

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

Govinda B. KC^{1,2†}, Giovanni Bocci^{3†}, Srijan Verma^{2,4}, Md Mahmudulla Hassan^{2,5}, Jayme Holmes³, Jeremy J. Yang³, Suman Sirimulla^{1,2,5,*}, and Tudor I. Oprea^{3,6,7,8,*}

¹ Computational Science Program, The University of Texas at El Paso, Texas 79968, USA.

² Department of Pharmaceutical Sciences, School of Pharmacy, The University of Texas at El Paso, Texas 79902, USA.

³ Translational Informatics Division, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, NM, USA.

⁴ Department of Pharmacy, Birla Institute of Technology and Science, Pilani, Pilani Campus, Rajasthan, 333031, India.

⁵ Department of Computer Science, The University of Texas at El Paso, Texas 79968, 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.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

* To whom correspondence should be addressed

E-mail: ssirimulla@utep.edu, toprea@salud.unm.edu

Manuscript Tracking Number: NATMACHINTELL-PI20082284A

Abstract

Strategies for drug discovery and repositioning are an urgent 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 the usage of eleven models spanning 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>). In addition, the web-app 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.

Currently, there is an urgent need to find effective drugs for treating coronavirus disease 2019 (COVID-19). Here, we present a suite of machine learning (ML) models termed “REDIAL-2020” that forecast activities for live viral infectivity, viral entry, and viral replication, specifically for SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). This application can serve the scientific community when prioritizing compounds for *in vitro* screening and may ultimately accelerate the identification of novel drug candidates for the COVID-19 treatment. REDIAL-2020 consists of eleven independently-trained ML models and includes a similarity/substructure 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, the Middle-East respiratory syndrome coronavirus (MERS-CoV) pseudotyped particle entry assay (MERS-PPE) and its counterscreen and the human fibroblast toxicity (hCYTOX) assay. Such assays represent five distinct categories: viral entry (CPE¹ and host cell cytotoxicity counterscreen²), viral replication (3CL enzymatic activity), live virus infectivity (AlphaLISA, TruHit counter screen, and ACE2 enzymatic activity)³, *in vitro* infectivity (coronavirus pseudotyped particle entry, PPE, with associated counter screens for two other coronaviruses, SARS-CoV and MERS) and human fibroblast cytotoxicity (hCYTOX), respectively, 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 analyzing the results.⁶ A more exhaustive description of each assay is provided in the Methods section.

For model development, three different types of descriptors were employed and a best model for each descriptor type was developed by employing various ML algorithms. Then, the three best models from each descriptor type were combined using a voting method to give an ensemble model. These ensemble ML models are integrated into a user-friendly web portal that allows input using three different formats: i) drug name, both as International Nonproprietary Name, INNs (e.g., remdesivir) or as trade name (e.g., Veklury); ii) PubChem CID,⁷ i.e., PubChem Compound ID number (e.g., 121304016 for remdesivir); or iii) using the chemical structure encoded in the SMILES (Simplified Molecular-Input Line-Entry System) format,⁸ respectively. 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. Since 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 ML models can also be used via the command line for large scale virtual screening. As new data sets become available in the public domain, we plan to tune the ML models 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 NCATS data associated with the aforementioned assays was 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 DrugCentral¹¹⁻¹³, to retrieve the chemical structure in SMILES format (see Methods section). For compounds that were not mapped in DrugCentral, the original SMILES strings were retained. Bioactivity data was 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 S1** and **S2**, 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.

For each assay, the data was labeled as positive and negative. The compounds with “LOW” activity class were treated as negative, whereas “HIGH” and “MODERATE” were treated as positive compounds. Finally, the following calculated physico-chemical 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 section). Upon use of the physico-chemical 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. Therefore, $\log P$ and $\log S$ filters were not applied for those datasets. Chemical structures were standardized in terms of

SMILES representation (see Methods section). 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. For each assay, several prediction models were developed, employing three categories of features and 22 distinct machine learning algorithms from the *scikit-learn* package.¹⁵ See **Methods** section for the complete description of features categories. **Supplementary Figure S1** shows the workflow of the model generation. The three different categories of features employed were based on chemical fingerprints, physicochemical descriptors and topological pharmacophore descriptors. Briefly, for fingerprint-based descriptors, 19 different RDKit fingerprints were tested. For physicochemical descriptors, Volsurf+ and RDKit descriptors were employed. For pharmacophore descriptors, Topological Pharmacophore Atom Triplets Fingerprints (TPATF) from Mayachemtools were used. For each model, input data was split into a 70% training set, 15% validation set, and 15% test set using a stratified sampling. **Supplementary Table S3** reports the number of compounds used in training, validation, and test sets for each model. Initial 6 assays (CPE, cytotox, AlphaLISA, TruHit, ACE2, and 3CL) were trained with 22 different classifiers available in scikit-learn (see **Methods**).¹⁷ However, some do not output probability estimates of the class labels (e.g., OneVsOne, Ridge, Nearest Centroid, Linear SVC, etc.). Since our “consensus based on probability” models rely 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 distinct features (see **Methods**) were trained across 11 assays, using “hypopt” for hyperparameter tuning.¹⁶

Applicability Domain. ML models have boundaries for predictability,¹⁷ traditionally called “applicability domain”, AD.¹⁸ AD is defined by the parameter space of the training set upon which ML models are built. ML predictions are deemed reliable when they fall within the AD of that specific model, and less reliable when outside AD. There are two categories of methods to determine AD for classification models: novelty detection and confidence estimation, respectively. Novelty detection defines AD in terms of molecular (feature) space, whereas confidence estimation defines AD in terms of expected prediction reliability.¹⁹ Since confidence estimation is more efficient in reducing error rate compared to novelty detection,¹⁹ we implemented this method for evaluating AD (see **Methods**). Confidence scores, which are averaged for each query molecule, as calculated by default using 3 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 Webpage. By accessing REDIAL-2020 (<http://drugcentral.org/Redial>) from any web browser, including

mobile devices, the submission page is displayed (**Figure 2**). The web server accepts SMILES, drug names, or PubChem CIDs as input. The User Interface (UI) at the top of the page allows users to navigate various options (**Figure 2**). The UI provides summary information about the models, such as model type, which descriptor categories were used for training, and the evaluation scores. The UI further depicts the processes of cleaning the chemical structures (encoded as SMILES) prior to training the ML models. Input queries such as drug name and PubChem CID are converted to SMILES prior to 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 11 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 webapp are summarized in **Supplementary Figure S2**.

Figure 3 shows an output panel example, which is loaded on the same web page. REDIAL-2020 links directly to DrugCentral¹¹⁻¹³ for approved drugs, and to PubChem for chemicals (where available), enabling easy access to additional information about 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; or active in the AlphaLISA assay and inactive in the TruHit assay while not blocking (inactive) ACE2; or active in CoV-PPE while inactive in its counterscreen (CoV-PPE_cs); or active in MERS-PPE while inactive in its counterscreen (MERS-PPE_cs); or active in the 3CL assay with any combination of the above. A schematic representation of the “best profile” that can be defined for a molecule, after running all the prediction models, is depicted in **Figure 4**.

Similarity Search. We used an ECFP4 bit vector fingerprint with 1024 bits, and Tanimoto coefficient (TC) calculations, for the fingerprints present in the database along with that of a query molecule, are computed on the fly. TC 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. TC values range from 0 to 1, with 1 corresponding to identical fingerprints. Thus, a fingerprint-based Tanimoto²² similarity search is conducted for each query molecule against training set molecules, based on NCATS COVID-19 portal⁵ data. The top 10 similar molecules to that of the query molecule, based on Tanimoto coefficient²³ scores, are displayed in the results page.

Discussion

Prior to developing ML models, unsupervised learning can detect patterns that might guide successive steps. Hence, upon definition of the experimental categories (see **Results** for details), 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; **Figure 1**). For CPE data, the majority of compounds showing high to moderate CPE activity are grouped in the right-hand of **Figure 1A**. At the same time, compounds with high to moderate cytotoxicity are grouped in the right-hand region of **Figure 1B**. By inspecting the loading score plot for VolSurf+ descriptors that are likely to contribute to these patterns, we identified membrane permeability, estimated using logP and water solubility, estimated using logS, as major contributors to the first latent variable (see **Supplementary Figure S3**). Compounds with low logP/high logS, 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 **Figures 1C** and **1D**, respectively (see also **Table 1**). For the AlphaLISA assay, although clustering is less pronounced with respect to CPE (**Figure 1A**), 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 ACE2 actives (**Figure 1E**). Thus, permeability and solubility are not the major determinants of this ACE2 inhibition assay.

This preliminary analysis can point to filtering data prior to machine learning. For example, the majority of compounds placed on the left side of the **Figure 1** PCA plot are inactive (except for ACE2). Therefore, prior to developing the ML models, we applied cutoff filters based on compounds calculated logP and logS using ALOGPS²⁶ to every dataset except for ACE2. These filters narrow the focus of ML models on features derived only from compounds for which simple property criteria (e.g., logP and logS) cannot be used to distinguish actives from inactives, specifically, the right-hand regions in **Figure 1**. As the fraction of active compounds excluded from the ACE2 dataset was quite high (34%), logP and logS filters were not applied for ACE2 inhibition.

For 3CL enzymatic activity, data from NCATS was retrieved separately. The initial set contained 12,263 data points. However, data wrangling identified 2,100 duplicates and 2,366 “inconclusive” entries, which were discarded. Additional 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 physico-chemical property filters were not applied. There were no significant activity clusters detected in the 3CL dataset via PCA-VolSurf+ (see **Supplementary Figure S4**).

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 re-evaluated 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. Since the fraction of active compounds filtered out upon applying physico-chemical property filters was over 15%, these filters were not applied for the five new datasets (see also **Table 1**).

With respect to actives vs. inactives, all 11 NCATS assays are highly unbalanced, with a disproportionate ratio of the active (few) compounds compared to inactive (many) compounds. For example, there were ~9 times more inactives than actives and ~3 times more non-cytotoxic compounds than cytotoxic compounds for the CPE and cytotoxicity assays, respectively. Thus, in order to avoid over-training 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, respectively.

To evaluate anti-SARS-CoV-2 activities of novel chemicals, we implemented 11 predictive models based on consensus methods. Of the two (voting-based and probability score-based) consensus methods evaluated, the voting-based consensus model showed better performance (see **Supplementary Figures S5-10**). The voting-based method was thus implemented in the REDIAL-2020 web-app. Consensus models were generated based on 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, based on physicochemical properties. **Supplementary Figure S11 (a-d)** summarizes our initial evaluation and the comparison between various features and ML algorithms. **Supplementary Figures S11a** and **S11b** compare the performance of each feature across 22 ML algorithms (classifiers) and 6 assays. **Supplementary Figures S11c** and **S11d** compare the performance of each classifier across 22 features and 6 assays (CPE, cytotoxicity, AlphaLISA, TruHit, ACE2, and 3CL). For example, the violin plot for the “Avalon” feature (see **Supplementary Figure S11a**) summarizes F1-scores from all 6 assays (and 22 classifiers). Among descriptors, VolSurf+ and LFCFP6 outperformed others, whereas the gradient boost and the multi-layer perceptron (MLP) classifiers performed better among ML algorithms. See **Supplementary Figures S12 and S13** for the comparison of each feature across 15 ML algorithms and 11 assays. **Supplementary Figures S14-47** depict more detailed comparisons across different features and ML algorithms with respect to individual models.

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 outperformed

VolSurf+ in cytotox, AlphaLISA, ACE2, 3CL, MERS-PPE_cs, CoV-PPE, CoV-PPE_cs, and hCYTOX, whilst VolSurf+ descriptors outperformed RDKit in CPE and hCYTOX along with similar results in MERS-PPE and TruHit based on the tested evaluation metrics such as Accuracy, F1-score, and AUC in validation sets (see **Supplementary Figures S48-58**). However, the situation slightly changed when considering consensus models. Inclusion of VolSurf+ yielded 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 Figures S5-10** show a comparison of the best models from each feature category. As the NCATS team released data for more compounds for the 6 initial assays plus 5 new assays in September 2020, we updated the initial 6 models and developed models for the 5 new assays. The comparison of models from each category for the new and updated models were shown in **Figures S53-57**. Among 11 assay models, the voting-based consensus model performed slightly better than individual feature type models based on validation F-1 score results; in 3 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, since RDKit is open-source software that can be ported and dockerized without restrictions. **Tables 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 experiments²⁷ to explore and download published *in vitro* COVID-19 bioactivity data for compounds which was reported in various recent papers.²⁸⁻³⁶ 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 S4**). Out of 27 compounds, 3 were excluded upon applying the logP/logS filters, and the remaining 24 were predicted by the CPE model. 16 compounds were correctly predicted as active by the consensus model i.e., at least two models (see **Supplementary Figure S59**), with 8 compounds predicted as inactive. Among those predicted to be inactive, the majority stem from the Ellinger et al. work, derived from Caco-2 cells for CPE experiments. There is a high degree of variability between these two CPE assays (Caco-2 vs. Vero E6), which explains the lack of predictivity using Vero E6-trained CPE models for Caco-2 data. The second dataset of 3CL (Mpro) inhibitors³⁶ identified 6 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), respectively (see **Supplementary Table S5**). Among these 6 inhibitors, our consensus 3CL model predicted 4 of them correctly as actives, and 5 of them as actives by at least one of the three 3CL ML models. Thus, the REDIAL-2020 suite of models correctly predicted 67% of the external compounds for CPE and 3CL inhibitors³⁶, respectively.

Although the external predictivity of CPE model appear to underestimate previous model performance in the validation and external sets (see **Supplementary Table S6**), it has been noted that CPE experiments are affected by significant intra- and inter-experiment 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 portal⁵ 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 ML models: CPE, cytotox, AlphaLISA, TruHit, ACE2, 3CL, SARS-CoV-PPE, SARS-CoV-PPE CS, MERS-CoV PPE, MERs-CoV PPE CS, and hCYTOX. Such 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 10 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 ML 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 ML models. For example by using the TMPRSS2 inhibition assay³⁷ 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 ML models and make these models available in future releases of REDIAL.

Methods

HTS Assays. The SARS-CoV-2 cytopathic effect (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). While 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. Since this confounds the interpretation of CPE assay results, masking the cyto-protective activity, a counter-screen to measure host (Vero E6) cell cytotoxicity is used to detect such compounds. Thus, a net, positive result from the combined CPE assays consist of a compound showing a protective effect but no cytotoxicity.

The Spike-ACE2 protein-protein interaction (AlphaLISA) assay measures a compound's ability to disrupt the interaction

between the viral Spike protein and its human receptor protein, ACE2 (angiotensin-converting enzyme type 2).³⁸ The surface of the ACE2 protein is the primary host factor recognized and targeted by SARS-CoV-2 virions.³⁹ 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 counter screen is used to determine false positives, i.e., 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) because 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 catalyzing the hydrolysis of angiotensin II (a vasoconstrictor peptide) into the vasodilator angiotensin (1-7).⁴⁰ While blocking the Spike-ACE2 interaction may stop viral entry, drugs effective in this manner could cause unwanted side-effects by blocking the endogenous vasodilating function of ACE2. Thus, the ACE2 assay 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 3C-like proteinase (3CL), also called “main protease” or Mpro,³⁶ which cleaves the two SARS-CoV-2 polyproteins into various proteins (e.g., RNA polymerases, helicases, and methyltransferases, etc.), which are essential to the viral life cycle. Since inhibiting the 3CL protein disrupts the viral replication process, this makes 3CL an attractive drug target.⁴¹ The SARS-CoV-2 3CL biochemical assay measures compounds' ability to inhibit recombinant 3CL cleavage of a fluorescently labeled peptide substrate.

In this category there are four assays: SARS-CoV pseudotyped particle entry and its counter screen, MERS-CoV pseudotyped particle entry, and its counter screen. 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. Instead, it uses pseudotyped particles that are generated by the fusion of the coronavirus spike protein with a murine leukemia virus core. Since 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.⁴² 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). Therefore, this assay is intended for discarding compounds that are likely to show high toxicity in human cells (i.e. 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),⁴³ by synonym (name), and by matching CAS (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 CAS 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. Whenever possible, SMILES were retrieved from DrugCentral. 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 RDKit⁴⁴ to transform input SMILES into standardized molecular representations. Five different filters were implemented via RDKit: i) input SMILES were standardized into canonical (isomeric where appropriate) SMILES strings. The input SMILES that failed to convert were discarded; ii) RDKit Salt Stripper was used to de-salt input compounds (i.e., remove the salt structures). The “doNotRemoveEverything” feature leaves the last salt structure when the entire canonical SMILES string is comprised of salts only; (iii) RDKit “Uncharger” neutralizes input molecules by adding/removing hydrogen atoms and setting formal charges to zero (except for e.g., quaternary ammonium cations); iv) Canonical SMILES were then formalized into specific tautomers using RDKit.

Molecular Features/Descriptors. A total of 22 features of three distinct types (fingerprints-based, pharmacophore-based, and 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.^{45,46} Circular fingerprints include the extended-connectivity fingerprints (ECFP_x) and feature-connectivity fingerprints (FCFP_x), where *x* is 0, 2, 4, and 6 are the bond length or diameter for each circular atom environment. ECFP consists of the element, number of heavy atoms, isotope, number of hydrogen atoms, and ring information, whereas FCFP consists of pharmacophore features.

Avalon and MACCS (Molecular ACCess System) 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, and augmented symbol path, etc. 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 RDK_x (where *x* is 5, 6, 7), topological torsion (TT), HashTT, atom pair (AP), and HashAP. The size of each fingerprint is 1024. The longer, 16,384-bits, versions of the fingerprint, marked by the prefix “L” (LAvalon, LECFP6, LECFP4, LFCFP6 and LFCFP4, respectively) were used for comparison.

Topological pharmacophore atomic triplets fingerprints (TPATF) were obtained using Mayachemtools.⁴⁷ TPATF describes the ligand sites that are necessary for molecular recognition of a macromolecule or a ligand, and passes that information to the ML 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 (*i*) triangle rule, i.e., the length of each side of a triangle cannot exceed the sum of the lengths of the other two sides; and (*ii*) 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), H (hydrophobic), and Ar (aromatic) were used during generation of TPATF.⁴⁸

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.⁴⁴ They are either experimental properties or theoretical descriptors, which are e.g. molar refractivity, logP, 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 3D map of the interaction energies between the molecule and chemical probes (GRID-based molecular interaction fields, or MIFs).⁴⁹ VolSurf+ compresses the information present in MIFs into numerical descriptors, which are simple to use and interpret.^{25,50}

Machine Learning Classifiers. Using assay data as input (specifically, CPE, cytotox, AlphaLISA, TruHit, ACE2, and 3CL) we trained ML models using the following 24 different classifiers: Complement Naive Bayes, Extreme Gradient Boosting, KNeighbors, Gradient Boosting, Perceptron, One Vs Rest , Extra-Tree, Ridge, One Vs One, Bagging, Random Forest, Output Code, Passive Aggressive, Linear SVC, Stochastic Gradient Descent, Logistic Regression, Extra Trees, Multinomial Naive Bayes, Ada Boost, Decision Tree, Nearest Centroid, Multi-layer perceptron, Support Vector Machines and Quadratic Discriminant Analysis, respectively. All these algorithms are implemented in the *scikit-learn* package.¹⁷ 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, respectively) that served as input to the ML classifiers are described above. All classifiers were trained on their default configurations. For hyperparameter tuning we used hypot¹⁶ and the best suited combination of classifiers and features (see **Supplementary Table S7**). 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.

Implementation and Accessibility

Web Portal

REDIAL-2020 is available online at <http://drugcentral.org/Redial>

Code Availability

All of the codes and the trained models are available at <https://github.com/sirimullalab/redial-2020>

Data Availability

All data used for the model described in this work are available at <https://github.com/sirimullalab/redial-2020>. The datasets were originally collected from the following links (warning: these data are subject to change without notice):

CPE <https://opendata.ncats.nih.gov/covid19/export/data/assay/14>

cytotox <https://opendata.ncats.nih.gov/covid19/export/data/assay/15>

AlphaLISA <https://opendata.ncats.nih.gov/covid19/export/data/assay/1>

TruHit <https://opendata.ncats.nih.gov/covid19/export/data/assay/2>
ACE2 <https://opendata.ncats.nih.gov/covid19/export/data/assay/6>
3CL <https://opendata.ncats.nih.gov/covid19/export/data/assay/9>
CoV-PPE <https://opendata.ncats.nih.gov/covid19/export/data/assay/22>
CoV-PPE_cs <https://opendata.ncats.nih.gov/covid19/export/data/assay/23>
MERS-PPE <https://opendata.ncats.nih.gov/covid19/export/data/assay/24>
MERS-PPE_cs <https://opendata.ncats.nih.gov/covid19/export/data/assay/25>
hCYTOX <https://opendata.ncats.nih.gov/covid19/export/data/assay/21>

Acknowledgement

We thank the High-Performance Computing support staff (Marc T. Hertlein and Leopoldo A. Hernandez) at The University of Texas at El Paso for assistance in using the Chanti cluster. We also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://www.tacc.utexas.edu>. Access to unpublished SARS-CoV-2 experimental data from Dr. Colleen Jonsson (University of Tennessee Health Sciences Center) and Dr. Steven Bradfute (University of New Mexico Health Sciences Center) is gratefully acknowledged.

Author Contributions

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

Funding

Dr. Sirimulla acknowledges support from the National Science Foundation through NSF-PREM grant #DMR-1827745. The DrugCentral component of this work is funded by NIH Common Fund U24 CA224370.

Corresponding Authors

ssirimulla@utep.edu

TOpera@salud.unm.edu

Conflict of Interest

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

Tables & Figures

Table 1. Number and percentage of compounds outside logP and logS criteria.

Assay	Actives (relative percentage)	Inactives (relative percentage)
CPE	44 (8%)	2913 (37%)
cytotox	193 (14%)	2764 (39%)
AlphaLISA	143 (19%)	1119 (49%)
TruHit	134 (16%)	1128 (51%)
ACE2	70 (38%)	1192 (41%)
3CL	81 (28%)	3330 (37%)
CoV-PPE	43 (27%)	881 (51%)
CoV-PPE_cs	247 (28%)	1085 (44%)
hCYTOX	81 (22%)	1306 (39%)
MERS-PPE	104 (20%)	1024 (49%)
MERS-PPE_cs	46 (24%)	1082 (45%)

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	0.70	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	0.75	0.75	0.75	0.775	0.755	0.777	0.84	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	0.7	0.661
hCYTOX	0.736	0.736	0.736	0.736	0.736	0.71	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	0.68	0.629	0.739	0.703

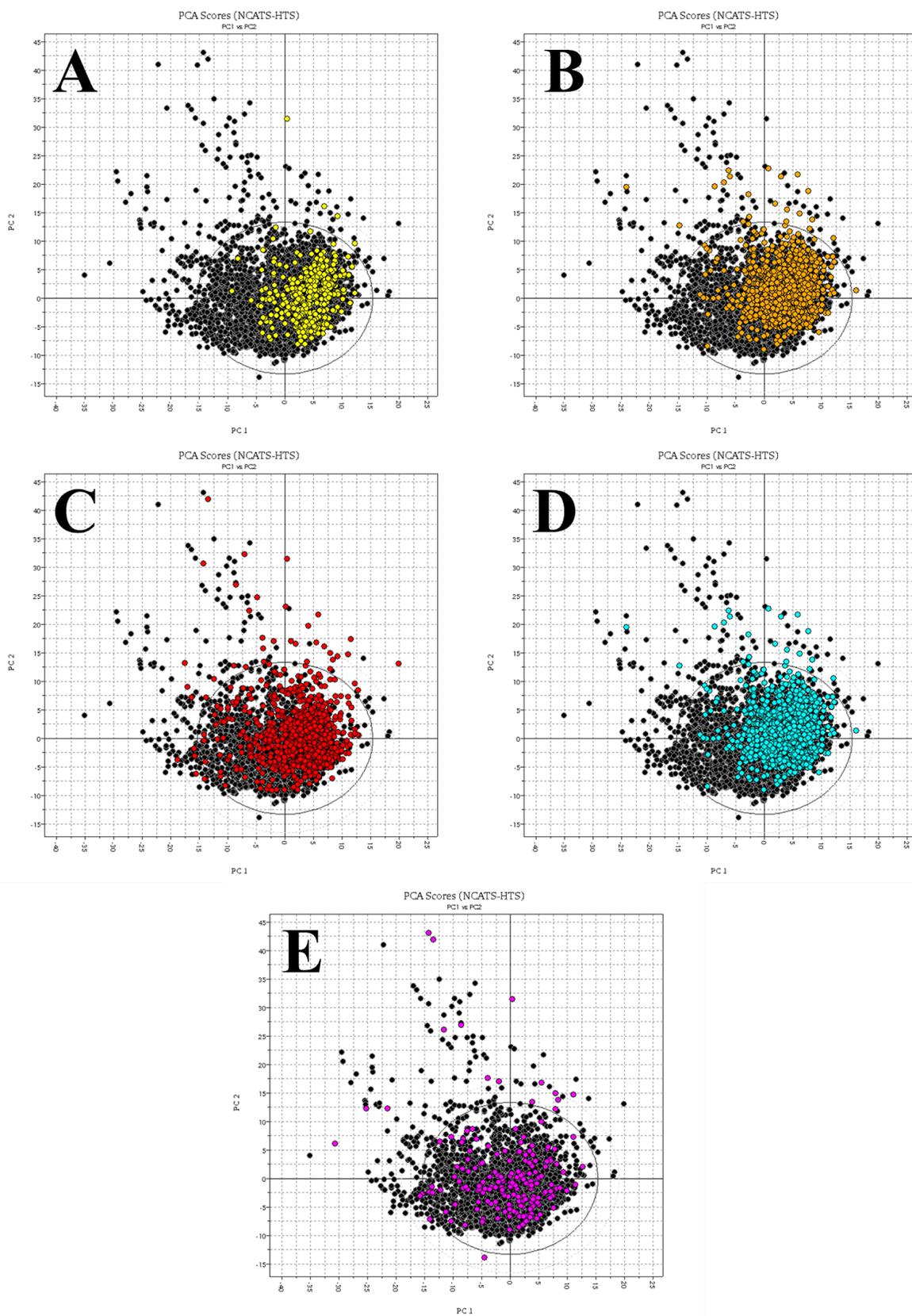
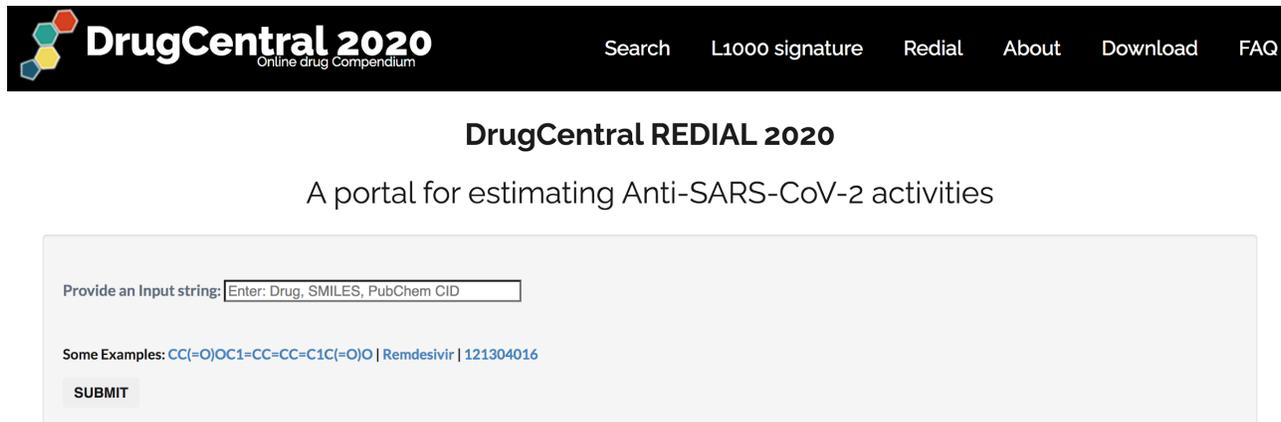


Figure 1. PCA scores plots of the molecules tested in NCATS SARS-CoV-2 experiments based on VolSurf+ descriptors. On each plot, the compound position is defined along the first and the second principal component, respectively PC1

and PC2. A) CPE compounds colored by CPE categories: high/moderate activity in yellow and low activity in black; B) cytotoxic compounds colored by cytotoxicity categories: high/moderate cytotoxic in orange and low (not) cytotoxic in black. C) AlphaLISA compounds colored by Spike-ACE2 interaction blockers categories: high/moderate (strong) blockers in red and low (weak) blockers in black. D) TruHit compounds, colored by AlphaLISA readout interfering categories: high/moderate interfering in cyan and low interfering in black. E) ACE2 compounds, colored by ACE2 inhibition categories: high/moderate (strong) inhibitors in magenta and low (weak) inhibitors in black.



DrugCentral 2020
Online drug Compendium

Search L1000 signature Redial About Download FAQ

DrugCentral REDIAL 2020

A portal for estimating Anti-SARS-CoV-2 activities

Provide an Input string:

Some Examples: [CC\(=O\)OC1=CC=CC=C1C\(=O\)O](#) | [Remdesivir](#) | [121304016](#)

SUBMIT

Figure 2. REDIAL-2020 submission webpage.

RESULTS

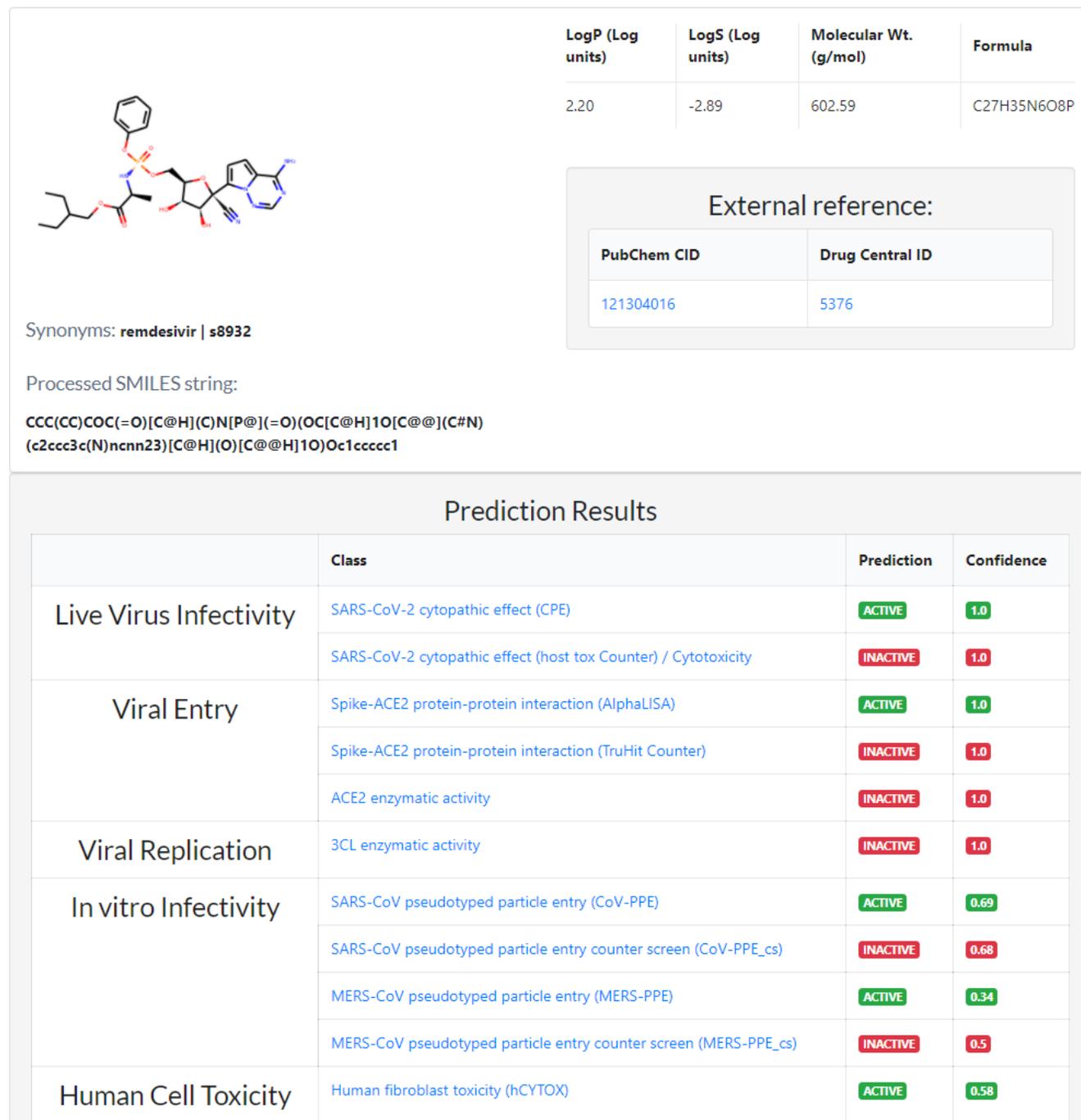


Figure 3. Screenshot of the webpage displaying the ML estimates and for a query molecule.

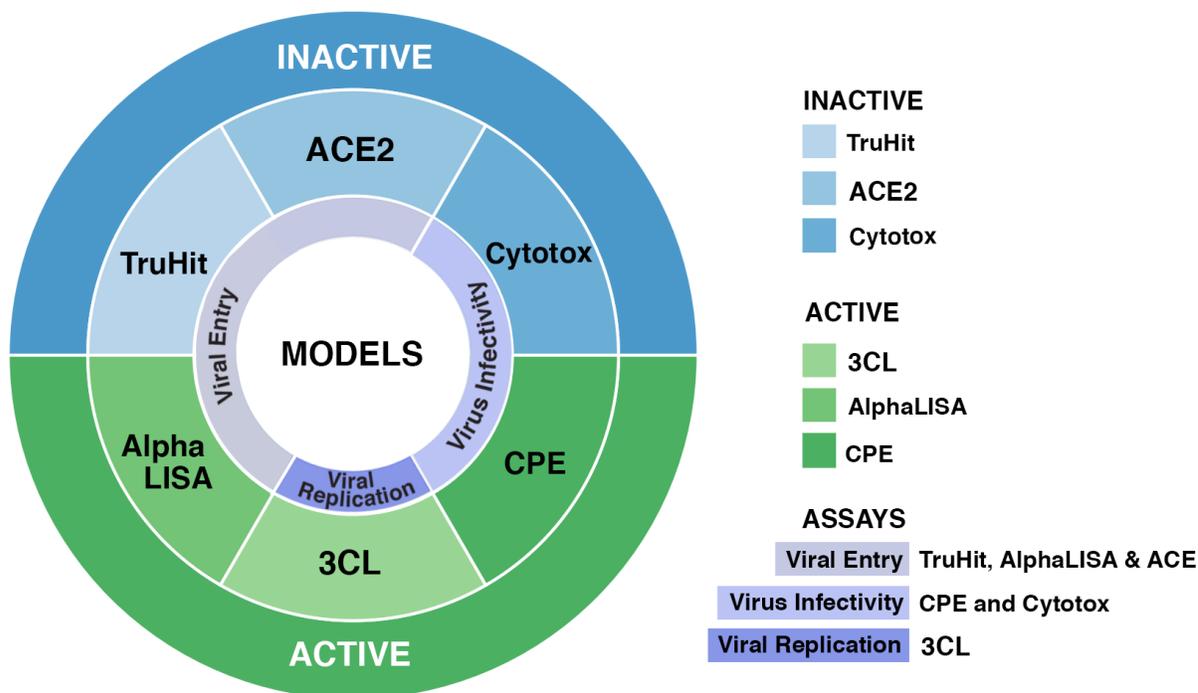


Figure 4. Schematic representation of the most desirable profile for anti-SARS-CoV-2 activities that can be observed via REDIAL-2020 predictions, based on the SARS-CoV-2 specific set of assays. The five additional assays (not depicted here) offer supporting evidence for the decision-making process and hit prioritization.

References

1. Gorshkov, K. *et al.* The SARS-CoV-2 cytopathic effect is blocked with autophagy modulators. *bioRxiv* (2020) doi:10.1101/2020.05.16.091520.
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.* (2020) doi:10.1021/acscptsci.0c00161.
4. Brimacombe, K. R. *et al.* An OpenData portal to share COVID-19 drug repurposing data in real time. *bioRxiv* (2020) doi:10.1101/2020.06.04.135046.
5. NCATS OpenData COVID-19. <https://opendata.ncats.nih.gov/covid19/assays>.
6. Huang, R. *et al.* Massive-scale biological activity-based modeling identifies novel antiviral leads against SARS-CoV-2.

bioRxiv (2020) doi:10.1101/2020.07.27.223578.

7. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44**, D1202–13 (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., Cebron, N., Dill, F. & Others. KNIME: the Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization. Springer: ISSN 1431–8814* (2007).
11. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic Acids Res.* **45**, D932–D939 (2017).
12. Ursu, O. *et al.* DrugCentral 2018: an update. *Nucleic Acids Res.* **47**, D963–D970 (2019).
13. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa997.
14. Seethala, R. & Zhang, L. *Handbook of Drug Screening.* (CRC Press, 2016).
15. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
16. Hypopt. <https://github.com/cgnorthcutt/hypopt>.
17. Oprea, T. I. & Waller, C. L. Theoretical and practical aspects of three-dimensional quantitative structure-activity relationships. in *Reviews in Computational Chemistry* 127–182 (John Wiley & Sons, Inc., 2007).
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. Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic Classification Methods and their Applicability Domain. *Mol. Inform.* **35**, 160–180 (2016).
20. Liu, R. & Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **59**, 181–189 (2019).
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 Suppl 2**, S29–39 (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 (N Y)* **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**, (2020).
 29. Weston, S., Haupt, R., Logue, J., Matthews, K. & Frieman, M. B. FDA approved drugs with broad anti-coronaviral activity inhibit SARS-CoV-2 in vitro. 2020.03.25.008482 (2020) doi:10.1101/2020.03.25.008482.
 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. *bioRxiv* (2020) doi:10.1101/2020.04.07.030734.
 32. Riva, L. *et al.* A Large-scale Drug Repositioning Survey for SARS-CoV-2 Antivirals. *bioRxiv* (2020) doi:10.1101/2020.04.16.044016.
 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. *bioRxiv* (2020) doi:10.1101/2020.05.27.117184.
 35. Riva, L. *et al.* Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* (2020) doi:10.1038/s41586-020-2577-1.
 36. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
 37. Shrimp, J. H. *et al.* An Enzymatic TMPRSS2 Assay for Assessment of Clinical Candidates and Discovery of Inhibitors as Potential Treatment of COVID-19. *bioRxiv* (2020) doi:10.1101/2020.06.23.167544.
 38. 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).
 39. 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).
 40. Keidar, S., Kaplan, M. & Gamliel-Lazarovich, A. ACE2 of the heart: from angiotensin I to angiotensin (1--7).

- Cardiovasc. Res.* **73**, 463–469 (2007).
41. 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).
 42. Millet, J. K. *et al.* Production of Pseudotyped Particles to Study Highly Pathogenic Coronaviruses in a Biosafety Level 2 Setting. *J. Vis. Exp.* (2019) doi:10.3791/59010.
 43. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 23 (2015).
 44. Landrum, G. & Others. RDKit: Open-source cheminformatics. (2006).
 45. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5**, 26 (2013).
 46. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
 47. Sud, M. MayaChemTools: An Open Source Package for Computational Drug Discovery. *J. Chem. Inf. Model.* **56**, 2292–2297 (2016).
 48. 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).
 49. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857 (1985).
 50. 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).