The COVID-19 Therapeutic Information Browser

Tonia Korves[1], Christopher Garay[1], Robyn Kozierok[1], Matthew Peterson[1], Ian Ross[2], Miron Livny[2], Theodosius Rekatsinas[2], Shanan E. Peters[2], and Lynette Hirschman[1]

[1]The MITRE Corporation, Bedford, MA, USA
[2]University of Wisconsin-Madison, Madison, WI, USA

***Abstract*— As the pace of COVID-19 research rapidly escalated during the pandemic, we built a platform to help biomedical experts easily discover published research about potential COVID-19 therapeutics and vaccines. The platform, COVID-TIB for short, uses rule-based natural language processing to identify documents about viruses, drugs, and vaccine types at scale. COVID-TIB displays numbers of journal papers, preprints, and clinical trials about particular drugs and SARS-CoV-2 and other viruses, binned by research stage. Users can apply multiple filters (e.g., recency, publication type, virus) to winnow results, select a set to see document metadata with links, and download search results. Users can also search for terms in the full text of papers and filter papers based on these search terms. In addition, users can link out to xDD's COSMOS system (https://xdd.wisc.edu/set_visualizer/sets/xdd-covid-19) to browse figures and tables related to a virus and therapeutic. COVID-TIB currently presents information on 13 viruses and nearly 3,000 therapeutics found in 26K paper abstracts and clinical trial summaries. COVID-TIB can be used to browse recent drug research, identify promising drug candidates, gather information for reviews and meta-analyses, and track research over time. COVID-TIB is available at https://covidtib.c19hcc.org/.**

Keywords—COVID-19, therapeutics, drugs, natural language processing, text mining

I. Introduction

With the urgent need for treatments during the COVID-19 pandemic, scientific literature about COVID-19 grew rapidly, at a rate of 10K papers per month (1). To help biomedical experts track and discover scientific results, we sought to develop a resource for easily finding publications and clinical trials about candidate COVID-19 therapeutics. This system had to be developed quickly to meet immediate needs and had to be built without the benefit of prior curated data for training algorithms. The system needed a pipeline to regularly process new documents at scale, and a dashboard to make these results easily accessible. The system also needed to be easy to adapt to meet the rapidly evolving research landscape, including changing virus nomenclature and new drugs.

Here we describe the platform we developed called the COVID-19 Therapeutic Information Browser, or COVID-TIB for short. This platform includes a natural language processing (NLP) pipeline to identify papers about therapeutics, viruses, vaccines, and research stages; a dashboard for easily exploring these results; and full-text search capabilities. In this report, we focus on the therapeutics portion.

II. Methods

A. Natural Language Processing

Prior to the pandemic, we developed a platform for identifying vaccine-related research for viruses. The platform used a rule-based NLP software called Reach (2), which we extended to extract information about viruses and vaccine types by creating a dictionary of virus names and writing rules for identifying types of vaccines, such as vector-based, protein subunit, and DNA vaccines. The initial platform also included a curation user interface for manually reviewing this extracted information and associated text evidence. When the SARS-CoV-2 outbreak started, we added reading for SARS-CoV-2 and related coronaviruses, therapeutics, and six research stages, such as clinical study, case report, and cell assay. The pipeline consists of the following steps:

1. *Select and download potentially relevant documents to process.* Sources include PubMed abstracts and metadata (https://pubmed.ncbi.nlm.nih.gov/); preprint abstracts and metadata from CORD-19 (3; https://www.semanticscholar.org/cord19); and clinical trial summaries and metadata from ClinicalTrials.gov (https://clinicaltrials.gov/). Documents for processing were identified by querying for a list of virus names and their synonyms.

2. *Extract entity information from sentences, utilizing Reach (2) rules, custom dictionaries for drugs and viruses, and keywords for vaccine information and research stages.* The virus dictionary uses NCBI Taxon Ids (https://www.ncbi.nlm.nih.gov/taxonomy). To create the drug dictionary, we started with a DrugBank dictionary (public domain DrugBank Vocabulary v5.1.5, released on January 3, 2020; https://go.drugbank.com/releases/5-1-5#open-data), removed irrelevant compounds such as allergens, and manually added drugs and synonyms using information from ClinicalTrials.gov metadata, the CORONA registry (https://cdcn.org/corona-data-viewer/), word embeddings from COSMOS (https://cosmos.wisc.edu), and other lists. We periodically manually update the dictionary by adding new drugs and synonyms found in ClinicalTrials.gov metadata and in papers. Each therapeutic has either a DrugBank or PubChem (https://pubchem.ncbi.nlm.nih.gov/) identifier where available. The dictionary has 34,506 synonyms for 12,286 therapeutics. For research stages, example keywords include "patient", "hospital*", "ICU", "first case", "in an adult", "model*", "IC50", and "Phase I".

3. *Assemble extracted entity information across sentences within documents and classify documents.* We classified

documents as being about a virus-drug pair and a research stage by creating rules and using thresholds of evidence. For virus-drug pairs, we required that 1) the drug (including any synonym) be mentioned at least once in the title, abstract, clinical trial summary, or ClinicalTrials.gov metadata Intervention field. and 2) the virus or viral disease be mentioned either once in the title or at least twice collectively in the title and abstract. For research stage classification, we defined eight stages: Clinical study, Emergency use and case report, Nonhuman primate, Small animal, Cell line, In vitro, AI/In silico, and Review. Documents that fell outside of classification criteria were classified as Not Identified and the In vitro category was used only in manual curation. Documents were classified as belonging to a research stage using rules involving numbers of keywords associated with stages within an abstract, incidence of particular phrases in the title, and PubMed paper type metadata. When rules pointed to multiple research stages, the most advanced research stage was assigned to the document. We implemented these rules in a series of R programming language scripts, which also cleaned the data and removed many preprint-journal paper duplicates.

We iteratively improved the entity extraction methods and classification rules over time after COVID-TIB's initial release. During the first several months, we manually reviewed and curated a portion of the results from the NLP pipeline with our curation user interface and used observations about erroneous classifications to inform changes to rules in the NLP pipeline. Earlier versions of COVID-TIB indicated which results were curated. In the current version of COVID-TIB, curation status is no longer marked and curated information now constitutes just 2% of the results presented.

We assessed precision for SARS-CoV-2-drug pair and research stage classifications by manually evaluating the NLP results available in COVID-TIB in spring 2021 and found that precision for SARS-CoV-2-drug pairs in papers and preprints classified with a research stage other than "Not Identified" was 0.96. We assessed recall for publications by comparing NLP results to a set of abstracts about COVID-19 and any named drug that we manually identified from the LitCovid Treatment set (1; https://www.ncbi.nlm.nih.gov/research/coronavirus/). We found a recall of 0.89; a third of the misses were due to drug names not being present in our dictionary. Additional information about precision and recall measures for the SARS-CoV-2-drug and research stage results and additional details about the NLP pipeline are available on the COVID-TIB website (https://covidtib.c19hcc.org/) in the data dictionary file, reached via the Data download button.

We further extended COVID-TIB to include full text search by integrating capabilities from the eXtract Dark Data (xDD) searchable digital library with full-text content provided by open access and partner publishers (https://xdd.wisc.edu/; COVID-19 instance: https://xdd.wisc.edu/set_visualizer/sets/xdd-covid-19). Information from text, tables, and figures was also retrieved using xDD's COSMOS system (https://cosmos.wisc.edu), a machine-learning based platform designed to serve as a technical assistant for discovering and extracting information from documents (4,5).

B. *The Dashboard*

The dashboard is an R Shiny application deployed using ShinyProxy. It includes a homepage with a description of COVID-TIB and a menu to navigate to other pages, including a viral therapeutics page, a vaccine page, descriptions of the research stages, and links to downloadable files that include a demo about how to use COVID-TIB.

On the Viral Therapeutics page, the table at the top shows the number of documents found for a virus (shown at the top of the table) about a therapeutic (row), in a specific research stage (columns below virus) (Fig. 1). The table is initially sorted by descending document abundance among therapeutics; users can sort by other columns or search for a particular therapeutic by typing in a name in the search box under "Therapeutic". Users can also use filters on the menu bar to limit by publication date, type of document, and virus. Clicking on a number in the table generates a new table with metadata about these documents (shown in Fig. 2); these results can be downloaded as a csv. Clicking a number also generates a link-out to COSMOS, which provides figures and tables from full text papers about SARS-CoV-2 and the chosen drug (Fig. 3). There is also a field to search for terms in the full text of journal papers via the xDD searchable digital library. This filters the results in the upper table to those that contain the search terms; the metadata table (Fig. 2) shows links to text snippets containing the search terms from these papers.

The initial version of the COVID-TIB dashboard was made publicly available on May 15, 2020. To date, the pipeline has processed over 270K documents. The dashboard currently displays information about nearly 3,000 different therapeutics involved in COVID-19 research found in nearly 26K paper abstracts, preprints, and clinical trials.

III. Use Cases and Evaluation Results

COVID-TIB can be used to find studies about a particular drug, to see the drugs that have been investigated for COVID-19 across research stages, and to find drugs investigated in the context of user-queried topics. To date, COVID-TIB has been used by researchers in gathering information for drug research studies and in report writing. Information from COVID-TIB has also been used to find papers more efficiently for a curation effort, for test cases for EMMAA models (https://emmaa.indra.bio/) of drug-molecular interactions, and for analyzing trends in COVID-19 drug research over time (see briefing for NIAID Data Science Seminar Series available on the COVID-TIB website).

In the BioCreative Track 4 evaluation, 11 reviewers with informatics or COVID-19 experimental research experience evaluated COVID-TIB by trying suggested use cases and then

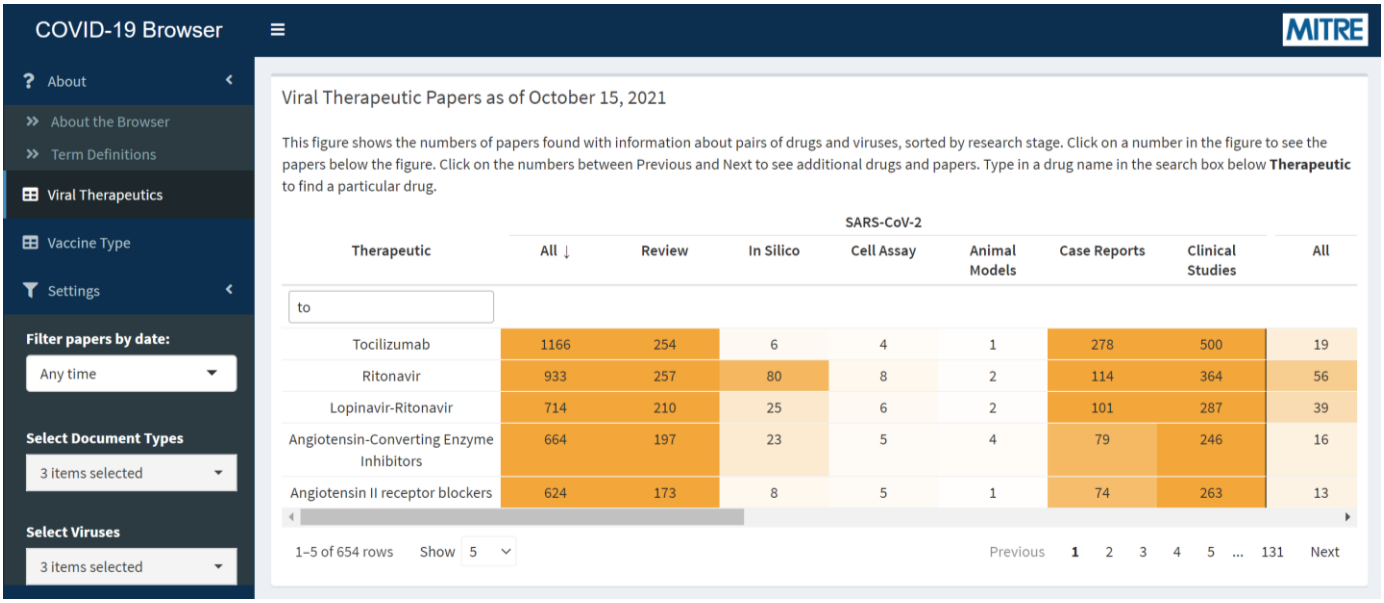

Fig. 1. Menu with filters and the Viral Therapeutics page, showing numbers of documents about therapeutics and research stages for SARS-CoV-2. Here the user has started to type in a drug name in the search box and sees drugs with "to" in the name.

Article ID	Title	Date	Journal Name	Research Stage	Search Snippets
33024963	LY-CoV555, a rapidly isolated potent neutralizing antibody, provides protection in a non-human primate model of SARS-CoV-2 infection.	2020-10-01	bioRxiv	Nonhuman Primate	See Snippets
33319649	Anti-SARS-CoV-2 neutralizing monoclonal antibodies: clinical pipeline.	2020-12-16	MAbs	Clinical Study	See Snippets
33359141	COVID-19 vaccines: The status and perspectives in delivery points of view.	2020-12-24	Adv Drug Deliv Rev	Review	See Snippets
33564771	The basis of a more contagious 501Y.V1 variant of SARS-COV-2.	2021-02-02	bioRxiv	Not Identified	See Snippets
33619479	501Y.V2 and 501Y.V3 variants of SARS-CoV-2 lose binding to Bamlanivimab in vitro.	2021-02-16	bioRxiv	Cell Line	See Snippets

Fig. 2. Document metadata table for Bamlanivimab with full text search for "variant" or "strain".

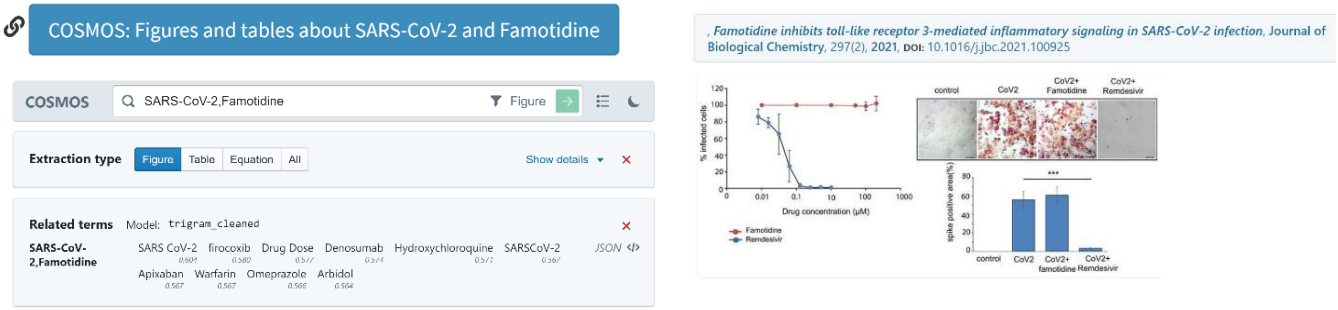

Fig. 3. Link out to COSMOS and xdd (https://xdd.wisc.edu/set_visualizer/sets/xdd-covid-19) to view figures and tables about a selected drug and SARS-CoV-2.

exploring COVID-TIB with their own questions. The reviewers then filled in a survey with open response and Likert scale questions. Nine of the 11 reviewers agreed with the statement "I thought the system was easy to use", and just one disagreed. Multiple reviewers commented that they liked that the system was easy to use, fast, and well-organized. However, three reviewers reported that some aspects did take at least a little time to figure out. One reviewer noted that the text snippets from full-text search were useful for assessing the relevance of articles. In response to the question, "How likely is it that you would recommend the system to a colleague performing COVID research?" eight reviewers responded positively (7 or higher on a 10-point Likert scale). When asked about applications for which the system could be useful, reviewers suggested the following: knowing the state of research for a type of vaccine, searching for information about bio-active molecules beyond COVID-19, finding information for systematic reviews, and clinical guideline development.

Multiple reviewers suggested that the system could be improved by adding a reset button to easily return to the original state after applying queries and filters. Two reviewers suggested improvements for finding drugs, including adding ability to search for synonyms and typo versions of the drug names in addition to the displayed names, and ability to sort the drug list alphabetically. A couple reviewers also suggested highlighting the selected number and/or corresponding result to make the connection between the user selection and the results more apparent. In addition, a few reviewers suggested improvements to the data available for download, including additional formats and larger batches of results in a single download.

Two comments suggest that comprehensiveness is a concern. One reviewer commented that the system could be very useful "if we trust that it is comprehensive." Another reviewer listed five drugs that they looked for and were unable to find. We corroborated that these drugs are not currently in our dictionary; they appear in PubChem and in COVID-19 abstracts in LitCovid (1) but are not at this time on the publicly available portion of the DrugBank website (https://go.drugbank.com/). We found these comments to be useful in directing us to specific areas for improvement.

IV. Discussion

Our experiences developing COVID-TIB have been valuable along several dimensions. First, we were able to stand up an end-to-end system that could be quickly retargeted to new applications without annotated data sets. Key ingredients were 1) availability of open, frequently updated document collections (PubMed, ClinicalTrials.gov, CORD-19, bioRXiv, MedRXiv, LitCovid) and tools to support document ingest and keyword search; 2) use of a light-weight rule-based NLP system, coupled with heuristic rules to assemble information at the abstract level; 3) reach-back to capability supporting full text search, with the ability to return snippets as well as tables and figures; and 4) creation of a flexible dashboard that provides an overview of the space, organized along dimensions relevant to the problem space, including ability to download results to a csv file.

This research has identified some key obstacles to address. First, there is a need for constant maintenance of rapidly evolving name spaces, such as for new drugs and for new virus nomenclature and synonyms. While we were able to make good use of existing resources (DrugBank; NCBI taxonomy), these are not complete, and regular updating is needed to keep pace with the emergence of new entities. Second, this research highlights the importance of keeping a resource like this up to date, including regular searches of new literature and, equally important, the need to keep terminology up to date. If users find that the system is incomplete or out of date, this will erode trust and therefore, the utility of the system.

Based on these experiences, one key area for further development is to semi-automate the extension of dictionaries and keyword lists, particularly the drug and virus dictionaries. Extension of the drug dictionary could be semi-automated by applying machine learning methods to identify candidate drug names in articles and by integrating information from multiple dictionaries; this would still require human review, but would facilitate keeping the drug list up to date, which is critical for ensuring the completeness of the resource

Overall, we have been greatly encouraged by the positive response to the COVID-TIB system. The existing framework can support a variety of use cases, such as conducting meta-analyses, writing reviews and synthesizing what is known, and detecting new trends. We are currently exploring other extensions and applications, taking advantage of the fact that the framework can be quickly retargeted to new applications.

Acknowledgment

We thank Eileen Chang, Lauren Quattrochi, Marta Bartlett, Tiffany Tsang, Sybil Russell, and Mihai Surdeanu for prior contributions and advice on this work. COVID-TIB development was supported by funding from the MITRE Innovation Program, the MITRE COVID-19 Coalition, and DARPA. This technical data deliverable was developed in part using contract funds under MITRE's Basic Contract No. W56KGU-18-D-0004. xDD and COSMOS work supported by DARPA ASKE HR00111990013. The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision.

References

1. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*, **49**, D1534-D1540.

2. Valenzuela-Escárcega,M.A., Babur,Ö., Hahn-Powell,G., Bell,D., Hicks,T., Noriega-Atala,E., Wang,X., Surdeanu,M., Demir,E. and Morrison C.T. (2018)

Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database*, **2018**, bay098.

3. 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. and Kohlmeier,S. (2020) CORD-19: The Covid-19 Open Research Dataset. *ArXiv*, 2004.10706v2.

4. Goswami,A., Bhat,A., Ohana,H. and Rekatsinas,T. (2020) Unsupervised relation extraction from language models using constrained cloze completion. *Findings of the Association for Computational Linguistics: EMNLP 2020*, **113**, 1263-1276.

5. Goswami,A., McGrath,J., Peters,S. and Rekatsinas,T. (2019) Fine-grained object detection over scientific document images with region embeddings. *ArXiv*, 1910.1246.