Check for updates

A machine learning platform to estimate anti-SARS-CoV-2 activities

Govinda B. KC [1,2,9], Giovanni Bocci [3,9], Srijan Verma[1,4], Md Mahmudulla Hassan [1,5], Jayme Holmes[3], Jeremy J. Yang[3], Suman Sirimulla [1,2,5,10] ✉ and Tudor I. Oprea [3,6,7,8,10] ✉

Strategies for drug discovery and repositioning are urgently need with respect to COVID-19. Here we present REDIAL-2020, a suite of computational models for estimating small molecule activities in a range of SARS-CoV-2-related assays. Models were trained using publicly available, high-throughput screening data and by employing different descriptor types and various machine learning strategies. Here we describe the development and use of eleven models that span across the areas of viral entry, viral replication, live virus infectivity, in vitro infectivity and human cell toxicity. REDIAL-2020 is available as a web application through the DrugCentral web portal (http://drugcentral.org/Redial). The web application also provides similarity search results that display the most similar molecules to the query, as well as associated experimental data. REDIAL-2020 can serve as a rapid online tool for identifying active molecules for COVID-19 treatment.

There is currently an urgent need to find effective drugs for treating coronavirus disease 2019 (COVID-19). Here we present REDIAL-2020, a suite of machine learning models that forecast activities for live viral infectivity, viral entry and viral replication, specifically for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), in vitro infectivity, and human cell toxicity. This application could serve the scientific community when prioritizing compounds for in vitro screening and may ultimately accelerate the identification of novel drug candidates for COVID-19 treatment. REDIAL-2020 consists of eleven independently trained machine learning models and includes a similarity search module that queries the underlying experimental dataset for similar compounds. These models were developed using experimental data generated by the following assays: the SARS-CoV-2 cytopathic effect (CPE) assay and its host cell cytotoxicity counterscreen, the Spike–ACE2 protein–protein interaction (AlphaLISA) assay and its TruHit counterscreen, the angiotensin-converting enzyme 2 (ACE2) enzymatic activity assay, the 3C-like (3CL) proteinase enzymatic activity assay, the SARS-CoV pseudotyped particle entry (CoV-PPE) assay and its counterscreen (CoV-PPE_cs), the Middle-East respiratory syndrome coronavirus (MERS-CoV) pseudotyped particle entry assay (MERS-PPE) and its counterscreen (MERS-PPE_cs), and the human fibroblast toxicity (hCYTOX) assay. Such assays represent five distinct categories: viral entry (CPE[1] and host cell cytotoxicity counterscreen[2]), viral replication (3CL enzymatic activity), live virus infectivity (AlphaLISA, TruHit counterscreen and ACE2 enzymatic activity)[3], in vitro infectivity (CoV-PPE with associated counterscreens for two other coronaviruses, SARS-CoV and MERS) and hCYTOX, as described in the National Center for Advancing Translational Sciences (NCATS) COVID-19 portal[4]. We retrieved these datasets from the NCATS COVID-19 portal[5]. The NCATS team is committed to performing a range of COVID-19-related

viral and host target assays, as well as analysing the results[6]. A more exhaustive description of each assay is provided in the Methods.

For model development, three different types of descriptors were employed and a best model for each descriptor type was developed by employing various machine learning algorithms. The three best models from each descriptor type were then combined using a voting method to give an ensemble model. These ensemble machine learning models are integrated into a user-friendly web portal that allows input using three different formats: (1) the drug name, both as the international non-proprietary name (for example, remdesivir) or as trade name (for example, Veklury); (2) the PubChem compound ID number (PubChem CID)[7] (for example, 121304016 for remdesivir); or (3) using the chemical structure encoded in the simplified molecular-input line-entry system (SMILES) format[8]. The workflow and output, regardless of input format, are identical and described below.

Drug repositioning requires computational support[9] and data-driven decision making offers a pragmatic approach to identifying optimal candidates while minimizing the risk of failure. As molecular properties and bioactivities can be described as a function of chemical structure, cheminformatics-based predictive models are becoming increasingly useful in drug discovery and repositioning research. Specifically, anti-SARS-CoV-2 models based on high-throughput data could be used as a prioritization step when planning experiments, particularly for large molecular libraries, thus decreasing the number of experiments and reducing downstream costs. REDIAL-2020 could serve such a purpose and help the scientific community reduce the number of molecules before experimental tests for anti-SARS-CoV-2 activity. This suite of machine learning models can also be used via the command line for large-scale virtual screening. As new datasets become available in the public domain, we plan to tune the machine learning models

[1]Department of Pharmaceutical Sciences, School of Pharmacy, The University of Texas at El Paso, El Paso, TX, USA. [2]Computational Science Program, The University of Texas at El Paso, El Paso, TX, USA. [3]Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM, USA. [4]Department of Pharmacy, Birla Institute of Technology and Science, Pilani, Pilani Campus, Rajasthan, India. [5]Department of Computer Science, The University of Texas at El Paso, El Paso, TX, USA. [6]Autophagy Inflammation and Metabolism Center of Biomedical Research Excellence, University of New Mexico Health Sciences Center, Albuquerque, NM, USA. [7]Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden. [8]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [9]These authors contributed equally: Govinda B. KC, Giovanni Bocci. [10]These authors jointly supervised this work: Suman Sirimulla, Tudor I. Oprea. ✉e-mail: ssirimulla@utep.edu; toprea@salud.unm.edu

further, add additional models based on SARS-CoV-2 assays and make these models available in future releases of REDIAL-2020.

Results

Data mining. All workflows and procedures were performed using the KNIME platform[10]. The NCATS data associated with the aforementioned assays were downloaded from the COVID-19 portal[4,5]. The files contained over 23,000 data points generated by high-throughput screening (HTS) experiments. When possible, each compound was cross-linked to drugs annotated in DrugCentral[11–13] to retrieve the chemical structure in SMILES format (see Methods), otherwise the original SMILES strings were retained. Bioactivity data were mined according to the curve class and maximum response parameters[14]. The activity class and a significance class were defined using criteria reported in Supplementary Tables 1 and 2, respectively. As a final data-wrangling step, all compounds were categorized and assay data grouped to have a unique record per molecule for each assay. When more than one assay was measured for the same molecule, only the datapoint with the best curve class was retained. At the end of this process, 4,954 unique molecules were stored.

The compounds were labelled as positive or negative for each assay. The compounds with a low-activity class were treated as negative, whereas compounds with high- and moderate-activity classes were treated as positive. Finally, the following calculated physicochemical property filters were applied: $\log[P] < 1$, $\log[P] > 9$, $\log[S] > -3$, $\log[S] < -7.5$, where $\log[P]$ is the $\log_{10}$ of the octanol/water partition coefficient and $\log[S]$ is the $\log_{10}$ of the aqueous solubility. These thresholds were initially used to maximize the number of inactive compounds removed while minimizing the number of active compounds excluded (see Discussion). Following use of the physicochemical property filters, each dataset was reduced in size (see Table 1). As shown in Table 1, certain datasets would have resulted in 15% or more of the active compounds being excluded; $\log[P]$ and $\log[S]$ filters were therefore not applied to those datasets. Chemical structures were standardized in terms of SMILES representation (see Methods). Following standardization, desalting, neutralizing and tautomer normalization, multiple input SMILES can resolve into the same output SMILES string. Hence, the final step was removal of duplicate chemical structures.

Model development. Several prediction models were developed for each assay, employing three categories of features and 22 distinct machine learning classification algorithms from the scikit-learn package[15] (see the Methods for the complete description of features categories; Supplementary Fig. 1 shows the workflow for model generation). The three different categories of features employed were based on chemical fingerprints, physicochemical descriptors and topological pharmacophore descriptors. Briefly, 19 different RDKit fingerprints were tested for fingerprint-based descriptors, Volsurf+ and RDKit descriptors were employed for physicochemical descriptors, and topological pharmacophore atom triplets fingerprints (TPATF) from Mayachemtools were used for pharmacophore descriptors. Input data were split into a 70% training set, 15% validation set and 15% test set for each model using a stratified sampling (Supplementary Table 3 reports the number of compounds used in training, validation and test sets for each model). Six assays (CPE, cytotox, AlphaLISA, TruHit, ACE2 and 3CL) were initially trained with 22 different classifiers available in scikit-learn (see Methods)[16]; however, some did not output probability estimates of the class labels (for example, OneVsOne, ridge, nearest centroid, linear SVC and so on). As our consensus based on probability models relies on predicted probability of each predicted label, only classifiers that output class probabilities were used for training. Two more classifiers—support vector machines and quadratic discriminant analysis—were evaluated. Finally, 15 classifiers and 22 features

Table 1 | Number and percentage of compounds outside of the $\log[P]$ and $\log[S]$ criteria

Assay	Actives (relative percentage)	Inactives (relative percentage)
CPE	44 (8%)	2,913 (37%)
cytotox	193 (14%)	2,764 (39%)
AlphaLISA	143 (19%)	1,119 (49%)
TruHit	134 (16%)	1,128 (51%)
ACE2	70 (38%)	1,192 (41%)
3CL	81 (28%)	3,330 (37%)
CoV-PPE	43 (27%)	881 (51%)
CoV-PPE_cs	247 (28%)	1,085 (44%)
hCYTOX	81 (22%)	1,306 (39%)
MERS-PPE	104 (20%)	1,024 (49%)
MERS-PPE_cs	46 (24%)	1,082 (45%)

of three distinct categories (see Methods) were trained across eleven assays, using hypopt for hyperparameter tuning[17].

Applicability domain. Machine learning models have boundaries for predictability[16], traditionally called the applicability domain[18]. The applicability domain is defined by the parameter space of the training set on which machine learning models are built. Machine learning predictions are deemed reliable when they fall within the applicability domain of that specific model and less reliable when outside of it[19]. There are two categories of methods to determine the applicability domain for classification models: novelty detection and confidence estimation. Novelty detection defines the applicability domain in terms of molecular (feature) space, whereas confidence estimation defines it in terms of expected prediction reliability[20]. As confidence estimation is more efficient at reducing the error rate than novelty detection[20], we implemented this method for evaluating applicability domain (see Methods). Confidence scores, which are averaged for each query molecule, as calculated by default using three different models, are incorporated along predictions in the results page. Confidence scores for each model can be examined by hovering over the confidence score value shown on the results webpage.

Submission web page. By accessing REDIAL-2020 (http://drugcentral.org/Redial) from any web browser, including mobile devices, the submission page is displayed (Fig. 1). The web server accepts SMILES, drug names or PubChem CIDs as input. The user interface at the top of the page allows users to navigate various options (Fig. 1). The user interface provides a summary of the models, such as model type, which descriptor categories were used for training and the evaluation scores. The user interface further depicts the processes of cleaning the chemical structures (encoded as SMILES) before training the machine learning models. Input queries such as drug name and PubChem CID are converted to SMILES before processing. Each SMILES string input is subject to four different steps, namely, converting the SMILES into canonical SMILES[21], removing salts (if present), neutralizing formal charges (except permanent ones) and standardizing tautomers. REDIAL-2020 predicts input compound activity across all eleven assays: CPE, cytotox, AlphaLISA, TruHit, ACE2, 3CL, CoV-PPE, CoV-PPE_cs, MERS-PPE, MERS-PPE_cs and hCYTOX. The workflow of operations performed on the submitted query SMILES through the redial web appplication are summarized in Supplementary Fig. 2.

Figure 2 shows an example of the output panel, which is loaded onto the same web page. REDIAL-2020 links directly to

DrugCentral 2021
Online drug Compendium

Search   Redial   About   Download   L1000 signature   FAQ
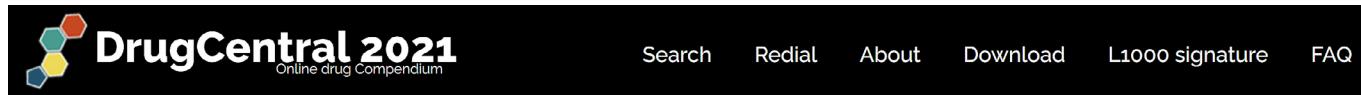
DrugCentral REDIAL 2020
A portal for estimating Anti-SARS-CoV-2 activities

Provide an Input string: Enter: Drug, SMILES, PubChem CID

Some Examples: CC(=O)OC1=CC=CC=C1C(=O)O | Remdesivir | 121304016

SUBMIT

Fig. 1 | Submission web page. A screenshot of REDIAL-2020's submission page.

Table 2 | Prediction metrics for the best models. ACC, accuracy; F1, F1 score; SEN, sensitivity; PREC, precision; AUC, area under the receiver operating characteristic curve

Model	Validation set results					Test set results				
	ACC	F1	SEN	PREC	AUC	ACC	F1	SEN	PREC	AUC
CPE	0.695	0.693	0.689	0.698	0.695	0.651	0.643	0.626	0.661	0.651
cytotox	0.782	0.780	0.773	0.787	0.782	0.688	0.700	0.727	0.675	0.688
AlphaLISA	0.824	0.831	0.863	0.801	0.823	0.790	0.787	0.777	0.798	0.790
TruHit	0.828	0.836	0.873	0.802	0.828	0.734	0.737	0.746	0.728	0.734
ACE2	0.755	0.750	0.750	0.750	0.775	0.755	0.777	0.840	0.724	0.753
3CL	0.804	0.808	0.837	0.782	0.804	0.712	0.705	0.681	0.731	0.713
CoV-PPE	0.771	0.761	0.732	0.793	0.771	0.665	0.658	0.643	0.674	0.665
CoV-PPE_cs	0.872	0.869	0.869	0.869	0.872	0.659	0.636	0.583	0.700	0.661
hCYTOX	0.736	0.736	0.736	0.736	0.736	0.710	0.713	0.719	0.706	0.710
MERS-PPE	0.813	0.823	0.875	0.777	0.814	0.696	0.698	0.698	0.698	0.696
MERS-PPE_cs	0.833	0.823	0.777	0.875	0.833	0.703	0.680	0.629	0.739	0.703

DrugCentral[11–13] for approved drugs and to PubChem for chemicals (where available), enabling easy access to further information on the query molecule. Using REDIAL-2020 estimates, promising anti-SARS-CoV-2 compounds would ideally be active in the CPE assay while inactive in cytotox and in hCYTOX; active in the AlphaLISA assay and inactive in the TruHit assay while not blocking (inactive) ACE2; active in CoV-PPE while inactive in CoV-PPE_cs; active in MERS-PPE while inactive in MERS-PPE_cs; or active in the 3CL assay with any combination of the above. After running all of the prediction models, a schematic representation of the best profile that can be defined for a molecule is depicted in Fig. 3.

Similarity search. A similiarity tool is implemented in the web portal. The similarity is determined using Tanimoto coefficient calculations with ECFP4 bit vector fingerprint of length 1,024. The Tanimoto coefficient represents the overlap of features between molecules as the ratio of the number of common features to the total number of features in each fingerprint. Tanimoto coefficient values range from 0 to 1, with 1 corresponding to identical fingerprints. A fingerprint-based Tanimoto[22] similarity search is thus conducted for each query molecule against training set molecules, based on NCATS COVID-19 portal[5] data. The top-ten similar molecules to that of the query molecule, based on Tanimoto coefficient[23] scores, are displayed in the results page.

Discussion
Before developing machine learning models, unsupervised learning can detect patterns that might guide successive steps. Hence, after establishing the experimental categories (see above), we inspected the data using principal component analysis (PCA)[24] on VolSurf+[25] descriptors. For both CPE and cytotox, clusters emerge along the first principal component (PC1; Fig. 4). For CPE data, the majority of compounds showing high-to-moderate CPE activity are grouped in the right-hand of Fig. 4a. At the same time, compounds with high-to-moderate cytotoxicity are grouped in the right-hand region of Fig. 4b. By inspecting the loading score plot for VolSurf+ descriptors that are likely to contribute to these patterns, we identified membrane permeability (estimated using log[P]) and water solubility (estimated using log[S]) as major contributors to the first latent variable (see Supplementary Fig. 3). Compounds with low log[P]/high log[S]—clustered in the left-hand region of the score plot—are less likely to be active in the CPE assay and more likely to be non-cytotoxic.

The distribution of actives was also visualized for AlphaLISA and TruHit compounds in Fig. 4c and Fig. 4d, respectively (see also Table 1). Although clustering is less pronounced for the AlphaLISA assay with respect to CPE (Fig. 4a), the right-hand part of the plot does capture most of the high/moderate-activity compounds. Such distribution of actives in the right-hand region was not observed for

RESULTS

LogP (Log units)	LogS (Log units)	Molecular Wt. (g/mol)	Formula
2.20	-2.89	602.59	C27H35N6O8P

External reference:

PubChem CID	Drug Central ID
121304016	5376
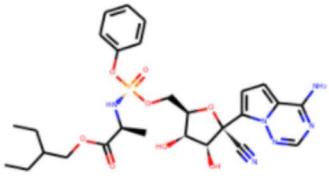
Synonyms: **remdesivir | s8932**

Processed SMILES string:

CCC(CC)COC(=O)[C@H](C)N[P@](=O)(OC[C@H]1O[C@@](C#N)
(c2ccc3c(N)ncnn23)[C@H](O)[C@@H]1O)Oc1ccccc1

Prediction Results

	Class	Prediction	Confidence
Live Virus Infectivity	SARS-CoV-2 cytopathic effect (CPE)	ACTIVE	1.0
	SARS-CoV-2 cytopathic effect (host tox Counter) / Cytotoxicity	INACTIVE	1.0
Viral Entry	Spike-ACE2 protein-protein interaction (AlphaLISA)	ACTIVE	1.0
	Spike-ACE2 protein-protein interaction (TruHit Counter)	INACTIVE	1.0
	ACE2 enzymatic activity	INACTIVE	1.0
Viral Replication	3CL enzymatic activity	INACTIVE	1.0
In vitro Infectivity	SARS-CoV pseudotyped particle entry (CoV-PPE)	ACTIVE	0.69
	SARS-CoV pseudotyped particle entry counter screen (CoV-PPE_cs)	INACTIVE	0.68
	MERS-CoV pseudotyped particle entry (MERS-PPE)	ACTIVE	0.34
	MERS-CoV pseudotyped particle entry counter screen (MERS-PPE_cs)	INACTIVE	0.5
Human Cell Toxicity	Human fibroblast toxicity (hCYTOX)	ACTIVE	0.58

Fig. 2 | Output panel example. A screenshot of the web page displaying the machine learning estimates and for a query molecule.

ACE2 actives (Fig. 4e); thus, permeability and solubility are not the major determinants of this ACE2 inhibition assay.

This preliminary analysis can point to filtering data before machine learning. For example, the majority of compounds placed on the left side of the Fig. 4 PCA plot are inactive (except for ACE2); therefore, before developing the machine learning models, we used ALOGPS[26] on every dataset except for ACE2 to apply cutoff filters on the basis of compounds's calculated $\log[P]$ and $\log[S]$ values. These filters narrow the focus of machine learning models on

features derived only from compounds for which simple property criteria (for example, $\log[P]$ and $\log[S]$) cannot be used to distinguish actives from inactives, specifically, the right-hand regions in Fig. 4. As the fraction of active compounds excluded from the ACE2 dataset was quite high (34%), $\log[P]$ and $\log[S]$ filters were not applied for ACE2 inhibition.

For 3CL enzymatic activity, data from NCATS were retrieved separately. The initial set contained 12,263 data points; however, data wrangling identified 2,100 duplicates and 2,366 inconclusive
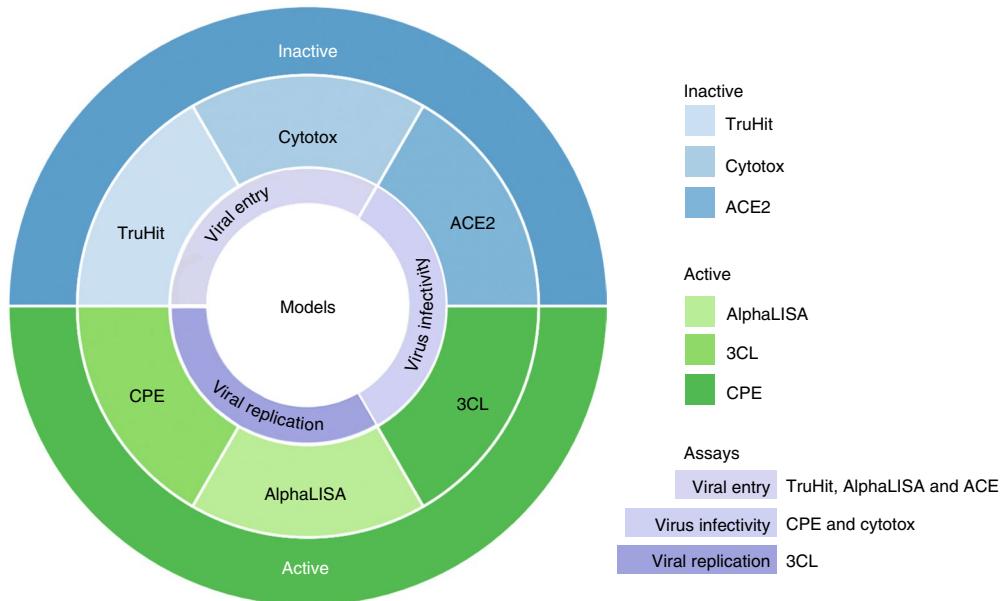
Fig. 3 | Best profile example. A schematic representation of the most desirable profile for anti-SARS-CoV-2 activities that can be observed via REDIAL-2020 predictions, on the basis of the SARS-CoV-2-specific set of assays. The five additional assays (not depicted here) offer supporting evidence for the decision making process and hit prioritization.

entries, which were discarded. More entries were removed during the desalting and physicochemical feature generation as VolSurf+ descriptors could not be computed for some of the compounds. The final 3CL dataset contains 7,716 entries, with 286 active and 7,430 inactive compounds. Given that the fraction of active 3CL compounds filtered would have been 30%, the physicochemical property filters were not applied. There were no notable activity clusters detected in the 3CL dataset via PCA-VolSurf+ (see Supplementary Fig. 4).

Furthermore, NCATS released data for five completely new HTS assays—and updated assay data for the other six after additional testing—between June and October 2020. Hence, we reevaluated the entire set of assays. The total number of compounds after data wrangling was 10,074. Our analysis showed that only the CPE and the cytotoxicity assays were enriched with more compounds. There were 2,354 more compounds, with 158 new actives in the CPE dataset and 2,332 more compounds (295 new actives) in the cytotox dataset. As the fraction of active compounds filtered out on applying physicochemical property filters was over 15%, these filters were not applied for the five new datasets (see also Table 1).

With respect to actives versus inactives, all eleven NCATS assays are highly unbalanced, with a disproportionate ratio of the active (few) compounds compared with inactive (many) compounds. For example, there were approximately nine times more inactives than actives and approximately three times more non-cytotoxic compounds than cytotoxic compounds for the CPE and cytotoxicity assays, respectively. Thus, to avoid overtraining for the dominant category, each model was derived using random selection wherein compounds from the majority class were selected in equal proportion to those of the minority class. Our balanced dataset numbers were as follows: 996 for CPE, 2,252 for cytotox, 1,260 for AlphaLISA,

1,668 for TruHit, 206 for ACE2, 572 for 3CL, 1,782 for CoV-PPE, 320 for CoV-PPE_cs, 760 for hCYTOX, 970 for MERS-PPE and 368 for MERS-PPE_cs.

We implemented eleven predictive models based on consensus methods to evaluate anti-SARS-CoV-2 activities of novel chemicals. Of the two consensus methods evaluated (voting-based and probability score-based), the voting-based consensus model exhibited better performance (see Supplementary Figs. 5–10) and was thus implemented in the REDIAL-2020 web application. Consensus models were generated on the basis of the top-three performing models trained on fingerprint, pharmacophore and physicochemical descriptors. First, we selected a fingerprint model from an initial evaluation of 19 different fingerprint descriptor methods; this was combined with a TPATF model. Finally, RDKit or VolSurf+ provided a third model, which was based on physicochemical properties. Supplementary Fig. 11a–d summarizes our initial evaluation and the comparison between various features and machine learning algorithms, Supplementary Fig. 11a,b compares the performance of each feature across 22 machine learning algorithms (classifiers) and six assays, and Supplementary Fig. 11c,d compares the performance of each classifier across 22 features and six assays (CPE, cytotoxicity, AlphaLISA, Truhit, ACE2, and 3CL). For example, the violin plot for the Avalon feature (see Supplementary Fig. 11a) summarizes F1 scores from all six assays (and 22 classifiers). Among descriptors, VolSurf+ and LFCFP6 outperformed others, whereas the gradient boost and the multilayer perceptron classifiers performed better among machine learning algorithms (see Supplementary Figs. 12 and 13 for comparisons of each feature across 15 machine learning algorithms and eleven assays; Supplementary Figs. 14–47 depict more detailed comparisons across different features and machine learning algorithms with respect to individual models).

Fig. 4 | PCA scores plots of the molecules tested in NCATS SARS-CoV-2 experiments based on VolSurf+ descriptors. On each plot, the compound position is defined along PC1 and PC2. **a**, CPE compounds coloured by CPE categories: high/moderate activity in yellow and low activity in black. **b**, Cytotoxic compounds coloured by cytotoxicity categories: high/moderate cytotoxicity in orange and low (not) cytotoxicity in black. **c**, AlphaLISA compounds coloured by Spike–ACE2 interaction blockers categories: high/moderate (strong) blockers in red and low (weak) blockers in black. **d**, TruHit compounds coloured by AlphaLISA readout interfering categories: high/moderate interference in cyan and low interference in black. **e**, ACE2 compounds coloured by ACE2 inhibition categories: high/moderate (strong) inhibitors in magenta and low (weak) inhibitors in black.

Two options for the consensus model were initially considered based on the potential overlap between VolSurf+ and RDKit descriptors: fingerprint+TPATF+RDKit and fingerprint+TPATF+VolSurf+, respectively. RDKit descriptors outperformed VolSurf+ in cytotox, AlphaLISA, ACE2, 3CL, MERS-PPE_cs, CoV-PPE, CoV-PPE_cs and hCYTOX, whereas
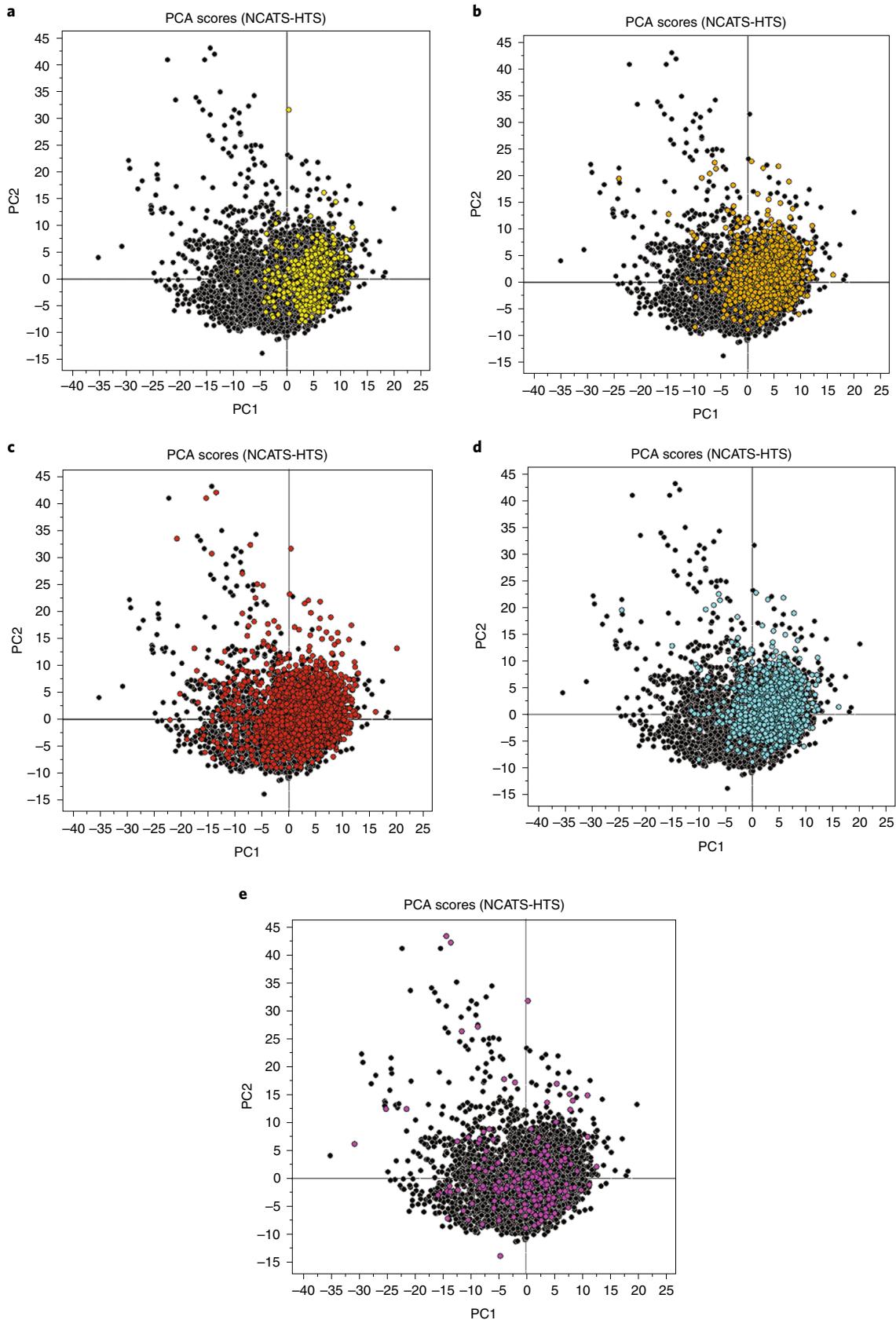
VolSurf+ descriptors outperformed RDKit in CPE and hCYTOX, along with similar results in MERS-PPE and TruHit for the tested evaluation metrics such as accuracy, F1 score and AUC in validation sets (see Supplementary Figs. 48–58). However, the situation slightly changed when considering consensus models. Inclusion of VolSurf+ yielded a better consensus model for the CPE, whereas including RDKit yielded better consensus models for the cytotox, 3CL, TruHit, AlphaLISA, MERS-PPE_cs, CoV-PPE and CoV-PPE_cs assays (Supplementary Figs. 5–10 compare the best models from each feature category). As the NCATS team released data for more compounds for the six initial assays plus five new assays in October 2020, we updated the initial six models and developed models for the five new assays (comparisons of models from each category for the new and updated models are shown in Supplementary Figs. 53–57). Among the eleven assay models, the voting-based consensus model performed slightly better than individual feature type models for validation F1 score results; in three assays (ACE2, MERS-PPE and hCYTOX), the voting-based consensus model was not the top performer, but its performance was close to the top performing model. For the web platform, we implemented voting-based consensus models for all eleven assay models using RDKit descriptors as opposed to Volsurf+ descriptors, as RDKit is open-source software that can be ported and dockerized without restrictions. Table 2 summarize the evaluation scores for all models implemented in REDIAL-2020.

To confirm the utility of our models, we collected three additional datasets from the literature and submitted these molecules (external to our training/validation/test sets) as input for prediction. First, we used a database for COVID-19 experiments[27] to explore and download recently published[28–37] in vitro COVID-19 bioactivity data of the reported compounds. After removing compounds already included in the NCATS experiments, we identified 27 external compounds active in anti-SARS-CoV-2 CPE assays (see Supplementary Table 4). Out of 27 compounds, three were excluded on applying the log[P]/log[S] filters, and the remaining 24 were predicted by the CPE model. Sixteen compounds were correctly predicted as active by the consensus model (that is, at least two models, see Supplementary Fig. 59), with eight compounds predicted as inactive. Among those predicted to be inactive, the majority stem from the work by Ellinger and colleagues[37], which were derived from Caco-2 cells for CPE experiments. There is a high degree of variability between these two CPE assays (Caco-2 versus Vero E6), which explains the lack of predictivity using Vero E6-trained CPE models for Caco-2 data. The second dataset of 3CL (Mpro) inhibitors[36] identified six inhibitors: ebselen (0.67 μM), disulfiram (9.35 μM), tideglusib (1.55 μM), carmofur (1.82 μM), shikonin (15.75 μM) and PX-12 (21.39 μM) (see Supplementary Table 5). Among these six inhibitors, our consensus 3CL model correctly predicted four of them as actives, and five of them as actives by at least one of the three 3CL machine learning models. The REDIAL-2020 suite of models therefore correctly predicted 67% of the external compounds for CPE and 3CL inhibitors[36]. Although the external predictivity of CPE model seems to underestimate previous model performance in the validation and external sets (see Supplementary Table 6), it has been noted that CPE experiments are affected by considerable intra- and interexperiment variability[27]. Hence, we cannot exclude the possibility that some of the experiments performed by other laboratories are not directly comparable with NCATS COVID-19 portal[5] results.

Conclusion

Here we described REDIAL-2020, an open-source, open-access machine learning suite for estimating anti-SARS-CoV-2 activities from molecular structure. By leveraging data available from NCATS, we developed eleven categorical machine learning models: CPE, cytotox, AlphaLISA, TruHit, ACE2, 3CL, CoV-PPE, CoV-PPE_cs, MERS-PPE, MERS-PPE_cs and hCYTOX. These

models are exposed on the REDIAL-2020 portal, and the output of a similarity search using input data as a query is provided for every submitted molecule. The top-ten most similar molecules to the query molecule from the existing COVID-19 databases, together with associated experimental data, are displayed. This allows users to evaluate the confidence of the machine learning predictions.

The REDIAL-2020 platform provides a fast and reliable way to screen novel compounds for anti-SARS-CoV-2 activities. REDIAL-2020 is available on GitHub and DockerHub as well, and the command-line version supports large-scale virtual screening purposes. Future developments of REDIAL-2020 could include additional machine learning models. For example by using the TMPRSS2 inhibition assay[38] data from the NCATS COVID-19 portal or additional NCATS data as they become available in the public domain. We will continue to update and enhance the machine learning models and make these models available in future releases of REDIAL.

Methods

HTS assays. The SARS-CoV-2 CPE assay measures the ability of a compound to reverse the cytopathic effect induced by the virus in Vero E6 host cells. As cell viability is reduced by a viral infection, the CPE assay measures the compound's ability to restore cell function (cytoprotection). Although this assay does not provide any information concerning the mechanism of action, it can be used to screen for antiviral activity in a high-throughput manner; however, there is the possibility that the compound itself may exhibit a certain degree of cytotoxicity, which could also reduce cell viability. As this confounds the interpretation of CPE assay results, masking the cytoprotective activity, a counterscreen to measure host (Vero E6) cell cytotoxicity is used to detect such compounds; thus, a net-positive result from the combined CPE assays consists of a compound showing a protective effect but no cytotoxicity.

The AlphaLISA assay measures a compound's ability to disrupt the interaction between the viral Spike protein and its human receptor protein, ACE2[39]. The surface of the ACE2 protein is the primary host factor recognized and targeted by SARS-CoV-2 virions[40]. This binding event between the SARS-CoV-2 Spike protein and the host ACE2 protein initiates binding of the viral capsid and leads to viral entry into host cells. Thus, disrupting the Spike–ACE2 interaction is likely to reduce the ability of SARS-CoV-2 virions to infect host cells. This assay has two counterscreens, as follows. The TruHit counterscreen is used to determine false positives, that is, compounds that interfere with the AlphaLISA readout in a non-specific manner, or with assay signal generation and/or detection. It uses the biotin–streptavidin interaction (one of the strongest known non-covalent drug–protein interactions) as other compounds are unlikely to disturb it. Consequently, any compound showing interference with this interaction is most likely a false positive. Common interfering agents are oxygen scavengers or molecules with spectral properties sensitive to the 600–700 nm wavelengths used in AlphaLISA. The second counterscreen is an enzymatic assay that measures human ACE2 inhibition to identify compounds that could potentially disrupt endogenous enzyme function. ACE2 lowers blood pressure by catalysing the hydrolysis of angiotensin II (a vasoconstrictor octapeptide) into the vasodilator angiotensin (1–7)[41]. Although blocking the Spike–ACE2 interaction may stop viral entry, drugs effective in this manner could potentially cause unwanted side-effects by blocking the endogenous vasodilating function of ACE2. The ACE2 assay thus serves to detect such eventualities and to de-risk such off-target events.

Following entry into the host cell, the main SARS-CoV-2 replication enzyme is 3CL, also called main protease or Mpro[36], which cleaves the two SARS-CoV-2 polyproteins into various proteins (for example, RNA polymerases, helicases, methyltransferases and so on), which are essential to the viral life cycle. As inhibiting the 3CL protein disrupts the viral replication process, this makes 3CL an attractive drug target[42]. The SARS-CoV-2 3CL biochemical assay measures the ability of compounds to inhibit recombinant 3CL cleavage of a fluorescently labelled peptide substrate.

In this category there are four assays: SARS-CoV pseudotyped particle entry and its counterscreen, and MERS-CoV pseudotyped particle entry and its counterscreen. The pseudotyped particle assay measures the inhibition of viral entry in cells but it does not require a BSL-3 facility (BSL-2 is sufficient) to be performed, as it does not use a live virus to infect cells. It instead uses pseudotyped particles that are generated by the fusion of the coronavirus Spike protein with a murine leukaemia virus core. As they have the coronavirus spike protein on their surface, the particles behave like their native coronavirus counterparts for entry steps. This makes them excellent surrogates of native virions for studying viral entry into host cells. The experimental protocol of such an assay is described in detail elsewhere[43]. The cell lines used are Vero E6 for SARS-CoV and Huh7 for MERS-CoV, respectively.

At the time of data extraction, compound data were available for one assay human fibroblast toxicity. With the human fibroblast toxicity assay, it is possible to assess the general human cell toxicity of compounds by measuring host cell

ATP content as a readout for cytotoxicity (similarly to what is done in the various counter screenings). This assay is therefore intended for discarding compounds that are likely to show high toxicity in human cells (that is, side effects in the organism). Hh-WT fibroblast cells are used in this assay and the highly cytotoxic drug bortezomib is used as a reference compound.

Data matching operations. The matching of NCATS compounds to DrugCentral was conducted in three sequential steps: by InChI (international chemical identifier)[44], by synonym (name), and by matching Chemical Abstracts Service registry numbers. First, NCATS molecules were matched by InChI. Molecules that did not match were then queried by drug name and associated synonyms, as annotated in DrugCentral. Finally, if not matched by either InChI or name, molecules were matched by Chemical Abstracts Service number. If none of the above steps resulted in a match, then the molecule in question was not classified as an approved drug. At the end of this process, 4,954 unique molecules (2,273 approved drugs and 2,681 chemicals) were stored. SMILES were retrieved from DrugCentral whenever possible, otherwise the original SMILES strings were retained.

SMILES standardization. Chemical structures were standardized to ensure rigorous deduplication, accurate counts and performance measures, and consistent descriptor generation, preserving stereochemistry, which is required for conformer-dependent descriptors. This workflow uses the MolStandardize SMARTS-based functionality in RDKit[45] to transform input SMILES into standardized molecular representations. Four different filters were implemented via RDKit: (1) input SMILES were standardized into canonical (isomeric where appropriate) SMILES strings. The input SMILES that failed to convert were discarded; (2) RDKit Salt Stripper was used to de-salt input compounds (that is, remove the salt structures). The donotRemoveEverything feature leaves the last salt structure when the entire canonical SMILES string is comprised of salts only; (3) RDKit Uncharger neutralizes input molecules by adding/removing hydrogen atoms and setting formal charges to zero (except for for example, quaternary ammonium cations); (4) canonical SMILES were then formalized into specific tautomers using RDKit.

Molecular features/descriptors. A total of 22 features of three distinct types (19 fingerprints-based, 1 pharmacophore-based and 2 physicochemical descriptors-based) were implemented. Fingerprints were converted into a bit vector of either 1,024 or 16,384 lengths. Pharmacophore type was also a bit vector of size 2,692, whereas RDKit and VolSurf+ descriptors were of length 200 and 128, respectively.

The fingerprints-based description includes the circular, path-based, and substructure keys[46,47]. Circular fingerprints include the extended-connectivity fingerprints (ECFPx) and feature-connectivity fingerprints (FCFPx), where x is 0, 2, 4, and 6 are the bond length or diameter for each circular atom environment. The ECFP consists of the element, number of heavy atoms, isotope, number of hydrogen atoms and ring information, whereas the FCFP consists of pharmacophore features.

Avalon and the molecular access system (MACCS) are two distinct types of substructure keys (fingerprints). The Avalon fingerprint, used here, is a bit vector of size 1,024. It includes feature classes such as atom count, atom symbol path, augmented atom, augmented symbol path and so on. MACCS structural keys are 166-bit structural key descriptors. Each bit here is associated with a SMARTS pattern and belongs to the dictionary-based fingerprint class. Path-based fingerprints include RDKx (where x is 5, 6, 7), topological torsion (TT), HashTT, atom pair (AP) and HashAP. The size of each fingerprint is 1,024. The longer, 16,384-bits, versions of the fingerprint, marked by the prefix L (LAvalon, LECFP6, LECFP4, LFCFP6 and LFCFP4) were used for comparison.

Topological pharmacophore atomic triplets fingerprints were obtained using Mayachemtools[48]; the TPATFs describe the ligand sites that are necessary for molecular recognition of a macromolecule or a ligand, and passes that information to the machine learning model to be trained. Ligand SMILES strings were passed through a Perl script to generate TPATF. The basis sets of atomic triplets were generated using two different constraints: (1) the triangle rule, that is, the length of each side of a triangle cannot exceed the sum of the lengths of the other two sides; and (2) elimination of redundant pharmacophores related by symmetry. The default pharmacophore atomic types hydrogen-bond donor (HBD), hydrogen-bond acceptor (HBA), positively ionizable (PI), negatively ionizable (NI), hydrophobic (H) and aromatic (Ar) were used during generation of TPATF[49].

The physicochemical description includes the RDKit molecular descriptors and VolSurf+ descriptors. For RDKit descriptors, a set of 200 descriptors were used, which were obtained from RDKit[45]. They are either experimental properties or theoretical descriptors, which are for example molar refractivity, log[P], heavy atom counts, bond counts, molecular weight, topological polar surface area.

A total of 128 descriptors were obtained using VolSurf+ software. VolSurf+ is a computational approach aimed at describing the structural, physicochemical and pharmacokinetic features of a molecule starting from a three-dimensional map of the interaction energies between the molecule and chemical probes (grid-based molecular interaction fields)[50]. VolSurf+ compresses the information present in molecular interaction fields into numerical descriptors, which are simple to use and interpret[25,51].

Machine learning classifiers. Using assay data as input (specifically, CPE, cytotox, AlphaLISA, TruHit, ACE2 and 3CL), we trained machine learning models using the following 24 different classifiers: complement naive Bayes, extreme gradient boosting, KNeighbors, gradient boosting, perceptron, OneVsRest, extra-tree, ridge, OneVsOne, bagging, random forest, output code, passive aggressive, linear SVC, stochastic gradient descent, logistic regression, extra trees, multinomial naive Bayes, AdaBoost, decision tree, nearest centroid, multilayer perceptron, support vector machines and quadratic discriminant analysis. All of these algorithms are implemented in the scikit-learn package[16]. The 22 types of features (ECFP0, ECFP2, ECFP4, LECFP4, ECFP6, LECFP6, FCFP2, FCFP4, LFCFP4, FCFP6, LFCFP6, RDK5, RDK6, RDK7, Avalon, LAvalon, MACCS, HashTT, HashAP, VolSurf+, TPATF and RDKit descriptors) that served as input to the machine learning classifiers are described above. All classifiers were trained on their default configurations. For hyperparameter tuning we used hypopt[17] and the best-suited combination of classifiers and features (see Supplementary Table 7). All models were optimized and selected based on the validation F1 score. The best-performing models were saved and used for the evaluation of external datasets.

Confidence scores. One way to calculate the certainty of prediction is provided by the classification algorithms framework applied here, as implemented in the scikit-learn package. The confidence estimate associated with predictions for each object (small molecule) recalls a basic feature of scikit-learn, predict_proba. For example, in the random forest classifier, votes are noted for each (sub)model; thus, for each class, predict_proba returns the number of votes divided by the number of trees in that particular forest (model). This confidence score, which estimates the model prediction's reliability, is used to gauge the applicability domain.

Data availability

All data used for the model described in this work are available at Zenodo (https://doi.org/10.5281/zenodo.4606720). These datasets were originally collected from the following links (please note that these data are subject to change without notice): CPE: https://opendata.ncats.nih.gov/covid19/assay?aid=14, cytotox: https://opendata.ncats.nih.gov/covid19/assay?aid=15, AlphaLISA: https://opendata.ncats.nih.gov/covid19/assay?aid=1, TruHit: https://opendata.ncats.nih.gov/covid19/assay?aid=2, ACE2: https://opendata.ncats.nih.gov/covid19/assay?aid=6, 3CL: https://opendata.ncats.nih.gov/covid19/assay?aid=9, CoV-PPE: https://opendata.ncats.nih.gov/covid19/assay?aid=22, CoV-PPE_cs: https://opendata.ncats.nih.gov/covid19/assay?aid=23, MERS-PPE: https://opendata.ncats.nih.gov/covid19/assay?aid=24, MERS-PPE_cs: https://opendata.ncats.nih.gov/covid19/assay?aid=25, hCYTOX: https://opendata.ncats.nih.gov/covid19/assay?aid=21.

Code availability

All of the codes and the trained models are available at Zenodo (https://doi.org/10.5281/zenodo.4606720).

References

1. Gorshkov, K., Chen, Z.C., Bostwick, R. et al. The SARS-CoV-2 cytopathic effect is blocked by lysosome alkalizing small molecules. *ACS Infect. Dis.* https://doi.org/10.1021/acsinfecdis.0c00349 (2021).
2. Sun, H., Wang, Y., Cheff, D. M., Hall, M. D. & Shen, M. Predictive models for estimating cytotoxicity on the basis of chemical structures. *Bioorg. Med. Chem.* **28**, 115422 (2020).
3. Hanson, Q. M. et al. Targeting ACE2–RBD interaction as a platform for COVID-19 therapeutics: development and drug-repurposing screen of an AlphaLISA proximity assay. *ACS Pharmacol. Transl. Sci.* **6**, 1352–1360 (2020).
4. Brimacombe, K. R. et al. An OpenData portal to share COVID-19 drug repurposing data in real time. Preprint at https://www.biorxiv.org/content/10.1101/2020.06.04.135046v1 (2020).
5. *SARS-CoV-2 Assays* (NCATS, accessed 25 September 2020); https://opendata.ncats.nih.gov/covid19/assays
6. Huang, R., Xu, M., Zhu, H. et al. Biological activity-based modeling identifies antiviral leads against SARS-CoV-2. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00839-1 (2021).
7. Kim, S. et al. PubChem substance and compound databases. *Nucl. Acids Res.* **44**, D1202–D1213 (2016).
8. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
9. Oprea, T. I. et al. Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Mol. Inform.* **30**, 100–111 (2011).
10. Berthold Michael, R. et al. in *Data Analysis, Machine Learning and Applications* 319–326 (Springer, 2007).

11. Ursu, O. et al. DrugCentral: online drug compendium. *Nucl. Acids Res.* **45**, D932–D939 (2017).
12. Ursu, O. et al. DrugCentral 2018: an update. *Nucl. Acids Res.* **47**, D963–D970 (2019).
13. Avram, S. et al. DrugCentral 2021 supports drug discovery and repositioning. *Nucl. Acids Res.* **49**, D1160–D1169 (2020).
14. Markossian, S. et al. (eds) *Assay Guidance Manual* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2004).
15. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learning Res.* **12**, 2825–2830 (2011).
16. Oprea, T. I. & Waller, C. L. in *Reviews in Computational Chemistry* Vol. 11, 127–182 (John Wiley and Sons, 2007).
17. *hypopt* (Github, accessed 24 July 2020); https://github.com/cgnorthcutt/hypopt
18. Eriksson, L. et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **111**, 1361–1375 (2003).
19. Liu, R. & Wallqvist, A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *J. Chem. Inf. Model.* **59**, 181–189 (2019).
20. Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic classification methods and their applicability domain. *Mol. Inform.* **35**, 160–180 (2016).
21. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
22. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115–1118 (1960).
23. Whittle, M., Gillet, V. J., Willett, P., Alex, A. & Loesel, J. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **44**, 1840–1848 (2004).
24. Carey, R. N., Wold, S. & Westgard, J. O. Principal component analysis. Alternative to referee methods in method comparison studies. *Anal. Chem.* **47**, 1824–1829 (1975).
25. Cruciani, G., Pastor, M. & Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **11**, S29–S39 (2000).
26. Tetko, I. V. et al. Virtual computational chemistry laboratory—design and description. *J. Comput. Aided Mol. Des.* **19**, 453–463 (2005).
27. Kuleshov, M. V. et al. The COVID-19 drug and gene set library. *Patterns* **1**, 100090 (2020).
28. Jeon, S. et al. Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs. *Antimicrob. Agents Chemother.* **64**, e00819-20 (2020).
29. Weston, S. et al. Broad anti-coronavirus activity of Food and Drug Administration-approved drugs against SARS-CoV-2 in vitro and SARS-CoV in vivo. *J. Virol.* **94**, e01218-20 (2020).
30. Touret, F. et al. In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. *Sci. Rep.* **10**, 13093 (2020).
31. Xing, J. et al. Reversal of infected host gene expression identifies repurposed drug candidates for COVID-19. Preprint at https://www.biorxiv.org/content/10.1101/2020.04.07.030734v1 (2020).
32. Riva, L. et al. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113–119 (2020).
33. Choy, K.-T. et al. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro. *Antiviral Res.* **178**, 104786 (2020).
34. Mirabelli, C. et al. Morphological cell profiling of SARS-CoV-2 infection identifies drug repurposing candidates for COVID-19. Preprint at https://www.biorxiv.org/content/10.1101/2020.05.27.117184v4 (2020).
35. Riva, L. et al. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* **586**, 113–119 (2020).
36. Jin, Z. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
37. Ellinger, B. et al. A SARS-CoV-2 cytopathicity dataset generated by high-content screening of a large drug repurposing collection. *Sci. Data* **8**, 70 (2021).
38. 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.* **5**, 997–1007 (2020).
39. 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).
40. Millet, J. K. & Whittaker, G. R. Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. *Virology* **517**, 3–8 (2018).
41. Keidar, S., Kaplan, M. & Gamliel-Lazarovich, A. ACE2 of the heart: from angiotensin I to angiotensin (1–7). *Cardiovasc. Res.* **73**, 463–469 (2007).
42. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. & Jung, S.-H. An overview of severe acute respiratory syndrome–coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. *J. Med. Chem.* **59**, 6595–6628 (2016).
43. Millet, J. K. et al. Production of pseudotyped particles to study highly pathogenic coronaviruses in a biosafety level 2 setting. *J. Vis. Exp.* https://doi.org/10.3791/59010 (2019).
44. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **7**, 23 (2015).
45. Landrum, G. et al. *RDKit: Open-source Cheminformatics Software* (RDKit, accessed 10 May 2020); https://www.rdkit.org/
46. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5**, 26 (2013).
47. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
48. Sud, M. MayaChemTools: an open source package for computational drug discovery. *J. Chem. Inf. Model.* **56**, 2292–2297 (2016).
49. Bonachéra, F., Parent, B., Barbosa, F., Froloff, N. & Horvath, D. Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *J. Chem. Inf. Model.* **46**, 2457–2477 (2006).
50. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857 (1985).
51. Zamora, I., Oprea, T., Cruciani, G., Pastor, M. & Ungell, A.-L. Surface descriptors for protein–ligand affinity prediction. *J. Med. Chem.* **46**, 25–33 (2003).

Acknowledgements

Author contributions

S.S. and T.I.O. designed the research study. G.B.K. and S.V. developed the prediction models. G.B. curated the public data. G.B.K., S.V., M.M.H., J.J.Y., J.H. and S.S. developed the web application. S.S., G.B.K., G.B. and T.I.O. wrote the paper. All authors read and approved the manuscript.

Competing interests

T.I.O. has received honoraria from or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth, and is on the Scientific Advisory Board of ChemDiv and InSilico Medicine.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00335-w.

Correspondence and requests for materials should be addressed to S.S. or T.I.O.

Peer review information *Nature Machine Intelligence* thanks Feixiong Cheng, Junmei Wang and Kemal Yelekçi for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.