BioKDE: a Deep Learning Powered Search Engine and Biomedical Knowledge Discovery Platform

Meng-Han Chung[1], Jun Zhou[1], Xiaodong Pang[1], Yuchuan Tao[1], Jinfeng Zhang[1,2,*]
[1]Insilicom LLC, Tallahassee, FL 32303, USA
[2]Department of Statistics, Florida State University, Tallahassee, FL 32306, USA

Abstract—**Search engines play important roles in scientific research by helping scientists find articles relevant to a set of keywords. As more and more papers are being published on a daily basis, the amount of the scientific literature has imposed a great challenge on current search engines for finding the most relevant articles to a given query. Despite significant progress made in the past two decades, the performance of current search engines is still not very satisfactory to many users. While search engines often return a large number of relevant articles, getting the relevant information out of the returned articles is another challenge for the researchers. Extracting the important information from the articles and presenting it in a user-friendly way will maximize the benefits of the search engines. In this paper, we introduce a new platform, BioKDE (Biomedical Knowledge Discovery Engine), which consists of a search engine for all the PubMed abstracts and a knowledge graph (KG) application with various functions (https://biokde.com). We show that the KG functions can be very helpful for researchers to quickly identify key information they are looking for. The returned KGs can also be manually curated and shared with other researchers, making them a great resource for structured biomedical knowledge freely accessible to all researchers. We demonstrated the KG editing functions by creating a KG for COVID-19 risk factors (https://biokde.com/KG/KG000367).**

Keywords— Search engine; knowledge graph; deep learning; natural language processing; search engine for biomedical literature; named entity recognition; relation extraction

I. INTRODUCTION

The number of biomedical research articles published in peer-reviewed journals has been increasing at an accelerated speed in recent years, and it has become harder and harder for scientists to keep track of the large amount of published literature. Missing important prior studies in a literature search when designing a new study can have serious consequences such as wasting resources and/or time. It can also result in making wrong conclusions or missing new discoveries when interpreting experimental results.

Most search engines nowadays use BM25-based similarity ranking algorithm [1]. BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document. There are some known issues of BM25 [2] and a number of variations have been proposed to improve it [3]. For biomedical literature, Lu and co-workers have used user feedback information and machine learning to develop an improved search engine at PubMed [4, 5]. A list of web-based tools with literature search functions has been compiled and continuously updated [6].

To improve the accuracy of a search engine, a major challenge is the availability of a large amount of labeled data, consisting of a large number of different queries and truly relevant articles matched to them. Generating such labeled data is very time and resource consuming. To tackle this challenge, we realized that the citation data from PubMed Central (PMC) full-text articles can be used for this purpose. In full-text articles, we can extract a sentence (with a citation) and the PubMed ID of the article the sentence cited, which form a pair of labeled data. The citing sentence can be used as a query and the article that the sentence sited can serve as the relevant document for the sentence. With this insight, we have built a deep learning model using millions of labeled cases obtained from PMC full-text articles. Additional changes were made to scale it up to handle all PubMed abstracts with good speed. We tested the search engine (https://biokde.com) using a large test dataset and some commonly searched keywords and found that it performed significantly better than other publicly available search engines, including Google Scholar, PubMed, and Semantic Scholar. A comparison page can be found at https://biokde.com/comparison/.

A picture is worth a thousand words. Displaying general biomedical knowledge written as raw text as an interactive picture can help researchers quickly explore and navigate the space of existing biomedical knowledge using the interactive tool. There are existing network visualization tools. However, they are not tied with a search engine. The search engines currently available do not have network visualization tools. To address the need of knowledge visualization, we implemented a knowledge graph (KG) module with associated functions. We use an interactive graph to show all the bio-entity (i.e. proteins, genes, diseases, pathways, gene ontology terms, chemical compounds, etc.) relationships in the returned articles from a query. Users can quickly visualize and explore the important relationships without manually reading all the articles. The nodes in the KGs are the bio-entities and the edges represent their relationships described in the returned articles. Clicking the edges will bring out the sentences containing the relationships and the corresponding articles containing the sentences. The KGs can be edited to add or remove nodes or edges. To demonstrate the editing functions, we created a KG for COVID-19 risk factors, which can be accessed at https://biokde.com/KG/KG000367.

In this paper, we will provide a basic introduction of the platform, the search engine, and the knowledge graph module and associated functions.

II. Methods and data

A. Method for the search engine

The method behind the search engine was based on an deep learning algorithm we developed recently [7] with some significant modifications to make it suitable for real time queries and the scale of the database with all PubMed abstracts.

B. Methods for named entity recognition

The entity types were tagged for several important types in biomedical sciences including diseases, genes/proteins, chemical compounds, species, cell lines, SNPs, mutations, etc. Currently, we are using the results from PubTator [8, 9], which will be replaced soon using the method we developed in BioCreative VII. The method is based on PubMedBERT with two main modifications: (1) we used a data augmentation strategy to enrich the training data with more positive cases; (2) we identified the full names of abbreviations and tagged all the same abbreviations in an article consistently as a single entity type. During the manual annotation of the COVID-19 knowledge graph, we added a few more entity types including demographics, a keynode type for COVID-19, and a miscellaneous type.

C. Methods for relation extraction

We are using co-occurrence method to display the entity relations in the knowledge graph to be inclusive for all possible entity relations in the returned articles. In the near future we will apply the relation extraction methods we developed in the past few years [10-19] and developed in the BioCreative VII to extract relations. The relation extraction method applied three different deep learning methods, which were then ensembled together. The three deep learning methods are: a fine-tuned BERT, a sentence BERT, and a T5 model.

D. Data

We have downloaded all the abstracts from PubMed, which are regularly updated. We are working on adding PubMed Central full text articles to the database.

III. Functions and services

A. Biomedical knowledge retrieval

On the homepage of BioKDE, https://biokde.com, users can use keywords to search articles relevant to certain research topics, author names or PMIDs. We recommend using more keywords in queries to define search topics. Our test has shown that BioKDE performs significantly better than other search engines such as Google Scholar and PubMed. Our patent search also performs much better than Google Patent search. A few examples are given at https://biokde.com/comparison/. Figure 1 shows the homepage of BioKDE, where we have added a tutorial video to walk users through the major functions of BioKDE.

Searching using keywords will bring a user to the search result page as shown in Figure 2, which displays the returned articles using the keywords "COVID-19 risk factors". On the search result page, users can use filters on the left panel to select subset of articles using publication date or article type as conditions. They can also sort the articles using relevance, publication date, or citations located above the search results. Clicking the Display Knowledge Graph button will bring users to the corresponding knowledge graph page (Figure 3).

B. Knowledge visualization functions

Users can visualize the set of selected articles on the knowledge graph page (Figure 3). Users can edit the display and save the knowledge graph. Users can also select articles, entity types, and edge types to be visualized. Clicking a node (entity) or an edge (relationship) will display the sentences and the articles (PMIDs) containing the corresponding entity or relationship in a table below the knowledge graph. Users can also search an entity on the search box above the graph. Nodes and edges can be moved around to adjust the topology of the graph. The entity types are tagged using deep learning-based name entity recognition methods. The relationships currently shown are extracted using co-occurrence method to be inclusive. Deep learning-based relationship extraction methods will be used to allow users to select subset of relationships with high probability to be true relationships.

C. COVID-19 specific functions

Most published COVID-19 literature is covered in the BioKDE search engine. We will add COVID-19 full-text articles to further increase the coverage. We have curated a COVID-19 risk factor knowledge graph to demonstrate how the tool can be used to help COVID-19 research (Figure 4). The link to this knowledge graph is https://biokde.com/KG/KG000362. It can also be accessed from the homepage.

D. Knowledge management

Users will be able to edit nodes and edges of a knowledge graph including adding a new node/edge, deleting a node/edge, and modifying the information about a node or edge. The modified knowledge graph can be saved and shared with other researchers.

IV. Discussion and future work

In this paper, we provided a brief description of the functions we have implemented at BioKDE. In the future we plan to make improvements in the following areas.

A. Data

We will download the PubMed Central full text data and add them to our database. In addition, we will also add preprint articles from biorxiv and Arxiv to our database to allow users to search the latest research discoveries.

This research was partially supported by a grant from the National Institute of General Medical Sciences of the National Institutes of Health, grant No. R01GM126558.

B. Methods

We will apply the latest NLP methods for both named entity recognition and relation extraction. The results will be directly used to construct knowledge graphs. We will implement functions to make knowledge graph computable so that users can make sophisticated inference using the knowledge graphs.

C. The knowledge graph project

We are in the process of initiating a crowdsourcing project: the Knowledge Graph Project. In this project, biomedical researchers will be able to curate knowledge graphs for a particular topic of interest. The curated knowledge graphs can be shared with other researchers and searchable at BioKDE search engine. We hope a large number of useful knowledge graphs can be created using this approach. Highly specific new article alerts can be created using the knowledge graphs to let system send alerts of certain entity relationships to a user.

D. Feedbacks from BioCreative Challenge VII

We participated in the BioCreative challenge VII track 4 for COVID-19 text mining tool interactive demo. All the participating tools were tested by potential users who have provided valuable feedback for further improvement of the tools. Overall, we obtained very positive feedbacks. There are several areas we need to improve: (1) the number of papers selected for displaying the KG and how to select them. We have the function available, but it is not obvious to new users. We will make the function accessible through multiple entry points and also put it at a location on the interface that makes it easier to find; (2) Output EndNote or BibTex citation format for selected publications; (3) Incorporating information extraction methods for relations and attach probabilities to the edges in the KGs.

We will work on the above areas and release a new version in the near future.

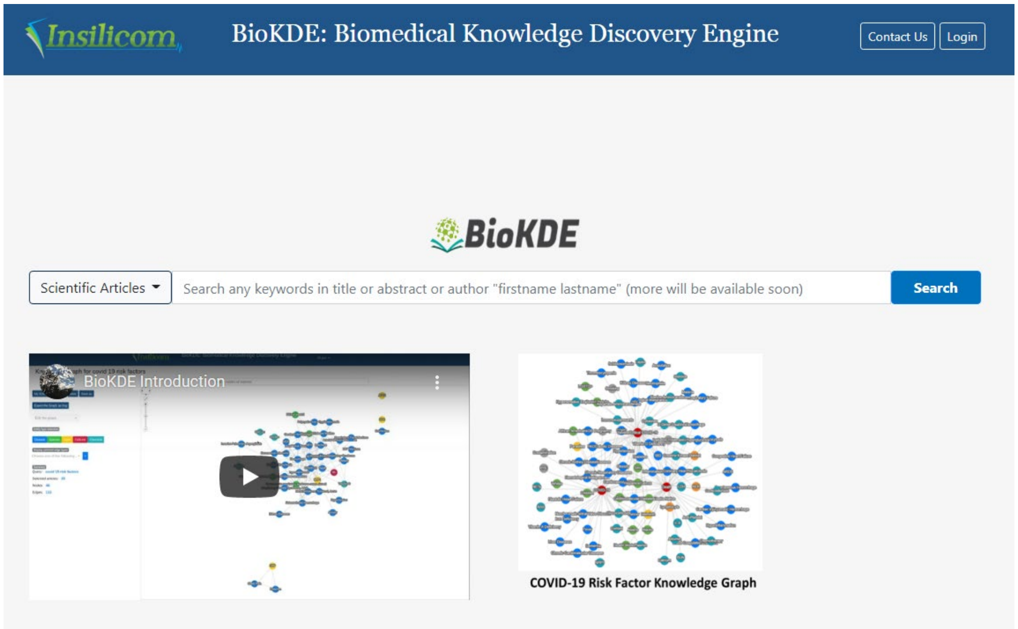
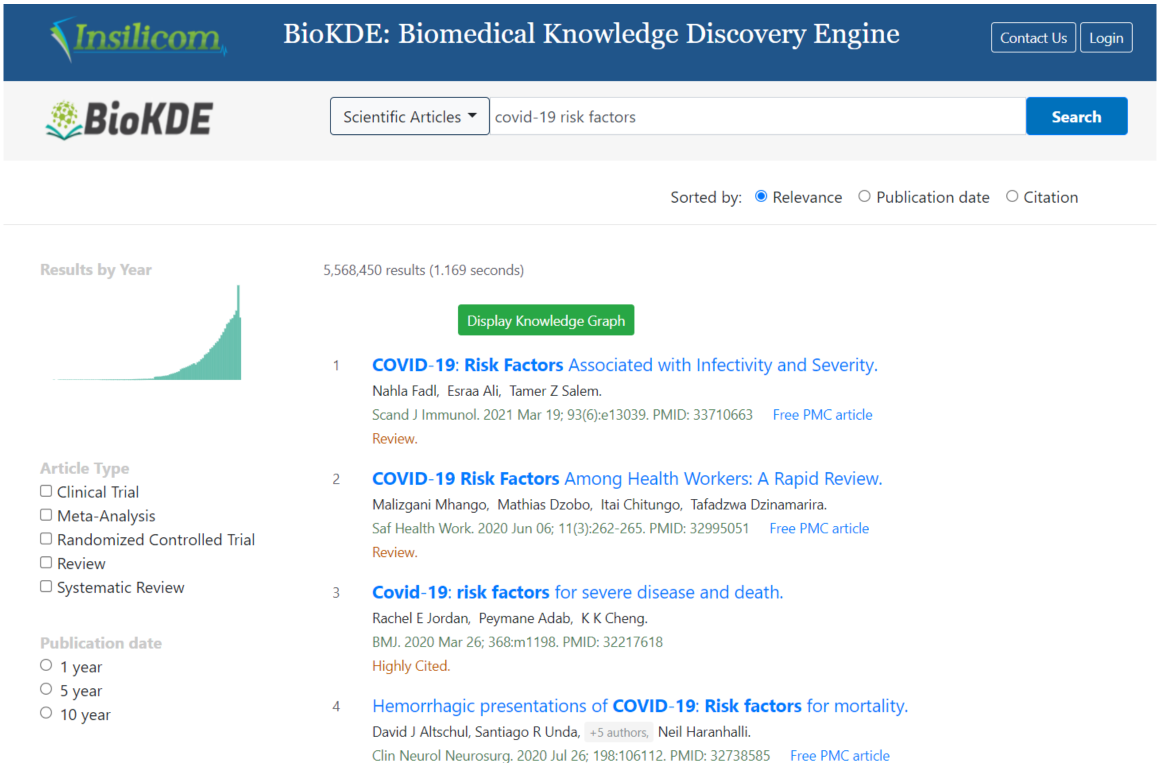

Insilicom
BioKDE: Biomedical Knowledge Discovery Engine
Contact Us
Login
BioKDE
Scientific Articles
Search any keywords in title or abstract or author "firstname lastname" (more will be available soon)
Search
BioKDE Introduction
COVID-19 Risk Factor Knowledge Graph

Figure 1. The landing page of BioKDE. Users can select either *Scientific Articles* or *Patents* from the dropdown menu on the left of the search box to choose to search either scientific articles or patents. Users can search several different fields including titles, abstracts, author names, PMIDs (patent IDs), affiliations, etc.



covid-19 risk factors

Sorted by: ● Relevance ○ Publication date ○ Citation

5,568,450 results (1.169 seconds)

Display Knowledge Graph

Article Type
☐ Clinical Trial
☐ Meta-Analysis
☐ Randomized Controlled Trial
☐ Review
☐ Systematic Review

Publication date
○ 1 year
○ 5 year
○ 10 year

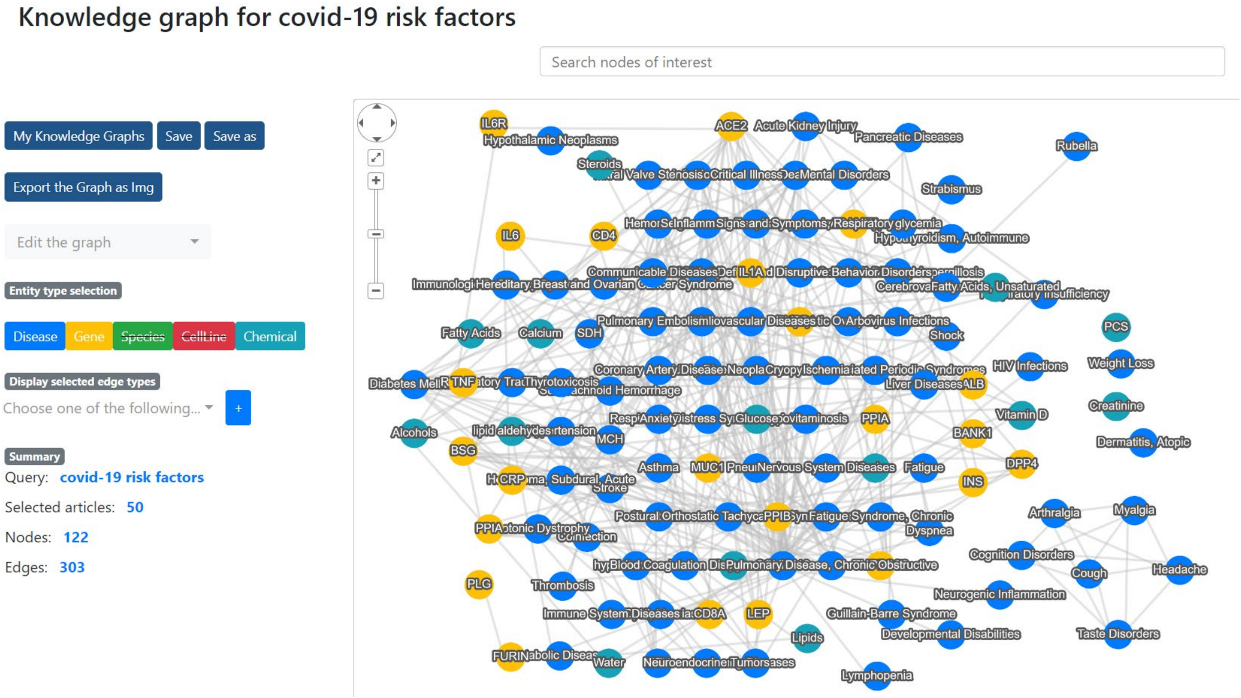
1 **COVID-19: Risk Factors** Associated with Infectivity and Severity.
Nahla Fadl, Esraa Ali, Tamer Z Salem.
Scand J Immunol. 2021 Mar 19; 93(6):e13039. PMID: 33710663 Free PMC article
Review.

2 **COVID-19 Risk Factors** Among Health Workers: A Rapid Review.
Malizgani Mhango, Mathias Dzobo, Itai Chitungo, Tafadzwa Dzinamarira.
Saf Health Work. 2020 Jun 06; 11(3):262-265. PMID: 32995051 Free PMC article
Review.

3 **Covid-19: risk factors** for severe disease and death.
Rachel E Jordan, Peymane Adab, K K Cheng.
BMJ. 2020 Mar 26; 368:m1198. PMID: 32217618
Highly Cited.

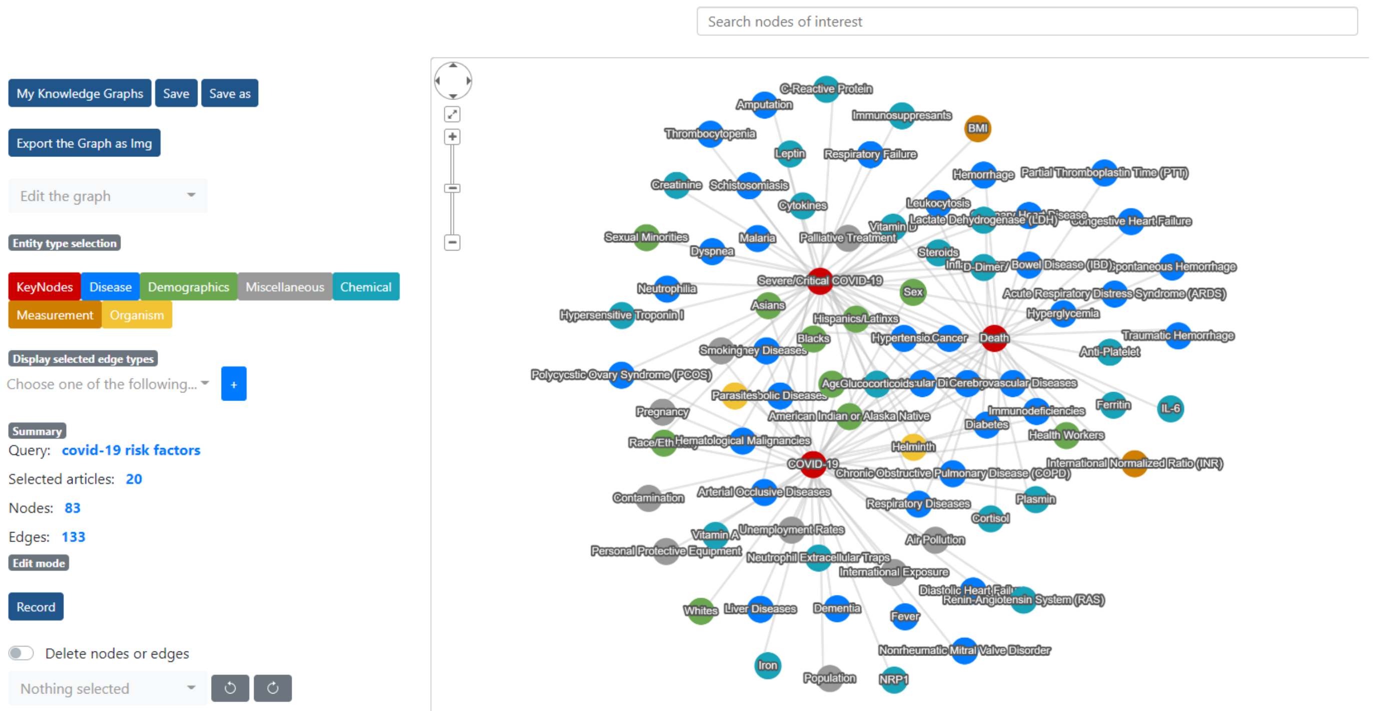
4 Hemorrhagic presentations of **COVID-19: Risk factors** for mortality.
David J Altschul, Santiago R Unda, +5 authors, Neil Haranhalli.
Clin Neurol Neurosurg. 2020 Jul 26; 198:106112. PMID: 32738585 Free PMC article

Figure 2. The search result page. The keywords used are "COVID-19 risk factors". Users can use the filters on the left panel to select the subset of articles using publication year and/or article type as filters. Users can also sort the result using relevance, publication date, or citations located above the search result. Users can click the Display Knowledge Graph button to see the knowledge graph plotted using selected articles (default selection is top 20).



Summary
Query: covid-19 risk factors
Selected articles: 50
Nodes: 122
Edges: 303

Figure 3. Knowledge graph page.



Query: covid-19 risk factors

Selected articles: 20

Nodes: 83

Edges: 133

Figure 4. Manually curated COVID-19 risk factor knowledge graph. There are three key phenotypes: COVID-19, severe COVID-19, and Death. Risk factors were annotated to be associated with one or more of these key phenotypes. The knowledge graph can be edited by a user and saved to a new knowledge graph.

ACKNOWLEDGMENT

We thank the organizers of BioCreative VII Track 4 for organizing this track, which has been very helpful for us to learn more about real users' opinions on our platform.

REFERENCES

[1] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval,* vol. 3, pp. 333-389, 2009.

[2] Y. Lv and C. Zhai, "When documents are very long, BM25 fails!," *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval,* 2011.

[3] A. Trotman*, et al.*, "Improvements to BM25 and Language Models Examined," presented at the Proceedings of the 2014 Australasian Document Computing Symposium, Melbourne, VIC, Australia, 2014.

[4] N. Fiorini*, et al.*, "Best Match: New relevance search for PubMed," *PLOS Biology,* vol. 16, p. e2005343, 2018.

[5] N. Fiorini*, et al.*, "How user intelligence is improving PubMed," *Nature Biotechnology,* vol. 36, pp. 937-945, 2018/11/01 2018.

[6] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database (Oxford),* vol. 2011, p. baq036, 2011.

[7] C.-c. Lo*, et al.*, "Developing a More Accurate Biomedical Literature Retrieval Method using Deep Learning and Citations in PubMed Central Full-text Articles," *bioRxiv,* p. 2021.10.21.465340, 2021.

[8] C. H. Wei*, et al.*, "PubTator central: automated concept annotation for biomedical full text articles," *Nucleic Acids Res,* vol. 47, pp. W587-W593, Jul 2 2019.

[9] C. H. Wei*, et al.*, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Res,* vol. 41, pp. W518-22, Jul 2013.

[10] J. Qu*, et al.*, "Triage of documents containing protein interactions affected by mutations using an NLP based machine learning approach," *Bmc Genomics,* vol. 21, p. 773, Nov 10 2020.

[11] P.-Y. Lung*, et al.*, "Extracting chemical-protein interactions from literature using sentence structure analysis and feature engineering," *Database (Oxford),* vol. bay138, 2019.

[12] K. Yu*, et al.*, "Automatic extraction of protein-protein interactions using grammatical relationship graph," *BMC Med Inform Decis Mak,* vol. 18, p. 42, Jul 23 2018.

[13] J. Zhang, "Automatic extraction of bio-entity relationships from literature. USPTO No. 8,886,522; No. 9,542,528," USA Patent, 2017.

[14] J. Qu*, et al.*, "Mining protein interactions affected by mutations using a NLP based machine learning approach " *Proceedings of BioCreative VI workshop,* pp. 130-135, 2017.

[15] R. Chowdhary*, et al.*, "Context-specific protein network miner--an online system for exploring context-specific protein interaction networks from the literature," *PLoS One,* vol. 7, p. e34480, 2012.

[16] S. Balaji*, et al.*, "IMID: integrated molecular interaction database," *Bioinformatics,* vol. 28, pp. 747-9, Mar 1 2012.

[17] L. Bell*, et al.*, "Mixture of logistic models and an ensemble approach for extracting protein-protein interactions," *ACM-BCB,* pp. 371-375, 2011.

[18] L. Bell*, et al.*, "Integrated bio-entity network: a system for biological knowledge discovery," *PLoS One,* vol. 6, p. e21474, 2011.

[19] R. Chowdhary*, et al.*, "Bayesian inference of protein-protein interactions from biological literature," *Bioinformatics,* vol. 25, pp. 1536-42, Jun 15 2009.