Overview of the COVID-19 Text Mining Tool Interactive Demo Track in BioCreative VII

Andrew Chatr-aryamontri[1], Lynette Hirschman[2], Karen E. Ross[3], Rose Oughtred[4], Martin Krallinger[5], Kara Dolinski[4], Mike Tyers[1], Tonia Korves[2], Cecilia N Arighi[6]

[1]Institute for Research in Immunology and Cancer (IRIC), University of Montreal, Montreal, QC, Canada. [2]MITRE Labs, The MITRE Corporation, Bedford, MA, USA. [3]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA. [4]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [5]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [6]Computer and Information Sciences Department, University of Delaware, Newark, DE, USA

Abstract-An **immediate consequence of the COVID-19 pandemic crisis was the necessity of communicating data in real time in order to provide information for the development of more effective medical treatments and public health policies. Unorthodox sources like preprint publications, articles not yet validated by peer review, became a crucial means of facilitating the dissemination of scientific findings. Several text mining systems were swiftly developed to support the retrieval and extraction of COVID-19 information or to organize the data in knowledge discovery systems. The BioCreative COVID-19 text mining tool interactive demo track, similar to the BioCreative InterActive Task (IAT), was created to gauge user-system compliance and establish a two-way communication channel between system developers and potential end users. The goal was to provide system designers of seven selected systems with useful feedback on the performance and usability of their tools and inform them of the need for additional features. Conversely to previous IAT editions, the exploratory nature of this track and the variety in scope of the competing tools meant no specific task was assigned and testers were not ad hoc matched to any specific system. More than 30 participants were involved in the task, covering a broad range of specialties including bench scientists, bioinformaticians, and biocurators. Users, who were given the opportunity to participate anonymously, were provided with video tutorials and documentation to evaluate the systems and were asked to complete a survey to formalize their evaluation. Additional feedback was also provided by system developers.**

URL:https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-4/

Keywords—text mining systems; usability; biocreative

I. INTRODUCTION

The emergence of COVID-19 as a global pandemic in early 2020 prompted intensive research efforts around the world. These pushes have yielded a wealth of new data about the biology of the SARS-CoV-2 coronavirus which causes COVID-19 and potential clinical treatments for this infectious disease. This crisis impacted how scientists share their work, placing greater emphasis on preprint publications, which are articles not yet validated by peer review, in order to speed up the dissemination of reported findings. More than 125,000 COVID-19 articles were published in 2020 of which over 30,000 were preprints, representing approximately 25% of the COVID-19 literature (1). Also, repositories for clinical trial registrations were not equipped to provide adequate support and functionalities to researchers and physicians during a global health emergency, making it extremely difficult for researchers to navigate and extract relevant information from the deluge of publications. In an unprecedented effort, a variety of bioinformatic resources were made available to support COVID-19 research, including fast genome and proteome annotation pipelines (NCBI, UniProt) (2-3) and tools for tracking and monitoring the spread of the disease (4). To support the biomedical community in their efforts and policy makers in setting up public health strategies, natural language processing (NLP) experts provided solutions for a wide range of tasks including the generation of COVID-19 specific corpora (5-6), with content mainly in English, but also in other languages (e.g., Spanish), along with the development of relevant information retrieval and extraction methods as well as knowledge discovery systems (7).

The main objective of BioCreative (Critical Assessment of Information Extraction in Biology) is to evaluate text mining approaches developed to address problems of importance in the biomedical field. Due to the multidisciplinary nature of the expertise involved in defining and solving these problems, the BioCreative InterActive Track (IAT) (8) was introduced to bring researchers together from different backgrounds such as NLP experts, bioinformaticians and bench scientists. A main aspect of the IAT is to allow two-way communication between system developers and a variety of end users. System functionalities are assessed based on agreed-upon standards and developers are provided with detailed feedback to plan further improvements and updates; end users can request new functionalities or propose different solutions for specific applications. With this in mind, the Text Mining Tool Interactive Demo for COVID-19-related tools was conceived as a text mining track analogous to the IAT. As the Interactive

Text Mining Track is a demonstration and collaborative task rather than a competitive evaluation, it allows for a more formal assessment of an assortment of resources based on text-mining applications that are devised to assist SARS-CoV-2 and COVID-19 research. The systems entered in the track varied widely in scope and implementation and covered a broad range of applications ranging from preprint aggregators to knowledge graph and reasoning system tools. The target audience was not limited to biomedical researchers but also included clinicians, pharmaceutical scientists, biocurators and policy makers as well as the public at large.

A crucial aspect of the Interactive Text Mining Track is the recruitment of appropriate testers which allows the tools to be exposed to a larger audience. In this context, participating teams were encouraged to provide some contacts to ensure that the target audience was represented in the group of users. Testers for the interactive task were recruited from bench scientists involved in COVID-19-related research, from the biocurator community with the goal of including curators across a broad range of expertise as well as clinicians working with COVID-19-related electronic health record (EHR) content and biomedical frontend text-mining tool developers. Curators volunteered from various databases, including those focused on chemical interactions, model organisms, and molecular interactions. The volunteers had relevant scientific knowledge of COVID-19 which allowed them to assess the interactive systems in relation to this area of focus. Each volunteer tested one or more text mining tools and provided feedback to the developers, thereby helping to improve the systems in terms of current projects such as capturing relevant data on model organisms, viral and viral-host protein interactions, and therapeutics in relation to COVID-19.

Our main goals in organizing this task were to provide developers with detailed user feedback about their interfaces, to expose users to new tools and to expand user adoption of text mining tools. In our experience, systems are not always developed with users in the loop and our hope is that the participation in the IAT task will help make the participants more aware of the community needs, and that any feedback received would serve to improve the user-system experience.

II. METHODS

A. Track design

The task was designed as an open exploration by end users who provided feedback about the text mining systems via a survey. Teams were invited to submit a document describing the proposed COVID-19 text mining system, including its main purpose and target user community, the tasks it could perform, data sources, interactivity features, and system performance metrics. Proposals were reviewed based on the relevance to COVID-19 research and the reported maturity of the system. Information for the track is available here: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-4/. Once approved, track organizers reviewed the systems, and the documentation and initial feedback were sent before advertising the track. In addition, teams were asked to provide a two-minute video describing the system for user recruitment purposes and a tutorial explaining the system's functionalities via examples. This information was included in the track page for users: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-4-users/. This user page contained all the information needed to start the activity (**Fig. 1**). For each system reviewed, users were asked to complete a guided activity and exploratory activities. The guided activity consisted of performing activities in a tutorial/guide proposed by the developers mostly to show the system capabilities. After this activity, users were asked to navigate the system and try their own examples, then report their findings via a survey.

The survey was anonymous and consisted of three sections. The first section was designed to learn about the user background: involvement in COVID-19-related research, type of institution (e.g., academic, pharma, healthcare), and their role (e.g., biocuration, clinical, experimental). The second section was designed to get feedback on their experience with the system through a set of open-ended questions. In the last section, we included the System Usability Scale (SUS), a post-test questionnaire containing 10 different questions that provides helpful information about a user's takeaways and overall experience (9). In this survey, the odd-numbered questions ask the user to agree/disagree with some positive aspect (e.g., I thought this system was easy to use), whereas even-numbered ones are about negative ones (e.g., I found this system unnecessarily complex). To calculate the SUS score, we first summed the score contributions from each item. Each question's score contribution ranged from 0 to 4. For odd-numbered questions the score contribution was the scale position minus 1. For even questions, the contribution was 5 minus the scale position. The sum of the scores was multiplied by 2.5 to obtain the overall value of SU. SUS scores have a range from 0 to 100 with 68 being the 50th percentile. Finally, we included a question about the overall impression of the system and another about the system meeting user expectations in the form of a Likert-scale response ranging from 1 (worst) to 5 (best). For each system, we calculated the percentage of users assigning scores <3 (considered negative), >3 (considered positive), and those selecting 3 (neutral). To display user feedback on aspects they liked or felt needed improvement, a word cloud generator was used (template used from https://venngage.com).

A separate survey was created to get feedback about the developer team's participation experience in this Track. One of the questions was whether teams obtained useful feedback from the users. This questions was in the form of a Likert-scale response ranging from 1 (worst) to 5 (best), and scores <3, >3 and =3 were considered as negative, positive or neutral, respectively. The user and the team's surveys were created using Google Forms.

To recruit users, we disseminated the information using Twitter, mailing lists (e.g., societies, university departments, and pharma), contacting the users suggested by the system developers, and also researchers found via a literature search.

Additional users were recruited from various scientific databases as well as through personal contacts.

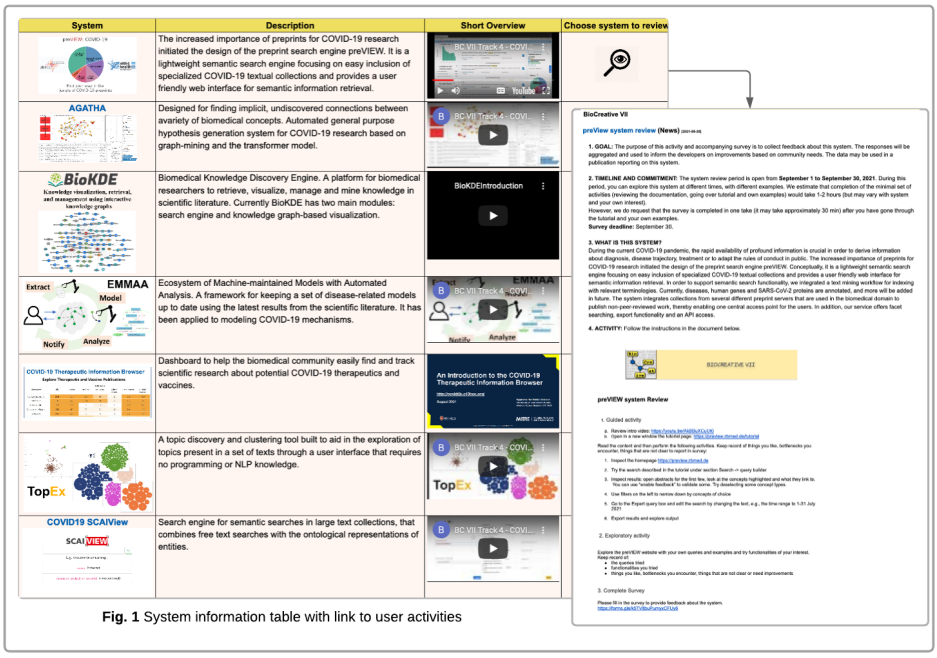

Fig. 1 System information table with link to user activities

III. Results and Discussion

This version of the interactive track differed significantly from previous ones in multiple aspects: the activity was not centered around a biocuration task, there was no human in the loop during development of the system in the interactive activity (except the feedback from organizers before exposure to users), the user interaction with the system was short (one to a few hours), and surveys were anonymous unless the user wanted to provide their information. As mentioned before, the purpose was to provide a first impression on some of the systems that were developed in response to the COVID-19 pandemic.

A. Summary of participating systems

Seven systems participated in this track. Although the common theme is COVID-19, the systems are very different. They vary in the type of tasks they perform (e.g., information retrieval, named entity recognition (NER), relation extraction, topic modeling, **Fig. 2, top**), their general purpose (e.g., search engine, knowledge graph, hypothesis generation); and type of text inputs (e.g., abstracts, preprints, clinical trials, tweets, **Fig. 2, bottom**). Most of the COVID-specific systems presented are specialized versions of systems designed for biomedical literature more generally.

preVIEW COVID-19 (10), COVID-19 SCAIview, and Therapeutic Information Browser (11) are primarily information retrieval systems. preVIEW and SCAIview highlight detected entities in the search results display. preVIEW COVID-19 detects diseases, human genes and proteins, and SARS-CoV-2 proteins. COVID-19 SCAIview normalizes entities to a variety of ontologies, including a custom COVID-19 ontology. The Therapeutic Information Browser uses rule-based NLP to identify abstracts and clinical trial summaries about drugs and selected viruses. Documents retrieved are classified according to the type of study (e.g., cell-based, animal, or clinical). From the system website, users can browse, search, and filter the results.

BioKDE (12) combines information retrieval with graph-based representation of search results. Scientific articles or patents can be searched by keyword using a custom search engine. Documents retrieved are subjected to deep learning based NER. Entities detected are represented as nodes in a knowledge graph; graph edges are based on co-occurrence.

AGATHA-C (13) and EMMAA COVID-19 (14) are primarily graph-based hypothesis generation systems. The original AGATHA system finds connections between biomedical concepts in general; AGATHA-C focuses on a subset of concepts relevant for COVID-19. It builds a semantic graph based on text and annotations (e.g., MeSH terms) from the literature. Embeddings are obtained for nodes in the graph and used to score the plausibility that two terms are connected. Topic modeling of sentences in the neighborhood of the connected terms is used to provide possible explanations for the connections, which are displayed in the AGATHA Visualizer. EMMAA COVID-19 performs NER and relation extraction on articles from the literature related to COVID-19 and assembles the information into statements. The statements are organized

into a knowledge graph. Mechanistic explanations for connections between concepts can be derived by following paths that connect the concepts in the network.

TopEx (15) is a topic modeling-based system that works on text corpora uploaded by the user. It makes vector representations of sentences from the corpus, clusters similar sentences, and identifies common topics in the clusters. It can be used to classify documents to facilitate information retrieval, and it also provides a graphical view of relationships among topics, enabling analyses such as trends in topics over time.

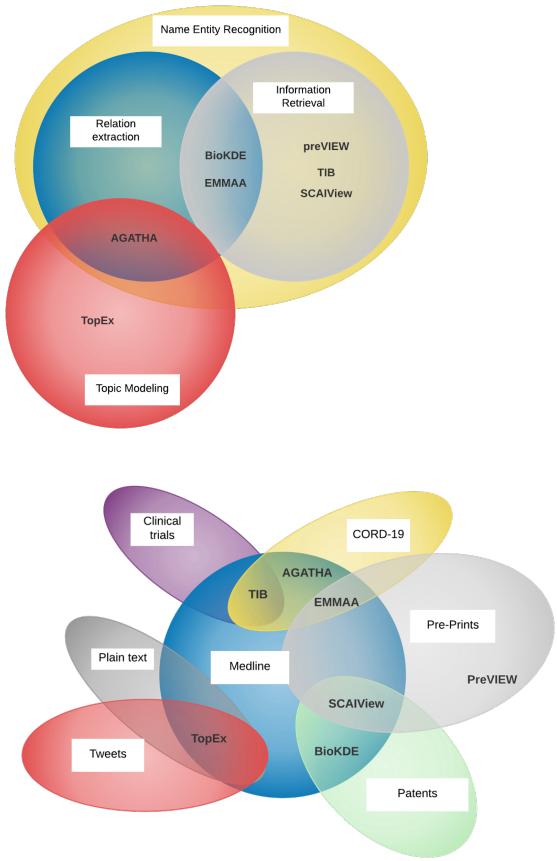

Fig. 2 Venn diagrams showing diversity of the tasks (top) and sources of text data for the participating systems (bottom).

B. Summary of users

All of the systems were developed with similar target user communities in mind: biomedical researchers, translational researchers, and clinicians with an interest in COVID-19. Some of the development teams also identified government agencies as potential users. For example, the Therapeutic Information Browser (TIB) team identified funding agency decision makers working on COVID-19 therapeutics, and the COVID-19 SCAIview team identified organizations, such as the WHO Pandemic Hub and the COVID-19 Data Portal. Overall, the systems were designed to be used by people beyond informatics or NLP experts. We were able to recruit 8-12 users per system (note that a user could sign up for multiple systems). The majority of recruited users were from academia, with a significant proportion working in biocuration and informatics domains. Approximately 60-75% of users worked in research related to COVID-19, except for testers of AGATHA, for whom only around 30% had coronavirus-related knowledge.

C. Survey highlights

Fig. 3 shows word clouds for each system highlighting some of the aspects the users liked most about the systems. Users like intuitive and easy to use systems, and also ones with familiar functionalities similar to other systems they use (e.g., filtering in PubMed). They also highlighted unique aspects of the systems they liked such as the comprehensive sources for preprints in preVIEW, highlighting concepts in semantic search engines like SCAIVIEW, filters and organization of data in TIB, graph capabilities and interactivity in BioKDE and EMMAA, topic paths in AGATHA and finding trends in TopEX.

It should be noted that the small number of users per system limits the analysis of the data. **Fig. 4** shows aggregated results from the Likert-based questions: the SUS plot (A), the Overall Impression (B) and "Met Expectations" (C) plots are a summary of the results from the users. By no means is there any claim of statistical significance, only a glimpse of first impressions from the users who tested them. Although a low score could be indicative of some issue with design that needs to be researched, it does not provide insight into the specific issue. Thus, the explanations provided by the users in free text form, like bottlenecks they encountered, are the most informative for the teams to improve their systems. Finally, one of the questions asked whether the users knew about any similar system before participating in BioCreative. It was encouraging to find that some of the users found the system they reviewed similar or better than the ones they used.

D. Challenges in this interactive track

- Diversity of the systems both in the type of task and technologies: It is interesting to note that the technologies observed in a set of the systems that participated this time (e.g., topic modeling, knowledge graphs) are not the traditional ones developed under BioCreative challenges and include approaches that do not require labeled data. Because of the diversity of systems we cannot directly compare across systems.
- Recruitment of users outside the biocuration domain: We asked developer teams about the number and background of users recruited for their systems. Five of the seven systems were satisfied with both, whereas teams from two systems would have liked more users with different backgrounds. Although we reached out to various research groups and pharma, the majority of participants were from the biocuration community. The question remains as to whether these other types of users are not interested in the tools or whether BioCreative needs to think of alternative strategies to engage these target communities. Regardless, biocurators work in a variety of settings and with

various types of data (literature, clinical, genomics) and are therefore well equipped to review the systems.

- Limitations in the analysis of results: Given the small number of users per system, the data are informative for first impressions but do not provide statistical significance. However, all teams indicated they obtained valuable feedback from users, i.e., all teams scored the question on useful feedback >3.

E. Suggestions from the developer teams

There were also several improvements suggested by the development teams:

- The developer teams appreciated the round of system review and feedback by the organizers prior to exposure to users; this helped to improve the systems.
  The teams would have liked a similar opportunity after receiving user feedback.
- Unexpectedly, the teams wanted their systems to be compared with other teams' systems and suggested having a common task for comparison purposes.

F. Value of interactive track to the developer teams

The development teams noted a number of positive outcomes of participating in the Track:

- Users with different types of expertise could provide useful feedback based on real-world projects.
- Tool developers could be connected with a broad field of potential users via a forum that has structured and consistent testing requirements (e.g. guided and open searches), thereby offering an easier and more efficient testing process with defined timelines.
- It facilitated an iterative process involving user-based testing of the text mining tools, which resulted in useful feedback for developers who could then optimize the tools.
- Users could suggest suggest additional improvements to the interface and functionalities so the tools could be tailored for specific scientific projects as well as for general search options.
- It required the developers to create clear written and visual documentation and think like a user.

G. Moving forward

Is organizing the IAT worth the effort? The overall positive feedback from the participating teams and users would motivate us to continue offering this track. Organizing the track is very resource-intensive and time-consuming so it is important to ensure that the output is maximally informative. Some changes need to take place to make this effort more "measurable" in terms of success with many questions to consider for the next round: How do we measure the success of this activity -- possibly by the number of modifications introduced due to feedback from users, or by an increased number of users or tool adoption over time? How can we provide a task that can accommodate very different systems yet still provide a basis for comparison? What are some possible strategies to engage a sufficient number of users for the tasks such that the results can be statistically significant in order to support a deeper analysis? What about sustainability of the systems that participate? Are they one-time experiments or will they be maintained going forward for the community to use? We hope to get some insight from participants on these aspects during the BioCreative workshop.

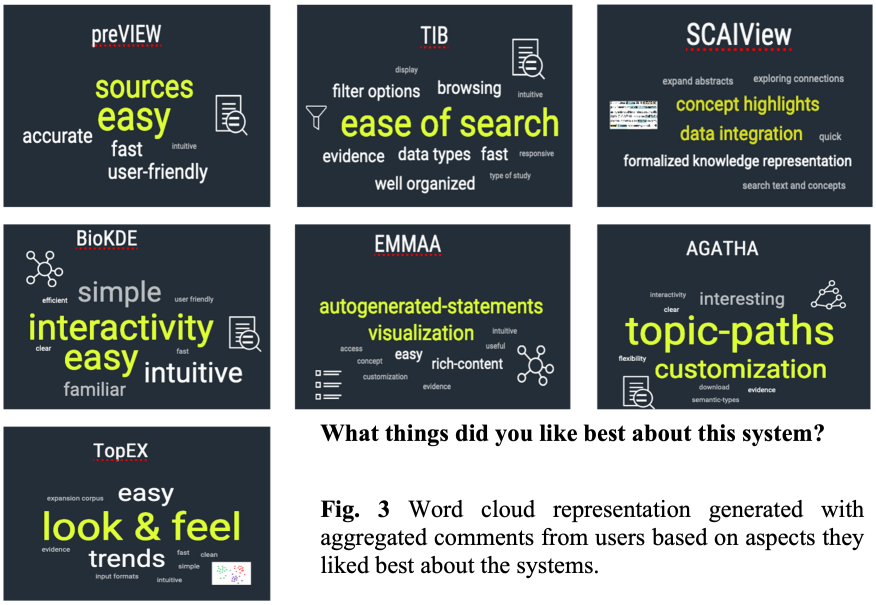

Fig. 3 Word cloud representation generated with aggregated comments from users based on aspects they liked best about the systems.

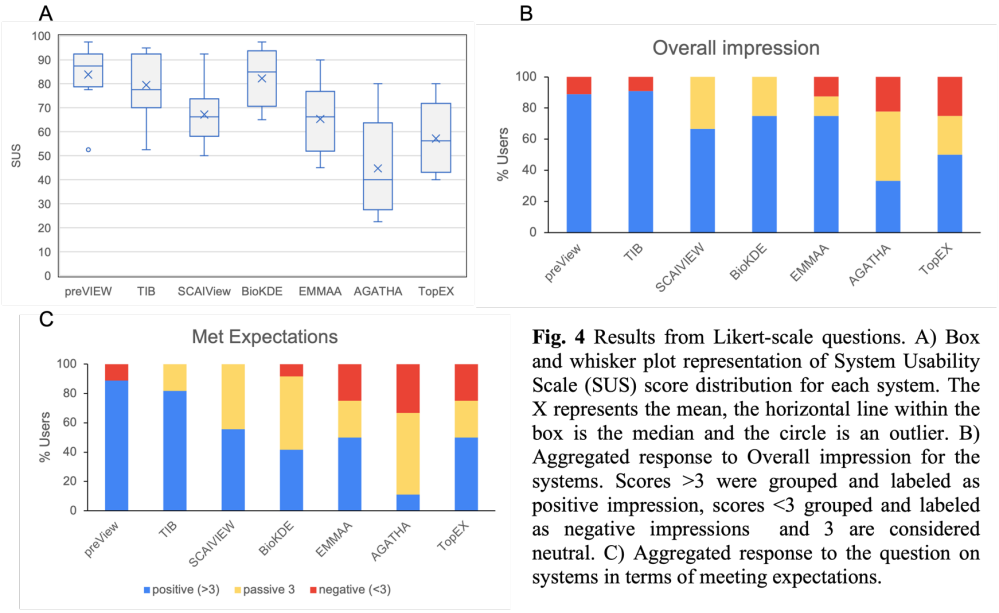

Fig. 4 Results from Likert-scale questions. A) Box and whisker plot representation of System Usability Scale (SUS) score distribution for each system. The X represents the mean, the horizontal line within the box is the median and the circle is an outlier. B) Aggregated response to Overall impression for the systems. Scores >3 were grouped and labeled as positive impression, scores <3 grouped and labeled as negative impressions and 3 are considered neutral. C) Aggregated response to the question on systems in terms of meeting expectations.

ACKNOWLEDGMENT

We would like to acknowledge the teams who participated in Track 4 and the users who took the time to review the systems and provide valuable feedback. We also would like to acknowledge the ISB and NLM who have done a great job in disseminating our track and helping us recruit participants. This work was supported by the National Institutes of Health Office of Research Infrastructure Programs [R01OD010929 to MT and KD]; the Canadian Institutes of Health Research [FDN-167277 to MT]; a Genome Canada and Genome Quebec Genomics Technology Platform Award [to MT and P. Thibault]; and a Canada Research Chair in Systems and Synthetic Biology [to MT]. This work has been partially supported by the National Institutes of Health grants 2U24HG007822-08 and 1R35 GM141873-01 [to KER and CNA]; the Spanish Plan for the Advancement of Language Technology (Plan TL) and Proyectos I+D+i 2020 - AI4PROFHEALTH (PID2020-119266RA-I00) [to MK]; and under Basic Contract No. W56KGU-18-D-0004 to MITRE [LH and TK]. 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. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biology*. 2021 Apr;19(4):e3000959.
2. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. 2021 Jan 8;49(D1):D480–9.
3. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 2018 Jan 4;46(D1):D8–13.
4. Hufsky F, Lamkiewicz K, Almeida A, Aouacheria A, Arighi C, Bateman A, et al. Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. *Briefings in Bioinformatics*. 2021 Mar 22;22(2):642–63.
5. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv. 2020 Apr 22;
6. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res. 2021 Jan 8;49(D1):D1534–40.
7. Wang LL, Lo K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*. 2021 Mar 22;22(2):781–99.
8. Arighi CN, Roberts PM, Agarwal S, Bhattacharya S, Cesareni G, Chatr-Aryamontri A, et al. BioCreative III interactive task: an overview. *BMC Bioinformatics*. 2011 Oct;12 Suppl 8(Suppl 8):S4.
9. Brooke J. "SUS-A quick and dirty usability scale." Usability evaluation in industry [Internet]. CRC Press; 1996. Available from: https://www.crcpress.com/product/isbn/9780748404605
10. Darms J, Langnickel L, Fluck J. Semantic Search Engine preVIEW COVID-19 - Evaluation in the BioCreative VII IAT Track. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop. 2021*.
11. Korves T, Hirschman L, Garay C, Kozierok R, Peters S, Peterson M, et al. The COVID-19 Therapeutic Information Browser. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. 2021.
12. Chung M-H, Zhou J, Pang X, Tao Y, Zhang J. BioKDE: a Deep Learning Powered Search Engine and Biomedical Knowledge Discovery Platform. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. 2021.
13. Tyagin I, Safro I. Interpretable Visualization of Scientific Hypotheses in Literature-based Discovery. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. 2021.
14. Gyori B, Bachman J, Kolusheva D. A self-updating causal model of COVID-19 mechanisms built from the scientific literature. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. 2021.
15. Olex A, French E, Burdette P, Sagijaru S, Neumann T, Gal T, et al. TopEx: Topic Exploration of COVID-19 Corpora - Results from the Biocreative VII Challenge Track 4. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. 2021.