JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Application Note

COVID-19 Knowledge Extractor (COKE): A Curated Repository of Drug–Target Associations Extracted from the CORD-19 Corpus of Scientific Publications on COVID-19

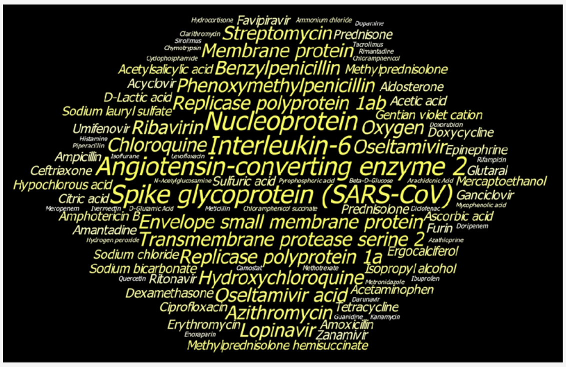
Daniel Korn, Vera Pervitsky, Tesia Bobrowski, Vinicius M. Alves, Charles Schmitt, Chris Bizon, Nancy Baker, Rada Chirkova, Artem Cherkasov, Eugene Muratov,* and Alexander Tropsha*

Cite This: https://doi.org/10.1021/acs.jcim.1c01285

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: 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 COVID-19 Knowledge Extractor (COKE), a web application to extract, curate, and annotate essential drug–target relationships from the research literature on COVID-19. SciBiteAI ontological tagging of the COVID Open Research Data set (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 the target protein and drug terms, and confidence scores were calculated for each entity pair. COKE processing of the current CORD-19 database identified about 3000 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. The COKE repository and web application can serve as a useful resource for drug repurposing against SARS-CoV-2. COKE is freely available at https://coke.mml.unc.edu/, and the code is available at https://github.com/DnlRKorn/CoKE.



■ INTRODUCTION

With over 180 million cases and over 4 million deaths worldwide as of June 2021, and no U.S. Food and Drug Administration (FDA) approved treatments against this virus except for several drugs authorized for emergency use, there are ongoing efforts to discover therapeutics 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 nomination of many drugs for clinical trials. Many databases collect data related to SARS-CoV-2;[3] however, the scientific literature concerning SARS-CoV-2 remains the largest repository of untapped biomedical data.[4,5] Indeed, hundreds of thousands of papers on COVID-19 and SARS-CoV-2 have appeared in the scientific literature since the beginning of the pandemic.[6]

Recently, the Allen Institute for AI, the National Institutes for Health (NIH), the White House, Georgetown University, and several other organizations collaborated to produce the COVID-19 Open Research Data set (CORD-19). As of June of 2021, this data set consisted of 660,000 scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses, including over 390,000 with full text.[6] SciBiteAI, a semantics research group based in the United Kingdom,[7] has curated an ontologically annotated version of the data set to identify biomedical terms within sentences of full papers or abstracts.[6]

Using this semantically annotated collection of research papers, we have developed the COVID-19 Knowledge Extractorr (COKE), a web application summarizing all drug–target–coronavirus relationships annotated in the research literature captured by the CORD-19 project. We have developed COKE by (i) detecting drug and protein literature co-occurrences within all manuscripts annotated in the CORD-19 corpus,[8] (ii) establishing a scoring system to rate the confidence of a co-occurrence pair, and (iii) highlighting specific sections of manuscripts where respective terms co-occur. COKE has been developed to provide the scientific community with data that could potentially contribute to COVID-19 drug repurposing efforts. The COKE portal (https://coke.mml.unc.edu/) provides data on

Received: October 21, 2021

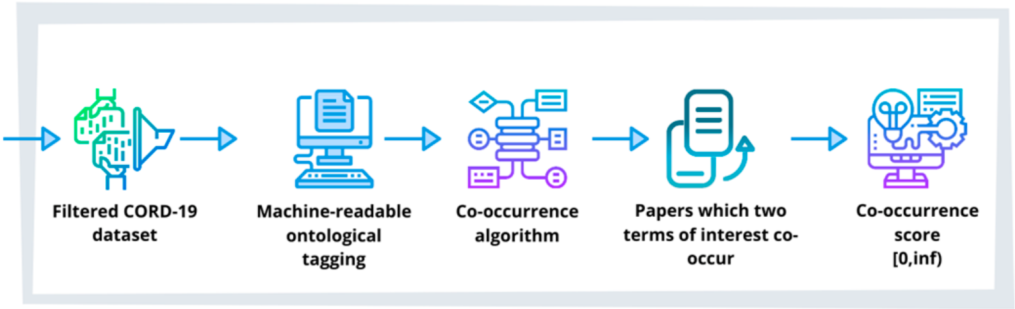

CORD-19 database
Filtered CORD-19 dataset
Machine-readable ontological tagging
Co-occurrence algorithm
Papers which two terms of interest co-occur
Co-occurrence score [0,inf)

Figure 1. Workflow for the COVID-19 Knowledge Extractor algorithm and co-occurrence score calculation.

human and viral proteins associated with SARS-CoV-2 and other coronaviruses as described in the CORD-19 corpus and indexed in UniProt as well as chemicals targeting those proteins that are indexed in the DrugBank.[9,10]

■ MATERIALS AND METHODS

The workflow employed for the development of COKE is summarized in Figure 1.

Data Set Collection. The SciBiteAI group has released, under open licenses, an ontologically annotated version of the CORD-19 database available publicly on their GitHub account.[6] The entire content of each paper within the CORD-19 data set is divided into paragraphs of plain text. Within each of these paragraphs, ontological terms divided into eight groups are assigned and matched to the sentences they occur in (cf. Table S1).

We extracted information on human proteins associated with coronavirus disease (host protein targets) or coronavirus proteins (viral protein targets) from the expert-curated UniProt database.[9] Additionally, we extracted information on synonyms of these proteins, the organisms from which they were derived, and their genome sequence length.

The DrugBank data set provides listings of chemical identifiers tagged as small molecule drugs as well as targets a drug is associated with. In addition, standardized naming of these compounds and their SMILES strings are also provided.[10] COKE makes these data available and organized such that users of COKE can easily examine relevant drug−target−SARS-CoV-2 relationships.

Data Curation and Integration. We found that many papers in the original CORD-19 data set were related to other viruses, such as Ebola and Zika, as well as epidemiological studies not relevant to drug repurposing for COVID-19. Thus, 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. Any paper in the data set, which was not annotated by one of the NCBITaxon tags for a coronavirus, was not considered.

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 in which a term appears 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,[11] and therefore, every term mentioned in the abstract is considered to be of greater significance in the context of the whole study.

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, (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 probability of mass function (pmf; eq 1) was calculated as

$$\text{pmf}\,(k) = \frac{\binom{\#\,\text{Occurences of Term 1}}{k}\binom{\#\,\text{Papers}-\#\,\text{Occurences of Term 1}}{\#\,\text{Papers}-\#\,\text{Occurences of Term 2}}}{\binom{\#\,\text{Papers}}{\#\,\text{Occurences of Term 2}}} \tag{1}$$

and the cumulative distribution function (CDF) for the co-occurrence score was calculated by summing all values between 0 and the overlap count

$$\text{CDF}\,(\#\,\text{Overlap of Term 1 and Term 2}) = \sum_{k=0}^{\#\,\text{Overlap}} \text{pmf}\,(k) \tag{2}$$

The SciPy implementation of these functions was used to enable calculations.[12] Many of the scores can be quite small and tightly clustered, so we used the logarithm of the CDF, which preserves the order of respective scores. To make the score more interpretable, the sign was flipped, so that the apparent 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.

We then filtered the large set of co-occurrence tuples only for CVPROT (COVID-19 UniProtKB, see the Table S1) 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 9500 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 data set. 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 chemical data in our data set (Figure 2). We cross-referenced all ChEMBL tags against

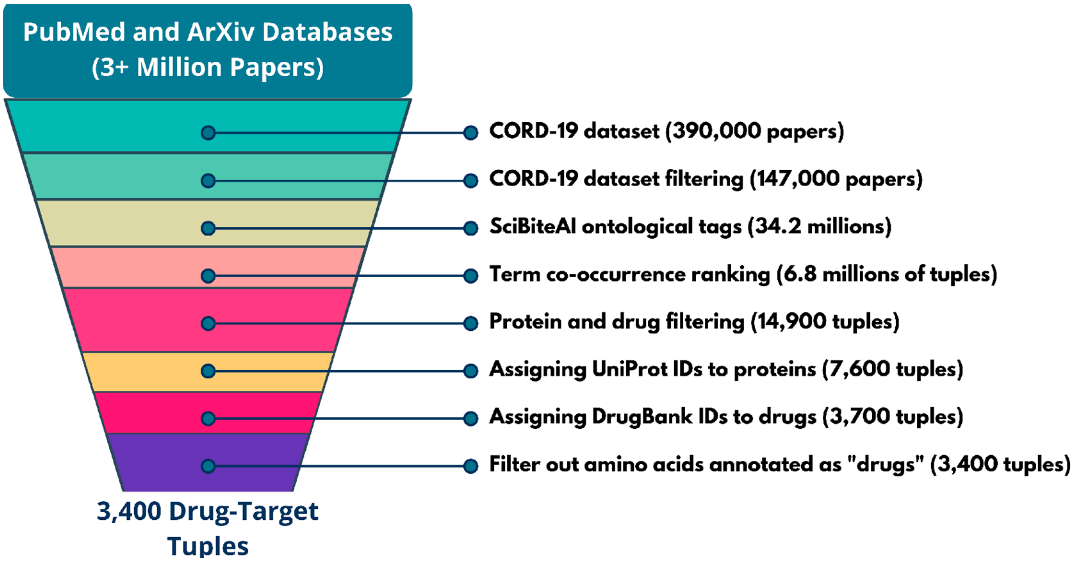

PubMed and ArXiv Databases (3+ Million Papers)
CORD-19 dataset (390,000 papers)
CORD-19 dataset filtering (147,000 papers)
SciBiteAI ontological tags (34.2 millions)
Term co-occurrence ranking (6.8 millions of tuples)
Protein and drug filtering (14,900 tuples)
Assigning UniProt IDs to proteins (7,600 tuples)
Assigning DrugBank IDs to drugs (3,700 tuples)
Filter out amino acids annotated as "drugs" (3,400 tuples)
3,400 Drug-Target Tuples

Figure 2. Data curation process leading to selection of drug−target associations.

DrugBank. Compounds that were not present in DrugBank were removed, resulting in 3711 drug/protein tuples. Additionally, we excluded amino acids, peptides, and proteins, so that only small molecules remained. Our final data set had 3368 drug−protein tuples. A histogram of scores for all drug−protein pairs is shown in Figure S1.

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.[13] 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.[14] The CPE assay initially contained 6988 chemicals with AC_{50} dose−response curves. The same collection was subjected to a counter screen to ensure compounds identified as active in the primary assay were not false positives because of their cytotoxicity. After curation, 4625 (165 actives, 4164 inactives, and 296 inconclusive) small molecules remained in the primary assay. In the Spike-ACE2 data set, 3406 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 were used to ensure that compounds were not false positives 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.[15−17] 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 "nontoxic" 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 data set, 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-lass 3 (CPE) or −3 (Spike-ACE2) as "active" because compounds with these curves were labeled as "low-quality actives" by NCATS.[13]

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 data set, and (iii) the highlighted sections of papers from CORD-19. 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 data set in which information relevant to the queried (drug−protein) tuple is presented to them. Using Python 3.7 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 DataTables jQuery library,[18] which aids in rendering dynamic complex HTML tables in the browser. We converted all the co-occurrence tuples into JSON files in DataTable's specified format. These JSON files are stored within COKE's web domain as static files. Then, each data table makes an AJAX request for their specified information when the user loads the table. 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.

RESULTS AND DISCUSSION

Comparison of Drug−Target Associations in COKE Data Set and Based on Bioactivity Screening in COVID-19 Assays. . Our drug list identified by COKE initially contained 499 molecules. After curation, 471 unique drugs were kept. From this list, there were 335 compounds in the Approved Drugs Collection evaluated in the SARS-CoV-2 CPE (13 actives, 304 inactives, and 17 inconclusive). Table 1 lists all 13 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.

Table 1. 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

Umifenovir (Arbidol) was also identified as active when predicted by the QSAR models developed by our group recently for the inhibitors of severe acute respiratory syndrome coronavirus (SARS-CoV) main protease (M^{pro}).[19] Umifenovir was found active against SARS-CoV-2 *in vitro*[20] as a binder to the spike glycoprotein of SARS-CoV-2 (UniProt ID P0DTC2).[21] Imatinib,[22−24] promethazine,[25] and fluoxetine[26] 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*.[27] In addition, imatinib is currently being studied in COVID-19 clinical trials[23] and has been shown to successfully treat COVID-19 in a case report.[28] Another study has shown that promethazine also has inhibitory activity against MERS-CoV *in vitro*.[29] Reserpine also demonstrated anti-SARS-CoV activity *in vitro*.[30,31] Although 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,[32] nonpeer-reviewed empirical evidence,[33] and computational studies.[34]

Among the compounds identified as active but cytotoxic, hexachlorophene is a topical antibacterial agent.[35] Nitazoxanide showed activity in the phenotypic screen,[36] and it has been included in prophylactic postexposure clinical trials.[37] In nonpeer-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*.[38] Chlorprothixene was shown to inhibit SARS-CoV replication *in vitro*.[39] Nelfinavir mesylate, an HIV protease inhibitor, was shown to inhibit M^{pro} in the computational analysis[40] and to have activity against the SARS-CoV-2 spike glycoprotein *in vitro*.[41,42] Tetrandrine is currently being explored in COVID-19 clinical trials.[43] In previous studies, chlorpromazine showed activity against MERS-CoV and SARS-CoV,[44,45] and it is being studied in clinical trials in hospitalized patients with COVID-19.[46] Past studies showed that amiodarone has *in vitro* antiviral activity against SARS-CoV by interfering with endocytosis and viral replication.[47,48] It is currently being studied in clinical trials against COVID-19,[49] and it was recently documented in a case report to successfully treat COVID-19.[50] In summary, COKE successfully highlighted compounds shown to be active against SARS-CoV-2 in the phenotypic assay. 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 (Table 1). 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. The complete list is available at https://github.com/DnlRKorn/CoKE.

Comparison of COKE Data Set 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 data set. To obtain a list of drugs already in clinical trials for COVID-19, we leveraged DrugBank's data set, which matched active clinical trials to DrugBank IDs.[51] 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 data set. This observation is surprising, if not shocking, as it means that 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 could escape peer review, a process commonly accepted by the global research community for validating research observations and hypotheses before exposing them to the broad research community.

Application of the COKE Web Portal. An example of using the COKE web portal can be seen in Figure S2. 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 assessed 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,[52] and it is currently being tested in clinical trials against SARS-CoV-2.[53,54] In the past studies, the synthetic steroid mifepristone demonstrated antiviral activity against human adenovirus,[55] Venezuelan equine encephalitis virus,[56] and HIV-1.[57] Flunarizine, an antimigraine drug, is known to arrest virus−membrane fusion for various hepatitis C virus genotypes.[57] A 1971 study by Follett and Pennington demonstrated that the antibiotic rifampicin could inhibit poxvirus replication;[58] more recent pieces of 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.[49] 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.[59] 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.

To understand which cutoff values of confidence score are optimal, we have benchmarked the different cutoffs for the confidence score for 20 randomly selected active and inactive drugs, four of which were not annotated in the CORD-19 database. The results, including the actual scores and confusion matrices are shown in Tables S2−S6. Overall, we have observed that if the cutoff is too strict (confidence score ≤ 0.1), then the recall is low, but precision is high. Use of moderate (confidence score ≤ 1; confidence score ≤ 2) cutoffs decreased the recall but increased the precision, and a loose cutoff (confidence score ≤ 5) leads to low precision and high recall. In other words, there is no optimal confidence score value, and its choice depends on the preferences of a researcher.

Data reported in COKE can be viewed as connections between biomedical entities, which could easily be incorporated into biomedical knowledge graphs such as ROBOKOP[60,61] 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 by natural language processing of research papers. 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.

■ DATA AND SOFTWARE AVAILABILITY

All data and software involved in the creation of this tool are free and open for all to access. The COVID Open Research Data set is freely available to download at https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. The SciBite ontological tagging of the CORD-19 data set is available at https://github.com/SciBiteLabs/CORD19. Finally, all the code for the generation of ranked tuples and the creation of the web portal is available at https://github.com/DnlRKorn/COKE.

■ CONCLUSIONS

We have built COKE, a web application to extract 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, 3368 drug−protein pairs were identified by COKE, including 29 unique proteins (22 viral targets and seven 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 for retrieving drug repurposing candidates annotated in the research literature, as well as for helping to triage drug repurposing candidates identified by computational methods such as virtual screening.

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.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01285.

Discussion of the development of the COKE web portal and figures and images that describe the utility of COKE as a predictive tool (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Eugene Muratov − *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, North Carolina 27599, United States*; orcid.org/0000-0003-4616-7036; Phone: (919) 966-2955; Email: murik@email.unc.edu; Fax: (919) 966-0204

Alexander Tropsha − *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, North Carolina 27599, United States;* orcid.org/0000-0003-3802-8896; Phone: (919) 966-2955; Email: alex_tropsha@unc.edu; Fax: (919) 966-0204

Authors

Daniel Korn − *Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States; 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, North Carolina 27599, United States*

Vera Pervitsky − *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, North Carolina 27599, United States*

Tesia Bobrowski − *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, North Carolina 27599, United States*

Vinicius M. Alves − *Office of Data Science, National Toxicology Program, NIEHS, Morrisville, North Carolina 27560, United States;* orcid.org/0000-0002-6182-1748

Charles Schmitt − *Office of Data Science, National Toxicology Program, NIEHS, Morrisville, North Carolina 27560, United States*

Chris Bizon − *Renaissance Computing Institute, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States*

Nancy Baker − *ParlezChem, Hillsborough, North Carolina 27278, United States*

Rada Chirkova − *Department of Computer Science, North Carolina State University, Raleigh, North Carolina 27606-5550, United States*

Artem Cherkasov − *Vancouver Prostate Centre, University of British Columbia, Vancouver, BC V6H 3Z6, Canada*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.1c01285

Funding

This work was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (OT2TR003441).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge the work of the Allen Institute of AI for putting together the CORD-19 data set. We also wish to recognize SciBiteAI that provided the ontological tagging for this project.

■ REFERENCES

(1) COVID-19 Map. *Johns Hopkins Coronavirus Resource Center.* https://coronavirus.jhu.edu/ (accessed Jun 21, 2021).

(2) Bobrowski, T.; Melo-Filho, C. C.; Korn, D.; Alves, V. M.; Popov, K. I.; Auerbach, S.; Schmitt, C.; Moorman, N. J.; Muratov, E. N.; Tropsha, A. Learning from History: Do Not Flatten the Curve of Antiviral Research! *Drug Discovery Today* **2020**, *25*, 1604−1613.

(3) Open-Access Data and Computational Resources to Address COVID-19. *Office of Data Science Strategy, National Institutes of Health.* https://datascience.nih.gov/covid-19-open-access-resources (accessed Jun 26, 2020).

(4) Hunter, L. E. Knowledge-Based Biomedical Data Science. *Data Sci.* **2017**, *1* (1−2), 19−25.

(5) Bakken, S. Informatics Is a Critical Strategy in Combating the COVID-19 Pandemic. *J. Am. Med. Informatics Assoc.* **2020**, *27*, 843−844.

(6) SciBiteLabs/CORD19: Annotated Data for the COVID-19 Open Research Dataset Challenge. *GitHub.* https://github.com/SciBiteLabs/CORD19 (accessed Jun 26, 2020).

(7) About us. *SciBite.* https://www.scibite.com/# (accessed Jun 26, 2020).

(8) Lu Wang, L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.; Liu, Z.; Merrill, W.; Mooney, P.; Murdick, D.; Rishi, D.; Sheehan, J.; Shen, Z.; Stilson, B.; Wade, A. D.; Wang, K.; Wilhelm, C.; Xie, B.; Raymond, D.; Weld, D. S.; Etzioni, O.; Kohlmeier, S. CORD-19: The Covid-19 Open Research Dataset. *arXiv Preprint.* arXiv:2004.10706, 2020.

(9) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515.

(10) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668−72.

(11) Cohen, K. B.; Johnson, H. L.; Verspoor, K.; Roeder, C.; Hunter, L. E. The Structural and Content Aspects of Abstracts versus Bodies of Full Text Journal Articles Are Different. *BMC Bioinf.* **2010**, *11*, 492.

(12) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C J; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Haggstrom, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T; Slavic, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schonberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodriguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kummerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vazquez-Baeza, Y. 1. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.

(13) OpenData, COVID-19. *National Center for Advancing Translational Sciences.* https://opendata.ncats.nih.gov/covid19/ (accessed Jun 26, 2020).

(14) Beaudet, L.; Rodriguez-Suarez, R.; Venne, M.-H.; Caron, M.; Bédard, J.; Brechler, V.; Parent, S.; Bielefeld-Sévigny, M. AlphaLISA Immunoassays: The No-Wash Alternative to ELISAs for Research and Drug Discovery. *Nat. Methods* **2008**, *5*, an8−an9.

(15) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11*, 535−535.

(16) 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*, 1189−1204.

(17) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56*, 1243−1252.

(18) DataTables, Table plug-in for jQuery. *DataTables.* https://datatables.net/ (accessed May 20, 2020).

(19) Alves, V. M.; Bobrowski, T.; Melo-Filho, C. C.; Korn, D.; Auerbach, S.; Schmitt, C.; Muratov, E. N.; Tropsha, A. QSAR Modeling of SARS-CoV Mpro Inhibitors Identifies Sufugolix, Cenicriviroc, Proglumetacin, and Other Drugs as Candidates for Repurposing against SARS-CoV-2. *Mol. Inf.* **2021**, *40*, 2000113 minf.202000113.

(20) Wang, X.; Cao, R.; Zhang, H.; Liu, J.; Xu, M.; Hu, H.; Li, Y.; Zhao, L.; Li, W.; Sun, X.; Yang, X.; Shi, Z.; Deng, F.; Hu, Z.; Zhong, W.; Wang, M. The Anti-Influenza Virus Drug, Arbidol Is an Efficient Inhibitor of SARS-CoV-2 in Vitro. *Cell Discovery* **2020**, *6*, 28.

(21) Arbidol, OpenDataPortal. *NCATS.* https://opendata.ncats.nih.gov/covid19/databrowser?q=arbidol (accessed Jul 20, 2020).

(22) The Safety & Efficacy of Imatinib for the Treatment of SARS-COV-2 Induced Pneumonia. *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04422678 (accessed Jun 26, 2020).

(23) Trial of Imatinib for Hospitalized Adults With COVID-19. *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04394416 (accessed Jun 26, 2020).

(24) Imatinib in COVID-19 Disease in Aged Patients (IMAGE-19). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04357613 (accessed Jun 26, 2020).

(25) Pharmacokinetics, Pharmacodynamics, and Safety Profile of Understudied Drugs Administered to Children per Standard of Care (POPS) (POPS or POP02). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04278404 (accessed Jun 26, 2020).

(26) Fluoxetine to Reduce Intubation and Death after COVID19 Infection. *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04377308 (accessed Jun 26, 2020).

(27) Pillaiyar, T.; Manickam, M.; Jung, S.-H. Middle East Respiratory Syndrome-Coronavirus (MERS-CoV): An Updated Overview and Pharmacotherapeutics. *Med. Chem. (Los Angeles, CA, U. S.)* **2015**, *5*, 361−372.

(28) Morales-Ortega, A.; Bernal-Bello, D.; Llarena-Barroso, C.; Frutos-Pérez, B.; Duarte-Millán, M. Á.; García de Viedma-García, V.; Farfán-Sedano, A. I.; Canalejo-Castrillero, E.; Ruiz-Giardín, J. M.; Ruiz-Ruiz, J.; San Martín-López, J. V. Imatinib for COVID-19: A Case Report. *Clin. Immunol.* **2020**, *218*, 108518.

(29) Liu, Q.; Xia, S.; Sun, Z.; Wang, Q.; Du, L.; Lu, L.; Jiang, S. Testing of Middle East Respiratory Syndrome Coronavirus Replication Inhibitors for the Ability To Block Viral Entry. *Antimicrob. Agents Chemother.* **2015**, *59*, 742−744.

(30) Wu, C.-Y.; Jan, J.-T.; Ma, S.-H.; Kuo, C.-J.; Juan, H.-F.; Cheng, Y.-S. E.; Hsu, H.-H.; Huang, H.-C.; Wu, D.; Brik, A.; Liang, F.-S.; Liu, R.-S.; Fang, J.-M.; Chen, S.-T.; Liang, P.-H.; Wong, C.-H. Small Molecules Targeting Severe Acute Respiratory Syndrome Human Coronavirus. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 10012−10017.

(31) Liang, P.-H. Characterization and Inhibition of SARS-Coronavirus Main Protease. *Curr. Top. Med. Chem.* **2006**, *6*, 361−376.

(32) Fitzgerald, P. J. Noradrenergic and Serotonergic Drugs May Have Opposing Effects on COVID-19 Cytokine Storm and Associated Psychological Effects. *Med. Hypotheses* **2020**, *144*, 109985.

(33) Zimniak, M.; Kirschner, L.; Hilpert, H.; Seibel, J.; Bodem, J. The Serotonin Reuptake Inhibitor Fluoxetine Inhibits SARS-CoV-2. *bioRxiv Preprint*, 2020. DOI: 10.1101/2020.06.14.150490.

(34) Beck, B. R.; Shin, B.; Choi, Y.; Park, S.; Kang, K. 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*, 784−790.

(35) U.S. Food and Drug Administration, Code of Federal Regulations 21. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm (accessed Jun 26, 2020).

(36) Wang, M.; Cao, R.; Zhang, L.; Yang, X.; Liu, J.; Xu, M.; Shi, Z.; Hu, Z.; Zhong, W.; Xiao, G. Remdesivir and Chloroquine Effectively Inhibit the Recently Emerged Novel Coronavirus (2019-NCoV) in Vitro. *Cell Res.* **2020**, *30*, 269−271.

(37) 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). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04343248 (accessed Jun 26, 2020).

(38) Swaim, C. D.; Perng, Y.-C.; Zhao, X.; Canadeo, L. A.; Harastani, H. H.; Darling, T. L.; Boon, A. C. M.; Lenschow, D. J.; Huibregtse, J. M. 6-Thioguanine Blocks SARS-CoV-2 Replication by Inhibition of PLpro Protease Activities. *bioRxiv Preprint*, 2020. DOI: 10.1101/2020.07.01.183020.

(39) Barnard, D. L.; Day, C. W.; Bailey, K.; Heiner, M.; Montgomery, R.; Lauridsen, L.; Jung, K. H.; Li, J. K. K.; Chan, P. K. S.; Sidwell, R. W. 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*, 105−113.

(40) Mittal, L.; Kumari, A.; Srivastava, M.; Singh, M.; Asthana, S. Identification of Potential Molecules against COVID-19 Main Protease through Structure-Guided Virtual Screening Approach. *J. Biomol. Struct. Dyn.* **2021**, *39*, 3662−3680.

(41) Ianevski, A.; Yao, R.; Fenstad, M. H.; Biza, S.; Zusinaite, E.; Reisberg, T.; Lysvand, H.; Løseth, K.; Landsem, V. M.; Malmring, J. F.; Oksenych, V.; Erlandsen, S. E.; Aas, P. A.; Hagen, L.; Pettersen, C. H.; Tenson, T.; Afset, J. E.; Nordbø, S. A.; Bjørås, M.; Kainov, D. E. Potential Antiviral Options against SARS-CoV-2 Infection. *Viruses* **2020**, *12*, 642.

(42) Musarrat, F.; Chouljenko, V.; Dahal, A.; Nabi, R.; Chouljenko, T.; Jois, S. D.; Kousoulas, K. G. 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**, *92*, 2087−2095.

(43) Tetrandrine Tablets Used in the Treatment of COVID-19 (TT-NPC). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04308317 (accessed Dec 13, 2020).

(44) Dyall, J.; Coleman, C. M.; Hart, B. J.; Venkataraman, T.; Holbrook, M. R.; Kindrachuk, J.; Johnson, R. F.; Olinger, G. G.; Jahrling, P. B.; Laidlaw, M.; Johansen, L. M.; Lear-Rooney, C. M.; Glass, P. J.; Hensley, L. E.; Frieman, M. B. Repurposing of Clinically Developed Drugs for Treatment of Middle East Respiratory Syndrome Coronavirus Infection. *Antimicrob. Agents Chemother.* **2014**, *58*, 4885−4893.

(45) Plaze, M.; Attali, D.; Petit, A.-C.; Blatzer, M.; Simon-Loriere, E.; Vinckier, F.; Cachia, A.; Chrétien, F.; Gaillard, R. [Repurposing of Chlorpromazine in COVID-19 Treatment: The ReCoVery Study]. *Encephale* **2020**, *46*, S35−S39.

(46) Repurposing of Chlorpromazine in Covid-19 Treatment (reCoVery). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04366739 (accessed Jul 27, 2020).

(47) Aimo, A.; Baritussio, A.; Emdin, M.; Tascini, C. Amiodarone as a Possible Therapy for Coronavirus Infection. *Eur. J. Prev. Cardiol.* **2021**, *28*, e16.

(48) Stadler, K.; Ha, H. R.; Ciminale, V.; Spirli, C.; Saletti, G.; Schiavon, M.; Bruttomesso, D.; Bigler, L.; Follath, F.; Pettenazzo, A.; Baritussio, A. Amiodarone Alters Late Endosomes and Inhibits SARS Coronavirus Infection at a Post-Endosomal Level. *Am. J. Respir. Cell Mol. Biol.* **2008**, *39*, 142−149.

(49) Amiodarone or Verapamil in COVID-19 Hospitalized Patients With Symptoms (ReCOVery-SIRIO). *U.S. National Library of Medicine.* https://clinicaltrials.gov/ct2/show/NCT04351763 (accessed Dec 10, 2020).

(50) Castaldo, N.; Aimo, A.; Castiglione, V.; Padalino, C.; Emdin, M.; Tascini, C. Safety and Efficacy of Amiodarone in a Patient With COVID-19. *JACC Case Reports* **2020**, *2*, 1307−1310.

(51) COVID-19 Information. *DrugBank*. https://www.drugbank.ca/covid-19 (accessed Jun 26, 2020).

(52) Xu, J.; Shi, P.-Y.; Li, H.; Zhou, J. Broad Spectrum Antiviral Agent Niclosamide and Its Therapeutic Potential. *ACS Infect. Dis.* **2020**, *6*, 909−915.

(53) Niclosamide in Moderate COVID-19. *U.S. National Library of Medicine*. https://clinicaltrials.gov/ct2/show/NCT04436458 (accessed Jul 27, 2020).

(54) Niclosamide for Mild to Moderate COVID-19. *U.S. National Library of Medicine*. https://clinicaltrials.gov/ct2/show/NCT04399356 (accessed Jul 27, 2020).

(55) Marrugal-Lorenzo, J. A.; Serna-Gallego, A.; González-González, L.; Buñuales, M.; Poutou, J.; Pachón, J.; Gonzalez-Aparicio, M.; Hernandez-Alcoceba, R.; Sánchez-Céspedes, J. Inhibition of Adenovirus Infection by Mifepristone. *Antiviral Res.* **2018**, *159*, 77−83.

(56) DeBono, A.; Thomas, D. R.; Lundberg, L.; Pinkham, C.; Cao, Y.; Graham, J. D.; Clarke, C. L.; Wagstaff, K. M.; Shechter, S.; Kehn-Hall, K.; Jans, D. A. Novel RU486 (Mifepristone) Analogues with Increased Activity against Venezuelan Equine Encephalitis Virus but Reduced Progesterone Receptor Antagonistic Activity. *Sci. Rep.* **2019**, *9*, 2634.

(57) Schafer, E. A.; Venkatachari, N. J.; 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*, 224−232.

(58) Follett, E. A. C.; Pennington, T. H. Antiviral Effect of Constituent Parts of the Rifampicin Molecule. *Nature* **1971**, *230*, 117−118.

(59) Lian, N.; Xie, H.; Lin, S.; Huang, J.; Zhao, J.; Lin, Q. Umifenovir Treatment Is Not Associated with Improved Outcomes in Patients with Coronavirus Disease 2019: A Retrospective Study. *Clin. Microbiol. Infect.* **2020**, *26*, 917−921.

(60) Bizon, C.; Cox, S.; Balhoff, J.; Kebede, Y.; Wang, P.; Morton, K.; Fecho, K.; Tropsha, A. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J. Chem. Inf. Model.* **2019**, *59*, 4968−4973.

(61) Korn, D.; Bobrowski, T.; Li, M.; Kebede, Y.; Wang, P.; Owen, P.; Vaidya, G.; Muratov, E.; Chirkova, R.; Bizon, C.; Tropsha, A. COVID-KOP: Integrating Emerging COVID-19 Data with the ROBOKOP Database. *Bioinformatics* **2021**, *37*, 586−587.