Exploration and Discovery of the COVID-19 Literature through Semantic Visualization

Jingxuan Tu[1], **Marc Verhagen**[1], **Brent Cochran**[2], **James Pustejovsky**[1]
[1]Brandeis University [2]Tufts University School of Medicine
{jxtu,verhagen,jamesp}@brandeis.edu
brent.cochran@tufts.edu

Abstract

We propose *semantic visualization* as a linguistic visual analytic method. It can enable exploration and discovery over large datasets of complex networks by exploiting the semantics of the relations in them. This involves extracting information, applying parameter reduction operations, building hierarchical data representation and designing visualization. We also present the accompanying COVID-SEMVIZ, a searchable and interactive visualization system for knowledge exploration of COVID-19 data to demonstrate the application of our proposed method.[1] In the user studies, users found that *semantic visualization*-powered COVID-SEMVIZ is helpful in terms of finding relevant information and discovering unknown associations.

1 Introduction

COVID-19 is the first global pandemic within a century. To facilitate the scientific and medical effort to stop this pandemic, most publishers are making full text of COVID-19 related manuscripts freely available.[2] However, every year, the number of published papers is growing at a rate that makes full use of these resources a daunting task (Johnson et al., 2018), and it is getting severer especially during the COVID-19 pandemic when new information is rapidly emerging.

To facilitate the research over these articles, many researchers also publish corpora of preprocessed and curated COVID-19 articles such as LidCovid (Chen et al., 2020) and CORD-19 (Wang et al., 2020). However, for most users and researchers, it is still challenging to fully explore such a corpus due to the complexity of scientific content it contains (for example, complicated pathways in biomedical field (Mercatelli et al., 2020)).

Finding connections among multiple corpora is another challenge. Even for corpora that are targeting a specific topic like *COVID-19*, they may contain information at different scale for different purposes. For example, one dataset provides parsed text and meta information of articles (Wang et al., 2020), and another provides detailed protein-protein interactions extracted from sentences (Gyori et al., 2017). It is difficult to gain full insight by looking either one of those individually. Although search engine is supported for some corpora and portals, this query-based and targeted search is limited in finding connections and patterns that are not obvious from individual articles or sentences (White and Roth, 2009).

To enhance the scientific discovery over complex corpora, we propose *semantic visualization*, a set of text processing and visualization techniques and accompanying tool COVID-SEMVIZ for enhanced knowledge exploration of COVID-19 data (Figure 1). *Semantic visualization* transforms large datasets of complex networks into rich semantic-aware text data; processes text data in a hierarchical manner; and provides visualizations for the indexed data.

The tool COVID-SEMVIZ allows for searchable and interactive visualization of data through word clouds, heat maps, graphs, etc. Unlike other work (See Section 4), we focus on constructing and navigating information from biomedical datasets in a unified hierarchical structure. For example, the activation relations between proteins and COVID-19 can be constructed as the functional type "COVID-19 activators". By reducing relations to a single functional type, it enables the visualization of higher order relations (*e.g.* relations between COVID-19 activators and other protein inhibitors) through a simple 2-dimensional heat map. Other types of visualizations will also appear on the side such as a word cloud of proteins that activate COVID-19, and a tabular form of evidencing sentences. All these visualizations compose a *habitat*

[1] https://www.semviz.org/

[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

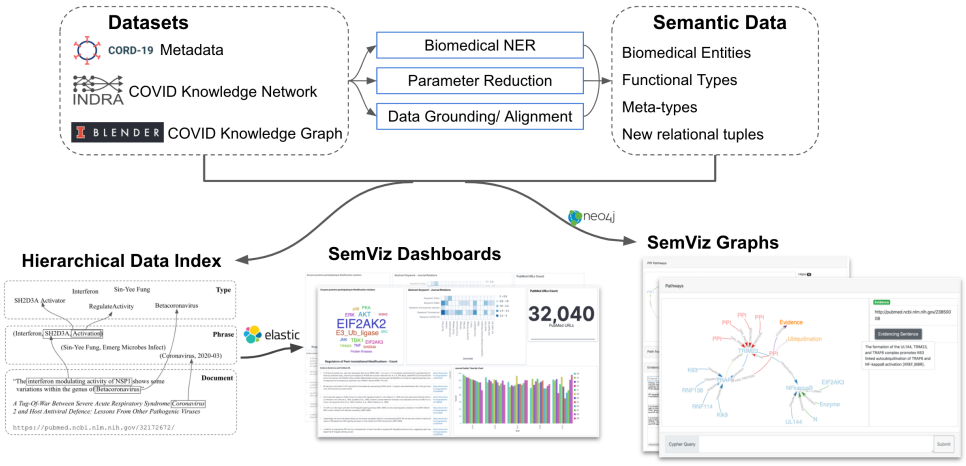

Figure 1: A system overview of COVID-SEMVIZ. The top shows the processing of raw corpora into semantic-aware data. At the bottom it shows the semantic data along with the original corpora are processed in the hierarchical manner and fitted in to the data index or transformed into graph data. Finally data is explored via dashboards and graphs.

of information about COVID-19. Through this, we aim to provide researchers with a global view of selected relationship subtypes drawn from hundreds or thousands of papers at a single glance. This enables the ready identification of novel relationships that would typically be missed by directed keyword searches.

We summarize our main contributions as the follows: (1) Proposed *semantic visualization*, a linguistic visual analytic method that enhances the exploration and visualization of scientific text datasets; (2) Implemented COVID-SEMVIZ, a working prototype for enhanced knowledge exploration of COVID-19 datasets; (3) User studies that evaluate the effectiveness of our system to the biomedical research community as well as future improvements.

2 Semantic Visualization

We propose *Semantic visualization* as a general linguistic visual analytic method for enabling exploration and discovery over large text datasets by exploiting the semantics of the relations in them. This involves (i) collecting data and applying NLP to extract named entities, relations and knowledge graphs from the original text; (ii) indexing the output and creating hierarchical representations for all relevant entities, relations and text that can be visualized in many different ways such as tag clouds, heat maps, graphs, etc.; (iii) applying parameter reduction operations to the extracted relations, creating *functional types* that can also be visualized using the same methods, allowing the visualization of multiple relations, partial graphs, and exploration across multiple dimensions.

2.1 Data collection and Extraction

The first step of semantic visualization involves collecting multiple text datasets of same domain and applying NLP techniques for information extraction to complement original data.

Recently, there is some important work that focuses on publishing new corpora and mining useful text from literature related to COVID-19. In our implementation, we choose to use the following three datasets of COVID-19 literature:

COVID-19 Open Research Dataset (CORD-19) is one of the most comprehensive resource of articles on COVID-19 (Wang et al., 2020). It contains metadata and parsed full text of each article collected from various sources.

Harvard INDRA CORD-19 causal assertions dataset (CKN) [3] contains over 320,000 causal assertions (CAs) extracted from the full text of CORD-19 articles by multiple machine reading systems including REACH (Valenzuela-Escárcega et al., 2018) and Sparser (Mcdonald, 1992). Extracted events were assembled by INDRA[4] and 24 relation types were defined (Gyori et al., 2017).

Blender lab Covid Knowledge Graphs (Blender KG) [5] contains knowledge including entities, re-

[3] https://emmaa.indra.bio
[4] https://github.com/sorgerlab/indra
[5] http://blender.cs.illinois.edu/covid19/

CKN DATASET	
Evidence:	**Ocrelizumab** `[Protein]` and Cladribine may **increase the risk of acquiring** `[Relation]` **COVID-19** `[Protein]`.
Relation:	*(ocrelizumab, COVID-19, Activation)*
BLENDER KG	
Evidence:	**10074-G5** `[Chemical]` results in **decreased expression** `[Relation]` of **MYC** `[Gene]` protein.
Relation:	*(10074-G5, MYC, Decrease Expression)*

Table 1: Example data from CKN and Blender KG.

lations, and events that are extracted from the CORD-19 dataset through deep learning methods (Lin et al., 2020).

Table 1 shows data samples from both the INDRA CKN and Blender KG. Each sample contains a biomedical relation and a corresponding evidencing sentence of that relation. For CORD-19, we extracted and normalized `PMID`, `Title`, `Abstract`, `Authors`, `Publish time` and `Journal` as the metadata for each article.[6] For the CKN dataset, we applied the ScispaCy NER model (Neumann et al., 2019) trained on the BIONLP13CG corpus (Pyysalo et al., 2013) on the original evidencing sentences to extract biomedical named entities, and constructed knowledge graph over encoded relations. For the Blender KG, we use the chemical-gene, chemical-disease, gene-disease relation extraction results. It has over 1,640,000 relations with evidencing sentences from biomedical articles.

In general, semantic visualization suggests information extraction of different granularity. General practice includes named entity recognition, relation extraction, document summarization and graph completion.

2.2 Parameter Reduction

Relational information usually is denoted as *(entity1, entity2, relation-type)* tuples. While individual relations can be visualized through 2-dimensional display techniques like heat maps, demonstrating how multiple relations relate to each other when chained together can be tricky to visualize, requiring cumbersome network visualization techniques (Mercatelli et al., 2020; Nelson et al., 2019). In the biomedical data we are processing, the large number of nodes and connections along with the heterogeneity of both node types (proteins, chemicals, diseases) and edges (structural, functional, and causal interactions) complicates the visualization (Agapito et al., 2013; Salazar et al., 2014; Baryshnikova, 2016).

Particularly for relational information from the data, we propose semantic parameter reduction, a method that reduces relations to *functional types*, allowing them to be treated as individuals. *Functional types* can show more capability and flexibility in terms of encoding information and visualization. The term "parameter reduction" has been used in computer science to refer to reducing model parameters (Kim et al., 2017; Glaws et al., 2020), and our proposed method has the same spirit that aims to reduce the complexity of multiple relations.

Formally, in our current model M, for any given relation tuple $(x,\ y,\ rel)$, we define the function of relation type rel as:

$$[\![rel]\!]^M\ =\ [\lambda y \in D_e.[\lambda x \in D_e.\ 1\ \textit{iff}\ (x,\ y,\ rel)\ \in\ M]] \quad (1)$$

where x and y denote the entities appear in this relation tuple; D_e denotes the set of all entities. Take the tuple *(ocrelizumab, COVID-19, Activation)* as an example, if we pass in *COVID-19* as the first argument to the relation function of *activation*, we will be able to get:

$$[\![\textit{activation}]\!]^M\ =\ [\lambda x \in D_e.\ 1\ \textit{iff}\ (x,\ \textit{COVID-19},\ \textit{activation})\ \in\ M] \quad (2)$$

Through the parameter reduction, we can get the functional type of Equation (2) such that:

$$[\![\textit{ocrelizumab}]\!]^M \in \textit{COVID-19 activator} \quad (3)$$

where the functional type *COVID-19 activator* can be treated as an individual entity instead of a relation. *ocrelizumab* is a member of the functional type in this example. We make the names of functional types both semantically and biologically meaningful based on the relation types, *e.g. activation→activator*, *phosphorylation→kinase*, etc.

Instead of visualizing relations in a heat map, the generated functional types can be visualized using single dimensional display techniques such as tag clouds as shown in Figure 2.

Functional types can also be arguments that will be passed into the relation function, enabling a chain of relations to be expressed in a conventional heat map visualization. For example, Equation (4) is the function of relations between an entity and the functional type *TNF regulator*:

$$[\![rel]\!]^M\ =\ [\lambda x \in D_e.\ 1\ \textit{iff}\ (x,\ \textit{TNF regulator},\ rel)\ \in\ M]] \quad (4)$$

[6]The release date of the dateset we use is 2020-7-5 to match the latest version of Blender KG. It contains over 180,000 scientific papers on COVID-19 and related historical coronavirus research. Download from `www.semanticscholar.org/cord19/download`.

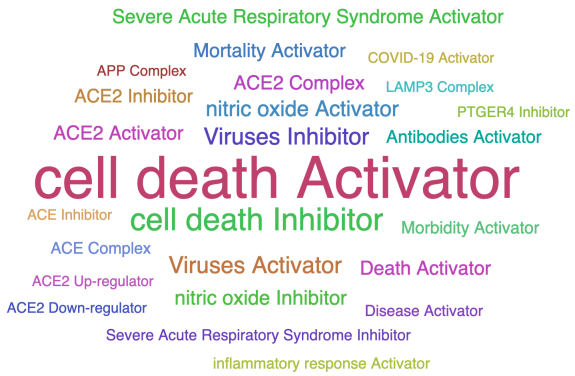

Figure 2: Functional Types as Regulators Tag Cloud from COVID-SEMVIZ.

Figure 3 illustrates such a dense heat map in the Blender KG dataset, where a functionally typed protein is implicated in a disease relation (e.g., "those proteins that are down regulators of TNF which are implicated in obesity")[7].

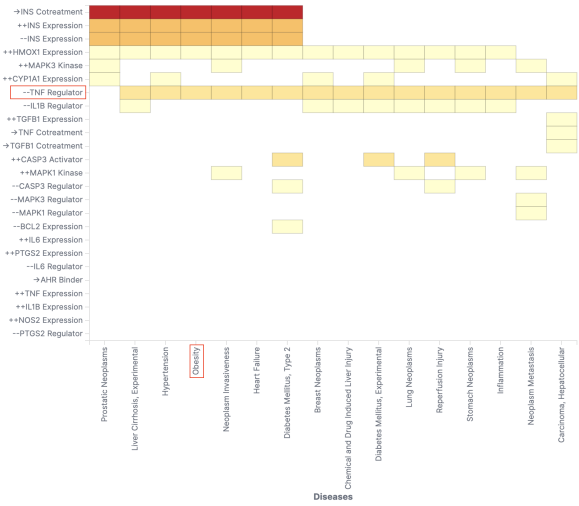

Figure 3: Regulatory Processes-Disease Interactions Heat Map from COVID-SEMVIZ.

2.3 Hierarchical Data Structure

Conceptually, semantic visualization suggests processing and representing data in a hierarchical manner. The resulting data structure composes of three different generic layers that enables better utility of information of various granularity and a global view of data. Although previous work has explored different text structure in data mining (Section 4), they didn't make a clear mapping from information in different layers to various visualization techniques. With the semantic parameter reduction, data can be also be passed and decomposed between different layers from the hierarchical structure.

Type-level layer Represents data that are entities or can be "parameter reduced" as functional types. In our data, individual arguments such as `COVID-19` and `MYC` that are involved in the relation (Table 1), can be seen as entities. In addition, the argument and predicate of a relation can be reduced as a functional type. The causal assertion *(ocrelizumab, COVID-19, Activation)* (Table 1) can be reduced to the entity `COVID-19 Activator`. Subsequently, it is implied that *ocrelizumab* is also included in the `COVID-19 Activators` set.

Phrase-level layer Represents data that can be transformed into "term tuples". A term tuple can be a natural relation that is identified in the datasets, e.g. the relation *(10074-G5, MYC, Decrease Expression)* in Table 1. It can also be built from entities and functional types. Term tuple *(COVID-19, Viruses)* contains the entity `COVID-19` that appears in the abstract of an article, and entity `Viruses` is the journal name where this article is from.

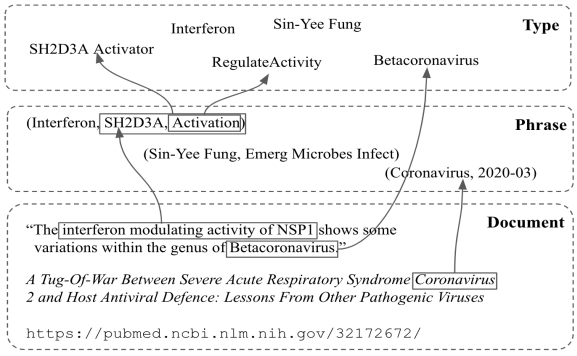

Figure 4: Hierarchical data representation for the datasets. Boxes from bottom to top show how data is represented in different layers. Arrows show how data is passed and decomposed between layers.

Document-level layer Represents data as documents that provide context information to the functional entities and term tuples. The document text is of variable length and it can be a phrase, sentence, or a whole paragraph. In our implementation, we index evidencing sentences, article titles and abstracts as documents. A clickable PubMed URL is also indexed to show the provenance of each

[7] We use the following symbols to indicate the "action" in each relation: "++" = increase, "−−" = decrease, "→" = affect.

evidencing sentence and article title.

Figure 4 shows how the data is processed into the hierarchical data representation. Arrows indicate some extracted relations and entities can be fitted into the other layers. For example, `Coronavirus` from document layer can be used to form a new term tuple with `2020-03` of type *(keyword in abstract, Publish time)*. In the phrase layer, the author name `Sin-Yee Fung` and journal name `Emerg Microbes Infect` that from the CORD-19 dataset can be processed into a new relational tuple. In the type layer, the generated entity `RegulateActivity` and the functional type `SH2D3A Activator` are all associated with a tuple in the phrase level.[8]

2.4 Visualization Techniques

We choose and apply multiple visualization techniques and combinations that are compatible with the hierarchical data representation and allows users to design and build semantically meaningful interactive visualization strategies. In practice, the following general visualization techniques are suggested to be considered: *Word Cloud* (a group of words), *Heat Map* (2D grid matrices for relational data), *Bar Chart* (for categorical data), *Line Chart* (for series of data), *Network* (graphs for complex pathways, KGs, etc), *Tabular Form* (tables for unstructured text) and *Indicator* (displays of the meta information of datasets).

2.5 COVID-SEMVIZ Overview

We release processed data and an implementation of COVID-SEMVIZ visualization system that has been applied with semantic visualization techniques. It contains three dashboards that use different subsets of the data. The Covid CA dashboard holds various visualizations designed principally for CKN dataset and CORD-19, and the Covid KGs dashboard contains visualizations designed for Blender KG and CORD-19. Covid Graph dashboard contains graph-based visualizations to show the all-connected knowledge graph and protein pathways. Due to the space limit, we will provide a detailed overview and technical aspects of the system in the extra page upon accepted.

[8] `RegulateActivity` is the parent relation of `Activation`.

3 User Studies and Evaluation

We present user studies from five researchers (**T1**-**T5**) by letting them interacting with COVID-SEMVIZ in their own research on coronaviruses.[9]

Finding supporting evidence and articles. Based on the search of anti-SARS CoV-2 antibodies, **T1** found most of the relevant literature and "allowed me to quickly zero in on the papers and evidencing sentences I would highlight." **T2** is interested in HEs activities in SARS CoV-2 and found "Many of the common and well known players were revealed in the word cloud".

Discovering unknown interactions. From the protein functional type word cloud, **T2** also found "`TTN Complex` that we had not previously considered." **T3** searched for AT2R and IL-6 inhibition and found the "linkage between those terms and respiratory distress", but the strategy in the linked literature "is not a viable therapeutic strategy in patients of certain conditions". **T4** also found "new links to follow up on, like glycosylation of the coronavirus M protein".

Raising new questions. Based on the search result for AT1R, **T3** found "AT2R activation may have a similar effect on IL-6 levels without impacting blood pressure", and "this is one that I can explore in my research". **T5** searched for TMPRSS2, and found TMPRSS4 appears in the same regulator word cloud. through the checking of linked evidence, **T5** found "Both TMPRSS2 and 4 can cleave the viral fusion protein. This raises the question whether the same is true for COVID-19".

Table 2 shows a summary of what levels of information from the hierarchical data structure that users have mentioned in their comments. We notice that all users find functional types are useful, suggesting the richness of information contained in the functional types from parameter reduction. Interestingly, only two users interacted with phrase-level information. This is probably due to the partial overlapping between phrases and functional types.

We also identify the limitations of our proposed system. One comes from the frequency-based method for displaying data, which means terms or relations that have larger counts are more "salient" in the visualizations (e.g. larger font in the word cloud or darker grids in the heat map). This might

[9] **T1** and **T5** study tumor virus and cancer cells; **T2**'s research focuses on the interface of chemistry, medicine and biology; **T3** studies medicine and nutrition and **T4** studies viral proteins.

User	Term	Functional Type	Phrase	Document
T1		✓		✓
T2	✓	✓	✓	
T3	✓	✓	✓	✓
T4	✓	✓		✓
T5	✓	✓		✓

Table 2: Summary of different levels of information that each user has interacted with.

lead to uncommon or less-studied topics unreachable unless the accurate term has been searched. Another limitation is from the integration of multiple datasets and tools. Artifacts that are in the original data or generated after processing might persist in the final visualizations.

4 Related Work

With the emerging of various COVID-19 data resources, many tools have been developed to enable the visualization and exploration of the large amount of articles that are growing everyday.

Hope et al. (2020) developed SciSight[10], a tool that can be used to visualize co-mentions of biomedical concepts such as genes, proteins and cells that are found in the articles related to COVID-19. It focuses more on displaying purely the association between entities that are mined from articles. IBM COVID-19 Navigator[11] supports the semantic search by building queries with the combination of general terms, UMLS (Unified Medical Language System) concepts, authors and boolean operators. It only provides term-level search and no visualization functionality. COVID-SEE[12], proposed by Verspoor et al. (2020), supports the search from CORD-19 dataset and visualization of article topics and relational concepts. Most other visualizations, however, relate to epidemiological statistics and the effects of Covid-19 on social and health factors[13].

Recent work has been mining useful data from biomedical text. Kordjamshidi et al. (2015) explored the text structure of biomedical data and used information from different levels of the structure as the features to automatically extract bacteria names. Liu et al. (2015) proposed a text mining system for identifying relationships between biomedical entities. It supports template-based queries for structured search and also provides key sentences as the provenance of identified relations. Fabregat et al. (2018) proposed a knowledge base of human pathways and reactions. It supports visualization of event hierarchy and pathway networks.

Linguistic visualization research in general is an emerging field of visual analytics for linguistics (Butt et al., 2020). Previous research in this field covers thematic text cluster analysis (Gold et al., 2015), NER-based document content analysis (El-Assady et al., 2017b), multi-party discourse analysis (El-Assady et al., 2017a) and topic modeling visualization (El-Assady et al., 2018). Butt et al. (2020) propose a web framework that consists of various linguistic visualization techniques. However, existing work in this field focuses on the analysis of corpora of conversational text and transcripts, and does not include approaches for analyzing and visualizing semantics of relations.

5 Conclusion

We have proposed *semantic visualization*, a linguistic visual analytic method of multiple steps involving data extraction, parameter reduction, hierarchical structure building and visualization design. It can facilitate the exploration over large and complex datasets by exploiting the semantics of the relations in them. We have also presented COVID-SemViz, a working prototype for the visualization and exploration of three COVID-19-related datasets. Our user studies indicate that COVID-SemViz is helpful to the biomedical community and the utility of *semantic visualization* techniques. Although we only demonstrated how to apply semantic visualization to COVID-related articles, our proposed method is generalizable enough to be applied to other text corpora. Future work includes addressing current limitations, applying to data from other domains and incorporating more and useful information extraction models in the pipeline. It is our hope that this semantic visualization environment will enable the discovery of novel inferences over relations in complex data that otherwise would go unnoticed.

Acknowledgments

Thanks to the NAACL SRW mentor Roy Schwartz and two anonymous reviewers for providing feedback on this paper. Thank you to Kyeongmin Rim and Kelley Lynch for providing various help on this project. This research is funded by DTRA grant

[10] https://scisight.apps.allenai.org/jnlpba/

[11] https://covid-19-navigator.mybluemix.net/search

[12] https://covid-see.com/

[13] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/data-visualization.htm

HDTRA1-16-1-0002, to Brandeis University.

References

Giuseppe Agapito, Pietro Hiram Guzzi, and Mario Cannataro. 2013. Visualization of protein interaction networks: problems and solutions. *BMC bioinformatics*, 14(S1):S1.

Anastasia Baryshnikova. 2016. Systematic functional annotation and visualization of biological networks. *Cell systems*, 2(6):412–421.

M. Butt, A. Hautli-Janisz, and V. Lyding. 2020. *LingVis: Visual Analytics for Linguistics*. CSLI lecture notes. CSLI Publications/Center for the Study of Language & Information.

Q. Chen, A. Allot, and Z. Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193.

Mennatallah El-Assady, Annette Hautli-Janisz, Valentin Gold, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2017a. Interactive visual analysis of transcribed multi-party discourse. In *Proceedings of ACL 2017, System Demonstrations*, pages 49–54, Vancouver, Canada. Association for Computational Linguistics.

Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, D. Keim, and C. Collins. 2017b. Nerex: Named-entity relationship exploration in multi-party conversations. *Computer Graphics Forum*, 36.

Mennatallah El-Assady, Fabian Sperrle, Rita Sevastjanova, M. Sedlmair, and D. Keim. 2018. Ltma: Layered topic matching for the comparative exploration, evaluation, and refinement of topic modeling results. *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pages 1–10.

A. Fabregat, S. Jupe, L. Matthews, Konstantinos Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, Florian Korninger, Bruce May, M. Milacic, C. Duenas, K. Rothfels, C. Sevilla, V. Shamovsky, Solomon Shorser, Thawfeek M. Varusai, G. Viteri, J. Weiser, Guanming Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. 2018. The reactome pathway knowledgebase. *Nucleic Acids Research*, 42:D472 – D477.

A. Glaws, P. Constantine, and R. Cook. 2020. Inverse regression for ridge recovery: a data-driven approach for parameter reduction in computer experiments. *Statistics and Computing*, 30:237–253.

Valentin Gold, Christian Rohrdantz, and Mennatallah El-Assady. 2015. Exploratory text analysis using lexical episode plots. In *EuroVis*.

Benjamin M. Gyori, John A. Bachman, Kartik Subramanian, Jeremy L. Muhlich, Lucian Galescu, and Peter K. Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. 2020. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*.

Rob Johnson, Anthony Watkinson, and Michael Mabe. 2018. *The STM report*. International Association of Scientific, Technical and Medical Publishers.

Juyong Kim, Yookoon Park, Gunhee Kim, and Sung Ju Hwang. 2017. Splitnet: Learning to semantically split deep networks for parameter reduction and model parallelization. In *ICML*.

Parisa Kordjamshidi, D. Roth, and Marie-Francine Moens. 2015. Structured learning for spatial information extraction from biomedical text: bacteria biotopes. *BMC Bioinformatics*, 16.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009.

Y. Liu, Yongjie Liang, and D. Wishart. 2015. Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Research*, 43:W535 – W542.

David Mcdonald. 1992. An efficient chart-based algorithm for partial-parsing of unrestricted texts. In *Proceedings of the 3d Conference on Applied Natural Language Processing*, pages 193–200.

Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. 2020. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430.

Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg, and Roded Sharan. 2019. To embed or not: network embedding as a paradigm in computational biology. *Frontiers in genetics*, 10:381.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the cancer genetics (cg) task of bionlp shared task 2013. In *BioNLP@ACL*.

Gustavo A Salazar, Ayton Meintjes, and Nicola Mulder. 2014. Ppi layouts: Biojs components for the display of protein-protein interactions. *F1000Research*, 3.

Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation*.

K. Verspoor, Simon vSuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Baldwin, A. J. Yepes, and D. Martínez. 2020. Covid-see: Scientific evidence explorer for covid-19 related research. *ArXiv*, abs/2008.07880.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Ryen W. White and Resa A. Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1:98.

A COVID-SEMVIZ Overview

Figure 5 shows the various visualization techniques that have been applied in COVID-SEMVIZ.

Technical Detail We store the processed hierarchical structured data as the JSON format, and store the generated COVID graph data into neo4j[14] database. The back-end text-based search functionality of COVID-SEMVIZ is built using Elasticsearch[15], and the back-end graph-based retrieval is supported by querying neo4j database. The front-end visualizations are build using Kibana[16] and D3.js[17]. Kibana supports a collection of visualization types. It can be directly applied on the data that has been indexed for Elasticsearch. Elements built from Kibana can be arranged as desired and visualizations will be updated in real-time when a search is performed. It can also provide quick insights into subset of data and enable users to drill down into details through a few clicks. We think our hierarchical indexed data can largely benefit from these features for interaction. D3.js is a JavaScript library that can be used to build customized interactive visualizations. We primarily used it to build graph-based visualizations.

Navigation The navigation of a dashboard from COVID-SEMVIZ is through clicking and searching. By clicking the functional type `CASP3 Activator` in the word cloud named "Regulatory Processes" (Figure 5), a constraint on the type and regulators of proteins is added. Correspondingly, all the other visualizations will be changed. For example, the "Subject Proteins" word will only contains protein entities that can activate CASP3; the "Evidence Sentences and PubMed URL" tabular form will display evidencing sentences that involve proteins that can activate CASP3 in the relations. The "Abstract Keyword - Journal Relations" heat map will form new color shade clusters based on the new set of articles that mentioned CASP3 or its regulators. One can also put a query into the search box to navigate the dashboard. Navigation through the Covid Graphs is similar. One can use searching and clicking to retrieve relevant sub-graphs and examine the context information of a node such as the relations it belongs to and its provenance. In addition, COVID-SEMVIZ supports abstracting graphs by reducing nodes to functional types and expanding node neighbors that are specifically for graphs.

The Covid Causal Assertions Visualization The Covid Causal Assertions (CA) dashboard contains a set of visualizations that are designed to enable users to discover novel inferences of protein-protein interactions and associated context information. Users can type in a query to search for relevant CA and context information. We include several kinds of visualizations: (1) tabular forms for tracing evidence associated with relations, (2) indicator panes to display the count of evidences and of unique articles, (3) word clouds and heat maps for some metadata, (4) type-level and phrase-level visualizations that enable users to drill down into the elements in the relations, (5) dense visualizations for functional types, and (6) visualizations of upstream regulators. We now elaborate on the last three of these.

Type-level and phrase-level visualizations. Each CA contains three elements: protein-A, relation type, and protein-B. We group the 24 relation types into two "metatypes": `RegulateActivity` and `Modification`. Furthermore, protein-A and protein-B involved in RegulateActivity relations are categorized into `Subject` and `Object`. Protein-A and protein-B involved in Modification relations are categorized into `Enzyme` and `Substrate`. We believe this categorization allows our visualizations to conform to biological convention. On the dashboard, we create words clouds for these categories. We also create a subject-object interaction heat map to show regulatory relationships, an enzyme-substrate interaction heat map to show protein modification relationships, and heat maps for some common relation types such as `Activation` and `Inhibition`. Finally, we include word clouds for entity types extracted with the NER model.

Visualizations for functional types. We also enable the visualization of CAs by applying parameter reduction, which is a critical step in semantic visualization. Given two CA tuples *(Protein-A, Activation, Protein-B)* and *(Protein-B, Activation, Protein-C)*, we create the functional type *Protein-C-Activator* with members *Protein-A* and *Protein-B*. We now have a word cloud for all functional types (see Figure 6) and a separate word cloud for the subject proteins associated with them. Clicking one of the functional types restricts the subject pro-

[14] https://neo4j.com/
[15] https://www.elastic.co/elasticsearch/
[16] https://www.elastic.co/kibana
[17] https://d3js.org/

Figure 5: Visualization techniques from COVID-SEMVIZ. First row: (1) Word cloud of functional types as regulatory processes; (2) Heat map represents the relations between article keywords and journal names; (3) Indicator of total number of articles. Second row: (4) Tabular form of evidencing sentences and provenance URL; (5) Bar chart shows the number of articles that are published in each month from year 2015 to 2019.

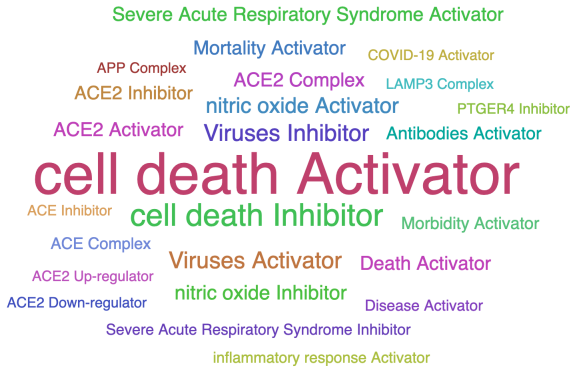

Figure 6: Sample regulators.

teins to just the ones involved in the functional type selected.

Visualizations for upstream regulators. One advantage of parameter reduction is that it can represent higher order relations so that those relations can be easily visualized with word clouds and heat maps . In the Covid CA dashboard, We present two types of second order CAs: one that has the same relation type as the functional type, and one that has the opposite relation type. In the dashboard, we add the "Upstream Regulators" word cloud and the "Opposite Upstream Regulators" word cloud to display second order relations. For example, with a functional type *Interferon-Activator* the "Upstream Regulators" word cloud would include all proteins X that activate one of the Interferon activators, thereby generating a novel inference from *X* to *Interferon*. Through navigation over the keywords in each word cloud, one can easily check the evidencing sentences of deeper CAs that are inferred through parameter reduction.

Formally, if we have identified `Protein-2 Activator` and have the opposite relation pair `Activation` and `Inhibition` in our dataset, we are interested in a set of `X` that `X` activate `Protein-2 Activator` or inhibit `Protein-2 Activator`. Thereby we are able to generate novel inference from `X` to `Protein-2`. `X` is also called the second order containers in our case. We pair the opposite relation types in our dataset and leave the others unchanged that can only have the same second order relations.

The Covid KGs Visualization The Covid KG dashboard contains a collection of visualizations that enable the discovery of the relationships among genes, chemicals and diseases that are related to COVID-19. This includes chemical-gene, chemical-disease and gene-disease relations, which are supported by the evidencing sentences not only from COVID-19 articles but also from various other medical articles. Thus, the most challenging part in the visualization is to simplify and unify the complex relations while displaying the information in breadth and depth.

We start by making the connections between chemical-gene and gene-disease relations using the same gene entries that appear in both sides. Then we index the new chemical-gene-disease relations and visualize them via chemical-gene sub-relation heat map and gen-disease sub-relation heat map. These two heat maps are designed to be interactive with each other to show the full chemical-gene-disease triplet relations, as well as to be flexible enough to be controlled by enabling or disabling arguments of the triplet relations.

Similar to the Covid CA dashboard, we build a tabular form that displays evidencing sentences and PubMed URLs, as well as word clouds of chemicals, genes and diseases from the relations. Users can navigate the dashboard to find relevant context information by filtering on entities from the word clouds. we also create a word cloud of gene functional types by grounding chemical-gene relations. For example, given a chemical-gene tuple (D014013, Decrease Reaction, CASP3), the functional type `-CASP3 Regulator` is generated.

The Covid Graph Visualization Covid Graph dashboard contains two graph-based visualizations: the all-connected knowledge graph and protein pathways. Figure 7 shows the knowledge graph visualization. The main window shows a color-coded graph of predefined nodes such as proteins, evidence and PPIs. Nodes are connected by different relationships based on the labels of nodes. The sidebar on the right displays the information of clicked node. For example, if an evidence node is clicked, it shows the content of the evidence and the article URL that contains this evidencing sentence. An input box on the bottom takes a Cypher query and generates the corresponding graph. The knowledge graph enables the visualization of data of different granularity in one place. It can also be context-aware by dynamically generating neighbors of a right-clicked node.

Figure 8 shows the interface of protein pathways visualization. A variable-length pathway can be retrieved by specifying the starting and ending proteins as well as the number of hops. We also apply parameter reduction operations on sub-paths of the whole pathway, compressing the graph without any semantic information loss, and provides the clear and dense visualization over complex graph. Specifically, given a sub-path of length 3 (e.g. SP-[decreseAmount]→ACE2-[Activates]→COVID-19), it can be compressed into a binary relation containing a functional type and an entity (e.g. ACE2 down-regulator-[Activates]→COVID-19). Each functional type like "ACE2 downRegulator" represents a set that can contain any protein down-regulating ACE2.

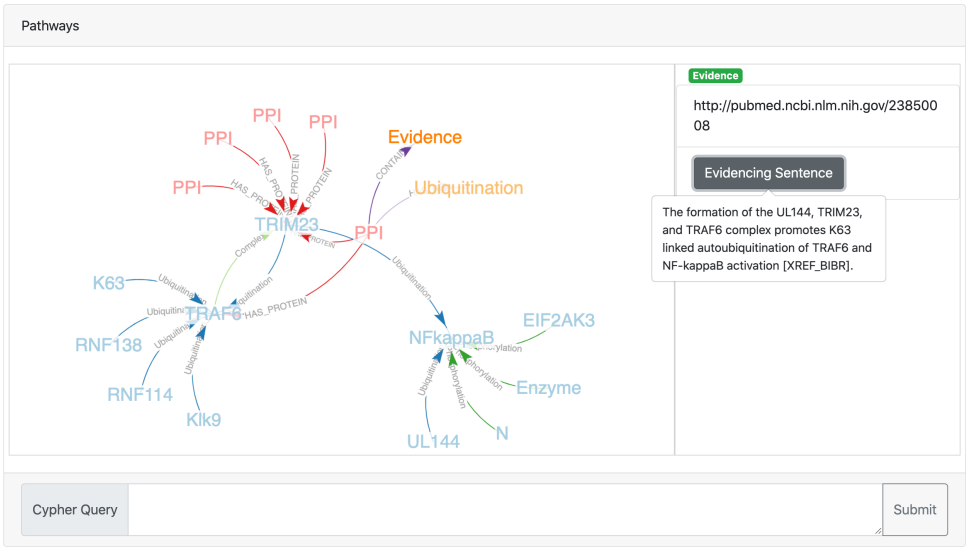

Figure 7: Interface of Covid knowledge graph visualization.

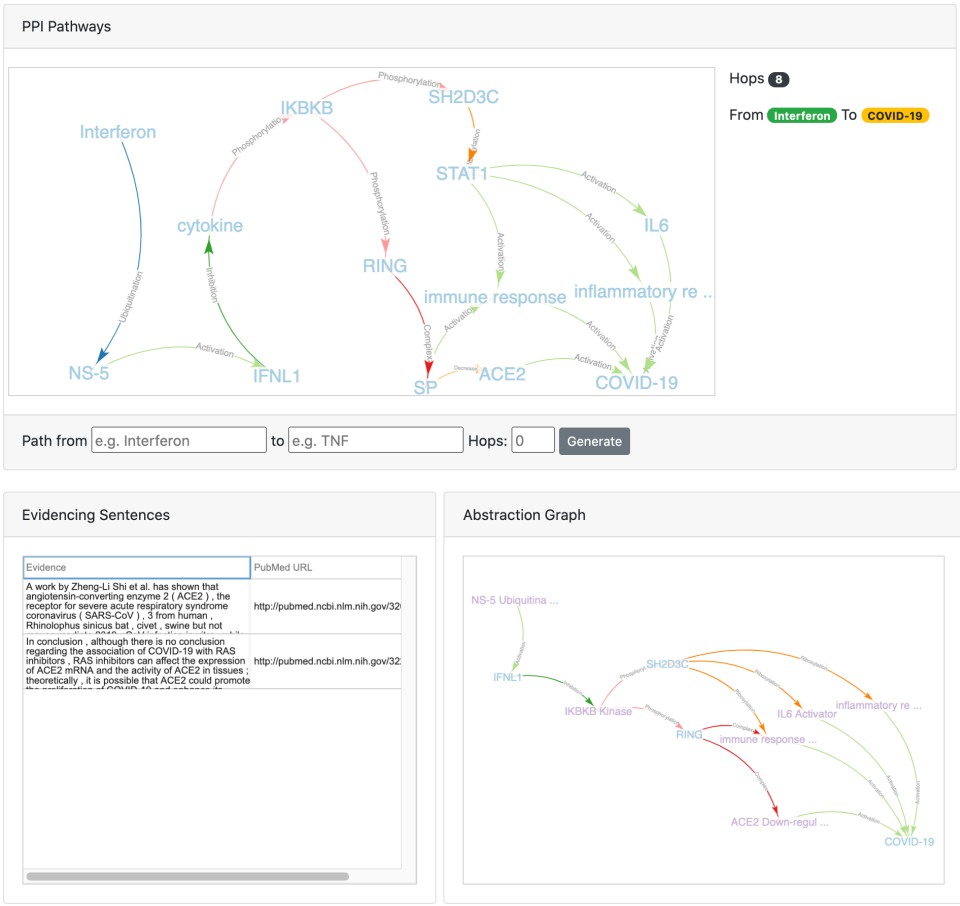

Figure 8: Interface of protein pathways visualization.