Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space

Joel Kowalewski\textsuperscript{a}, Anandasankar Ray\textsuperscript{a,b,}\textsuperscript{*}

\textsuperscript{a} Interdepartmental Neuroscience Program, University of California, Riverside, CA 92521, USA
\textsuperscript{b} Department of Molecular, Cell and Systems Biology, University of California, Riverside, CA 92521, USA

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Microbiology
Virology
Toxicology
Computer-aided drug design
Viruses
Viral disease
Structure activity relationship
SARS-CoV-2
Covid-19
Chemical informatics
Machine learning
Drug discovery
ACE2

\textbf{ABSTRACT}

There is an urgent need for the identification of effective therapeutics for COVID-19 and we have developed a machine learning drug discovery pipeline to identify several drug candidates. First, we collect assay data for 65 target human proteins known to interact with the SARS-CoV-2 proteins, including the ACE2 receptor. Next, we train machine learning models to predict inhibitory activity and use them to screen FDA registered chemicals and approved drugs (~100,000) and ~14 million purchasable chemicals. We filter predictions according to estimated mammalian toxicity and vapor pressure. Prospective volatile candidates are proposed as novel inhaled therapeutics since the nasal cavity and respiratory tracts are early bottlenecks for infection. We also identify candidates that act across multiple targets as promising for future analyses. We anticipate that this theoretical study can accelerate testing of two categories of therapeutics: repurposed drugs suited for short-term approval, and novel efficacious drugs suitable for a long-term follow up.

1. Introduction

SARS-CoV-2 is a novel coronavirus that is responsible for the COVID-19 disease which is a rapidly evolving global pandemic. Coronaviruses primarily target the upper respiratory tract and the lungs, with varying degrees of severity. Related coronaviruses such as the SARS-CoV emerging in China in 2002 and the MERS-CoV in the Middle East in 2012 result in severe respiratory conditions. The SARS-CoV-2 also produces similarly severe respiratory conditions, albeit at a lower rate but with a higher contagion factor \cite{1}. Alarming, infected individuals may be asymptomatic carriers, presumably harboring the viral infection in the upper airway tract, increasing the likelihood of infecting populations that are most susceptible to severe complications \cite{2, 3}.

Although the mechanisms underlying SARS-CoV-2 infection are not completely understood, select human proteins are targets for the virus including ACE2 \cite{4}. The SARS-CoV-2 receptor binding domain (RBD) interacts strongly with the human ACE2 receptor and TMPRSS2 to enter a human cell \cite{5}. In addition to ACE2, a recent systems-level analysis of protein-protein interaction with peptides encoded in the SARS-CoV-2 genome identified ~300 additional human proteins, of which, 66 were considered suitable candidates for identification of therapeutics \cite{6}. Gordon et al. performed an in vitro assay with human cells expressing 26 SARS-CoV-2 proteins, which was followed by an analysis for high-confidence interactions. Of the 100s of reported interactions 66 were prioritized, and the authors subsequently mined and tested FDA approved drugs that were known or suspected to target these human proteins. Most of the human target proteins are overexpressed in the respiratory tract. Of particular note is the entry receptor ACE2 which is expressed at high levels in a few cell types of the nasal epithelium, as well as elsewhere \cite{6, 7}. This could be an unusual opportunity for volatile inhaled therapeutics and prophylactics that will have direct access to the cells that are infected by the virus.

The Gordon et al study also identified FDA-approved drugs that have known activity against these human protein targets or are structurally related to chemicals with known activity on the targets. While these drugs have not been comprehensively tested on the virus, another study performed high-throughput testing of ~12,000 FDA-approved or clinical stage drugs on viral replication in cell lines \cite{8}. This study identified at least 6 potential leads that include a kinase inhibitor, a CCR1 inhibitor...

\textsuperscript{*} Corresponding author.
E-mail address: anand.ray@ucr.edu (A. Ray).

https://doi.org/10.1016/j.heliyon.2020.e04639
Received 15 May 2020; Received in revised form 16 June 2020; Accepted 3 August 2020
2405-8440/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
and 4 cysteine protease inhibitors that are candidates for testing in clinical trials.

Since the regulatory process for the approval of new drugs can take several years, the repurposing of FDA approved drugs for COVID-19 offers a potential fast-track to approval. One of the more promising candidates being tested is the antiviral Remdesivir, which has been effective in vitro [9] as well as in non-human primates [10], with human trials currently ongoing. The other drug being tested is the antimalarial, hydroxychloroquine, which showed some promise alongside the antibiotic, azithromycin, in small clinical trials [11, 12]. However, hydroxychloroquine has shown less promise in larger trials for treating COVID-19 [13].

While drug repurposing is expedient, it is possible that drugs designed for other diseases will not be as well suited to respiratory organs, where a large percentage of putative human proteins targeted by the virus are enriched [6], or to the nervous system, implicated by neurological symptoms as well as prior evidence that coronaviruses can cross the blood brain barrier [14, 15]. Drug-development strategies are also often guided by minimizing off-target interactions. Repurposed drugs might have to be used in combination, and the side effects and interactions that this entails are presently not well defined. While there are recent efforts exploring novel, directed therapies from small molecule libraries [16], it is desirable to identify 100–1000s of putative chemicals as the majority may be difficult to synthesize in mass, prove toxic at therapeutic concentrations, or yield inconsistent benefits across patients due to genetic variability. These shortcomings have significantly increased the demand for additional drugs or small molecules that might interfere with viral entry and replication. Additionally, if prophylactics or non-toxic, easy-to-use therapeutics were available even for mild cases that do not require hospitalization and experimental drug treatments, contracting the virus may nevertheless impact long-term health and community transmission [17].

There are subsequently unmet needs in COVID-19 research, including identification of compounds that target the relevant SARS-CoV-2 human proteins from (1) approved drugs, (2) FDA registered chemicals or (3) a large repository of ~14 million purchasable chemicals from the ZINC 15 database [18], which we computed additional properties for such as mammalian toxicity, vapor pressure, and logP. For 65 human protein targets that SARS-CoV-2 interacts with that had publicly available bioassay and chemical data [6], we first generated a database of predictions based on structural similarity to chemicals that interact with the targets and then machine learning models (34). Many chemicals we have identified have little or no known biological activities and are predicted to have low toxicity in addition to a wide range of vapor pressures. These data are a resource to rapidly identify and test novel, safe treatment strategies for COVID-19 and other diseases where the target proteins are relevant.

2. Results

2.1. Identification of important structural features from known inhibitors of human target proteins

In order to test whether there is a structural basis for inhibitors of the target proteins identified previously [5, 6], we used two complementary approaches to evaluate each target’s training set of compounds with known activity, compiled from the literature. First, we performed an exhaustive search for maximum common substructures among active chemicals. In some cases, enriched substructures were apparent among known ligands, with slight variation in the substructure based on the sensitivity to the targets, suggesting physicochemical features may be relevant in predicting activity against these targets (Supplementary Table 1). Next, we used a machine learning pipeline for predicting chemicals that interfere with SARS-CoV-2 targets. It involves selection of important physicochemical features for each target, followed by fitting support vector machines (SVM) with these features and then evaluating the predictions using various computational validation methods (Figure 1A). The chemical features that best predicted activity for the different targets included simple 2D information, describing the type and number of bonds, but also more abstract 3D geometries (Tables 1 and 2). Identification of each target-specific feature set provides a foundation to better understand the physicochemical basis of the activity. To that end, Supplementary Tables 2-3 include more comprehensive rank ordered lists of the physicochemical features that optimally predict activity against the targets (details about the feature ranking algorithms in Materials and Methods).

2.2. Machine learning models can successfully predict activity from chemical structure

We identified 24 targets with training sets large enough to model the log IC50, Ki, or AC50 (Figure 2A). Rigorous computational validation was performed and the results on training (Figure 2B, left) and test data that had been set aside (Figure 2C, left) indicated good overall performance according to the average mean absolute error (MAE) and the correlation between predicted and observed assay measures (MAE = 0.48; R = 0.62). Predictions of log K for the viral entry receptor, ACE2, were also accurate (test set R = 0.92; test set mean absolute error (MAE) = 0.53) (Figure 2C, left; Supplementary Information 1).

For some of the viral targets, we noticed that assay data included additional inhibitory measurements or descriptions of general activity against the targets. Some of the available data such as % inhibition, for instance, are less quantitative. However, to include as much of the available data as possible, we created models to identify physicochemical features that might broadly contribute to inhibition or activity against the targets. We therefore assigned binary, active and inactive, labels to the chemicals, then trained models as outlined before (Figure 2A; Materials and Methods). The models that were developed using this classification approach similarly proved successful, validating over partitions of the training data (avg. AUC = 0.87, avg. Shufle AUC = 0.50, p < 10^-13) (Figure 2B, right), as well as over sets of external test chemicals (avg. AUC = 0.83, avg. Shuffle AUC = 0.51, p < 10^-6) (Figure 2C, right) (Supplementary Information 1). Collectively, these results suggested the models provided accurate predictions and could be used to screen approved drug libraries as well as databases of commercially available chemicals for novel therapeutics.

2.3. Predicting candidates for repurposing of FDA-approved drugs

Repurposing of existing FDA approved drugs offers a path towards rapid deployment of therapeutics against SARS-CoV-2. Approved drugs may have activity that extend beyond the original target protein. Accordingly, we used the machine learning models to predict activities of ~100,000 FDA registered chemicals (UNII database) [19] as well as the DrugBank [20] and Therapeutic Targets [21, 22] databases, which include information on drug interactions, pathways, and approval status. Interestingly, some of the approved drugs are predicted to have high activity against the SARS-CoV-2 targets (Figure 3A). In order to identify more efficacious candidates, we isolated the drugs scoring in the top 25 for multiple targets and found a few of high priority (Figure 3B). The structural analysis suggested that hits visually display 2D similarity to known active chemicals as well. (Supplementary Information 2).

2.4. Predicting volatile drug candidates from a large ~14M chemical space

Given that many of the human target proteins are overexpressed in the respiratory tract, including the entry receptor ACE2 in only a few cells types of the nasal epithelium, the upper airways and lungs [7, 23], we reasoned that volatile chemicals may offer a unique opportunity as inhaled therapeutics that will have direct access to the cells and tissues that are infected by the virus. We used the machine learning models to search a large database of ~14 million commercially available chemicals
Figure 1. Machine learning pipeline to identify chemicals that interfere with SARS-CoV-2 targets. a) Overview of the pipeline to predict chemicals for 65 SARS-CoV-2 human targets selected from Gordon et al., 2020 and using bioassay data from publicly available databases. b) Graphically depicts the pipeline details. Available bioassay data on the viral targets were mined for information to use in machine learning or structural analysis. This resulted in 24 targets that could be modeled using values for the most abundant inhibitory assay measure (e.g. Ki or IC50) and 21 targets modeled by classifying broad inhibition or activity against the proteins (34 unique targets in total). The remaining targets with limited data were funneled into a structural similarity analysis, which aids in developing more bioassay data and helps clarify the chemical features contributing to bioactivity. For targets modeled with supervised machine learning, optimal chemical features were identified on subsets of training data. The top features were sampled by support vector machines (SVM). These models were then aggregated. In certain cases, the Random Forest algorithm was included to improve the fit. External chemicals were used to verify successful predictions. Models trained for the 34 targets predicted large chemical databases including FDA registered chemicals and approved drugs, as well as 10+ million purchasable chemicals from the ZINC database. Top scoring predicted chemicals were subsequently assigned theoretical toxicity, log vapor pressure, and MLOGP, which estimates membrane permeability.
Table 1. Important chemical features for regression models. Top three physicochemical features for the viral targets with known bioassay activities.

| Feature | Target | Description |
|---------|--------|-------------|
| GATS5s | ABCC1  | Geary autocorrelation of lag 5 weighted by I-state |
| RDF055m| ABCC1  | Radial Distribution Function - 055/weighted by mass |
| SpMax_B(s)| ABCC1 | leading eigenvalue from Burden matrix weighted by I-State |
| CATS2D_08_AA| BRD2 | CATS2D Acceptor-Acceptor at lag 08 |
| RDF035s| BRD2  | Radial Distribution Function - 035/weighted by I-state |
| SpDiam_X|  | spectral diameter from chi matrix |
| HATS8p| BRD4  | leverage-weighted autocorrelation of lag 8/weighted by polarizability |
| R5i+ | BRD4  | R maximal autocorrelation of lag 5/weighted by ionization potential |
| RDF035m| BRD4  | Radial Distribution Function - 035/weighted by mass |
| Eig02_EA(bo) | CSNK2A2 | eigenvalue n. 2 from edge adjacency mat. weighted by bond order |
| Eig05_EA(bo) | CSNK2A2 | eigenvalue n. 5 from edge adjacency mat. weighted by bond order |
| SpMax2_Bh(m) | CSNK2A2 | largest eigenvalue n. 2 of Burden matrix weighted by mass |
| CATS2D_04_AA | CSNK2B | CATS2D Acceptor-Acceptor at lag 04 |
| SHED_DN | CSNK2B | SHED Donor-Negative |
| SpMin1_Bh(m) | CSNK2B | smallest eigenvalue n. 1 of Burden matrix weighted by mass |
| DISPm | DCTPP1 | displacement value/weighted by mass |
| HATS7u | DCTPP1 | leverage-weighted autocorrelation of lag 7/unweighted |
| Mor31s | DCTPP1 | signal 31/weighted by I-state |
| MATS1e | DNMT1 | Moran autocorrelation of lag 1 weighted by Sanderson electronegativity |
| Mor23m | DNMT1 | signal 23/weighted by mass |
| TDB06u | DNMT1 | 3D Topological distance based descriptors - lag 6 unweighted |
| GATS4m | GFER  | Geary autocorrelation of lag 4 weighted by mass |
| Mor14m | GFER  | signal 14/weighted by mass |
| RSi | GFER | R autocorrelation of lag 5/weighted by ionization potential |
| DISPp | HDAC2 | displacement value/weighted by polarizability |
| IC2 | HDAC2 | Information Content index (neighborhood symmetry of 2-order) |
| P_VSA_MR.5 | HDAC2 | P_VSA-like on Molar Refractivity, bin 5 |
| F04[C-C] | IMPDH2 | Frequency of C - C at topological distance 4 |
| HOMA | IMPDH2 | Harmonic Oscillator Model of Aromaticity index |
| VE1_B(s) | IMPDH2 | coefficient sum of the last eigenvector (absolute values) from Burden matrix weighted by I-State |
| Eig02_AEA(dm) | ITGB1 | eigenvalue n. 2 from augmented edge adjacency mat. weighted by dipole moment |
| SHED_AA | ITGB1 | SHED Acceptor-Acceptor |
| SpMax2_Bh(s) | ITGB1 | largest eigenvalue n. 2 of Burden matrix weighted by I-state |
| F01(C-C) | MARK2 | Frequency of C - N at topological distance 10 |
| nPyrroles | MARK2 | number of Pyrroles |
| SanNH | MARK2 | Sum of sanNH E-states |
| max_conj_path | MARK3 | maximum number of atoms that can be in conjugation with each other |
| SanNH | MARK3 | Sum of sanNH E-states |
| VE1_H2 | MARK3 | coefficient sum of the last eigenvector (absolute values) from reciprocal squared distance matrix |
| GATS3s | NSD2 | Geary autocorrelation of lag 3 weighted by I-state |
| HOMA | NSD2 | Harmonic Oscillator Model of Aromaticity index |
| Mor16s | NSD2 | signal 16/weighted by I-state |
| H7m | PABC1 | H autocorrelation of lag 7/weighted by mass |
| JGI7 | PABC1 | mean topological charge index of order 7 |
| P_VSA_MR.2 | PABC1 | P_VSA-like on Molar Refractivity, bin 2 |
| GATS4m | PLAT | Geary autocorrelation of lag 4 weighted by mass |
| Mor04s | PLAT | signal 04/weighted by I-state |
| R6p+ | PLAT | R maximal autocorrelation of lag 6/weighted by polarizability |
| nPyrroles | PRKACA | number of Pyrroles |
| RDF040v | PRKACA | Radial Distribution Function - 040/weighted by van der Waals volume |
| SpMin3_Bh(m) | PRKACA | smallest eigenvalue n. 3 of Burden matrix weighted by mass |
| Eig02_EA(bo) | PSEN2 | eigenvalue n. 2 from edge adjacency mat. weighted by bond order |
| nArX | PSEN2 | number of X on aromatic ring |
| VE1sign_D/Dt | PSEN2 | coefficient sum of the last eigenvector from distance/detour matrix |
| SHED_DL | PTGES2 | SHED Donor-Lipophilic |
| VE2sign_G | PTGES2 | average coefficient of the last eigenvector from geometrical matrix |
| VE3sign_G | PTGES2 | logarithmic coefficient sum of the last eigenvector from geometrical matrix |
| CATS3D_08_AL | RPK1 | CATS3D Acceptor-Lipophilic BIN 08 (8.000-9.000 Å) |

(continued on next page)
3. Discussion

SARS-CoV-2 is a significant world health crisis. The full scope of COVID-19 disease and any long-term health complications following infection remain unclear. Although vaccines are the best long-term solution, treatments will be necessary to mitigate disease severity in the short term. What is concerning is that several repurposed drugs have already been tested in some form of clinical trial, and only one drug, Remdesivir, has shown a clear benefit in randomized clinical trials. Additionally, there is not yet sufficient evidence that an effective vaccine can be found for the SARS-CoV-2 virus, and therefore drug candidate pipelines are extremely important for the long-term research effort against COVID-19. A vaccine against SARS-CoV-2 would likely need to stimulate local immunity, since the infection is limited to mucosal surfaces, and these could be short-lived immunities.

We have therefore taken a comprehensive approach to try and provide a pipeline for short and long-term use, and for a potentially local application route via inhalation. Existing FDA-approved drugs that target a single protein important for viral replication and host entry are currently the highest priority for repurposing as new COVID-19 drugs. However, we think that there are compelling reasons to create pipelines to explore many putative targets, and chemical spaces that are far larger and more diverse than the known approved drugs. We have therefore screened ~10 million potentially purchasable compounds from the ZINC database and also predicted toxicity values for the numerous candidates. In addition, we have identified chemicals that are predicted to affect more than one of the host proteins, suggesting these may have more efficacy. One unusual category we have emphasized is volatiles, as these compounds may be biologically sourced, and therefore microbes could be genetically engineered to produce them in mass [24]. This would subsequently reduce the strain on global supply chains for chemicals that are necessary in synthesizing certain pharmaceuticals. These chemicals are also intriguing options for drug cocktails. If present in metabolic pathways, they possibly already interact in vivo. Therefore, short-term therapeutic concentrations may be better tolerated in humans.

It is nevertheless important to note that machine learning depends on available data. Because the size and diversity of publicly available bioassay data are limited, caution is required in interpreting the predictions. It is common to find past bioassays focused on similarly shaped chemicals, limiting the scope of the machine learning approach to find new chemistries. Importantly, apart from ACE2, the other human proteins that were identified to interact with SARS-CoV-2 are yet to be tested in vivo for efficacy. And although some of the candidate chemicals we identified may be biologically sourced, the concentrations are not well defined or unknown, nor is there any understanding of a therapeutic range of other diseases and conditions. We therefore anticipate that the AI-based predictions of purchasable compounds from 10+ million chemicals will accelerate drug discovery in general and facilitate research on these chemicals in the future for a number of diseases. In general, the use of AI-driven tools could provide additional valuable solutions for tackling Covid-19 [25].

4. Materials and methods

4.1. Data sources for machine learning

4.1.1. ZINC

ZINC is a free database comprised of 230 million chemicals for in silico analyses. It was developed as a resource for non-commercial research. Chemicals predicted here are from a purchasable subset; however, availability is subject to change and pricing may vary widely [18, 26].

4.1.2. Bioassay data

Bioassay data was retrieved from ChEMBL 25 using the associated Python module, which enables access to the API services via Python [27, 28]. The various inhibitory measures/endpoints, whenever possible, are standardized to nM units; the logarithm of the standardized values was

### Table 1 (continued)

| Feature       | Target | Description                                                                                                                                 |
|---------------|--------|---------------------------------------------------------------------------------------------------------------------------------------------|
| MATS5i        | RIPK1  | Moran autocorrelation of lag 5 weighted by ionization potential                                                                                     |
| VEGA116M_HG   | RIPK1  | Logarithmic coefficient sum of the last eigenvector from reciprocal squared geometrical matrix                                                     |
| BLA96         | SKGMAR1| Verhaar Algae base-line toxicity from MILGCP (mmol/l)                                                                                       |
| F10(C-C)      | SKGMAR1| Frequency of C - C at topological distance 10                                                                                                  |
| TPSA(Tot)     | SKGMAR1| Topological polar surface area using N,O,S,P polar contributions                                                                             |
| Eig01_AEA(dm) | TBK1   | Eigenvalue n. 1 from augmented edge adjacency mat. weighted by dipole moment                                                                  |
| HAT54i        | TBK1   | Leverage-weighted autocorrelation of lag 4/weighted by ionization potential                                                                  |
| SdscC         | TBK1   | Sum of dsC E-states                                                                                                                           |
| AROM          | VCP    | Aromaticity index                                                                                                                             |
| Elm           | VCP    | 1st component accessibility directional WHIM index/weighted by mass                                                                          |
| MATS5m        | VCP    | Moran autocorrelation of lag 5 weighted by mass                                                                                               |
| H5s           | ACE2   | H autocorrelation of lag 5/weighted by I-state                                                                                                 |
| Mor10m        | ACE2   | Signal 10/weighted by mass                                                                                                                    |
| Mor17m        | ACE2   | Signal 17/weighted by mass                                                                                                                    |
Table 2. Important chemical features for classification models. Top three physicochemical features for viral targets where the models classified chemicals as active vs inactive relative to broad inhibition or activation rather than a specific assay value (e.g. $K_i$, $IC_{50}$, and $AC_{50}$).

| Feature | Target | Description |
|---------|--------|-------------|
| Mor18s | BRD4   | signal 18/weighted by I-state |
| SpMAD.G/D | BRD4   | spectral mean absolute deviation from distance/distance matrix |
| SpMax3_Bh(i) | BRD4   | largest eigenvalue n. 3 of Burden matrix weighted by polarizability |
| _P_VSA_LogP_3 | HDAC2  | _P_VSA-like on LogP, bin 3 |
| SHED.DA | HDAC2  | SHED Donor-Acceptor |
| SHED.DL | HDAC2  | SHED Donor-Lipophilic |
| G(N..N) | IDE    | sum of geometrical distances between N..N |
| SM1_Dz(i) | IDE    | spectral moment of order 1 from Barysz matrix weighted by ionization potential |
| Wap    | IDE    | all-path Wiener index |
| CATS2D_08_DA | TBK1   | CATS2D Donor-Acceptor at lag 08 |
| F08[N-N] | TBK1   | Frequency of N - N at topological distance 8 |
| _P_VSA_e_3 | TBK1   | _P_VSA-like on Sanderson electronegativity, bin 3 |
| H7m    | PRKACA | H autocorrelation of lag 7/weighted by mass |
| H7s    | PRKACA | H autocorrelation of lag 7/weighted by I-state |
| RDF060m | PRKACA | Radial Distribution Function - 060/weighted by mass |
| GATS6e | MARK3  | Geary autocorrelation of lag 6 weighted by Sanderson electronegativity |
| GATS6m | MARK3  | Geary autocorrelation of lag 6 weighted by mass |
| Mor02m | MARK3  | signal 02/weighted by mass |
| CATS2D_02_DL | IMPDH2 | CATS2D Donor-Lipophilic at lag 02 |
| CATS3D_07_DL | IMPDH2 | CATS3D Donor-Lipophilic BIN 07 (7.000–8.000 Å) |
| NaasC  | IMPDH2 | Number of atoms of type aasC |
| C-039  | ABCCl  | Ar-(C–X-R) |
| VE2sign_Dz(p) | ABCCl | average coefficient of the last eigenvector from Barysz matrix weighted by polarizability |
| VE3sign_Dz(v) | ABCCl | logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by van der Waals volume |
| Mor31s | ABHD12 | signal 31/weighted by I-state |
| RT1+   | ABHD12 | R maximal index/weighted by ionization potential |
| VE3sign_Dz(p) | ABHD12 | logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability |
| E2m    | BRD2   | 2nd component accessibility directional WHIM index/weighted by mass |
| GATS2m | BRD2   | Geary autocorrelation of lag 2 weighted by mass |
| TDB08i | BRD2   | 3D Topological distance based descriptors - lag 3 weighted by ionization potential |
| MAXDP  | COMT   | maximal electrotopological positive variation |
| nDB    | COMT   | number of double bonds |
| _P_VSA_MR_2 | COMT   | _P_VSA-like on Molar Refractivity, bin 2 |
| CATS2D_02_AL | DNMT1 | CATS2D Acceptor-Lipophilic at lag 02 |
| Mor04s | DNMT1  | signal 04/weighted by I-state |
| VE3sign_Dt | DNMT1 | logarithmic coefficient sum of the last eigenvector from detour matrix |
| Chia_8(k) | EJF4H  | average Randic-like index from Burden matrix weighted by ionization potential |
| F05[C-O] | EJF4H  | Frequency of C - O at topological distance 5 |
| NaasC  | EJF4H  | Number of atoms of type aasC |
| CENT   | LOX    | centralization |
| EE_G   | LOX    | Estrada-like index (log function) from geometrical matrix |
| VE2_D/Dt | LOX    | average coefficient of the last eigenvector (absolute values) from distance/detour matrix |
| Eta_D.beta | MARK2  | eta measure of electronic features |
| Mor29v | MARK2  | signal 29/weighted by van der Waals volume |
| SpPosA_B(i) | MARK2 | normalized spectral positive sum from Burden matrix weighted by ionization potential |
| CATS2D_07_AL | NEK9   | CATS2D Acceptor-Lipophilic at lag 07 |
| CATS2D_08_AL | NEK9   | CATS2D Acceptor-Lipophilic at lag 08 |
| TDB05p | NEK9   | 3D Topological distance based descriptors - lag 5 weighted by polarizability |
| CATS2D_06_DL | NEU1   | CATS2D Donor-Lipophilic at lag 06 |
| TDB04i | NEU1   | 3D Topological distance based descriptors - lag 4 weighted by ionization potential |
| X3A    | NEU1   | average connectivity index of order 3 |
| nR06   | RHOA   | number of 6-membered rings |
| R8s+   | RHOA   | R maximal autocorrelation of lag 8/weighted by I-state |
| SpMin1_Bh(m) | RHOA | smallest eigenvalue n. 1 of Burden matrix weighted by mass |
| CATS3D_08_NL | SIRT5  | CATS3D Negative-Lipophilic BIN 08 (8.000–9.000 Å) |
| O-057  | SIRT5  | phenol, enol, carboxyl OH |
| SpMax2_Bh(s) | SIRT5 | largest eigenvalue n. 2 of Burden matrix weighted by I-state |
| CATS2D_04_AL | TK2    | CATS2D Acceptor-Lipophilic at lag 04 |

(continued on next page)
Table 2 (continued)

| Feature | Target | Description |
|---------|--------|-------------|
| JGI3 | TK2 | mean topological charge index of order 3 |
| MATSI1 | TK2 | Moran autocorrelation of lag 1 weighted by ionization potential |
| P_VSA_e_3 | VCP | P.VSA-like on Sanderson electronegativity, bin 3 |
| RDF020p | VCP | Radial Distribution Function - 020/weighted by polarizability |
| SpMaxA_AEA(dm) | VCP | normalized leading eigenvalue from augmented edge adjacency mat. weighted by dipole moment |

Figure 2. Models of chemical features accurately predict inhibitors of SARS-CoV-2 targets. a) Pipeline for fitting and validating models that predict IC50, Ki, or AC50 or a classification score, which reflects broad inhibitory activity against the listed viral targets. b) Left, mean absolute error (MAE) in predicting the log transformed endpoints (IC50, Ki, AC50). Right, classification of chemicals for broad inhibition or activity against targets, validating using the area under the receiver operating characteristic (ROC) curve (AUC). Plots are for 10-fold cross validation, repeated 5 times. The model predictions are from an ensemble of three support vector machines (SVM), trained on different chemical feature sets or in some cases SVM and Random Forest. c) Left, external test set performance for regression models, where possible. Right, external test set performance for classification models, where possible. More comprehensive performance data in Supplementary Information 1.
used for machine learning. Regression models were fit for a single endpoint. For classification machine learning models, however, ‘active’ class chemicals were defined using the deposited activity comments such as for assays of general activity against proteins, and added active labels for endpoints with values up to 10,000 nM (K<sub>i</sub> and IC<sub>50</sub>) and for the semi-quantitative % inhibition, greater than 10%. The majority class was downsampled during the training and model tuning phases to adjust for possible class imbalances. Because the class labels were assigned using

![Figure 3. Approved drugs with putative activity against SARS-CoV-2 targets. a) The best predicted activity against SARS-CoV-2 targets among databases of approved drugs. Viral targets with few promising candidates are omitted. Comprehensive table in Supplementary Information 2. b) Network showing drugs that are among the top 25 for multiple viral targets (drugs: black nodes; viral targets: red nodes).](image-url)
arbitrary cutoffs and the predicted activities for classification models from various assay endpoints are not clearly defined, we also compared each model fit to shuffled labels. Training for the regression and classification approaches was done on 85% of the total data. Notably, in a small number of cases the remaining 15% was insufficient to effectively estimate performance using an external test set. To reduce bias, feature selection (recursive feature elimination (RFE) algorithm) was always run on 85% of the data over 250–300 different partitions (iteratively running the 10-fold cross validation 25–30 times). However, for these cases, the held-out portion (15%) was then incorporated back into the dataset.

Figure 4. Predicting activity against SARS-CoV-2 targets among theoretical volatile chemicals. a) Left, count of chemicals per target after initially filtering based on predicted scores. Right, chemical counts across all viral targets for the models predicting general inhibitory or activity against (Classification) and those for specific inhibitory endpoints (Regression) (e.g. IC50). b) Pipeline for further prioritizing chemical sets according to estimated log vapor pressure and low mammalian toxicity (LD50). c) Top ranking predictions of general inhibition or activity against targets (Score) and/or specific inhibitory endpoints (Predicted Assay Value) against SARS-CoV-2 targets from the ZINC database, filtered to the highest estimated log vapor pressures.
better estimate performance of the trained model by 10-fold cross-validation (repeated 5 times) and obtain a better fit. We also fit 3 different radial basis function (RBF) support vector machine (SVM) models, wherein the chemical features (predictors) were randomly sampled (50%) from the top 70. This makes the performance estimates more conservative (see Key Resources Table for machine algorithm source files). However, the structural diversity and size of the datasets imply some bias in the performance estimates.

4.1.3. Toxicity data
Training and testing data are curated by various government agencies and provided freely to the general public as databases (see Key Resources Table) [29, 30, 31].

4.1.4. Vapor pressure data
Training and testing data are from EPI Suite [32], which is developed and maintained by the Environmental Protection Agency (EPA) (see Key Resources Table). Methods for fitting these models are as outlined in the Figure 1 pipeline. To compare the vapor pressure model predictions with respect to different machine learning methods as well as EPI suite, data were split into train/test partitions as defined in a previous study [33].

4.2. Selecting optimally predictive chemical features

4.2.1. Optimizing chemical structures
Chemical features were computed with ~5300 AlvaDesc descriptors, from the developers of DRAGON software, and 3D coordinates and optimization performed using RDKit in Python [34].

4.2.2. Chemical feature ranking and importance

4.2.2.1. Cross-validated recursive feature elimination (CV-RFE). Recursive feature elimination iteratively selects subsets of features to identify optimal sets. The algorithm is a “wrapper” and therefore relies on an additional algorithm to supply predictions and quantify importance. We used two different algorithms, depending on the size and composition of data: (1) Random Forest and (2) Support Vector Machine (SVM). Random forest determines the importance in relation to the % increase in error when permuting a feature or predictor. There is no equivalent method for computing importance with the SVM. Accordingly, the importance is based on fitting a model between the response and each predictor or feature as compared to null. If the response is numeric, importance is derived from the pseudo R2 (non-linear regression). If, however, the response is binary, the AUC is instead computed for each predictor or feature (see Key Resources Table for algorithm source files).

Including cross-validation with the recursive feature elimination (RFE) partitions the training data into multiple folds. This step avoids biasing performance estimates but results in lists of top predictors over the cross-validation folds such that importance of a predictor is based on a selection rate.

4.2.2.2. Selection bias. Selecting features or predictors on the same dataset used for cross validation results in models that have already “seen” possible partitions of the data and therefore performance metrics will be biased. Selection bias [35] was addressed by bootstrapping and cross validation, which ensure some separation between predictor/feature selection and model-fitting/validation. In addition to these methods, we used hidden test sets or more generally performed the feature selection on a portion of the data.

4.3. Selecting optimal machine learning algorithms
The support vector machine (SVM) with the radial basis function kernel (RBF) outperformed regularized Random Forest (regRF) or performed comparably. Rather than utilize many different approaches, we aggregated multiple SVM models to improve generalizability. However, in the case of the classification model for EIF4H, we included the regularized random forest algorithm, as the aggregated prediction (SVM and
Figure 6. Predictions of SARS-CoV-2 targets among chemicals lacking odorant properties. a) Sample of ZINC chemicals scoring highly for activity against the viral targets (classification or regression models, Score). Comprehensive tables in Supplementary Information 4, detailing the model type and predicted assay endpoint.
Enriched cores were analyzed using RDKit through Python [34]. The algorithm performs an exhaustive search for a maximum common substructure among a set of chemicals. In practice, larger sets often yield fewer substantive cores. To remedy this, the algorithm includes a threshold parameter that relaxes the proportion of chemicals containing the core. We used a threshold of 0.55, which ensures that the majority of the chemicals contained the core.

The kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The radial basis function (RBF) or Gaussian kernel enables the learning of more complex, non-linear boundaries. It is therefore well suited for problems in which the biologically active chemicals cannot be properly classified as a linear function of physicochemical properties. This kernel computes the similarity for each chemical ($x$) and a set of landmarks ($l$), where $\sigma^2$ is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals ("Active" vs. "Inactive").

\[
\text{Gaussian Kernel} = \exp\left(-\frac{(x - l_i)^2}{2\sigma^2}\right)
\]

### 4.5. Chemical fingerprinting

Extended Connectivity Fingerprints (ECFP) are a class of chemoinformatic algorithms that iteratively combine chemical features that are present within a predefined radius/diameter, representing them by a set of integer values. Typically, the fingerprint is converted into a binary string of fixed length using a hash function. Here, the bit length was set at 1024 and a radius of 2 (diameter = 4 or ECFP4). This structural representation was preferred as it is strongly associated with activity [39]. Accordingly, it is a suitable alternative to identify drug candidates in the absence of machine learning models. We used the ECFP algorithm in RDKit (Morgan or circular fingerprint) [34]. The similarity between the fingerprints of chemicals with known activity against the SARS-CoV-2 targets and prospective chemicals was computed using the Tanimoto index. This index is a similarity coefficient (0–1; 1 = max similarity). It is the overlap of the “on-bits” divided by the sum of the unique “on-bits”. Notably, coefficients of 1 need not imply identical chemicals.

\[
sim(AB) = \frac{c}{a + b - c}
\]

where $c = \text{overlapping "on-bits"}; a = \text{"on-bits" in A}; b = \text{"on-bits" in B}$.

### 4.6. Support vector machine (SVM)

Training the support vector machine (SVM) involves identifying a set of parameters that optimize a cost function, where cost 1 and cost 0 correspond to training chemicals labeled as "Active" and "Inactive", respectively. $\theta$ is the scoring function or output of the support vector machine. If the output is $\leq 0$, the prediction is "Active." The function ($f$) is a kernel function.

\[
\text{SVMCost} = \min_C \sum_{i=1}^{n} y_i \cos(\theta_f(x_i)) + (1 - y_i) \cos(\theta_f(x_i)) + \frac{1}{2} \sum_{i>j} \theta_f^2
\]

The kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The radial basis function (RBF) or Gaussian kernel enables the learning of more complex, non-linear boundaries. It is therefore well suited for problems in which the biologically active chemicals cannot be properly classified as a linear function of physicochemical properties. This kernel computes the similarity for each chemical ($x$) and a set of landmarks ($l$), where $\sigma^2$ is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals ("Active" vs. "Inactive").

\[
\text{Gaussian Kernel} = \exp\left(-\frac{(x - l_i)^2}{2\sigma^2}\right)
\]

#### 4.6.1. Model performance metrics

The Area under the ROC Curve (AUC) assesses the true positive rate (TPR or sensitivity) as a function of the false positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score ($x$) is greater than the threshold ($T$), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at AUC = 0.5.

\[
\text{TPR}(T) = \int_{T}^{\infty} f_1(x) dx
\]

\[
\text{FPR}(T) = \int_{T}^{\infty} f_0(x) dx
\]

where $T$ is a variable threshold and $x$ is a probability score.

However, we generated classifiers that are more authentic than theoretical random classification, shuffling the chemical feature values in the models and statistically comparing the mean AUCs across multiple partitions of the data. This controls against optimally tuned algorithms predicting well simply because of specific predictor attributes (e.g. range, mean, median, and variance) or models that are of a specific size (number of predictors) performing well even with shuffled values. Additionally, biological data sets are often small, with stimuli or chemicals that—rather than random selection—reflect research biases, possibly leading to optimistic validation estimates without the proper controls.

We used the AUC for evaluating classification models. For the classification-based training, we initially converted the inhibitory data into a binary label (Active/Inactive). For predictions of quantitative bioassay measures (e.g. $K_c$, $IC_{50}$, $AC_{50}$, Log $LD_{50}$), we computed the mean absolute error (MAE), the correlation coefficient (R) and the squared correlation coefficient (R2). MAE: Mean absolute error is the mean of the absolute difference between predicted and observed (% usage). It therefore assigns equal weight to all prediction errors, whether large or small.
where, 

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \]

where, \( TP \) = True Positive and \( FN \) = False Negative 

\[ \text{Specificity} = \frac{TN}{TN + FP} \]

where, \( TN \) = True Negative and \( FP \) = False Positive. 

Declarations 

Author contribution statement 

Joel Kowalewski: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. 

Manasankar Ray: Conceived and designed the experiments; Wrote the paper. 

Funding statement 

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. 

Competing interest statement 

J.K. and A.R. are listed as inventors in patents submitted by the University of California Riverside. A.R. is also founder of Sensorygen Inc. 

Additional information 

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e04639. 

References 

[1] S. Sanche, Y.T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, R. Ke, High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2, Emerg. Infect. Dis. J. 26 (2020). 

[2] Y. Bai, L. Yao, T. Wei, F. Tian, D.Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of COVID-19, JAMA - J. Am. Med. Assoc. 323 (14) (2020) 1406–1407. 

[3] M. Day, Covid-19: four fifths of cases are asymptomatic, China figures indicate, BMJ 369 (2020) m1375. 

[4] Y. Wan, J. Shang, R. Graham, R.S. Baric, F. Li, Receptor recognition by novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS, J. Virol. 94 (7) (2020). 

[5] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2, Science 367 (6485) (2020) 1444–1448. 

[6] D.E. Gordon, G.M. Jang, M. Bouhadou, J. Xu, K. Obenier, K.M. White, M.J. O’Meara, V.V. Rezeli, J.Z. Guo, D.L. Swaney, T.A. Tummino, R. Huettenhain, R.M. Kaske, A.L. Richards, B. Tantucoungho, H. Fournard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B.J. Polacco, H. Braberg, J.M. Fabius, M. Eckhardt, M. Soucheron, M.J. Bennett, M. Cakir, M.J. McGregor, Q. Li, B. Meyer, F. Roeesch, T. Vallet, A. Mac Kain, L. Miorin, E. Moreno, Z.Z.C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I.T. Kirby, J.E. Meñy, J.S. Charba, K. Lou, S.A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lys, C.J.F. Mathy, T. Perica, K.B. Pilla, S.J. Ganesan, D.J. Saltzberg, R. Rakesh, X. Liu, S.B. Rosenthal, L. Calvisiello, S. Venkataramanan, J. Libby-Lugo, Y. Lin, X.P. Huang, Y.F. Liu, S.A. Wankowicz, M. Bohn, M. Safari, F.S. Ugar, C. Koh, N.S. Savar, Q.D. Tran, D. Shengguider, S.J. Fletcher, M.C. O’Neal, Y. Cai, J.C.S. Chang, D.J. Broadhurst, S. Klippsten, P.P. Sharp, N.A. Wenzell, D. Kuzouglou, H.Y. Wang, R. Trenker, J.M. Young, D.A. Caveno, J. Hiatt, T.L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R.M. Stredulinski, A.D. Frankel, O.S. Rosenberg, K.A. Verba, D.A. Agard, M. Ott, M. Emerman, N. Jura, M. von Zastrau, E. Verdin, A. Ashworth, O. Schwartz, C. d’Enfert, S. Mukherjee, M. Jacobson, H.S. Malik, D.G. Fujimori, T. Ideker, C.S. Craik, S.N. Floor, J.S. Fraier, J.D. Gross, A. Sali, B.L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. García-Sastre, K.M. Shokat, B.K. Shoichet, N.J. Krogan, A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, Nature 583 (2020) 459–468. 

[7] W. Sungnak, N. Huang, C. Bécair, M. Berg, R. Queen, M. Litvivukova, C. Talavera-Lopez, H. Maatz, D. Reichert, F. Sampatizitis, B.B. Worlock, M. Yoshida, J.L. Barnez, N.E. Banovich, P. Barbary, A. Braxma, J. Collin, T.J. Desai, T.E. Duong, O. Eckelberg, C. Falk, M. Farzan, I. Glass, R.K. Gupta, M. Hanifia, P. Horvath, N. Hubner, D. Huang, N. Kaminia, M. Kranz, J.A. Kroppski, M. Kuhlmann, M. Lako, H. Lee, S. Leroy, S. Linnarson, J. Lundeberg, K.B. Meyer, Z. Miao, A. V. Misharin, M.C. Nawijn, M.Z. Nikolich, M. Noseda, J. Orlovsky-Montanes, G.Y. Oudit, D. Pe’er, J. Powell, S. Quale, J. Rajagopala, P.R. Tata, E.L. Rawlins, A. Regev, P.A. Reyfman, O. Rozenblatt, B.K. Shoichet, N.J. Krogan, A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, Nature 583 (2020) 459–468. 

[8] L. Riva, S. Yuan, X. Yin, L. Martin-Sancho, N. Matsunaga, S. Burgstaller, E. Herner, J. Powell, J. Rajagopal, P.R. Tata, E.L. Rawlins, A. Regev, P.A. Reyfman, O. Rozenblatt, B.K. Shoichet, N.J. Krogan, A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, Nature 583 (2020) 459–468.
