A self-updating causal model of COVID-19 mechanisms built from the scientific literature

Benjamin M. Gyori*, John A. Bachman, Diana Kolusheva
Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA
*Corresponding author: benjamin_gyori@hms.harvard.edu

***Abstract*—**With the emergence of COVID-19, scientific publications about the disease and the SARS-CoV-2 virus causing it have exploded and continue to appear at a rate of more than 300 new papers each day. We developed a self-updating model of COVID-19 mechanisms within the Ecosystem of Machine-maintained Models with Automated Analysis (EMMAA, emmaa.indra.bio), a framework for keeping a set of disease-related models up to date using the latest results from the scientific literature. The model integrates causal and mechanistic relations extracted by multiple text mining systems from literature relevant to COVID-19, and through a process of knowledge assembly, integrates new findings into the model from novel publications each day. The model is also subjected to causal path-based analysis to systematically explain drug effects on viruses, and is available for interactive querying by users. The EMMAA COVID-19 model is available at https://emmaa.indra.bio/dashboard/covid19.

Keywords*—systems biology; modeling; text mining; knowledge assembly; COVID-19*

I. Introduction

There are over 1.2 million publications appearing in biomedicine per year, or around 3,200 each day, more than any human scientist or clinician could keep track of. The effective and timely response to global health crises such as COVID-19 crucially depends on being able to make use of knowledge embedded in past literature combined with findings reported in new publications as they appear. Publications on COVID-19 and the SARS-CoV-2 virus causing it have exploded since early 2020 and continue to appear at a rate of more than 300 new papers each day. Automated text mining is a scalable way to process the content of new publications and extract concepts and relationships in a structured format. However, while text mining can extract large amounts of findings from text, it does not address the problem of (i) resolving relationships between extracted findings to detect novelty and (ii) determining, in a principled way, whether the new findings have a meaningful effect on our understanding of the underlying system or on explaining empirical observations. In other words, text mining has to be coupled to modeling and analysis workflows and interactive user-facing interfaces to produce actionable insight.

We have developed the Ecosystem of Machine-maintained Models with Automated Analysis (emmaa.indra.bio), a framework for keeping a set of disease-related models up to date using the latest results from the scientific literature, and applied it to modeling COVID-19 mechanisms. The key challenge our system aims to tackle is automatically monitoring and interpreting publications relevant for COVID-19, and assembling a causal model of the information extracted from them to turn published knowledge into an actionable form, opening it up for data analysis and user interaction. We expect our system will be of interest to biologists studying molecular mechanisms underlying COVID-19, as well as translational scientists or clinicians aiming to understand drug mechanisms and identify potential repurposing opportunities.

II. Results

A. A self-updating COVID-19 model based on text mining and knowledge assembly

The EMMAA COVID-19 model taps into multiple literature sources to find relevant publications. First, it integrates the CORD-19 corpus [1] which contains a mixture of prior literature on coronaviruses and newly published COVID-19 research, providing full-text access whenever possible. In addition, EMMAA searches PubMed each day to identify the latest publications using "COVID-19" or "SARS-CoV-2" as keywords, and downloads the underlying full text content or abstract, depending on availability. Finally, EMMAA is integrated with the xDD framework (https://xdd.wisc.edu/) through which it obtains the latest bioRxiv and medRxiv pre-prints. The EMMAA COVID-19 model has overall processed around 800 thousand publications and typically processes between 300-600 new publications a day.

To process literature content, EMMAA invokes multiple text mining systems via the Integrated Network and Dynamical Reasoning Assembler (INDRA) [2]. The Reach [3], Sparser [4], and TRIPS/DRUM [5] systems define biology-specific extraction rules over dependency- or semantic parses of sentences to reconstruct causal relations with mechanistic detail (protein modification sites, activity conditions, etc.) described in text. Each system also implements named entity recognition and normalization of biomedical entities to a suite of biomedical ontologies which we have adapted to capture relevant novel terms for studying COVID-19 (e.g., synonyms for SARS-CoV-2 polyprotein fragments). We also integrated the Eidos system [6], a general purpose causal relation extraction system, which we coupled to Gilda [7] to produce relations between high-level concepts (complementary to the previously described systems which focus on molecular mechanisms) grounded to biomedical

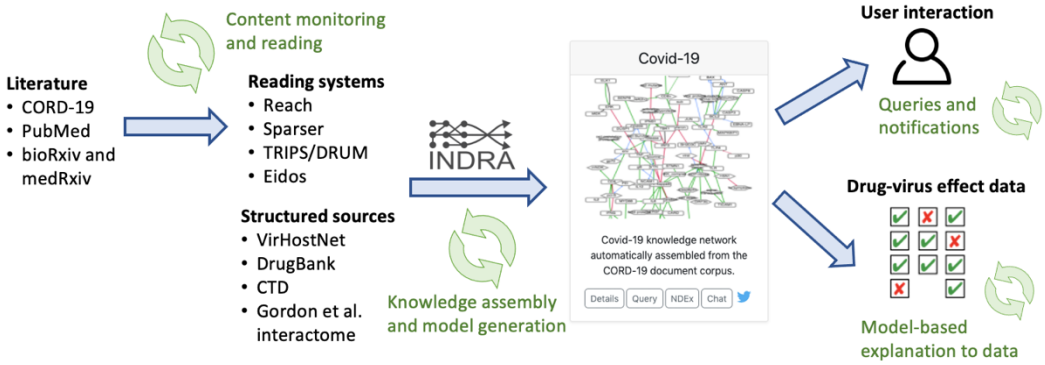

Figure 1. The EMMAA COVID-19 model workflow. Literature content is processed with machine reading systems and combined with structured sources in a process of knowledge assembly. Networks models derived from the assembled knowledge are used to construct explanations for drug-virus effects and to respond to user queries. Users also receive notifications about relevant updates. Parts of the modeling workflow that are self-updating on a daily bases are shown with green cycle arrows.

ontologies. The EMMAA COVID-19 model also combines mechanisms from relevant structured sources including VirHostNet [8], DrugBank [9] and the Comparative Toxicogenomics Database [10]. The model also integrates interactions between SARS-CoV-2 proteins and proteins of human host cells made available in Gordon et al. [11]. This workflow is summarized in Fig. 1.

Extractions from the above sources are normalized to INDRA Statements which can represent a variety of molecular mechanisms (complex formation, phosphorylation, metabolic conversion, amount regulation, etc.) and more abstract causal relations (e.g., activation, inhibition), as described in Gyori et al. [2]. Each Statement contains one or more Agents as arguments which represent normalized entities (i.e., are identified as specific entries in biomedical ontologies). Each Statement is also associated with a set of Evidences which describe the provenance (extraction system, source publication identifiers, evidence sentence, etc.) supporting the Statement.

Statements collected from knowledge sources are subjected to a knowledge-assembly pipeline which standardizes Agent identifiers using a graph of cross-references between different ontologies (crucial, since different sources ground to inconsistent ontologies), and finds exact and partial redundancies between Statements. Here, an exact redundancy means that two Statements (extracted from two different sentences/publications or extracted from the same sentence by multiple reading systems) represent equivalent relationships between Agents. Groups of such equivalent Statements are merged during assembly into a single Statement and their Evidences are aggregated in a list. Partially redundant statements, which, for instance, represent the same relationship at different levels of detail (e.g., one involving the gene MAPK1 and the other, the protein family ERK, one of whose members is MAPK1) are recognized and are exploited when determining the support for a given Statement. During assembly, INDRA also calculates a belief score based on the overall evidence supporting a statement using a probabilistic model of the random and systematic error characteristics of each knowledge source.

Using the above logic, new Statements extracted from the body of COVID-19 literature on a given day are aligned with the existing assembled Statement set to determine what new knowledge has been obtained. A new Statement can be (i) equivalent to an existing Statement, (ii) partially related to an existing statement, or (iii) novel with no equivalent Statement having been reported before. Thus, the EMMAA framework allows for determining which newly reported Statements in the context of the COVID-19 literature represent meaningfully novel knowledge.

After each daily update, a new version of the assembled EMMAA COVID-19 model is made available. As of 10/10/2021, the model represents over 430 thousand unique Statements. Fig. 2 shows the number of assembled Statements in the model over time between March 2020 and October 2021, with each dot representing the model's Statement count after a daily update.

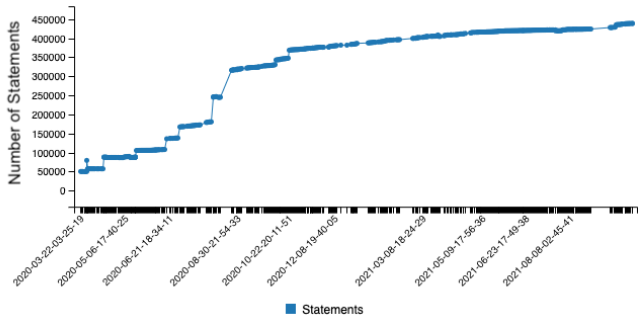

Figure 2. Number of Statements over time constituting the EMMAA COVID-19 model.

Statements in the EMMAA COVID-19 model represent relations between a variety of types of entities including human and non-human proteins (ant their families, complexes, chains and fragments), small molecules (including drugs and metabolites), biological processes, diseases, and organisms (such as different types of viruses). Table 1 summarizes the number of unique grounded entities (represented as Agent arguments of Statements) by type in the model.

TABLE 1. ENTITY TYPES REPRESENTED IN THE EMMAA COVID-19 MODEL

Entity type	# entities
Human protein	11,137
Small molecule	7,836
Non-human protein	5,654
Disease	3,933
Biological process	2,776

Experimental factor	1,361
Organism	796
Anatomical region	679
Human protein family/complex	548
Human RNA	297
Cellular location	243
Non-human protein chain/fragment	23
Human protein chain/fragment	19
Other	1,701

*as of 11/10/2021

The EMMAA dashboard allows exploring and curating (marking as correct or incorrect) the Statements contained in the COVID-19 model, with each Statement shown in the context of the specific sentences from which it was derived, and linked back to the underlying literature.

B. Automated model analysis and validation against empirical observations

The assembled Statements constituting the EMMAA COVID-19 model can be further generated into causal graph structures that allow efficient path finding subject to constraints. This approach can be used to construct hypotheses of mechanistic paths by which, for instance, a given drug might be beneficial against COVID-19.

We curated a set of publications on *in vitro* screening for drugs that inhibit coronavirus replication to identify 165 observations of the form "Mefloquine inhibits SARS-CoV-2 replication". We also imported 2,573 observations of similar form compiled in the MITRE COVID-19 Therapeutic Information Browser (https://covidtib.c19hcc.org) about the effectiveness of certain drugs against a set of coronaviruses. For each such observation, two graph exports (signed and unsigned) of the EMMAA COVID-19 model are used to find paths between the drug perturbation and the readout - subject to a set of semantic constraints - to construct mechanistic hypotheses. An example path is shown in Fig. 3 which hypothesizes SMPD1 as an intermediate in the effect of the drug toremifene on SARS-CoV-2 replication.

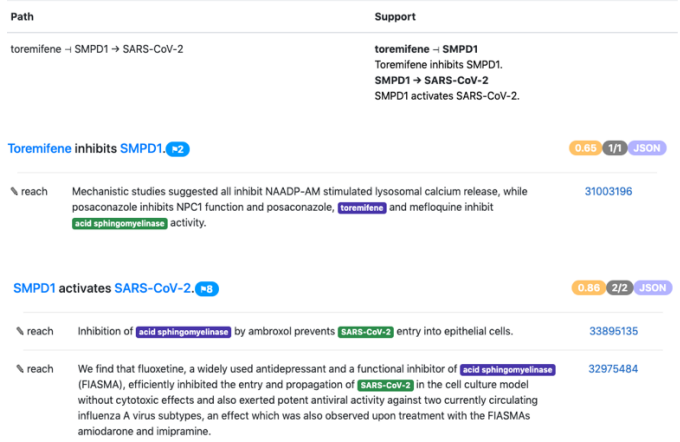

Figure 3. An automatically constructed causal path explanation for the observation that "Toremifene inhibits SARS-CoV-2" showing SMPD1 as an intermediate. Each link in the causal path is supported by textual evidence from publications.

With each daily model update the set of paths can change, for instance, a newly published discovery might allow EMMAA to explain a new drug-readout observation. As of 10/10/2021, the COVID-19 EMMAA model using a signed graph export could provide explanations for 2,262 of these observations. Users can browse these results on the EMMAA dashboard with the ability to drill down into specific evidences supporting each link in an explanation path, and curate any incorrect ones.

C. User interaction and notifications

Users can interact with the EMMAA COVID-19 model through the EMMAA dashboard at https://emmaa.indra.bio which links to the model-specific page at https://emmaa.indra.bio/dashboard/covid19. A tutorial at https://emmaa.readthedocs.io/en/latest/tutorial/index.html demonstrates interacting with the model through a walkthrough example.

The COVID-19 model page is divided into multiple tabs each providing access to a different facet of the model. The page provides an overview of the number and types of Statements in the model over time (e.g., the Statement representing the fact that ACE2 binds SARS-CoV-2), the most frequently appearing Agents (e.g., "inflammatory response" representing the Gene Ontology term GO:0006954), and allows browsing (i) new Statements added in the last update as well as (ii) all Statements in the model filtered and sorted by different criteria. Importantly, the dashboard allows curating (i.e., marking as correct or incorrect) Statements. Fig. 4 shows a Statement in the model representing that Dexamethasone (CHEBI:41879) inhibits IL6 (HGNC:6018) with some of its supporting evidences, and also exposing the curation interface.

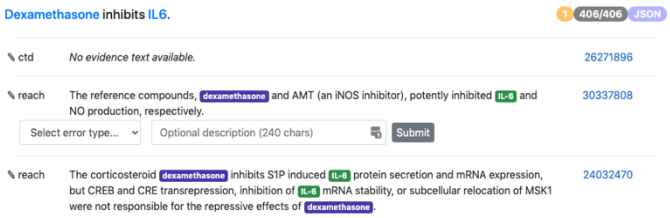

Figure 4. An example Statement in the EMMAA COVID-19 model with three (out of 406) of its evidences shown. The names of the two Agents link out to respective outside pages describing them on the ChEBI and HGNC websites. On the right, the gray badge shows the overall number of evidences for the Statement, and the orange badge shows the belief score, and the JSON badge allows downloading the Statement and its evidences in a machine-readable JSON format. Each row under the header represents a distinct Evidence supporting the Statement, including the knowledge source (here CTD or Reach), the evidence sentence (if available), and the source publication's identifier, linking out to PubMed or another appropriate source. Clicking on the pencil icon on the left reveals the curation form where the given Evidence can be marked as correctly supporting or incorrectly being associated with the given Statement.

The dashboard also provides an interface to explore explanations generated by the COVID-19 model (as described in Section II/B), including the number of "test" statements (i.e., experimental observations) explained over time, the specific causal path for each explanation, and the underlying evidence. Further tabs provide a window into the model from the perspective of specific Agents or specific publications.

A key feature of EMMAA is its support for user queries against models (available at https://emmaa.indra.bio/query). Through a structured query interface, users can run queries representing questions such as "how does losartan affect ACE2?" or "what are the upstream small-molecule inhibitors of

This work was funded under the DARPA Automating Scientific Knowledge Extraction program, ARO grant HR00111990009.

TMPRSS2?" against the COVID-19 model. EMMAA generates mechanistic paths to answer these queries, with each relation in the path annotated with the underlying literature or database evidence. Users can register specific queries to get daily email updates on any new results pertaining to their queries, they can also subscribe to general updates from the model and be notified about all new Statements and model-derived explanations.

The EMMAA COVID-19 model is also integrated with Twitter (https://twitter.com/covid19_emmaa) where it provides updates about its progress with links to relevant pages on the EMMAA dashboard. With each daily update, it summarizes the number of papers it has read and the number of meaningfully new statements (i.e., ones not matching any previously known ones) added to the model from these. It also tweets about new explanations it was able to construct against one of the drug-readout observations described in Section II/B, linking to the page where users can examine and curate the new explanations.

Further, we integrated a human-machine dialogue feature with EMMAA which allows for exploring the assembled knowledge in an intuitive way, using natural language dialogue. A key feature of dialogue is that the machine maintains context and can interpret references to previously mentioned results, thereby allowing sequential exploration of knowledge (Fig. 5).

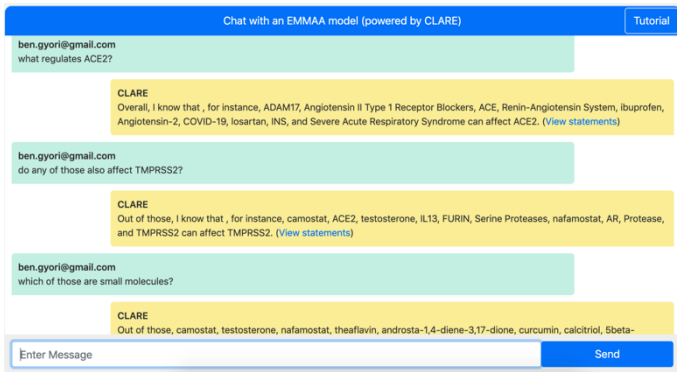

Figure 5. Human-machine dialogue to explore the EMMAA COVID-19 model.

Finally, the COVID-19 EMMAA model can be downloaded in a JSON format for downstream processing and integration, and is also available on NDEx (ndexbio.org) [12] for network-oriented browsing.

D. User feedback

As part of BioCreative VII, the EMMAA COVID-19 model was tested by a total of eight expert users from academic and governmental institutions. Users completed a tutorial and then used the system to answer scientific questions. User surveys were then collected by the BioCreative VII organizers and shared with participant teams.

In their feedback, users pointed to several applications EMMAA could be useful for: knowledge monitoring in future pandemics, building disease models, helping identify papers and interactions to curate, and aiding hypothesis testing and drug discovery. Among the things they liked most about the system, users highlighted the overall concept of self-updating models, the fact that text mining is integrated with modeling and analysis in a single framework, and the ability to see the results of automated analysis and user queries in a way that is linked to supporting literature.

Users made a number of useful suggestions for improvement ranging from conceptual to technical aspects of the system. For instance, users pointed out some interaction types (e.g., palmitoylation of viral proteins) and named entities (e.g., FT-IR) not being recognized correctly, suggesting potential improvements to the text mining systems feeding into the EMMAA COVID-19 model. Users also suggested further speeding up the loading of various statistics and displays (e.g., the page displaying all statements in the model). Finally, users pointed out the usefulness of extending the documentation and tutorials to cover more details.

Though satisfaction among users varied, when asked "Please rate your overall impression with the system" (on a scale of 1-5, 1 being very negative and 5 being very positive) six of the eight users gave a score of at least 4. Similarly, to the question "How likely is it that you would recommend this system to a colleague performing COVID-19 related research?" (on a scale of 1-10 with 1 being not at all likely and 10 being extremely likely), four of the eight users gave a score of 9 with two further users giving a score of 7 and 8, respectively.

III. Conclusion

Here we presented the EMMAA COVID-19 model which is automatically built and updated daily from newly published literature. The model incorporates text mined extractions by four different machine reading systems from a total of around 800 thousand relevant publications, and combines these with statements from structured sources through a process of knowledge assembly provided by the INDRA system. The model is then applied automatically to explain a set of experimentally observed drug-virus relationships using causal path finding. Users can sign up for notifications about general model updates and specific new results for queries they register through the EMMAA dashboard. This constitutes a novel approach to monitoring the COVID-19 scientific literature that goes beyond what typical search or recommendation engines offer: it alerts scientists about specific new discoveries being reported which - in the context of a model of prior knowledge - can provide a meaningfully new answer to a scientific question they are interested in.

References

1. L. L. Wang et al., "CORD-19: The COVID-19 Open Research Dataset," ArXiv200410706 Cs, Jul. 2020, Accessed: Jun. 07, 2021. [Online]. Available: http://arxiv.org/abs/2004.10706

2. B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger, "From word models to executable models of signaling networks using automated assembly," Mol. Syst. Biol., vol. 13, no. 11, p. 954, 24 2017

3. M. A. Valenzuela-Escárcega et al., "Large-scale automated machine reading discovers new cancer-driving mechanisms," Database J. Biol. Databases Curation, vol. 2018, Jan. 2018, doi: 10.1093/database/bay098.

4. D. D. McDonald, S. E. Friedman, A. Paullada, R. Bobrow, and M. H. Burstein, "Extending Biology Models with Deep NLP over Scientific Articles.," 2016.

5. J. F. Allen, W. de Beaumont, L. Galescu, and C. M. Teng, "Complex Event Extraction using DRUM," 2015. doi: 10.18653/v1/W15-3801.

6. R. Sharp et al., “Eidos, INDRA, & Delphi: From Free Text to Executable Causal Models,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, Minnesota, Jun. 2019, pp. 42–47. doi: 10.18653/v1/N19-4008.

7. B. M. Gyori, C. T. Hoyt, A. Steppi, “Gilda: biomedical entity text normalization with machine-learned disambiguation as a service,” bioRxiv preprint, doi: 10.1101/2021.09.10.459803, 2021

8. Guirimand, Thibaut, Stéphane Delmotte, and Vincent Navratil. "VirHostNet 2.0: surfing on the web of virus/host molecular interactions data." Nucleic acids research 43.D1 (2015): D583-D587.

9. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research, 46(D1), D1074-D1082.

10. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., Wiegers, J., Wiegers, T.C. and Mattingly, C.J., 2021. Comparative toxicogenomics database (CTD): update 2021. Nucleic acids research, 49(D1), pp.D1138-D1143.

11. Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O’Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L. and Tummino, T.A., 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature, 583(7816), pp.459-468.

12. Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S. and Stojmirovic, A., 2015. NDEx, the network data exchange. Cell systems, 1(4), pp.302-305