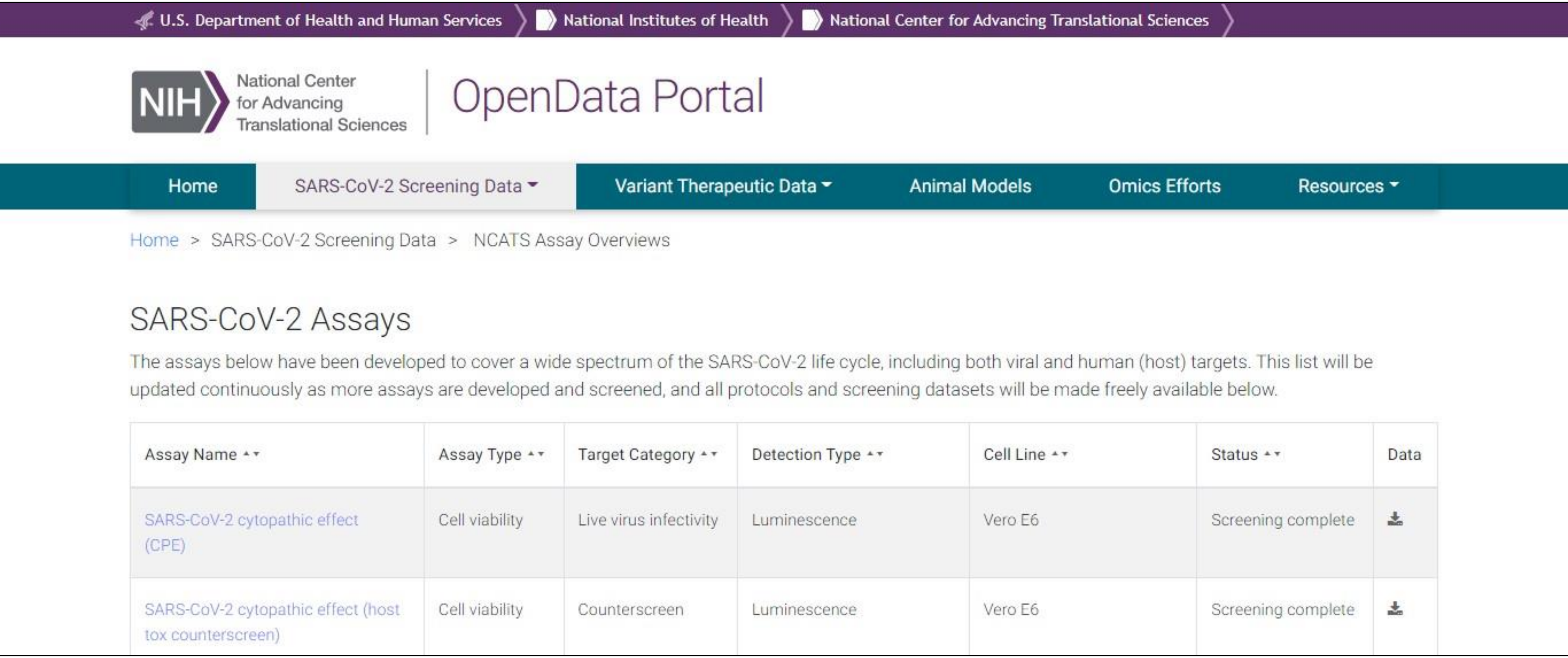
(iii) Reverse the cytopathic effect of host cells (caused by SARS-CoV-2 virion).

(iv) Show toxic effects against normal human/host cells.

- **Data pre-processing:** 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**".

- Only those molecules for which all the descriptor/fingerprint values are calculated have been used for ML-based models training, validation and further analysis.

A Screenshot from NCATS Open Data Portal



U.S. Department of Health and Human Services 〉 National Institutes of Health 〉 National Center for Advancing Translational Sciences

NIH National Center for Advancing Translational Sciences | OpenData Portal

Home | SARS-CoV-2 Screening Data | Variant Therapeutic Data | Animal Models | Omics Efforts | Resources

Home > SARS-CoV-2 Screening Data > NCATS Assay Overviews

SARS-CoV-2 Assays

The assays below have been developed to cover a wide spectrum of the SARS-CoV-2 life cycle, including both viral and human (host) targets. This list will be updated continuously as more assays are developed and screened, and all protocols and screening datasets will be made freely available below.

Assay Name	Assay Type	Target Category	Detection Type	Cell Line	Status	Data
SARS-CoV-2 cytopathic effect (CPE)	Cell viability	Live virus infectivity	Luminescence	Vero E6	Screening complete	
SARS-CoV-2 cytopathic effect (host tox counterscreen)	Cell viability	Counterscreen	Luminescence	Vero E6	Screening complete	

Table 1. A brief description of the nine assays used for training and evaluation of the prediction models.

Assay ID	Assay name	Assay description
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	Contains the therapeutic molecules that can potentially disrupt the interaction between SARS-CoV-2 Spike protein and the human host ACE2 receptor.
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Helps identify false-positive compounds that interfere with the AlphaLISA readout in a non-specific manner.
6	ACE2 enzymatic activity	Measures the ACE2 inhibitory potential of compounds to prevent the disruption of endogenous enzyme function.
8	TMPRSS2 enzymatic activity	Specific to measure the TMPRSS2 inhibitory potential of compounds.
9	3CL enzymatic activity	Measures the inhibitory potential of molecules against viral 3-chymotrypsin like protease (3CLpro), the main protease of SARS-CoV-2.
14	SARS-CoV-2 cytopathic effect (CPE)	Determines the ability of a compound to reverse the cytopathic effect caused by SARS-CoV-2 in Vero E6 host cells.
15	SARS-CoV-2 cytopathic effect (host tox counterscreen)	To measure the toxicity of compounds against host (Vero E6) cells is used for the detection of such compounds.
20	HEK293 cell line toxicity	To measure the general toxicity of compounds against HEK293 cell line.
21	Human fibroblast toxicity	To measure the general toxicity of compounds against Hh-Wt cell line.

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

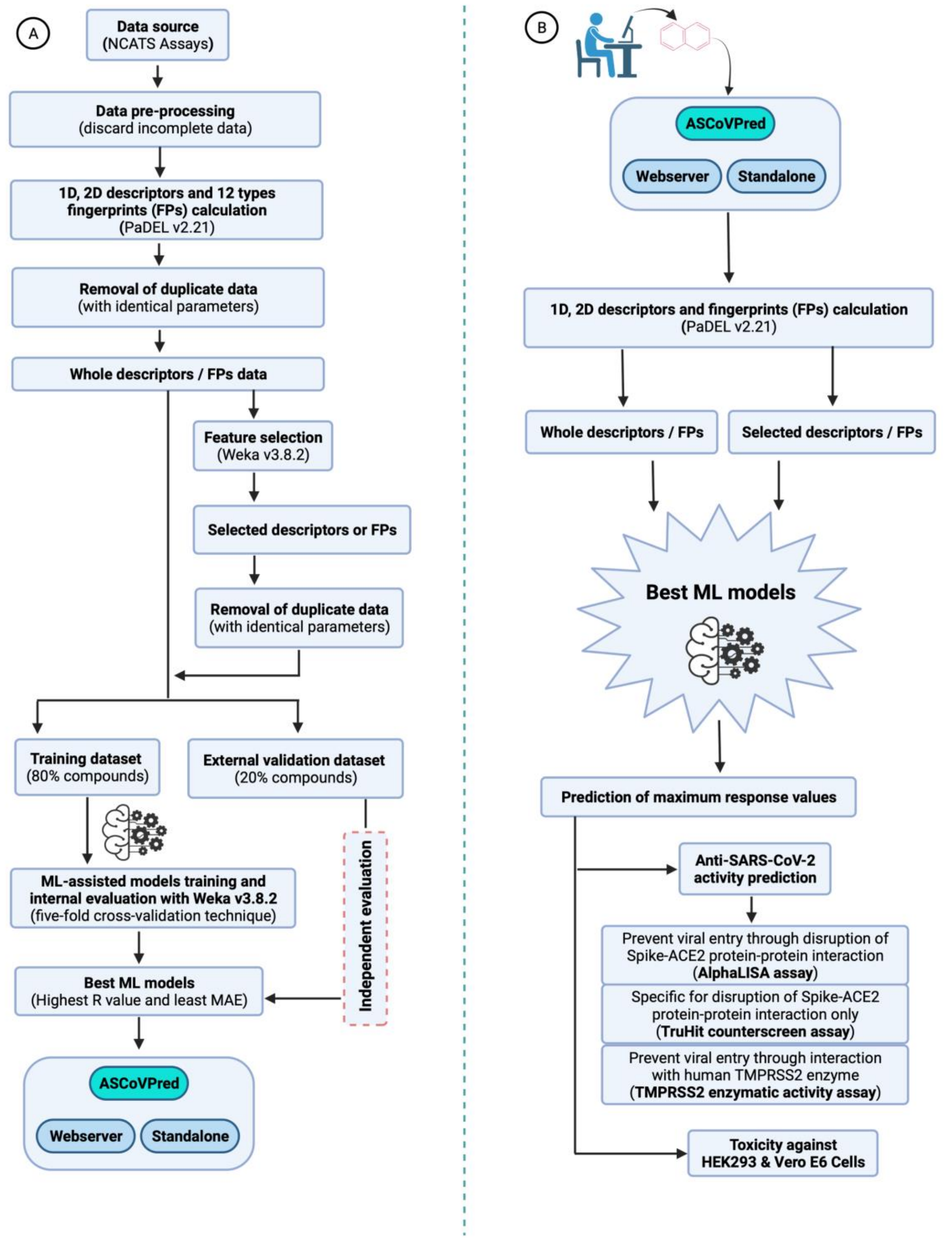
Figure 4. A systematic computational approach used for building ML-based QSAR prediction models and their usage by users.

(A). Flow diagram depicting the overall strategy used to train, evaluate and build the ML-based QSAR prediction models.

(B). The best prediction models can be used by users to predict the anti-SARS-CoV-2 activity and human cell toxicity of compounds.

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 a **external validation dataset (20% molecules)**.



- The training datasets are used for training and internal validation (through five-fold cross-validation technique) of the ML-based models, while external validation datasets are kept separate for the final or external validation of the developed models.

- **Descriptors or feature selection:** A feature selection technique in WEKA v3.8.2 is applied to determine the most relevant descriptors and fingerprints associated with the biological activity of the molecules.

- "**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.

- **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.

- **Cross-validation technique used:** Selection of the best models is made through the five-fold cross-validation technique.

Formulae used to evaluate the models' performance: The in-built functions available with WEKA (v3.8.2), such as Pearson Correlation Coefficient (R), mean absolute error (MAE) and root mean squared error (RMSE), 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.**

$$R = \frac{\left(\sum Xi\,Yi - \frac{\sum Xi \sum Yi}{N}\right)}{\sqrt{\left(\sum Xi^2 - \frac{(\sum Xi)^2}{N}\right)\left(\sum Yi^2 - \frac{(\sum Yi)^2}{N}\right)}} \qquad (1)$$

$$\text{MAE} = \frac{\sum_{i=1}^{N} |Yi - Xi|}{N} \qquad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Yi - Xi)^2} \qquad (3)$$

For i^{th} compound, Yi and Xi represent predicted and actual maximum response value, respectively. N is total number of compounds. **The value of R is used to measure the quality of 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, better will be the quality of model in terms of the predicted maximum response value of the compounds.**

- We have developed machine learning (ML) models for the rapid discovery of molecules potentially inhibitory to SARS-CoV-2, with negligible or no human cell toxicity.

- The ML QSAR models were trained and optimized with features (descriptors and fingerprints) based on the activity assays of experimentally validated SARS-CoV-2 inhibitory compounds.

- The feature selection for selecting the best descriptors for ML training helped identify a set of decisive training descriptors and fingerprints that correlate positively or negatively with the anti-SARS-CoV-2 activity and toxicity of the compounds.

- The selected features were used to train thousands of different ML models. The best-optimized models are deployed as **ASCoVPred webserver and standalone software that provides easy and free access to the models.**

Performance of ML models with training and external validation datasets

Table 2. The number of molecules used for training and evaluation of the 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

Table 3. Results of the evaluation for best prediction models.

•SubstructureFingerprintCount; #Not deployed on ASCoVPred webserver

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	R^2	MAE	RMSE	WEKA technique used for training the model with training dataset of compounds	R	R^2	MAE	RMSE
1	**Spike-ACE2 protein-protein interaction (AlphaLISA)**	**SubFPC•**	**307**	**0.67**	**0.44**	**14.96**	**21.22**	**weka.classifiers.meta.AdditiveRegression (RandomForest)**	**0.66**	**0.42**	**17.41**	**24.95**
2	**Spike-ACE2 protein-protein interaction (TruHit Counterscreen)**	**1D & 2D**	**50**	**0.72**	**0.51**	**13.51**	**18.83**	**weka.classifiers.meta.AdditiveRegression (RandomForest)**	**0.74**	**0.55**	**14.44**	**20.50**
6	ACE2 enzymatic activity	ExtendedFingerprinter#	28	0.32	0.10	14.31	37.09	weka.classifiers.trees.RandomForest	0.57	0.32	16.68	44.88
8	**TMPRSS2 enzymatic activity**	**SubFPC•**	**307**	**0.52**	**0.24**	**11.12**	**36.47**	**weka.classifiers.trees.RandomForest**	**0.73**	**0.36**	**14.99**	**44.64**
9	3CL enzymatic activity	SubFPC•,#	307	0.40	0.16	5.90	11.31	weka.classifiers.meta.RandomCommittee (RandomForest)	0.45	0.19	5.11	9.14
14	SARS-CoV-2 cytopathic effect (CPE)	1D & 2D#	1444	0.50	0.25	7.43	14.59	weka.classifiers.meta.RandomSubSpace (RandomForest)	0.43	0.18	8.23	14.90
15	**SARS-CoV-2 cytopathic effect (host tox counterscreen)**	**1D & 2D**	**1444**	**0.66**	**0.43**	**13.12**	**21.13**	**weka.classifiers.meta.AdditiveRegression (RandomForest)**	**0.65**	**0.42**	**14.11**	**21.93**
20	**HEK293 cell line toxicity**	**SubFPC•**	**307**	**0.66**	**0.44**	**26.68**	**34.16**	**weka.classifiers.meta.RandomCommittee (RandomForest)**	**0.68**	**0.46**	**26.41**	**33.87**
21	Human fibroblast toxicity	1D & 2D#	45	0.43	0.18	12.61	19.69	weka.classifiers.meta.RandomCommittee (RandomForest)	0.51	0.26	11.99	18.63

Table 4. The desired activity prediction profile for an ideal multi-target hit molecule.

Assay ID	Assay Name	Target Category	Predicted maximum response value (PMRV) threshold	Activity Class
1	Spike-ACE2 protein-protein interaction (AlphaLISA)	Viral entry	< -66	High
2	Spike-ACE2 protein-protein interaction (TruHit Counterscreen)	Counterscreen	> -33	Low
8	TMPRSS2 enzymatic activity	Viral entry	< -66	High
15	SARS-CoV-2 cytopathic effect(host tox counterscreen)	Counterscreen	> -33	Low
20	HEK293 cell line toxicity	Counterscreen	> -33	Low

Deployment of the best ML models on ASCoVPred web-server and standalone

Figure 5. Screenshots of ASCoVPred webserver usage. Website link: http://192.168.5.81/ascovpred/index.html

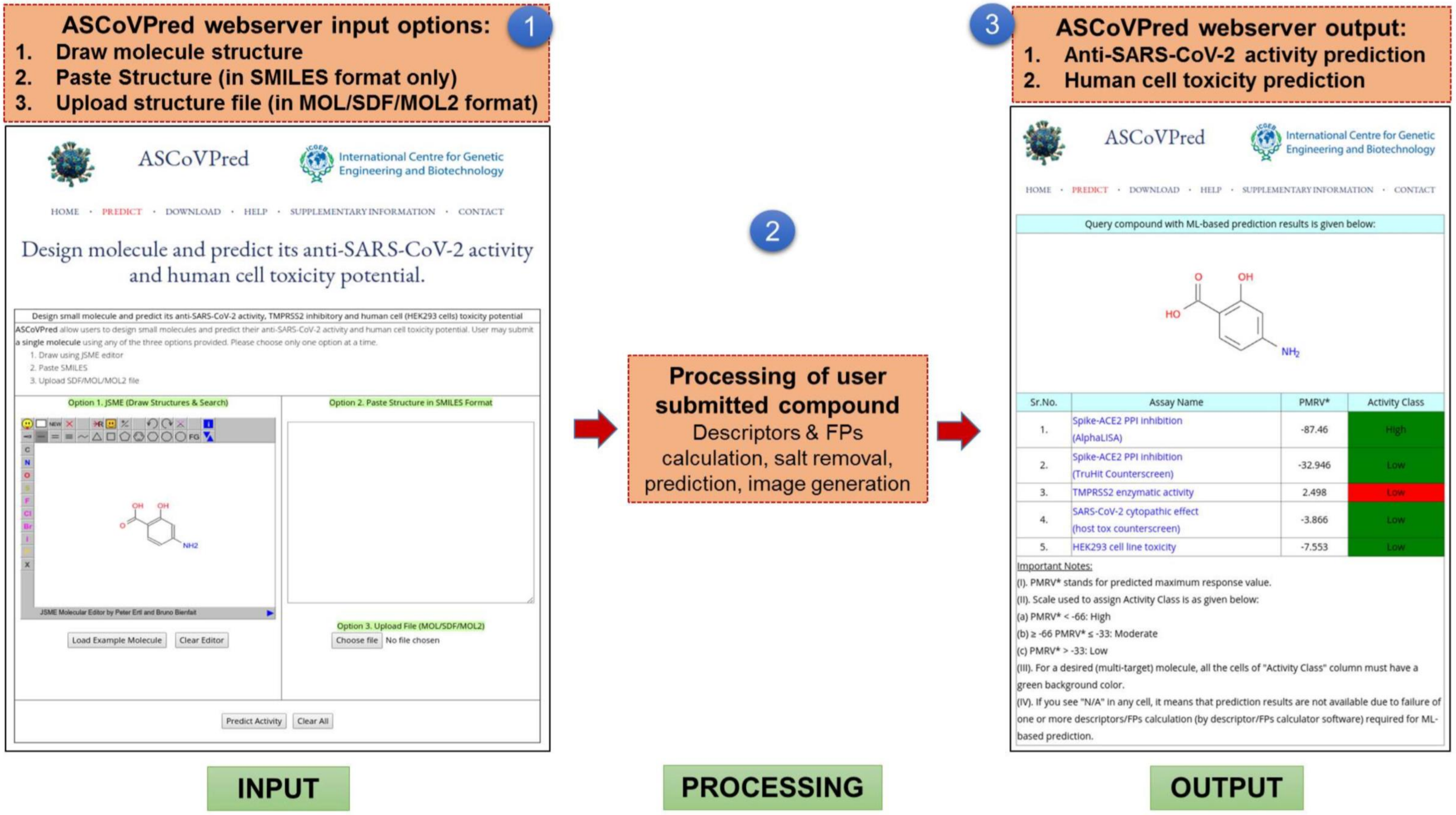
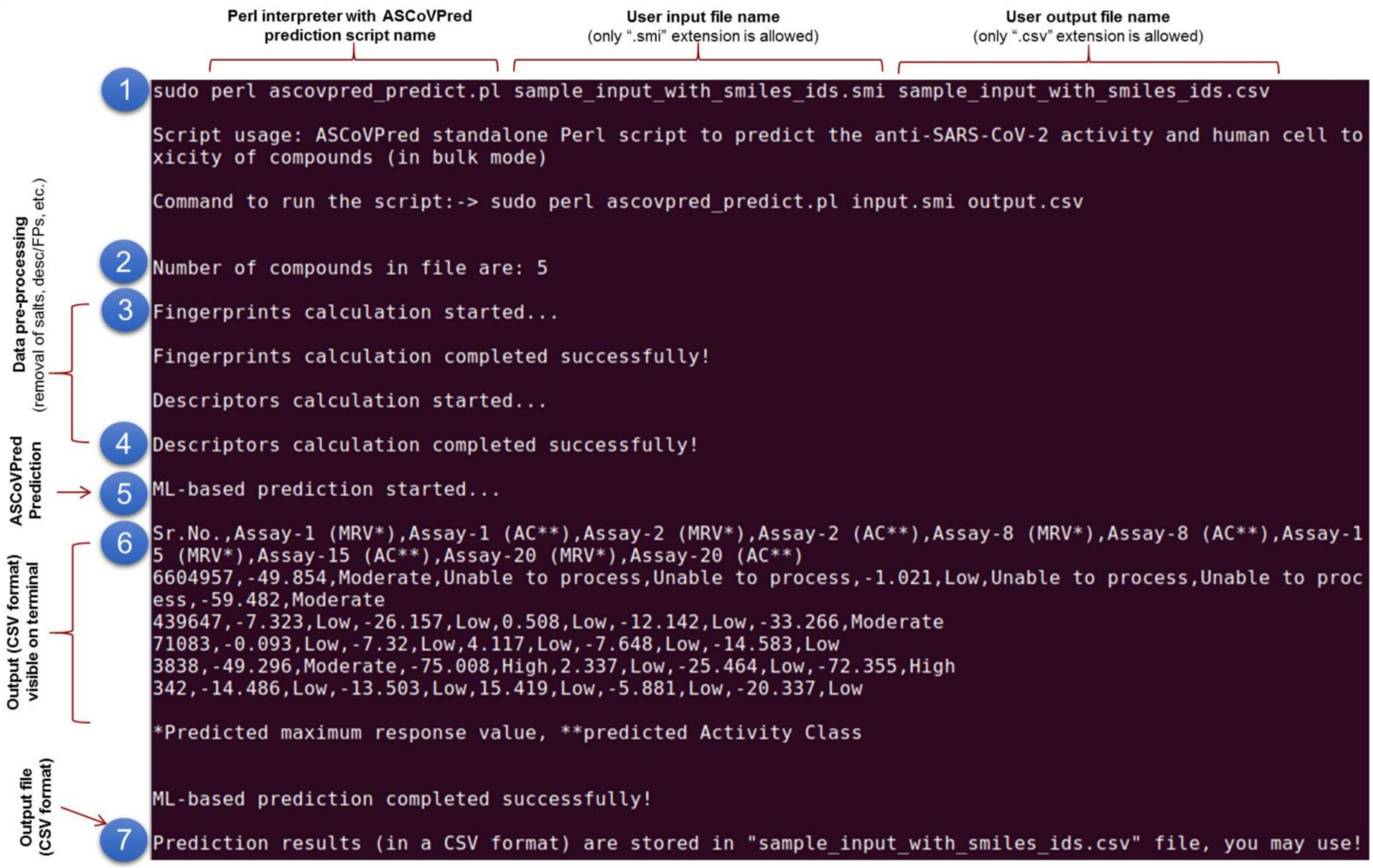


Figure 6. Screenshots of ASCoVPred standalone software usage. Website link: http://192.168.5.81/ascovpred/index.html



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.

- Functional groups associated with the anti-SARS-CoV-2 activity of the molecules may provide better insights while designing the better lead 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.

Acknowledgements

Dr. Dinakar M. Salunke
Director, ICGEB New Delhi,
Group Leader, Structural Immunology,
International Centre for Genetic Engineering and Biotechnology (ICGEB),
ICGEB Campus, Aruna Asaf Ali Marg- 110 067 New Delhi, India
E-mail: icgeb.director@gmail.com
Tel: +91-11-26742317

Dr. Dinesh Gupta
Group Leader,
Translational Bioinformatics Group,
International Centre for Genetic Engineering and Biotechnology (ICGEB),
New Delhi, India-110067.
Tel: +91-11-26742184
E-mail address: dinesh@icgeb.res.in

Department of Biotechnology (DBT),
6th-8th Floor, Block 2, CGO Complex,
Lodhi Road, New Delhi, India -110003

Hands-on demonstration

By

Mr. Neeraj Chaturvedi,
TBG, ICGEB, New Delhi.

Website link: http://192.168.5.81/ascovpred/index.html


Thank you