COVID-19 Knowledge Extractor (COKE): a tool and a web portal to extract drug - target protein associations from the CORD-19 corpus of scientific publications on COVID-19

Daniel Korn[1,2], Vera Pervitsky[2], Tesia Bobrowski[2], Vinicius M. Alves[3], Charles Schmitt[3], Chris Bizon[4], Nancy Baker[5], Rada Chirkova[6], Artem Cherkasov[7], Eugene Muratov[2*], Alexander Tropsha[2,*].

[1] Department of Computer Science, the University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA.

[2] Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, the University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA.

[3] Office of Data Science, National Toxicology Program, NIEHS, Morrisville, NC, 27560, USA.

[4] Renaissance Computing Institute, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7568, USA.

[5] ParlezChem, 123 W Union Street, Hillsborough, NC, 27278, USA.

[6] Department of Computer Science, North Carolina State University, Raleigh, NC, 27606-5550.

[7] Vancouver Prostate Centre, University of British Columbia, Vancouver, BC, Canada.

Corresponding Authors

* Address for correspondence: 100K Beard Hall, UNC Eshelman School of Pharmacy, the University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-2955; FAX: (919) 966-0204; E-mail: murik@email.unc.edu and alex_tropsha@unc.edu.

Keywords: Data curation, text mining, natural language processing.

Word count: 3888 words.

Abstract

Objective: The COVID-19 pandemic has catalyzed a widespread effort to identify drug candidates and biological targets of relevance to SARS-COV-2 infection, which resulted in large numbers of publications on this subject. We have built the **CO**VID-19 **K**nowledge **E**xtractor (COKE), a web application to extract, curate, and annotate essential drug-target relationships from the research literature on COVID-19 to assist drug repurposing efforts.

Materials and Methods: SciBiteAI ontological tagging of the COVID Open Research Dataset (CORD-19), a repository of COVID-19 scientific publications, was employed to identify drug-target relationships. Entity identifiers were resolved through lookup routines using UniProt and DrugBank. A custom algorithm was used to identify co-occurrences of protein and drug terms, and confidence scores were calculated for each entity pair.

Results: COKE processing of the current CORD-19 database identified about 3,000 drug-protein pairs, including 29 unique proteins and 500 investigational, experimental, and approved drugs. Some of these drugs are presently undergoing clinical trials for COVID-19.

Discussion: The rapidly evolving situation concerning the COVID-19 pandemic has resulted in a dramatic growth of publications on this subject in a short period. These circumstances call for methods that can condense the literature into the key concepts and relationships necessary for insights into SARS-CoV-2 drug repurposing.

Conclusion: The COKE repository and web application deliver key drug - target protein relationships to researchers studying SARS-CoV-2. COKE portal may provide comprehensive and critical information on studies concerning drug repurposing against COVID-19. COKE is freely available at https://coke.mml.unc.edu/ and the code is available at https://github.com/DnlRKorn/CoKE.

BACKGROUND AND SIGNIFICANCE

With over 35 million cases and over 1 million deaths worldwide as of the beginning of October, 2020, and no Food and Drug Administration (FDA) approved drug treatments or vaccines against this virus, there are unprecedented global efforts to discover critical therapeutic treatments against COVID-19 [1]. These efforts already resulted in the identification and characterization of many SARS-CoV-2 proteins essential for virus replication [2] and the pathogenesis of COVID-19 [3] and nomination of many drugs for clinical trials. Within a few months of the outbreak, thousands of papers on COVID-19 and SARS-CoV-2 have appeared in the scientific literature [4]. There are many databases collecting data related to SARS-CoV-2 [5]; however, the scientific literature concerning SARS-CoV-2 remains the largest repository of untapped biomedical data [6,7].

Recently, the Allen Institute for AI, the NIH, the White House, Georgetown University, and several other organizations collaborated to produce the COVID-19 Open Research Dataset (CORD-19). This dataset consists of, at the time of this study, 129,000 full-text scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses [4]. SciBiteAI, a semantics research group based in the UK [8], curated an ontologically annotated version of the dataset to identify biomedical terms within sentences of full papers or abstracts [4].

Arguably, the information about biological targets implicated in COVID-19 and drugs acting at these targets is of greatest value to scientists interested in drug repurposing. However, current tools do not provide a user-friendly way to retrieve such information from the research literature. To address this gap, we have developed the COVID-19 Knowledge Extractor (COKE), a web application tool that provides the scientific community with (i) up-to-date data on human and viral proteins associated with SARS-CoV-2 and other coronaviruses that are indexed in UniProt and (ii) chemicals reported to target these proteins. We have created a database summarizing all drug-target-coronavirus triangle relationships annotated in the research literature by (i) creating an approach to detect drug and protein literature co-occurrences within all manuscripts annotated in the CORD-19 corpus [9]; (ii) establishing a scoring system to rate the confidence of a co-occurrence pair; and (iii) implementing an algorithm that allows users of COKE to highlight specific sections of the manuscript where respective terms co-occur.

OBJECTIVES

COKE has been developed to provide the scientific community with data that could potentially contribute to current COVID-19 drug repurposing efforts. COKE portal provides data on human and viral proteins associated with SARS-CoV-2 and other coronaviruses as described in the CORD-19 corpus and that are indexed in UniProt as well as chemicals targeting those proteins that are indexed in the DrugBank [10,11].

MATERIALS AND METHODS

Dataset Collection

The SciBiteAI group has published an ontologically annotated version of the CORD-19 database available publicly on their GitHub account.[4] These ontology citations and custom vocabularies were released under open licenses, which allow for unrestricted use and further development. The entire content of each paper within the CORD-19 dataset is divided into paragraphs of plain text. Within each of these paragraphs, ontological terms divided into nine groups are assigned and matched to the sentences they occur in. Table 1 shows the nine groups from which the assigned keys are: SPECIES, GOONTOL, INDICATION, COUNTRY, HPO, GENE, DRUG; and two custom vocabularies created for this dataset: SARSCOV and CVPROT.

The UniProt database provides information on over half of a million proteins, the majority of which has been reviewed and curated by experts in the field of proteomics.[10] We extracted information on manually reviewed proteins that were either human proteins associated with coronavirus disease (host protein targets) or coronavirus proteins (viral protein targets). Additionally, we extracted information on synonyms of these proteins, the organisms from which they were derived, and their genome sequence length.

The DrugBank dataset provides listings of chemical identifiers tagged as small molecule drugs. Standardized naming of these compounds and their SMILES strings are also provided in a clean, relational format.[11] DrugBank provides the number of other targets a given small molecule is associated with. COKE makes these data available and organized in such a way that users of COKE can easily examine relevant drug-target-disease relationships related to SARS-CoV-2.

Data Curation and Integration

The original CORD-19 dataset contained more than 129,000 papers. We found that many papers in this original dataset were related to other viruses, such as Ebola and Zika, as well as epidemiological studies not relevant to drug repurposing for COVID-19. For this reason, a major curation step in our protocol was to incorporate only papers where COVID-19 related terms were explicitly mentioned. For this task, we employed MeSH (Medical Subject Headings) IDs related to SARS-CoV-2, COVID-19, and coronaviruses. This filtering was performed by leveraging SciBiteAI's ontological tagging. Any paper in the dataset, which was not annotated by one of the NCBITaxon tags for a coronavirus, was not considered. The curation protocol was performed on every new version of CORD-19. At the time of this analysis, 94,000 papers remained in our dataset.

Next, a custom algorithm to detect the co-occurrence of two terms within a specified paper was employed. Our inputs were the sentence-level annotations provided by SciBiteAI. For every biomedical term observed, the number of publications that a term individually appears in was determined via a simple count. Then, each publication in CORD-19 received a vote if the terms co-occurred. This vote was "yes" if either of the following two conditions were met: (i) both terms appear in the abstract or (ii) both terms appear in a single sentence of the publication. The reason for this distinction is that abstracts are considered as significantly more information-dense [12] and, therefore, every term mentioned in the abstract is considered to be of greater significance in the context of the whole study. In contrast, many authors were quite verbose in the body of a publication and may mention terms within a section that were unrelated to each other.

The COKE portal provides the user with a scoring function that rates the confidence of co-occurrence pairs. This scoring function was created by implementing a hypergeometric distribution with the following parameters: (i) a population size equal to the number of publications which meet our curation standard (~94,000 at the time of writing), (ii) the number of successes in the population equal to the number of Term 1 occurrences, (iii) the number of samples drawn equal to the number of Term 2 occurrences, and (iv) the number of observed successes equal to the co-occurrence votes as described above. The cumulative distribution function (CDF) of hypergeometric distribution with the parameters was calculated; then, this score can be used to easily compare different pairings. The SciPy implementation of these functions was used for these calculations [13]. Many of the scores can be quite small and tightly clustered, so the logarithm of the CDF was calculated. Since logarithms are monotonic values, the ordering of the tuples was maintained. To make the score more interpretable, the sign was flipped, so that the score ranged between 0 and infinity. This score helps the user judge how strongly two terms are connected, i.e., the closer the score is to zero, the higher the degree of connectivity between any two terms (see Figure 1).

We then filtered the large set of co-occurrence tuples only for CVPROT to DRUG relationships. To provide users with a more reliable curated set of relationships, we leveraged the identifiers provided by the SciBiteAI tagging. As a result of this filtering, we were left with 9,500 tuples. For additional filtering, we cross-referenced UniProt identifiers from both SciBiteAI's tagging and our UniProt data. Any proteins that had not been marked as reviewed were purged from the dataset. This filtered out under a dozen tuples for minor proteins. We sought to only use proteins that have been hand-reviewed by UniProt.

We also sought to clean the chemicals in our dataset. We cross-referenced all ChEMBL tags against DrugBank. Compounds that were not present in DrugBank were removed resulting in 4,700 drug/protein tuples. Additionally, we excluded amino acids, peptides, and proteins, so that only small molecules remained. Because we did not require the user to load DrugBank's large central XML file into memory, we used the BeautifulSoup web scraping library to isolate the file and parse out both of the DrugBank datasets (original and curated) we needed to perform this processing. Our final dataset had 2,000 drug-protein tuples. Details of this filtering process can be found on the GitHub and in Figure 2.

Development of the COKE Web Portal

The COKE web portal provides the user with the ability to view (i) the co-occurrence tuples on protein and drug on tables separated by targets or all tuples in a large table; (ii) the aggregated information on targets in our dataset; and (iii) the highlighted sections of papers from CORD-19 (Figure 3). COKE contains 18 tables with hundreds of rows each consisting of various forms of information related to the tuple. The data are stored as a JSON object in the same domain as the web portal.

Additionally, COKE provides the user with the ability to view selected publications from the CORD-19 dataset in which information relevant to the queried (drug-protein) tuple is presented to them. Using Python3.7 and the Flask web development framework, we developed a dynamic web API for highlighting respective sections in the CORD-19 papers hosted at https://coke.mml.unc.edu. This API takes in three parameters: the CORD-19 identifiers of a publication, the drug, and the protein as formatted by SciBiteAI. Then, the publication is checked for the co-occurrence of drug and protein names in the abstract and any sentences in the body of the text. The part of the text in which a co-occurrence is found is highlighted by displaying the entire section in bold and with increased font size. This is then rendered as HTML and the user's web view is automatically taken to highlighted text. Links to these highlighted papers are included in each COKE table.

To allow faster rendering of the web portal, we utilized the Data Tables jQuery library [14], which aids in rendering dynamic complex HTML tables in the browser. We converted all the co-occurrence tuples into JSON files in Data Table's specified format. These JSON files are stored within COKE's web domain as static files. Then, when the user loads the table, each Data Table makes an AJAX request for their specified information. By separating the data from the website, we provide the user with an interactive display that works significantly faster than a monolithic website, due to the parallel loading of the data for each table.

Curation of chemical bioactivity data from experimental screening assays for COVID-19

To assess the value of drug-target linkage identified in the COKE database, we explored quantitative high-throughput screening data for compounds in the Approved Drugs Collection from the NCATS OpenData Portal on COVID-19 [15]. This collection was screened for the SARS-CoV-2 cytopathic effect (CPE) assay (a phenotypic assay) and an AlphaLISA assay that measures the antiviral effect as the ability of a small molecule to disrupt the spike-ACE2 protein-protein interaction.[16] The CPE assay initially contained 6,988 chemicals with AC_{50} dose-response curves. The same collection was subjected to counter screen to ensure compounds identified as active in the primary assay were not cytotoxic, i.e., that they did not merely kill the host cell. After curation, 4,625 (165 actives, 4164 inactives, and 296 inconclusive) small molecules remained in the primary assay. In the Spike-ACE2 dataset, 3,406 data points were collected. After curation, 3030 (352 actives, 2099 inactives, and 579 inconclusive) small molecules remained in the primary assay. The counter screen data was used to ensure that compounds were not false positive because of interfering with the AlphaLISA readout. Both counter screens were used to look up the experimental results for compounds identified by COKE.

The structures of the compounds tested in the CPE assay were obtained from the NCATS OpenData Portal and curated following a protocol previously developed by our group [17–19]. Salts and solvents were stripped from all compounds, and large organic mixtures and inorganic compounds were removed. Chemotypes were standardized using the ChemAxon "Standardizer" software (v. 20.8.0). Compounds with replicate runs were analyzed. Replicates that had contradictory classifications were removed completely. For the CPE assay, compounds are labeled as "active" if the associated assay report shows a Hill slope equal to 1.1, 1.2, 2.1, 2.2, or 3 and an associated pAC_{50} higher than 4.9. Compounds with dose-response curve class 4.0 are considered inactive, while the remaining ones are inconclusive. Compounds that inhibit host cell growth in the counter screen assay are cytotoxic. Therefore, compounds were labeled as "non-toxic" if the dose-response curve class was 4.0, and other compounds were considered potentially toxic, even if they were labeled as inconclusive. In the Spike-ACE2 dataset, compounds labeled as "active" reported a Hill slope equal to -1.1, -1.2, -2.1, -2.2, or -3 and had an associated pAC50 higher than 4.9. We decided to keep compounds with curve-class 3 (CPE) or -3 (Spike-ACE2) as "active" because compounds with these curves were labeled as "low-quality actives" by NCATS.[15]

RESULTS

Comparison of the drug-target associations in the COKE dataset and bioactivity screening data on COVID-19

Our drug list identified by COKE initially contained 499 drugs. After curation, 471 unique drugs were kept. From this list, there were 335 compounds in the Approved Drugs Collection tested in the SARS-CoV-2 CPE (14 actives, 304 inactives, and 17 inconclusive). Table 2 lists all 14 active compounds. From these, five compounds were shown to be inactive in the counter screen indicating that they were true positives in the CPA assay: umifenovir, imatinib, promethazine, fluoxetine, and reserpine.

Umifenovir (arbidol) was also identified as active in the QSAR models for severe acute respiratory syndrome coronavirus (SARS-CoV) M^{pro} our group described recently [20]. Umifenovir was found active against SARS-CoV-2 in vitro [21] as a binder to the spike glycoprotein of SARS-CoV-2 (UniProt ID P0DTC2).[22] However, a study in humans showed that patients in the group receiving umifenovir had a longer hospital stay than patients in the control group. No deaths or severe adverse reactions were found in either group.[23] Imatinib,[24–26] promethazine,[27] and fluoxetine[28] are being tested in clinical trials.

Previous studies have shown that imatinib inhibits both SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) *in vitro*.[29] In addition, imatinib is currently being studied in COVID-19 clinical trials,[25] and has been shown to successfully treat COVID-19 in a case report.[30] Another study has shown that promethazine also has inhibitory activity against MERS-CoV *in vitro*.[31] Reserpine also demonstrated anti-SARS-CoV activity *in vitro*.[32,33] Though currently there is no literature on the antiviral activity of fluoxetine against SARS-CoV-2 *in vitro*, this compound has been suggested as a possible antiviral drug candidate against the virus based on scientific reasoning,[34] non-peer-reviewed empirical evidence,[35] and computational studies.[36]

Among the compounds identified as active but cytotoxic, hexachlorophene is a topical antibacterial agent [37]. Nitazoxanide showed activity in the phenotypic screen [38], and it has been included in prophylactic post-exposure clinical trials [39]. In non-peer-reviewed evidence, tioguanine was shown to inhibit SARS-CoV-2 papain-like protease by viral protein cleavage catalysis and to prevent replication of SARS-CoV-2 *in vitro* [40]. Chlorprothixene was shown to inhibit SARS-CoV replication *in vitro* [41]. Nelfinavir mesylate, an HIV protease inhibitor, was shown to inhibit M^{pro} in the computational analysis [42] and to have activity against the SARS-CoV-2 spike glycoprotein *in vitro* [43,44]. Tetrandrine is currently being explored in COVID-19 clinical trials [45]. In previous studies, chlorpromazine showed activity against MERS-CoV and SARS-CoV [46,47], and it is being studied in clinical trials in hospitalized patients with COVID-19 [48]. Past studies showed that amiodarone has *in vitro* antiviral activity against SARS-CoV by interfering with endocytosis and viral replication [49,50]. It is currently being studied in clinical trials against COVID-19 [51] and it was recently documented in a case report to successfully treat COVID-19 [52]. In summary, COKE successfully highlighted compounds shown to be active against SARS-CoV-2 in the phenotypic assay. This observation demonstrates the potential of COKE to rapidly identify unique compounds reported to have antiviral activity against SARS-CoV-2. Most importantly, the linkage between drug-target pairs identified by COKE and the results of drug bioactivity screening reported in NCATS OpenData Portal explored in this study illustrates the importance of validation, by the experimental data, of the functional significance of drug-target co-occurrences identified in the research literature.

The literature score of compounds (described in the Materials and Methods section) shows how strongly two terms are connected, i.e., the closer the score is to zero, the higher is the degree of connectivity between any two terms of interest. The results are shown in Table 2. We observe that active and inconclusive compounds that did not appear to be cytotoxic in the counter screen assays have substantially stronger associations (lower scores) than compounds labeled as inactive in the CPE assay (fluoxetine, umifenovir, imatinib, promethazine), except for reserpine. In COKE, lower scores are associated with compounds that have appeared more frequently in the literature co-occurring with COVID-19 related targets. The complete list is available at https://github.com/DnlRKorn/CoKE.

Comparison of COKE dataset and Clinical Trials for COVID-19

We also sought to know how many drugs in CORD-19 are already under investigation in clinical trials. We performed a simple cross-reference check of all drugs in the CORD-19 dataset. To obtain a list of drugs already in clinical trials for COVID-19, we leveraged DrugBank's dataset, which matched active clinical trials to DrugBank IDs [53]. Of the 435 entries found in DrugBank, 271 were small molecules and not amino acids. We were only able to identify 155 of these compounds in the curated COKE dataset. This observation is surprising if not shocking as it means that apparently nearly half of all drugs that went into clinical trials were not examined in the open research literature in the context of COVID-19. This leaves one doubt as to why compound nomination for clinical trial escaped peer-review, a process commonly accepted by the global research community for validating research observations and hypotheses before exposing them to broad research community.

DISCUSSION

Application of the COKE Web Portal

Quantitative structure-activity relationship (QSAR) modeling has been used in the past to discover potential drugs to treat viral diseases such as COVID-19 [54–58]. Similarly, molecular docking studies of targets associated with the disease have been used to the same effect. Virtual screening hits generated by QSAR models or resulting from docking studies could be run against drugs annotated in COKE (and/or reported in the NCATS OpenData Portal) to help select the promising candidates for drug repurposing. Furthermore, the highlighted paper sections where drug-target co-occurrences are reported, allow for the quick discovery of possible mechanistic reasons for strong virtual screening hits. For example, nitazoxanide (found in COKE) was predicted to be an inhibitor of the main protease of SARS-CoV-2 in a recent QSAR study by our group, which is reasonable since past literature shows that this drug has anti-coronaviral activities [59].

An example of using the COKE web portal can be seen in Figure 3. Here, we show the drugs with linkages to the spike glycoprotein of SARS-CoV-2 (UniProt ID: P0DTC2), ranked by their score (as described in the Materials and Methods section). The current COKE version identified 153 unique drugs (143 after curation). COKE output overlapped with 90 drugs tested in Spike-ACE2 protein-protein interaction (AlphaLISA) by NCATS with nine compounds (umifenovir, hexachlorophene, chlorprothixene, nicardipine, mifepristone, rifampicin, flunarizine, niclosamide, and trypan blue free acid) labeled as active. Umifenovir and hexachlorophene were also active in the phenotypic screen, but only umifenovir was not also cytotoxic (*vide supra*). Niclosamide, an anthelmintic drug, has shown broad-spectrum antiviral activity against a wide array of viruses, including SARS-CoV-1 and MERS-CoV,[60] and it is currently being tested in clinical trials against SARS-CoV-2.[61,62] In past studies, the synthetic steroid mifepristone demonstrated antiviral activity against human adenovirus [63], Venezuelan equine encephalitis virus [64], and HIV-1 [65]. Flunarizine, an antimigraine drug, is known to arrest virus-membrane fusion for various hepatitis C virus genotypes [65]. A 1971 study by Follett and Pennington demonstrated that the antibiotic rifampicin could inhibit poxvirus replication[66]; more recent evidence for the drug's possible activity against other viruses is lacking.

Neither nicardipine nor trypan blue free acid had antiviral activities reported in the literature; in fact, trypan blue, a commonly used dye, is a known carcinogen and teratogen [51]. Unfortunately, all these drugs were found to interact with the AlphaLISA in the counter screen assay, meaning that these compounds could be false-positive inhibitors of viral entry. As discussed above, umifenovir was active in CPE, but not shown to be effective in humans [23]. Nevertheless, this exercise shows how the COKE web application allows for quick gathering and sorting of protein/drug connections that can be further explored by targeted analysis of the data reported in the NCATS OpenData Portal.

Data reported in COKE can be viewed as connections between biomedical entities, which could easily be incorporated into biomedical knowledge graphs such as ROBOKOP [67,68] to enable exploration of the linkages between COVID-19 and other biomedical entities. Additional integration of other biomedical information would allow for a more detailed exploration of these connections, leveraging other information about the drugs or proteins to enable more dynamic research.

In summary, valuable information about drugs and targets that could be implicated in COVID-19 can be gained from the utilization of natural language processing. We have listed the publications in which we find co-occurrence between drugs and targets at the straightforward sentence level. A more targeted processing of these specified papers may yield (*subject*, *object*, *predicate*) triples from those papers, providing more insight and possibly, higher confidence in the functional significance of the identified drug-protein associations.

CONCLUSIONS

We have built COKE, a web application to extract, condense, and prioritize the key drug-target relationships in the current literature concerning SARS-CoV-2. COKE is based on the CORD-19 literature collection, ontological tagging of papers in this collection by SciBiteAI, and entity identifiers derived from UniProt and DrugBank. Co-occurrences of protein and drug terms as well as the confidence scores for each pair were calculated using a custom algorithm specially designed for COKE. Overall, ca. 3,000 drug-protein pairs were identified by COKE, including 29 unique proteins (22 viral targets and 7 host targets) and 500 unique investigational, experimental, and approved drugs, some of which are currently undergoing clinical trials for COVID-19. At the same time, surprisingly, nearly half of the drugs nominated for or in clinical trials already, were not reported in the COVID-19 research literature as annotated in the CORD-19 database. We have demonstrated that COKE could be useful not only for direct identification of drug repurposing candidates but also for informing the final selection of drugs identified by other methods. In summary, COKE makes drug-protein relationships reported in the literature relevant to SARS-CoV-2 readily available to researchers and has the potential to provide important insights into drug repurposing efforts against COVID-19. COKE is implemented as a web platform that is freely available at https://coke.mml.unc.edu/; the code is available at https://github.com/DnlRKorn/COKE. The COKE web portal will be updated monthly with the latest data.

ACKNOWLEDGMENTS

We would like to acknowledge the work of the Allen Institute of AI for putting together the CORD-19 dataset. We also wish to recognize SciBiteAI that provided the ontological tagging for this project. Co-authors contributed to this manuscript as follows. AC, RC, EM, and AT conceived and designed the study. DK curated CORD-19 data and developed the COKE webserver. VP performed the comparison of COKE and clinical trial data. VMA performed the cheminformatics analysis comparing COKE results and NCATS data and prepared the figures. DK, TB, VP, and VMA wrote the first draft of the manuscript. All authors analyzed and discussed the data. All authors have read, edited, and given approval to the final version of the manuscript.

COMPETING INTERESTS

The authors declare no conflicts of interest.

FUNDING

This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health [OT2TR003441], and National Institutes of Health [1U01CA207160].

FIGURE LEGENDS

Figure 1. Co-occurrence algorithm.

Figure 2. General overview of the data processing in COKE.

Figure 3. An example response from the COKE Web Portal.

TABLES

Table 1. CORD-19 Ontological Mappings.

NAME	Ontologies Referenced
SPECIES	NCBITaxon (https://www.ncbi.nlm.nih.gov/taxonomy)[69]
GOONTOL	Gene Ontology (http://geneontology.org/)[70]
INDICATION	MeSH (https://www.ncbi.nlm.nih.gov/mesh/)[71]
COUNTRY	Country Codes[72]
HPO	Human Phenotype Ontology (https://hpo.jax.org/app/)[73]
DRUG	ChEMBL (https://www.ebi.ac.uk/chembl/)[74]
SARSCOV*	NCBITaxon[69] or Gene Ontology[70] or Malaria Ontology (https://bioportal.bioontology.org/ontologies/IDOMAL)[75]
CVPROT*	COVID-19 UniProtKB (https://covid-19.uniprot.org/)[10] and SWISS-MODEL (https://swissmodel.expasy.org/repository)[76]

*these vocabularies were custom-built for the CORD-19 dataset and can be accessed at https://github.com/SciBiteLabs/CORD19/tree/master/vocabularies-CORD-19

Table 2. List of compounds identified by COKE as active validated by NCATS in CPE assay.

Drug name	NCATS ID	Score	Counter screen cytotoxicity
Fluoxetine	NCGC00015428-15	0.010	Safe
Umifenovir	NCGC00246387-06	0.02	Safe
Imatinib	NCGC00159456-06	0.02	Safe
Promethazine	NCGC00015817-14	0.42	Safe
Reserpine	NCGC00015888-06	1.88	Safe
Tioguanine	NCGC00094792-18	0.0006	Cytotoxic
Nelfinavir	NCGC00090782-17	0.0007	Cytotoxic
Tetrandrine	NCGC00017376-12	0.03	Cytotoxic
Hexachlorophene	NCGC00091195-08	0.04	Cytotoxic
Chlorpromazine	NCGC00015273-19	0.31	Cytotoxic
Chlorprothixene	NCGC00013683-06	0.31	Cytotoxic
Nitazoxanide	NCGC00090774-05	2.57	Cytotoxic
Amiodarone	NCGC00015096-17	5.11	Cytotoxic

REFERENCES

1 John Hopkins University and Medicine. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. John Hopkins Coronavirus Resour. Cent. 2020;:1.https://coronavirus.jhu.edu/map.html (accessed 1 Jun 2020).

2 Bobrowski T, Melo-Filho CC, Korn D, *et al.* Learning from history: do not flatten the curve of antiviral research! *Drug Discov Today* 2020;:S1359-6446(20)30285-3. doi:10.1016/j.drudis.2020.07.008

3 Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J Microbiol Immunol Infect* 2020;**in press**. doi:10.1016/j.jmii.2020.03.022

4 SciBiteAI. Annotated Data for the COVID-19 Open Research Dataset Challenge. https://github.com/SciBiteLabs/CORD19 (accessed 12 May 2020).

5 National Institutes of Health: Office of Data Science Strategy. Open-Access Data and Computational Resources to Address COVID-19. 2020.https://datascience.nih.gov/covid-19-open-access-resources

6 Hunter LE. Knowledge-based biomedical Data Science. *Data Sci* 2017;**1**:19–25. doi:10.3233/ds-170001

7 Bakken S. Informatics is a critical strategy in combating the COVID-19 pandemic. *J Am Med Inform Assoc* 2020;**27**:843–4. doi:10.1093/jamia/ocaa101

8 SciBite. SciBite | About us. https://scibite.brumpton.me/about-us/company/ (accessed 26 May 2020).

9 Wang LL, Lo K, Chandrasekhar Y, *et al.* CORD-19: The COVID-19 Open Research Dataset. *ArXiv* Published Online First: 22 April 2020.http://arxiv.org/abs/2004.10706

10 The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15. doi:10.1093/nar/gky1049

11 Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668-72.

doi:10.1093/nar/gkj067

12 Cohen KB, Johnson HL, Verspoor K, *et al.* The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* 2010;**11**:492. doi:10.1186/1471-2105-11-492

13 Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;**17**:261–72. doi:10.1038/s41592-019-0686-2

14 DataTables.net. DataTables | Table plug-in for jQuery. DataTables. 2014.https://datatables.net/ (accessed 18 May 2020).

15 National Center for Advancing Translational Sciences. OpenData | COVID-19. OpenData Portal. 2020.https://opendata.ncats.nih.gov/covid19/

16 Beaudet L, Rodriguez-Suarez R, Venne M-H, *et al.* AlphaLISA immunoassays: the no-wash alternative to ELISAs for research and drug discovery. *Nat Methods* 2008;**5**:an8–9. doi:10.1038/nmeth.f.230

17 Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol* 2015;**11**:535–535. doi:10.1038/nchembio.1881

18 Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 2010;**50**:1189–204. doi:10.1021/ci100176x

19 Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model* 2016;**56**:1243–52. doi:10.1021/acs.jcim.6b00129

20 Alves V, Bobrowski T, Melo-Filho C, *et al.* QSAR modeling of SARS-CoV Mpro inhibitors identifies Sufugolix, Cenicriviroc, Proglumetacin and other drugs as candidates for repurposing against SARS-CoV-2. *Mol Inform* 2020;**in press**. doi:10.1002/minf.202000113

21 Wang X, Cao R, Zhang H, *et al.* The anti-influenza virus drug, arbidol is an efficient

inhibitor of SARS-CoV-2 in vitro. *Cell Discov* 2020;**6**:28. doi:10.1038/s41421-020-0169-8

22 NCATS. Arbidol | OpenDataPortal.

23 Lian N, Xie H, Lin S, *et al.* Umifenovir treatment is not associated with improved outcomes in patients with coronavirus disease 2019: a retrospective study. *Clin Microbiol Infect* 2020;**in press**. doi:10.1016/j.cmi.2020.04.026

24 U.S. National Library of Medicine. The Safety & Efficacy of Imatinib for the Treatment of SARS-COV-2 Induced Pneumonia. ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04422678 (accessed 15 Jun 2020).

25 U.S. National Library of Medicine. Trial of Imatinib for Hospitalized Adults With COVID-19. ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04394416 (accessed 15 Jun 2020).

26 U.S. National Library of Medicine. IMATINIB IN COVID-19 DISEASE IN AGED PATIENTS. (IMAGE-19). ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04357613 (accessed 15 Jun 2020).

27 U.S. National Library of Medicine. Pharmacokinetics, Pharmacodynamics, and Safety Profile of Understudied Drugs Administered to Children Per Standard of Care (POPS) (POPS or POP02). ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04278404 (accessed 15 Jun 2020).

28 U.S. National Library of Medicine. Fluoxetine to Reduce Intubation and Death After COVID19 Infection. ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04377308 (accessed 15 Jun 2020).

29 Pillaiyar T, Manickam M, Jung S-H. Middle East Respiratory Syndrome-Coronavirus (MERS-CoV): An Updated Overview and Pharmacotherapeutics. *Med Chem (Los Angeles)* 2015;**5**:361–72. doi:10.4172/2161-0444.1000287

30 Morales-Ortega A, Bernal-Bello D, Llarena-Barroso C, *et al.* Imatinib for COVID-19: A case report. *Clin Immunol* 2020;**218**:108518. doi:10.1016/j.clim.2020.108518

31 Liu Q, Xia S, Sun Z, *et al.* Testing of Middle East Respiratory Syndrome Coronavirus Replication Inhibitors for the Ability To Block Viral Entry. *Antimicrob Agents Chemother* 2015;**59**:742 LP – 744. doi:10.1128/AAC.03977-14

32 Wu C-Y, Jan J-T, Ma S-H, *et al.* Small molecules targeting severe acute respiratory syndrome human coronavirus. *Proc Natl Acad Sci U S A* 2004;**101**:10012–7. doi:10.1073/pnas.0403596101

33 Liang P-H. Characterization and inhibition of SARS-coronavirus main protease. *Curr Top Med Chem* 2006;**6**:361–76. doi:10.2174/156802606776287090

34 Fitzgerald PJ. Noradrenergic and serotonergic drugs may have opposing effects on COVID-19 cytokine storm and associated psychological effects. *Med Hypotheses* 2020;**144**:109985. doi:10.1016/j.mehy.2020.109985

35 Zimniak M, Kirschner L, Hilpert H, *et al.* The serotonin reuptake inhibitor Fluoxetine inhibits SARS-CoV-2. *bioRxiv* 2020;:2020.06.14.150490. doi:10.1101/2020.06.14.150490

36 Beck BR, Shin B, Choi Y, *et al.* Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput Struct Biotechnol J* 2020;**18**:784–90. doi:10.1016/j.csbj.2020.03.025

37 U.S. Food and Drug Administration. Code of Federal Regulations. Code Fed. Regul. Title 21. 2020;:Sec. 250.https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=250.250

38 Wang M, Cao R, Zhang L, *et al.* Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res* 2020;**30**:269–71. doi:10.1038/s41422-020-0282-0

39 U.S. National Library of Medicine. Trial to Evaluate the Efficacy and Safety of Nitazoxanide (NTZ) for Post-Exposure Prophylaxis of COVID-19 and Other Viral

Respiratory Illnesses in Elderly Residents of Long-Term Care Facilities (LTCF). ClinicalTrials.gov. 2020.https://clinicaltrials.gov/ct2/show/NCT04343248 (accessed 15 Jun 2020).

40 Swaim CD, Perng Y-C, Zhao X, *et al.* 6-Thioguanine blocks SARS-CoV-2 replication by inhibition of PLpro protease activities. *bioRxiv Prepr Serv Biol* 2020;:2020.07.01.183020. doi:10.1101/2020.07.01.183020

41 Barnard DL, Day CW, Bailey K, *et al.* Is the anti-psychotic, 10-(3-(dimethylamino)propyl)phenothiazine (promazine), a potential drug with which to treat SARS infections?. Lack of efficacy of promazine on SARS-CoV replication in a mouse model. *Antiviral Res* 2008;**79**:105–13. doi:10.1016/j.antiviral.2007.12.005

42 Mittal L, Kumari A, Srivastava M, *et al.* Identification of potential molecules against COVID-19 main protease through structure-guided virtual screening approach. *J Biomol Struct Dyn* 2020;**Just accep**:1–19. doi:10.1080/07391102.2020.1768151

43 Ianevski A, Yao R, Fenstad MH, *et al.* Potential Antiviral Options against SARS-CoV-2 Infection. *Viruses* 2020;**12**. doi:10.3390/v12060642

44 Musarrat F, Chouljenko V, Dahal A, *et al.* The anti-HIV drug nelfinavir mesylate (Viracept) is a potent inhibitor of cell fusion caused by the SARSCoV-2 spike (S) glycoprotein warranting further evaluation as an antiviral against COVID-19 infections. *J Med Virol* 2020;**Just accep**. doi:10.1002/jmv.25985

45 Tetrandrine Tablets Used in the Treatment of COVID-19 - Full Text View - ClinicalTrials.gov. ClinicalTrials.gov. 2020.

46 Dyall J, Coleman CM, Hart BJ, *et al.* Repurposing of Clinically Developed Drugs for Treatment of Middle East Respiratory Syndrome Coronavirus Infection. *Antimicrob Agents Chemother* 2014;**58**:4885–93. doi:10.1128/AAC.03036-14

47 Plaze M, Attali D, Petit A-C, *et al.* [Repurposing of chlorpromazine in COVID-19 treatment: the reCoVery study]. *Encephale* 2020;**46**:S35–9. doi:10.1016/j.encep.2020.04.010

48 Repurposing of Chlorpromazine in Covid-19 Treatment (reCoVery). ClinicalTrials.gov. 2020.

49 Aimo A, Baritussio A, Emdin M, *et al.* Amiodarone as a possible therapy for coronavirus infection. *Eur J Prev Cardiol* 2020;:204748732091923. doi:10.1177/2047487320919233

50 Stadler K, Ha HR, Ciminale V, *et al.* Amiodarone alters late endosomes and inhibits SARS coronavirus infection at a post-endosomal level. *Am J Respir Cell Mol Biol* 2008;**39**:142–9. doi:10.1165/rcmb.2007-0217OC

51 Amiodarone or Verapamil in COVID-19 Hospitalized Patients With Symptoms (ReCOVery-SIRIO). ClinicalTrials.gov. 2020.

52 Castaldo N, Aimo A, Castiglione V, *et al.* Safety and Efficacy of Amiodarone in a Patient With COVID-19. *JACC Case Reports* 2020;**2**:1307–10. doi:10.1016/j.jaccas.2020.04.053

53 DrugBank. COVID-19 Information. DrugBank. https://www.drugbank.ca/covid-19 (accessed 20 Jun 2020).

54 Capuzzi SJ, Sun W, Muratov EN, *et al.* Computer-Aided Discovery and Characterization of Novel Ebola Virus Inhibitors. *J Med Chem* 2018;**61**:3582–94. doi:10.1021/acs.jmedchem.8b00035

55 Muratov EN, Artemenko AG, Varlamova E V, *et al.* Per aspera ad astra: application of Simplex QSAR approach in antiviral research. *Future Med Chem* 2010;**2**:1205–26. doi:10.4155/fmc.10.194

56 Rajput A, Kumar A, Kumar M. Computational Identification of Inhibitors Using QSAR Approach Against Nipah Virus. *Front Pharmacol* 2019;**10**:71. doi:10.3389/fphar.2019.00071

57 Muratov EN, Varlamova E V, Artemenko AG, *et al.* QSAR analysis of [(biphenyloxy)propyl]isoxazoles: agents against coxsackievirus B3. *Future Med Chem* 2011;**3**:15–27. doi:10.4155/fmc.10.278

58 Bhargava S, Patel T, Gaikwad R, *et al.* Identification of structural requirements and prediction of inhibitory activity of natural flavonoids against Zika virus through molecular docking and Monte Carlo based QSAR Simulation. *Nat Prod Res* 2019;**33**:851–7. doi:10.1080/14786419.2017.1413574

59 Bobrowski T, Alves V, Melo-Filho CC, *et al.* Computational Models Identify Several FDA Approved or Experimental Drugs as Putative Agents Against SARS-CoV-2. *ChemRxiv Prepr* Published Online First: April 2020. doi:doi.org/10.26434/chemrxiv.12153594.v1

60 Xu J, Shi P-Y, Li H, *et al.* Broad Spectrum Antiviral Agent Niclosamide and Its Therapeutic Potential. *ACS Infect Dis* 2020;**6**:909–15. doi:10.1021/acsinfecdis.0c00052

61 Niclosamide In Moderate COVID-19. ClinicalTrials.gov. 2020.

62 Niclosamide for Mild to Moderate COVID-19. ClinicalTrials.gov. 2020.

63 Marrugal-Lorenzo JA, Serna-Gallego A, González-González L, *et al.* Inhibition of adenovirus infection by mifepristone. *Antiviral Res* 2018;**159**:77–83. doi:10.1016/j.antiviral.2018.09.011

64 DeBono A, Thomas DR, Lundberg L, *et al.* Novel RU486 (mifepristone) analogues with increased activity against Venezuelan Equine Encephalitis Virus but reduced progesterone receptor antagonistic activity. *Sci Rep* 2019;**9**:2634. doi:10.1038/s41598-019-38671-y

65 Schafer EA, Venkatachari NJ, Ayyavoo V. Antiviral effects of mifepristone on human immunodeficiency virus type-1 (HIV-1): targeting Vpr and its cellular partner, the glucocorticoid receptor (GR). *Antiviral Res* 2006;**72**:224–32. doi:10.1016/j.antiviral.2006.06.008

66 FOLLETT EAC, PENNINGTON TH. Antiviral Effect of Constituent Parts of the Rifampicin Molecule. *Nature* 1971;**230**:117–8. doi:10.1038/230117a0

67 Bizon C, Cox S, Balhoff J, *et al.* ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J Chem Inf Model* 2019;**59**:4968–73.

doi:10.1021/acs.jcim.9b00683

68 Korn D, Bobrowski T, Li M, *et al.* COVID-KOP: Integrating Emerging COVID-19 Data with the ROBOKOP Database. *Bioinformatics* 2020;:In press. doi:10.26434/chemrxiv.12462623

69 Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res* 2012;**40**:D136–43. doi:10.1093/nar/gkr1178

70 Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: Tool for the unification of biology. *Nat Genet* 2000;**25**:25–9. doi:10.1038/75556

71 Lipscomb CE. Medical Subject Headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265–6.

72 Murphy CN, Yates J. *The International Organization for Standardization (ISO): Global governance through voluntary consensus*. Routledge 2008. doi:10.4324/9780203884348

73 Robinson PN, Köhler S, Bauer S, *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet* 2008;**83**:610–5. doi:10.1016/j.ajhg.2008.09.017

74 Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7. doi:10.1093/nar/gky1075

75 Topalis P, Mitraka E, Bujila I, *et al.* IDOMAL: An ontology for malaria. *Malar J* 2010;**9**:230. doi:10.1186/1475-2875-9-230

76 Schwede T, Kopp J, Guex N, *et al.* SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003;**31**:3381–5. doi:10.1093/nar/gkg520