Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.202000113

QSAR Modeling of SARS-CoV M^{pro} Inhibitors Identifies Sufugolix, Cenicriviroc, Proglumetacin, and other Drugs as Candidates for Repurposing against SARS-CoV-2

Vinicius M. Alves[+],[a] Tesia Bobrowski[+],[b] Cleber C. Melo-Filho,[b] Daniel Korn,[b, c] Scott Auerbach,[d] Charles Schmitt,[a] Eugene N. Muratov,*[b, e] and Alexander Tropsha*[b]

Abstract: The main protease (M^{pro}) of the SARS-CoV-2 has been proposed as one of the major drug targets for COVID-19. We have identified the experimental data on the inhibitory activity of compounds tested against the closely related (96% sequence identity, 100% active site conservation) M^{pro} of SARS-CoV. We developed QSAR models of these inhibitors and employed these models for virtual screening of all drugs in the DrugBank database. Similarity searching and molecular docking were explored in parallel, but docking failed to correctly discriminate between experimentally active and inactive compounds, so it was not relied upon for prospective virtual screening. Forty-two compounds were identified by our models as consensus computational hits. Subsequent to our computational studies, NCATS reported the results of experimental screening of their drug collection in SARS-CoV-2 cytopathic effect assay (https://opendata.ncats.nih.gov/covid19/). Coincidentally, NCATS tested 11 of our 42 hits, and three of them, cenicriviroc (AC_{50} of 8.9 μM), proglumetacin (tested twice independently, with AC_{50} of 8.9 μM and 12.5 μM), and sufugolix (AC_{50} 12.6 μM), were shown to be active. These observations support the value of our modeling approaches and models for guiding the experimental investigations of putative anti-COVID-19 drug candidates. All data and models used in this study are publicly available via Supplementary Materials, GitHub (*https://github.com/alvesvm/sars-cov-mpro*), and Chembench web portal (*https://chembench.mml.unc.edu/*).

Keywords: SARS-CoV-2 · drug repurposing · cheminformatics · virtual screening · SARS-CoV-2 M^{pro}.

Introduction

On December 8th, 2019, the Chinese health authorities in Wuhan detected the first case of an infection caused by a novel coronavirus named SARS-CoV-2.[1,2] On January 31, less than two months later, the World Health Organization declared the SARS-CoV-2 outbreak a global health emergency.[3] SARS-CoV-2 is in the same family as the notorious human coronaviruses SARS-CoV (severe acute respiratory syndrome coronavirus) and MERS-CoV (Middle East respiratory syndrome coronavirus), which have reported fatality rates of 15% and 35%, respectively.[4,5] Current (as of July 7, 2020) estimates of the fatality rate of COVID-19 still vary per age cohort, but most recent estimates agree with an average of 0.6% fatality rate for the total population and 5.6% for people aged 65 and older.[6] To date, the virus is estimated to have infected over ten million people,[7] though these statistics are likely underrepresentative due to established asymptomatic transmission of the disease, or underreporting or lack of testing by health authorities.[8] While the fatality rate of SARS-CoV-2 is estimated to be less than that of SARS and MERS-CoV, it has been shown to be highly transmissible, infecting the first 1.000 patients in only 48 days, whereas SARS took 130 days and MERS took 2.5 years to infect a similar number of people.[9] The initial velocity of the spread of SARS-CoV-2 was enough to hint at pandemic potential at the start of the outbreak, and now millions of cases and over half a million deaths have been reported worldwide despite strict

[a] *V. M. Alves,[+] C. Schmitt*
Office of Data Science,National Toxicology Program, NIEHS, Morrisville, NC, 27560, USA
[b] *T. Bobrowski,[+] C. C. Melo-Filho, D. Korn, E. N. Muratov, A. Tropsha*
Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, Beard Hall, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA
Telephone: (919) 966-2955
FAX: (919) 966-0204
E-mail: murik@email.unc.edu
alex_tropsha@unc.edu
[c] *D. Korn*
Department of Computer Science, University of North Carolina, Chapel Hill, NC, 27599, USA
[d] *S. Auerbach*
Toxinformatics Group, National Toxicology Program, NIEHS, Morrisville, NC, 27560, USA
[e] *E. N. Muratov*
Department of Pharmaceutical Sciences, Federal University of Paraiba, Joao Pessoa, PB, Brazil
[+] *These authors contributed equally.*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202000113

quarantine and travel protocols set in place in many countries.

As of this writing, no vaccines exist against SARS-CoV-2 or past epidemic betacoronaviruses, which represents a larger-scale paucity of data on this genus of viruses. Genomic sequences of the SARS-CoV-2 continue to be uploaded to GenBank, hosted by the National Center for Biotechnology Information (NCBI), and there are over 16,244 distinct sequences listed there to date.[10] An early study investigating compounds with anti-SARS-CoV-2 activities tested seven compounds total and reported four hits, most notably remdesivir and chloroquine.[11] Other early studies have reported other compounds with anti-SARS-CoV-2 activities, such as ivermectin[12] and β-D–N4-hydroxycytidine (NHC, EIDD-1931).[13] A myriad of COVID-19 clinical trials are being performed to repurpose existing experimental nucleoside analogs such as remdesivir, ribavirin, and favipiravir, which have all demonstrated antiviral activities in the past.[15] In May 2020, the FDA approved remdesivir for use in patients with COVID-19 on the basis of these results, stating that "it is reasonable to believe that the known and potential benefits of remdesivir[14] outweigh the known and potential risks of the drug for the treatment of patients hospitalized with severe COVID-19."[15] More recently, dexamethasone, a corticosteroid with anti-inflammatory activity, has been shown to reduce deaths by one-third in patients receiving invasive mechanical ventilation.[16]

Past research has identified several targets for coronavirus drug development, namely nonstructural protein 14 (nsp14-ExoN) and the proteins involved in the coronaviral RNA replication process (replicase polyprotein 1ab and M^{pro}).[17] Replicase polyprotein 1ab is responsible for the synthesis of the large, functional polyproteins pp1a and pp1ab, which are precursors of 16 non-structural proteins that are important in the replication of coronavirus RNA.[18–20] The replicase polyprotein 1ab (CHEMBL5118) is a precursor of 16 non-structural proteins,[21] such as RNA polymerase, helicase, 3′–5′ exonuclease, and 2′-O-ribose methyltransferase. The polyprotein 1ab along with polyprotein 1a are precursors of all proteins that form the viral replication complex (e.g., 1ab has 7,095 aminoacids). These are not functional until viral proteases (M^{pro} and papain-like proteinase) cleave them into 16 distinct proteins.[20] M^{pro} is integral to the proteolytic processing of these polyproteins and is highly conserved in coronaviruses, as are the cleavage sites and lengths of the polyproteins themselves.[19,20,22] The first protein crystal structure for SARS-CoV-2 – the SARS-CoV-2 main protease (also known as 3 C-like protease or M^{pro}) in complex with an inhibitor N3 (PDB ID: 6LU7)[23] – was deposited in the Protein Data Bank in February 2020. This target has been considered before in the design of anti-coronaviral compounds, as demonstrated in a 2012 study by Kim et al.[24], which reported *in vitro* inhibition of SARS-CoV replication by M^{pro} inhibitors.[19]

Since the outbreak began, a massive influx of computational papers identifying possible antiviral drug repurposing candidates have been published in both peer-reviewed and arXiv preprint servers but for the most part, without any experimental confirmatory studies. Herein, we curated available open-source data on SARS-CoV-2 and SARS-CoV and employed both structure- and ligand-based computational approaches to select a set of compounds with the potential to inhibit SARS-CoV-2 replication by inhibiting M^{pro}. In this initial investigation, we have exclusively focused on FDA approved medications and experimental/investigational compounds because these could be quickly repurposed as COVID-19 treatments if their experimental validation were successful. Luckily, upon completion of our computational modeling studies, NCATS tested 11 of our 42 hits, and three of them, cenicriviroc (AC_{50} of 8.9 μM), proglumetacin (tested twice independently, with AC_{50} of 8.9 μM and 12.5 μM), and sufugolix (AC_{50} 12.6 μM), were shown to be active, supporting the value of our modeling approaches and models for guiding the experimental investigations of putative anti-COVID-19 drug candidates.

Materials and Methods

Our study design is shown in Figure 1.

Quantitative Structure-activity Relationship (QSAR) Modeling

(a) Data Collection and Curation

(i) M^{pro} SARS-CoV Data

We collected 201 data points for the SARS-CoV M^{pro} (ChEMBL Assay ID: CHEMBL3927[25]) containing IC_{50} and K_i values. After curation, 91 compounds (27 actives and 64 inactives, considering a threshold of 10 μM) were kept. Then, we found 22 additional compounds in the Protein Data Bank[26] (13 actives and 9 inactives) that were not available in ChEMBL. In the end, 113 compounds (40 actives and 73 inactives) were kept for model development. The modelability index[27] for this dataset was 0.75, indicating that we could progress with binary QSAR model development, despite the heterogenicity of the data.[28] All chemical structures and corresponding biological information were carefully standardized using Standardizer v.20.8.0 (ChemAxon, Budapest, Hungary, http://www.chemaxon.com) according to the protocols proposed by Fourches and colleagues.[29,30] Briefly, inorganics, counterions, metals, organometallic compounds, and mixtures were removed. In addition, specific chemotypes such as aromatic rings and nitro groups were normalized. Furthermore, we excluded duplicates as follows: (i) if duplicates had different biological activity, both entries were excluded; and (ii) if the reported outcomes for the duplicates were the same, one entry was retained in the dataset and the other excluded.

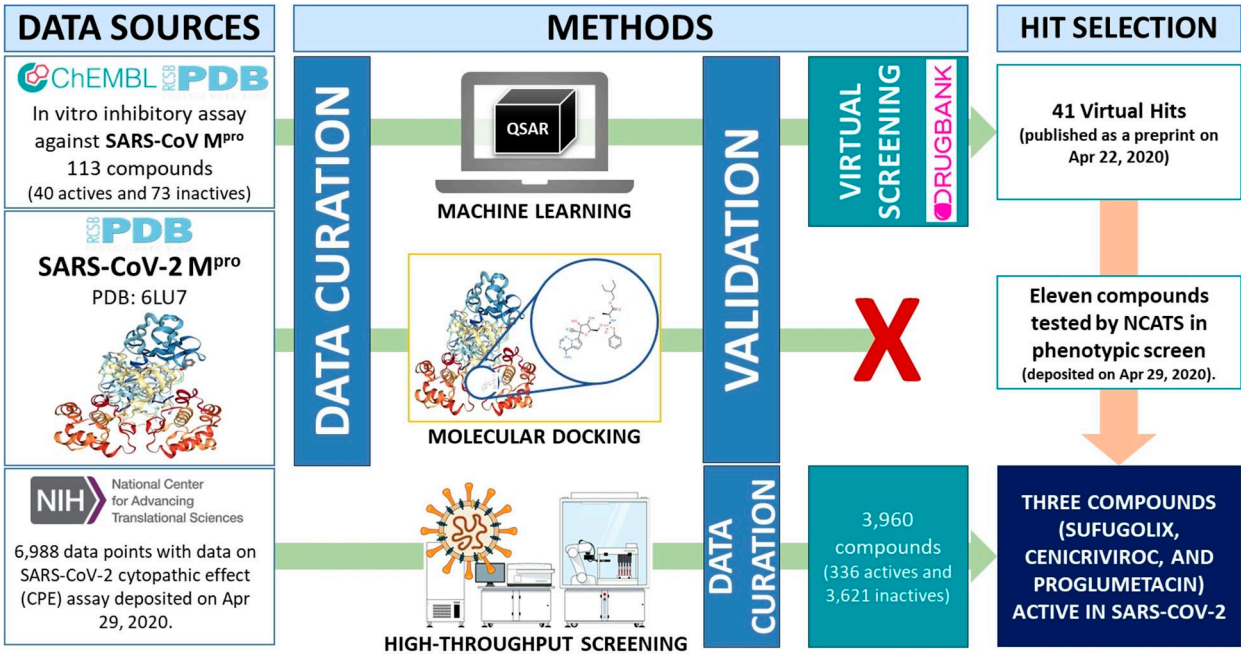

Figure 1. Study design.

The curated dataset is available in the Supplementary Materials (Table S1), GitHub (*https://github.com/alvesvm/sars-cov-mpro*), Chembench web portal (*https://chembench.mml.unc.edu/*).[31]

(ii) SARS-CoV-2 Cytopathic Effect Data

While this manuscript was under review, the National Center for Advancing Translational Sciences (NCATS) released SARS-CoV-2 cytopathic effect (CPE) assay data for the NCATS Pharmaceutical Collection, containing 6,988 data points for compounds that have been approved for clinical use in the U.S., European Union, Japan, Australia, and Canada authorities and also are suitable for high-throughput screening.[32] Per the reviewer's suggestion, we analyzed these data and compared them with the predictions from our models. We also collected and curated the counter screen data to make sure compounds identified as active in the primary assay were not cytotoxic, *i.e.*, that their antiviral effect was not due to killing the host cell. At the NCATS OpenData Browser, compounds with AC_{50} below 12.6 μM ($-\log AC_{50} = 4.9$) and curve-class 3 are considered "low quality actives". We applied the same rule to define active/inactives. After curation, 4,625 (459 actives and 4,166 inactives) small molecules remained in the primary assay. It is worth noting that, the active classification was done solely based on $-\log AC_{50}$ value ($-\log AC_{50} >= 4.9$ active, $-\log AC_{50} < 4.9$ inactive), not fully analyzing the dose-response curves. Compounds with dose-response curve-class 3 are single-point actives and, therefore, additional experiments are needed to confirm their activity.[33] Compounds shown to be cytotoxic to the host cell in the counter screen assay were removed. In the end, 3,957 (336 actives and 3,621 inactives) tested in a phenotypic screen against SARS-CoV-2 were kept for analysis.

(b) Molecular Descriptors

The QSAR models were developed using three types of descriptors: Morgan fingerprints,[34] 2D Simplex Representation of Molecular Structure (SiRMS) descriptors[35], and Dragon (v.7 Kode Chemoinformatics srl – Pisa, Italy). The open-source Morgan fingerprints with 2048 bits and an atom radius of 3 calculated in RDKit (http://www.rdkit.org) using Python 3.6. SiRMS were calculated using HiT QSAR[36] at the 2D level. SiRMS descriptors account not only for the atom type, but also for other atomic characteristics that may impact molecular bioactivity, e.g., partial charge, lipophilicity, refraction, and atom ability for being a donor/acceptor in hydrogen-bond formation (H-bond). A detailed description of HiT QSAR and SiRMS can be found elsewhere.[36] Dragon descriptors were calculated at the 2D level as well. For both SiRMS and Dragon, descriptors with less than 0.01 variance were removed. Correlated descriptors were also removed.

(c) Model Generation

QSAR models were built and rigorously validated following the best practices in the field.[37,38] These models were built using the Random Forest (RF) algorithm[39] implemented in scikit-learn (*http://scikit-learn.org*). Random Forest hyper-parameters were tuned using the GridSearchCV module implemented in scikit-learn. Trees were decorrelated by randomly bootstrapping compound instances used in modeling with replacement and selecting a random sample of the root(N)-many features for each tree, where N is the total number of features available. Trees were configured to evaluate features on classification accuracy at the median value and to use Gini as the split criterion.

A 5-fold external cross-validation procedure was performed using the following protocol. The full set of compounds with known experimental activity is randomly divided into five subsets of equal size. One of these subsets (20% of all compounds) is set aside as the external validation set, while the remaining four sets form the modeling set (80% of all compounds). This procedure is repeated five times, allowing each of the five subsets to be used as an external validation set. Models are built using the training set only, and it is essential to emphasize that compounds are never simultaneously part of both the training and external validation set.

Two types of consensus predictions were performed: (i) assigning compound class (i.e., active or inactive) based on the majority vote across three independent models developed with Morgan, SiRMS, and Dragon descriptors, i.e., assigning a compound a class that at least two models agreed on. (ii) Consensus AD, i.e., the same as above but in addition, requiring that compounds were within the applicability domain of each model. The local (tree) applicability domain approach[40] setting a threshold of 70% was used for all RF models developed in this study.

Molecular Docking

Molecular docking experiments were performed using the structure of M^{pro} from SARS-CoV-2 (PDB ID: 6LU7). To enable these calculations, the structure was processed using the Protein Preparation Wizard module of Maestro v.12.0.12[41] under pH 7.0 ± 2.0 and optimized with OPLS3e force field. All ligands were prepared under the same conditions in the LigPrep module and submitted to molecular docking using Glide[42] with the standard precision (SP) option.

Similarity Search

A similarity search was performed in the KNIME platform (https://www.knime.com/) using Morgan fingerprints. Three compounds described by Wang et al.[11] as actives in the phenotypic screen (remdesivir, chloroquine, and nitazoxanide) were employed as queries. A Tanimoto similarity threshold of 75% was employed to select compounds from DrugBank as putative actives.

Results and Discussion

The main goal of this study was to find drugs that could be repurposed for SARS-CoV-2. To this end, we curated open-source data on M^{pro} inhibitors for both SARS-CoV-2 and SARS-CoV. We also employed both structure- and ligand-based computational approaches to select a set of compounds that may have the potential to inhibit SARS-CoV-2 replication by inhibiting M^{pro}. In this initial investigation, we have exclusively focused on FDA approved medications or experimental/investigational compounds because these could be quickly repurposed as COVID-19 treatments if their experimental validation is successful. Before submitting this manuscript for peer-review, we deposited a preprint version online on April 22, 2020[43], and later, when NCATS released screening data on April 29, 2020, we had the opportunity to validate our predictions as reported in this paper.

As shown in Figure 1, we employed three different computational strategies to screen DrugBank to find the drug repurposing candidates against SARS-CoV-2: QSAR models, docking, and similarity searching. We started by collecting all publicly available data on SARS-CoV-2 and other coronaviruses and focused on M^{pro} as a critical target for SARS-CoV-2 replication. Using Basic Local Alignment Search Tool (BLAST) available in UniProt (https://www.uniprot.org/blast/),[44] we observed that the primary sequences of M^{pro} in both SARS-CoV and SARS-CoV-2 had 96% identity (Figure 2a). The crystal structure of SARS-CoV-2 M^{pro} was recently elucidated and superposition of the respective 3D protein structures (PDB IDs: 5N19, 6LU7) revealed a conserved binding site around the co-crystallized inhibitors including the catalytic dyad represented by His41 and Cys145 (Figures 2b and 2c).[23] This level of conservation makes M^{pro} a particularly attractive target, as compounds inhibiting this protease could possibly serve as antivirals against other pathogenic coronaviruses such as SARS-CoV-1 and MERS-CoV..

The 113 compounds (40 actives and 73 inactives) kept after curation were used for binary QSAR modeling. It is worth noting that PubChem has a large library of 290.893 compounds tested in QFRET-based primary biochemical high throughput screening assay to identify inhibitors of the SARS-CoV M^{pro} (PubChem AID: 1706). A recent paper has modeled these data[45] showing high accuracy (recall = 64% and specificity = 84%). We collected the 140 data points from seven confirmatory assays for this run IC_{50} values (PubChem AIDs: 1890, 488958, 488999, 588771, 588786, 602486). After curation, 135 compounds (60 actives and 75 inactives, considering a threshold of 10 μM) were kept. There was just one compound matching the original dataset, presenting good agreement with IC_{50} of 6.2 and

(a)

```
SARS-CoV-2   1 SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDML
SARS-CoV     1 SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDML

SARS-CoV-2  51 NPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPK
SARS-CoV    51 NPNYEDLLIRKSNHSFLVQAGNVQLRVIGHSMQNCLLRLKVDTSNPKTPK

SARS-CoV-2 101 YKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGF
SARS-CoV   101 YKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNHTIKGSFLNGSCGSVGF

SARS-CoV-2 151 NIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAGTDTTI
SARS-CoV   151 NIDYDCVSFCYMHHMELPTGVHAGTDLEGKFYGPFVDRQTAQAAGTDTTI

SARS-CoV-2 201 TVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDIL
SARS-CoV   201 TLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDIL

SARS-CoV-2 251 GPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQC
SARS-CoV   251 GPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQC

SARS-CoV-2 301 SGVTFQ
SARS-CoV   301 SGVTFQ
```

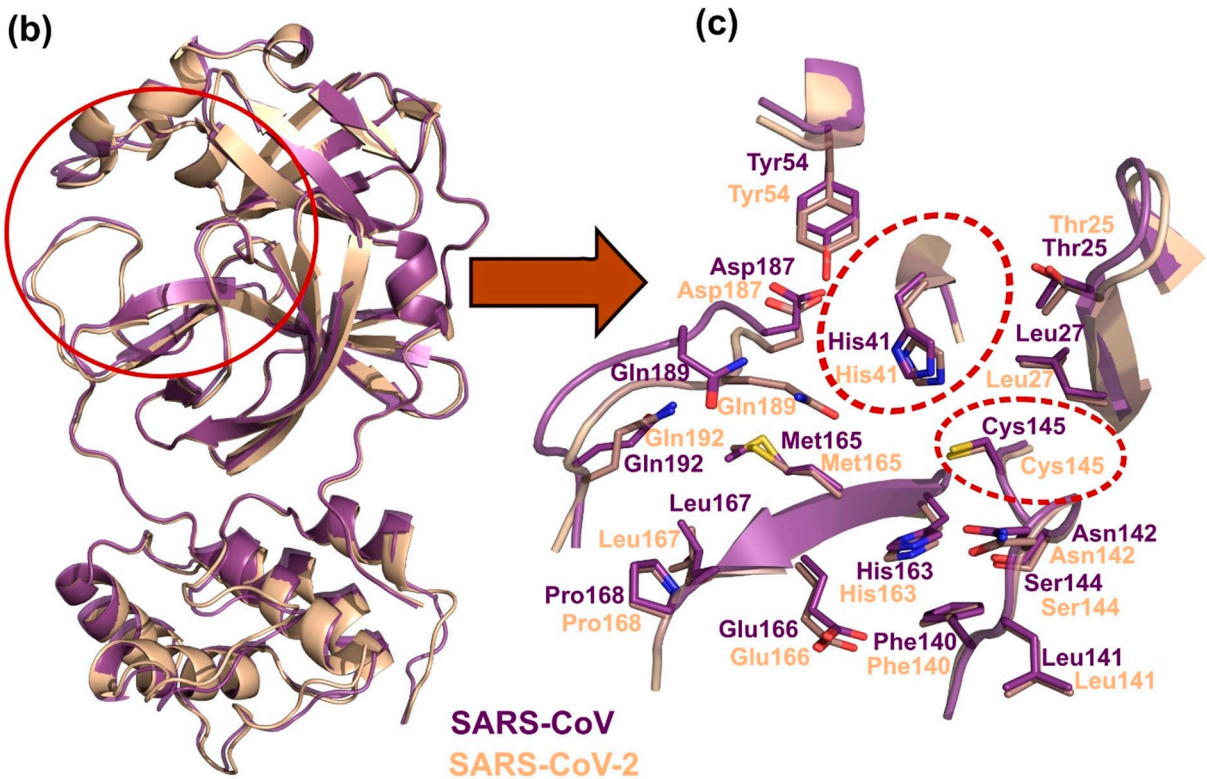

Figure 2. Alignment of SARS-CoV and SARS-CoV-2 M^{pro} monomers. **(a)** Primary sequence alignment highlighting the conserved residues in bold font. The binding site residues are shown in red and the catalytic dyad, represented by His41 and Cys145, is marked with asterisks. **(b)** Alignment of M^{pro} monomers available in PDB (IDs: 5N19, 6LU7). **(c)** Visualization of the overlap between residues at the M^{pro} active site for SARS-CoV and SARS-CoV-2. The red dashed circles show the conserved catalytic dyad and the remarkable conservation of the binding site of M^{pro} between the coronaviruses.

6.3, respectively. In the end, 59 additional actives and 75 inactives were found. However, this data was shown to not be modelable either alone or in addition to the data collected from ChEMBL and PDB, presenting sensitivity lower than 60% and CCR around 50% in all the cases.

Therefore, QSAR models were developed employing only ChEMBL and PDB data. The statistical characteristics of our QSAR models are shown in Table 1. Models were validated using 5-fold external cross-validation protocols. They achieved external correct classification rate (CCR) of 71–83% (sensitivity = 55–72%, positive predicted value [PPV] = 72–100%, specificity = 88–100%, negative predicted value [NPV] = 78–85%). Models were generated with the entire (unbalanced) dataset. Although the models were biased towards the larger inactive class and their sensitivity was barely acceptable[39] for all but Dragon models, high PPV values demonstrated their utility for virtual screening. Overall, the statistical characteristics of the developed models suggested that a small number of hits would be found, but there was also a high confidence that they would be active.

Table 1. Statistical characteristics of QSAR models for SARS-CoV M^{pro} assessed by 5-fold external validation.

Model	CCR	Sensitivity	PPV	Specificity	NPV	Coverage
Morgan	0.78	0.65	0.81	0.92	0.83	1.00
Morgan AD	0.80	0.62	0.94	0.98	0.85	0.69
SiRMS	0.76	0.65	0.72	0.86	0.82	1.00
SiRMS AD	0.83	0.72	0.86	0.93	0.85	0.61
Dragon	0.71	0.55	0.71	0.88	0.78	1.00
Dragon AD	0.78	0.56	1.00	1.00	0.87	0.54
Consensus	0.74	0.60	0.73	0.88	0.80	1.00
Consensus (AD)	0.78	0.62	0.86	0.95	0.83	0.77

As one can see in Table 1, all the models had similar statistical characteristics. Due to the diversity and small size of the dataset, we decided to build QSAR models using different types of descriptors, and then, use these different models in consensus QSAR modeling.[46,47] This approach was shown to have, on average, higher reliability than any of the individual models.[27,48–51] In addition, this approach was shown to be helpful for chemogenomics data curation.[30,52] Despite similar statistical accuracy of predictions, the contributing models had a certain disagreement in assigning the activity class when used for virtual screening. This is sensible due to the different features of these descriptors. SiRMS[35] are fragment-based descriptors that also consider several physical-chemical parameters, such as atoms' lipophilicity and partial charges, while Dragon[53] represents a collection of whole-molecule descriptors (constitutional, topological, geometric, etc.). Lastly, Morgan descriptors are a type of Extended-Connectivity Fingerprints (ECFPs) that encode fragments as binary bit strings.[54] These descriptors have become very popular due to their presence in RDKit packages for both Python and KNIME, widely used cheminformatics workbenches. However, these descriptors are prone to fragment collisions that can result in different molecules having the same descriptors, ultimately adding noise to and affecting the model negatively.[55] We, and others, have shown that model accuracy is influenced much stronger by data quality and accuracy, rather than by the choice of molecular descriptors and machine learning algorithms.[35] To increase the accuracy of the predictions, as expected in consensus QSAR modeling, we regarded compounds as predicted active when at least two models agreed on their activity class assessment. In addition, if two models agreed as active with low majority voting of local trees from RF (< 70%), while the third model predicted the compound as inactive with high confidence (> 70%), then this compound was assigned as inactive. This led to identifying 42 compounds as predicted active hits in the DrugBank library.

Wang et al.[11] demonstrated that remdesivir and chloroquine were highly active; nitazoxanide was moderately active; and ribavirin, penciclovir, nafamostat, faviparir were inactive against SARS-CoV-2 in phenotypic assays *in vitro*. The SiRMS models predicted remdesivir and ribavirin as active, while Dragon predicted ribavirin only. Currently, there is no evidence that any of these compounds act on M^{pro}; remdesivir is a known RNA-dependent RNA polymerase inhibitor.[56]

Jin et al.[57] submitted a library of ~10.000 compounds to high-throughput screening (HTS) and identified six inhibitors of SARS-CoV-2 M^{pro}, namely, ebselen, disulfiram, tideglusib, carmofur, shikonin, and PX-12. After additional phenotypic assays, only ebselen inhibited *in vitro* viral replication. Despite the large number of compounds tested in HTS, only the activity of those six inhibitors was reported, so there was no publicly available data on SARS-CoV-2 $M^{pro,}$ yet that could enable the development of QSAR models.

Due to the small amount of publicly available SARS-CoV-2 M^{pro} assay data and the high level of similarity (96% sequence identity between M^{pro} of SARS-CoV and SARS-CoV-2, including fully conserved active site (see above), we hypothesized that compounds predicted to be active in the SARS-CoV M^{pro} assay[58] (using models developed with data from this assay) are likely to be active against SARS-CoV-2.

In addition, we also predicted M^{pro} activity for twenty-three compounds reported to be undergoing clinical trials (as of March 23, 2020)[49] (See Table S1 in Supplementary Materials). Of these compounds, lopinavir, ritonavir, tetrandrine, cobicistat, losartan, ribavirin, remdesivir, aviptadil, and danoprevir were predicted as active by SiRMS models. Lopinavir was also predicted as active by models built with Dragon descriptors. None of the molecules were predicted as active by models based on Morgan descriptors. Lopinavir is an established protease inhibitor approved for use in HIV patients and is usually used in combination with ritonavir, another protease inhibitor.[59] Lopinavir and lopinavir/ritonavir have been tested previously on SARS-CoV[60] and

MERS-CoV,[61] but some evidence from clinical trials suggests that the drug combination is not as successful as hoped for treating COVID-19.[62]

Since no data was available to build models for SARS-CoV-2 M^{pro}, and due to the high degree of similarity between this protein and its analog in SARS-CoV, we decided to employ models built with SARS-CoV M^{pro} data to virtually screen the curated DrugBank dataset and consider hits identified with these models as active against SARS-CoV-2 M^{pro}. Applying our models to screen this dataset of 9,615 compounds yielded 42 compounds predicted as actives using Consensus AD models.

In parallel, we have also conducted molecular docking experiments using the structure of M^{pro} from SARS-CoV-2 (PDB ID: 6LU7).[23] Before using docking as a virtual screening tool, it is crucial to validate the approach with known experimental data. Therefore, known inhibitors and non-inhibitors of M^{pro} were used to evaluate if the docking score was capable of ranking active compounds better than inactives. For this purpose, the curated dataset (CHEMBL3927[26]), used for QSAR modeling, and three compounds described by Wang et al.[10] as active against SARS-CoV-2, were employed in a docking validation run. Then, compound ranking by the docking score was compared with ranking by activity in the ChEMBL assay. We found that docking scores were poorly correlated with the binding affinity as indicated by the area under the receiver operating characteristic (ROC) score of 0.49. Additionally, the early enrichment was poor with a sensitivity of only 0.11 for the top 10% ranked compounds, i.e., actives were ranked poorly while inactives were occupying the top of the list of virtual hits. The top 15% also presented poor sensitivity (0.14). Only after the top 69% of the list was considered, the sensitivity reached reasonable values (0.70). Based on these results, docking was discarded as a viable virtual screening approach as applied to M^{pro}.

We also employed a similarity search using three compounds described by Wang et al.[11] as active in the phenotypic screen (remdesivir, chloroquine, and nitazoxanide). We found that only the following 13 compounds from the curated DrugBank dataset had Tanimoto similarity coefficient higher than 75% to any of those three drugs: anhydrovinblastine, GS-6620, hydroxychloroquine, lurbinectedin, quinacrine, quinacrine mustard, rifalazil, vinblastine, vincristine, vindesine, vinflunine, vinorelbine, and 3′′-(beta-chloroethyl)-2′′,4′′-dioxo-3, 5′′-spiro-oxazolidino-4-deacetoxy-vinblastine. Most of these compounds are vinca alkaloids. The literature on this class of alkaloids concerns cancer biology since many are chemotherapy drugs, but other classes of alkaloids have been noted to have antiviral activities.[54–57]

In total, we selected 42 hits from DrugBank predicted by at least two out of the three QSAR models built independently with SiRMS, Dragon, and Morgan descriptors; this list included four compounds also identified by similarity search: lurbinectedin, rifalazil, vinblastine and 3′′-(beta-chloroethyl)-2′′,4′′-dioxo-3, 5′′-spiro-oxazolidino-4-deacetoxy-vinblastine as active. Our hits have been found among commercially available compounds listed in the ZINC database[63] and the vendors selling these compounds were identified using our in-house ZINC Express software (https://zincexpress.mml.unc.edu/) (Table S1 and Table S2 in Supplemental Materials).

A preprint version of this manuscript, including a selection of computational hits, was deposited online in ChemRxiv on April 22, 2020.[43] A week later, on April 29, 2020, NCATS released, via OpenData Portal,[32] the quantitative HTS data on drugs approved for clinical use tested in the SARS-CoV-2 CPE assay. This assay measures live virus infectivity, and thus, it is expected to be sensitive to M^{pro} inhibitors. We downloaded the SARS-CoV-2 cytopathic effect (CPE) tsv file from https://opendata.ncats.nih.gov/covid19/assays and standardized the chemical structures following our traditional workflow.[29] We found that 11 out of 42 compounds identified by our models as computational hits were tested by NCATS. Three of them, sufugolix (annotated as NCGC00509985-02), cenicriviroc, and proglumetacin, were shown to be active. Sufugolix and cenicriviroc had AC_{50} of 12.6 μM and 8.9 μM, respectively, and proglumetacin, which was independently tested twice and had two associated records with AC_{50} of 12.6 and 8.9 μM. The remaining nine compounds (atazanavir, barasertib, indinavir, lurbinectedin, navitoclax, tilmicosin, venetoclax, and vinblastine) were inactive. The summary of results reported by NCATS for these 11 compounds is given in Table 2. We also used our binary models to predict activity classes for all remaining compounds from the NCATS CPE assay. Formally, the overall model prediction accuracy was 1% for the experimentally active compounds and 99% for inactive compounds (See Table 3). The additional compound predicted correctly as active, and not present in DrugBank was LDN-57444, an inhibitor of ubiquitin carboxyl-terminal hydrolase isozyme L1 (human). One should keep in mind that our models were built to predict inhibitors of M^{pro} whereas the experiments conducted used a phenotypic assay, which does not indicate activity against a specific protein target, only the presence or absence antiviral activity. This assay is expected to be sensitive to M^{pro} inhibition. Still, the screening results in these two assays may disagree, as is often observed when comparing enzymatic and phenotypic assay results. Thus, our working hypothesis, which can be evaluated if, and when, testing results for M^{pro} inhibition become available, is that the observed phenotypic effects are due to the M^{pro} mediated mechanism. However, it is important to emphasize that observing activity in the phenotypic assay is a highly valuable result prompting further investigation of hit compounds *in vivo*. Normalized chemical structures, predictions, and experimental results for all the non-cytotoxic compounds obtained from NCATS OpenData Portal[32] are available in the Supplementary Materials (Table S3).

Table 2. List of 11 hits selected by QSAR models and tested by NCATS.

Drug NAME	NCATS ID	DrugBank recorded status	CPE (μM)
Cenicriviroc	NCGC00685392-02	investigational	8.9
Sufugolix	NCGC00509985-02	investigational	12.6
Proglumetacin	NCGC00183024-01	experimental	12.5 and 8.91
Atazanavir	NCGC00182552-02	approved; investigational	Inactive
Barasertib	NCGC00378734-08	investigational	Inactive
Indinavir	NCGC00159460-01	approved	Inactive
Lurbinectedin	NCGC00510477-01	investigational	Inactive
Navitoclax	NCGC00188344-07	investigational	Inactive
Tilmicosin	NCGC00348375-01	investigational; vet_approved	Inactive
Venetoclax	NCGC00345789-05	approved; investigational	Inactive
Vinblastine	NCGC00263548-23	approved	Inactive

Table 3. Confusion matrix showing performance of consensus QSAR models of SARS-CoV M^{pro} to predict compounds with antiviral activities against SARS-CoV-2.

	Inactive	Active	Total
Inactive	3586	35	3621
Active	332	4	336
Total	3917	40	3957

Conclusions

High similarity of the M^{pro} active site in both SARS-CoV and SARS-CoV-2 is in good agreement with high conservation of M^{pro} among coronaviruses that has been noted in the previous studies.[15] Therefore, we utilized previous experimental data on SARS-CoV M^{pro} to develop a QSAR model for virtual screening of the DrugBank database in the search for candidates for drug repurposing against SARS-CoV-2. Despite the availability of a crystal structure of SARS-CoV-2 M^{pro} in complex with inhibitor N3, molecular docking was not sufficient to discriminate between experimental actives and inactives and was ultimately not used to select hits. As a result of the virtual screening of DrugBank library, we identified 42 virtual hits including several compounds currently being tested in clinical trials such as lopinavir and ritonavir.

The 42 virtual hits were analyzed for availability and price using our in-house ZINC Express software (https://zincexpress.mml.unc.edu/) (see Table S2) and described in the preprint of this paper.[43] One week after we deposited our predictions online, NCATS released their OpenData Portal including the quantitative HTS data of approved drugs tested in the SARS-CoV-2 CPE assay.[32] Eleven computational hits resulting from our studies were coincidentally tested by NCATS; three of them were active. Sufugolix and cenicriviroc exhibited AC_{50} of 12.6 μM and 8.9 μM, respectively, and proglumetacin was annotated with two data records reporting AC_{50} of 12.5 and 8.9 μM. In addition, our models formally predicted the inactive compounds in CPE assay with 98% accuracy. Overall, our results indicate that QSAR models developed using SARS-CoV M^{pro} data can be used to identify compounds active against SARS-CoV-2.

All collected and curated data, models, and virtual screening results are publicly available in the Supplementary Materials of this paper and at GitHub (*https://github.com/alvesvm/sars-cov-mpro*). The curated data are also available in the Chembench web portal (*https://chembench.mml.unc.edu/*).

Conflict of Interest

None declared.

Acknowledgments

This study was inspired by "Calling all coronavirus researchers" Nature editorial.[27]. It was supported in part by NIH grants 1R01 GM114015 and 1U01CA207160.

References

[1] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, L. Zhang, *Lancet* **2020**, *395*, 507–513.

[2] World Health Organization, "Naming the coronavirus disease (COVID-19) and the virus that causes it," Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it, Accessed Jun 26, 2020, **2020**.

[3] World Health Organization, "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)," Available at: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), Accessed Jun 26, 2020, **2020**.

[4] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, R. Hilgenfeld, *Science* **2003**, DOI 10.1126/science.1085658.
[5] C. A. Donnelly, M. R. Malik, A. Elkholy, S. Cauchemez, M. D. Van Kerkhove, *Emerging Infect. Dis.* **2019**, *25*, 1758–1760.
[6] S. Mallapaty, *Nature* **2020**, *582*, 467–468.
[7] "COVID-19 Map - Johns Hopkins Coronavirus Resource Center," Available at: https://coronavirus.jhu.edu/map.html, Accessed Jun 26, 2020, **2020**.
[8] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, J. Schubert, J. Bania, T. Khosrawipour, *J. Microbiol. Immunol. Infect.* **2020**, *53*, 454–458.
[9] D. Beasley, K. Kelland, "Comparing outbreaks: How the new virus compares to previous coronavirus outbreaks," Available at: https://graphics.reuters.com/CHINA-HEALTH-VIRUS-COMPARISON/0100B5BY3CY/index.html, Accessed Jun 26, 2020, **2020**.
[10] NCBI, "SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences," Available at: https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/, Accessed Jun 26, 2020, **2020**.
[11] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong, G. Xiao, *Cell Res.* **2020**, *30*, 269–271.
[12] L. Caly, J. D. Druce, M. G. Catton, D. A. Jans, K. M. Wagstaff, *Antiviral Res.* **2020**, 104787.
[13] T. P. Sheahan, A. C. Sims, S. Zhou, R. L. Graham, A. J. Pruijssers, M. L. Agostini, S. R. Leist, A. Schäfer, K. H. Dinnon, L. J. Stevens, J. D. Chappell, X. Lu, T. M. Hughes, A. S. George, C. S. Hill, S. A. Montgomery, A. J. Brown, G. R. Bluemling, M. G. Natchus, M. Saindane, A. A. Kolykhalov, G. Painter, J. Harcourt, A. Tamin, N. J. Thornburg, R. Swanstrom, M. R. Denison, R. S. Baric, *Sci. Transl. Med.* **2020**, *12*, eabb5883.
[14] NIH, "NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19 | NIH: National Institute of Allergy and Infectious Diseases," Available at: https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19, Accessed Jun 26, 2020.
[15] Food and Drug Administration, "Remdesivir EUA Letter of Authorization," Available at: https://www.fda.gov/media/137564/download, Accessed Jun 26, 2020, **2020**.
[16] P. Horby, W. S. Lim, J. Emberson, M. Mafham, J. Bell, L. Linsell, N. Staplin, C. Brightling, A. Ustianowski, E. Elmahi, B. Prudon, C. Green, T. Felton, D. Chadwick, K. Rege, C. Fegan, L. C. Chappell, S. N. Faust, T. Jaki, K. Jeffery, A. Montgomery, K. Rowan, E. Juszczak, J. K. Baillie, R. Haynes, M. J. Landray, R. C. Group, *medRxiv* **2020**, 2020.06.22.20137273.
[17] E. C. Smith, H. Blanc, M. Vignuzzi, M. R. Denison, *PLoS Pathog.* **2013**, *9*, e1003565.
[18] G. Heusipp, C. Gro, J. Herold, S. G. Siddell, J. Ziebuhr, *J. Gen. Virol.* **1997**, *78*, 2789–2794.
[19] J. Ziebuhr, S. G. Siddell, *J. Virol.* **1999**, *73*, 177–85.
[20] S. G. Fang, H. Shen, J. Wang, F. P. L. Tay, D. X. Liu, *Virology* **2008**, *379*, 175–180.
[21] Y. Chen, Q. Liu, D. Guo, *J. Med. Virol.* **2020**, *92*, 418–423.
[22] X. Deng, S. E. StJohn, H. L. Osswald, A. O'Brien, B. S. Banach, K. Sleeman, A. K. Ghosh, A. D. Mesecar, S. C. Baker, *J. Virol.* **2014**, *88*, 11886–11898.
[23] X. Liu, B. Zhang, Z. Jin, H. Yang, Z. Rao, "The crystal structure of COVID-19 main protease in complex with an inhibitor N3," DOI 10.2210/PDB6LU7/PDBAvailable at: http://www.rcsb.org/structure/6LU7, Accessed Jun 26, 2020, **2020**.
[24] Y. Kim, S. Lovell, K.-C. Tiew, S. R. Mandadapu, K. R. Alliston, K. P. Battaile, W. C. Groutas, K.-O. Chang, *J. Virol.* **2012**, *86*, 11754–11762.
[25] ChEMBL, "Target Report Card - CHEMBL3927," Available at: https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3927/, Accessed Jun 26, 2020.
[26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–42.
[27] A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, *J. Chem. Inf. Model.* **2014**, *54*, 1–4.
[28] A. A. Lagunin, A. Geronikaki, P. Eleftheriou, P. V. Pogodin, A. V. Zakharov, *J. Chem. Inf. Model.* **2019**, *59*, 713–730.
[29] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, *50*, 1189–204.
[30] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2016**, *56*, 1243–52.
[31] S. J. Capuzzi, I. S.-J. Kim, W. I. Lam, T. E. Thornton, E. N. Muratov, D. Pozefsky, A. Tropsha, *J. Chem. Inf. Model.* **2017**, *57*, 105–108.
[32] NCATS, "SARS-CoV-2 cytopathic effect (CPE)," Available at: https://opendata.ncats.nih.gov/covid19/assay?aid=14, Accessed Jun 26, 2020, **2020**.
[33] NCATS, "NCGC CurveFit," Available at: https://tripod.nih.gov/curvefit/, Accessed Jul 1, 2020.
[34] J. Figueras, *J. Chem. Inf. Model.* **1993**, *33*, 717–718.
[35] E. N. Muratov, A. G. Artemenko, E. V Varlamova, P. G. Polischuk, V. P. Lozitsky, A. S. Fedchuk, R. L. Lozitska, T. L. Gridina, L. S. Koroleva, V. N. Sil'nikov, A. S. Galabov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, V. E. Kuz'min, *Future Med. Chem.* **2010**, *2*, 1205–26.
[36] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, L. Gorb, M. Qasim, J. Leszczynski, *J. Comput.-Aided Mol. Des.* **2008**, *22*, 747–59.
[37] A. Tropsha, *Mol. Inf.* **2010**, *29*, 476–488.
[38] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rev.* **2020**, DOI 10.1039/D0CS00098 A.
[39] L. E. O. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[40] A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. N. Muratov, E. V Varlamova, A. V Kuz'mina, L. G. Gorb, A. Golius, F. C. Hill, J. Leszczynski, A. Tropsha, *SAR QSAR Environ. Res.* **2011**, *22*, 575–601.
[41] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, W. Sherman, *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221–34.
[42] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin, *J. Med. Chem.* **2004**, *47*, 1739–1749.
[43] T. Bobrowski, V. Alves, C. C. Melo-Filho, D. Korn, S. S. Auerbach, C. Schmitt, E. Muratov, A. Tropsha, *ChemRxiv* **2020**, DOI 10.26434/chemrxiv.12153594.
[44] A. Bateman, *Nucleic Acids Res.* **2019**, *47*, D506–D515.
[45] S. Ekins, M. Mottin, P. R. P. S. Ramos, B. K. P. Sousa, B. J. Neves, D. H. Foil, K. M. Zorn, R. C. Braga, M. Coffee, C. Southan, A. C. Puhl, C. H. Andrade, *Drug Discovery Today* **2020**, *25*, 928–941.
[46] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, I. V Tetko, *J. Chem. Inf. Model.* **2008**, *48*, 766–84.
[47] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, Q. Song, *J. Chem. Inf. Model.* **2005**, *45*, 786–799.
[48] X. S. Wang, H. Tang, A. Golbraikh, A. Tropsha, *J. Chem. Inf. Model.* **2008**, *48*, 997–1013.
[49] V. E. Kuz'min, E. N. Muratov, A. G. Artemenko, E. V. Varlamova, L. Gorb, J. Wang, J. Leszczynski, *QSAR Comb. Sci.* **2009**, *28*, 664–677.

[50] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell.* **2002**, *20*, 269–76.
[51] A. V. Zakharov, T. Zhao, D. T. Nguyen, T. Peryea, T. Sheils, A. Yasgar, R. Huang, N. Southall, A. Simeonov, *J. Chem. Inf. Model.* **2019**, *59*, 4613–4624.
[52] D. Fourches, E. Muratov, A. Tropsha, *Nat. Chem. Biol.* **2015**, *11*, 535–535.
[53] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-WCH, New York, **2009**.
[54] RDKit, "Morgan Fingerprints," Available at: http://rdkit.org/docs/GettingStartedInPython.html#morgan-fingerprints-circular-fingerprints, Accessed Jun 26, 2020, **2020**.
[55] C. Bologa, T. K. Allu, M. Olah, M. A. Kappler, T. I. Oprea, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 625–635.
[56] G. Li, E. De Clercq, *Nat. Rev. Drug Discovery* **2020**, *19*, 149–150.
[57] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, H. Yang, *Nature* **2020**, *582*, 289–293.
[58] ChEMBL, "Target Report Card - CHEMBL612575," Available at: https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL612575/, Accessed Jun 26, 2020.
[59] D. J. Kempf, J. D. Isaacson, M. S. King, S. C. Brun, Y. Xu, K. Real, B. M. Bernstein, A. J. Japour, E. Sun, R. A. Rode, *J. Virol.* **2001**, *75*, 7462–7469.
[60] C. M. Chu, *Thorax* **2004**, *59*, 252–256.
[61] J. F.-W. Chan, Y. Yao, M.-L. Yeung, W. Deng, L. Bao, L. Jia, F. Li, C. Xiao, H. Gao, P. Yu, J.-P. Cai, H. Chu, J. Zhou, H. Chen, C. Qin, K.-Y. Yuen, *J. Infect. Dis.* **2015**, *212*, 1904–1913.
[62] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai, M. Wei, X. Li, J. Xia, N. Chen, J. Xiang, T. Yu, T. Bai, X. Xie, L. Zhang, C. Li, Y. Yuan, H. Chen, H. Li, H. Huang, S. Tu, F. Gong, Y. Liu, Y. Wei, C. Dong, F. Zhou, X. Gu, J. Xu, Z. Liu, Y. Zhang, H. Li, L. Shang, K. Wang, K. Li, X. Zhou, X. Dong, Z. Qu, S. Lu, X. Hu, S. Ruan, S. Luo, J. Wu, L. Peng, F. Cheng, L. Pan, J. Zou, C. Jia, J. Wang, X. Liu, S. Wang, X. Wu, Q. Ge, J. He, H. Zhan, F. Qiu, L. Guo, C. Huang, T. Jaki, F. G. Hayden, P. W. Horby, D. Zhang, C. Wang, *N. Engl. J. Med.* **2020**, *382*, 1787–1799.
[63] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177–82.
[64] PubChem, "CID 16760696," Available at: https://pubchem.ncbi.nlm.nih.gov/compound/16760696, Accessed Jul 7, 2020.

Received: May 10, 2020
Accepted: July 28, 2020
Published online on August 24, 2020

