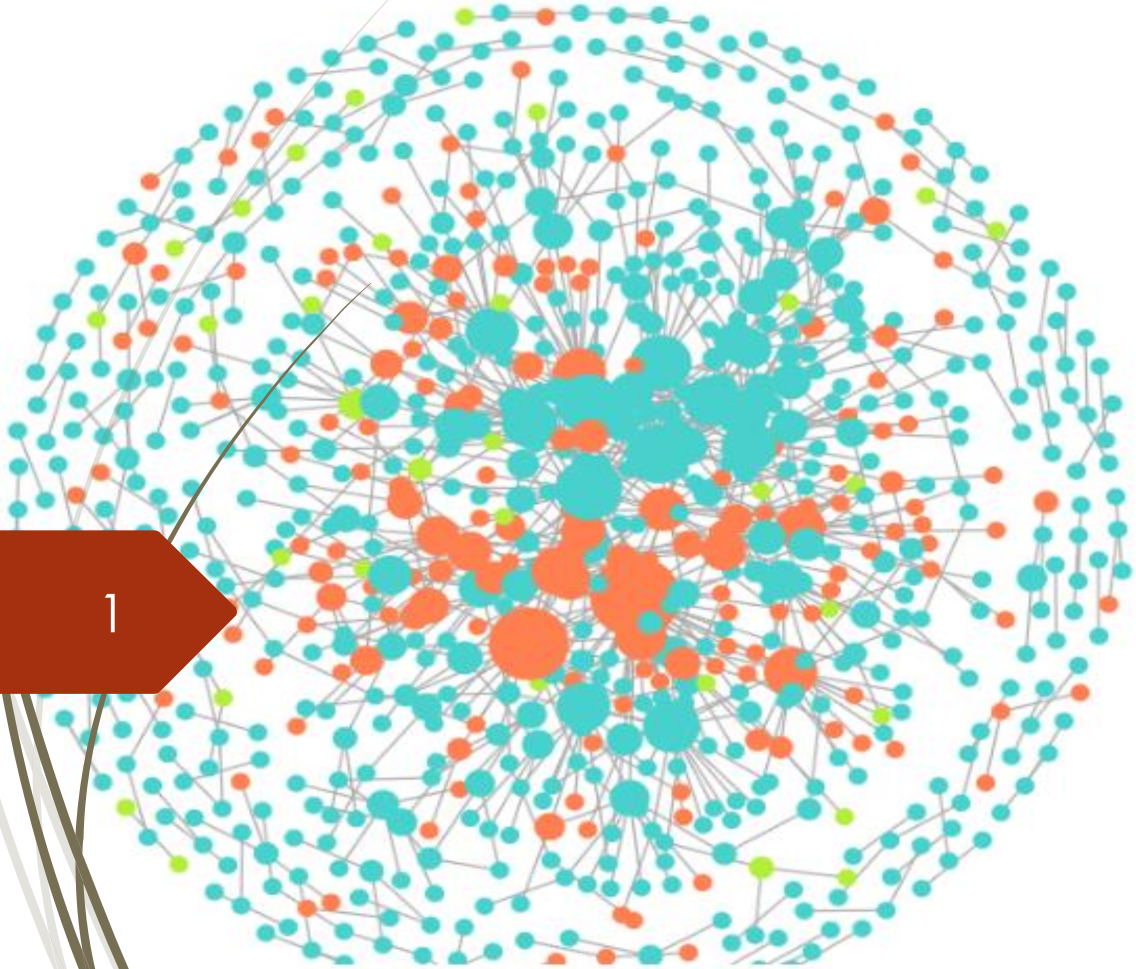
Anveshak : A data mining tool to enable researchers combat Covid-19

अन्वेषक- explorer



Team: Pymetrics

Proposed by :
1. Ankit Barai , Btech CSE, IIIT Nagpur
2. Dr. Amol Shinde, MBBS MD

Contents

Anveshak

- About the Team
- Problem statement and Inspiration
- Impact of solution
- What's special about solution
- Anveshak
    - Q&A model
    - Knowledge graph based solution
- Some screenshots & Demo of Anveshak
- Conclusion
- Future directions
- References

About the team

We are here Team Pymetrics striving to enable researchers combat Covid-19.

Our Team members are from diverse backgrounds :

- Ankit K Barai , B.Tech(CSE), Indian Institute of Information Technology, Nagpur
    - Data science Intern at Jio AI CoE, Intern offer from A-STAR Singapore.
    - Linkedin : https://www.linkedin.com/in/ankit-barai-708941144/

- Dr. Amol Shinde, MBBS MD, Assistant Professor at JN Medical College, Sawangi, Wardha
    - Linkedin: https://www.linkedin.com/in/dr-amol-shinde-2510b767/

- Prof. Jitendra Tembhurne, PhD, Assistant Professor at Indian Institute of Information Technology, Nagpur
    - Bio: https://iiitn.ac.in/faculty-cse-jvt.php

Prior to this work, our team members have also submitted our Research article to a journal on SIR modelling :a study of Lockdown measures on the spread of Covid-19 in India.

Problem Statement and Inspiration

- "*Nothing happens quite by chance. It's a question of accretion of* ***information*** *and experience"* ~ Jonas Salk

- A **literature review is probably the most important step** in understanding a particular research topic or question.

- Since the onset of COVID-19 pandemic, more than **44K research articles** are published already and the numbers are growing. The huge **information** from these articles **exceeds the analyzing capacity** of any researcher, which **critically limit full benefit of the findings** in these articles.

Problem Statement and Inspiration

- Artificial Intelligence(AI), data mining(DM) and state of the art Natural Language Processing(NLP) techniques can be used to **empower the medical research community** by helping them analyze and get important **insights from thousands of research articles** and **extract relevant information** precisely and rapidly.

- Text mining makes **relationships between genes, diseases, biomarkers and chemicals, <u>computationally mappable</u>** and thus helps to dig out **"<u>hidden</u>"** relationships which **aren't distinctly published** in literature.

Impact of the solution in India and rest of the world

- Knowledge graphs that we have created from the 36K research article in CORD-19 dataset will enable researchers to get **direct to the point answers** for the questions directly dealing with Medical terminology like Drugs and Gene names.

- This tool will help:
  1. To know the **pathophysiology of Covid-19.**
  2. For **drug discovery and repurposing.**
  3. To develop **vaccine candidates.**

- Using a streamlined process, our solutions can easily be **updatable to include new publications** as they are published, giving researchers easy access to the latest insights. In addition, the questions we present here are formulated considering the **needs of clinical researchers**, and such questions may be frequently encountered in future disease outbreaks.

- In Addition, if the researcher wants to get complete paper for reference, our Q&A model can fetch them **Top N candidate articles** and **highlight probable answers** to their question.

What’s special about the solution?

- Our Knowledge graph based solution gives :

1. Rapid Information Retrieval
2. Supports huge number of documents
3. Very simple model
4. Very intuitive/explainable : The way Human brains perceive information.

- We propose a knowledge graph based solution (Inspired by the way our Brains stores the information) which **eliminates the need** to go through complete knowledge extraction process for each query.

- We have **separated Knowledge extraction and querying** knowledge graph. Former one is to be done once only. Once done, it can be stored in relevant format. For each query then this stored information is used, **minimizing Inference time** to the time required for finding relevant data from the stored knowledge base.

- Most **conventional solutions doesn't recognize** the names of Genes, diseases, chemicals and simply ignore them, BUT this is where our approach focuses on. Our approach revolves around finding these Biomedical entities and relation between them.

Methodology

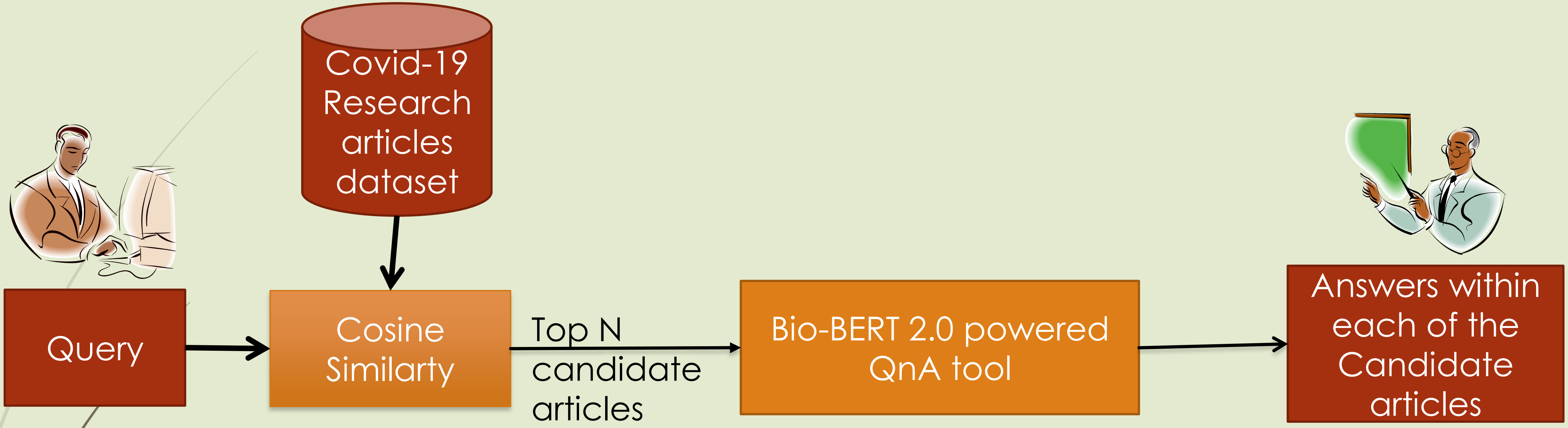
Two-way solution based on needs:

- **Q&A model** : to get probable answers within Top N candidate articles. The researcher then needs to go through the candidate articles manually.
- **Knowledge graph** : To get **direct to the point answers** and specifically designed for **Biomedical entities like Gene, Disease and Chemical**.

- **Q&A tool**

1. Firstly, we extract meta information of all articles related to covid 19 search keywords(Covid19, SARS COV2 etc.) from the CORD-19 open research dataset in a csv file.
2. Based on TF-IDF similarity score of question with the document's title, abstract and text. we select top N candidate articles based on the similarity score.
3. We then used the fine-tuned(BioBERT's Transfer learning on SQUAD2.0) model to get the answer for that question from these top N relevant articles.

Covid-19 Q&A tool



Covid-19 Research articles dataset

Query

Cosine Similarty

Top N candidate articles

Bio-BERT 2.0 powered QnA tool

Answers within each of the Candidate articles

- Most Q&A tools utilises BERT, which will simply consider important BioMedical entities like Chemical, Disease and Gene name as out of vocabulary.

- Instead of using BERT based QnA tool, we used Bio-BERT 2.0 which is obtained by Fine tuning the Bio-BERT model on SQUAD 2.0 (Stanford question answering dataset).

- Note that BioBERT is itself fine tuned for Medical terminologies corpus.

Details on Q&A model

- Dataset used : **SQUAD 2.0** (Stanford question answering dataset)

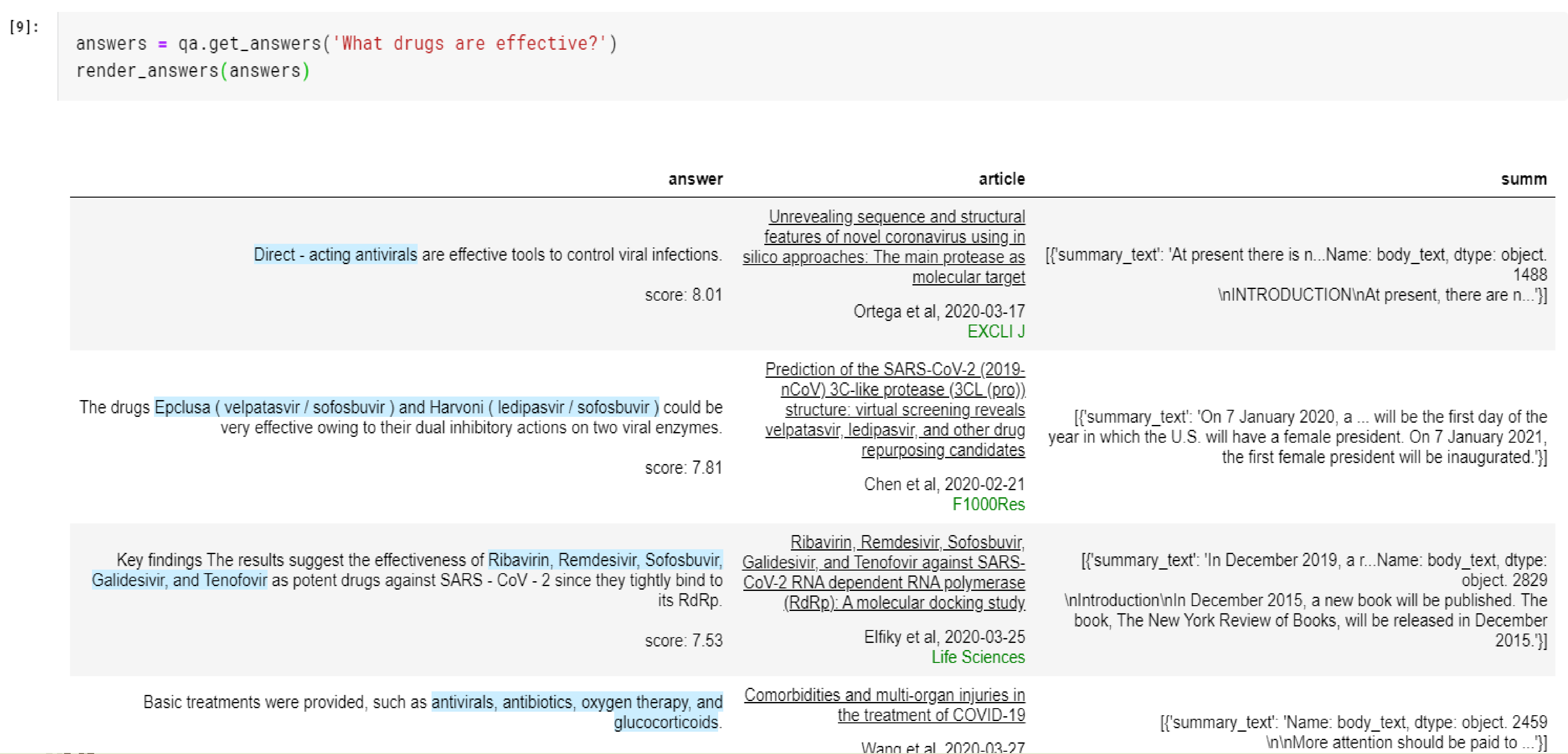
- Fine tuning **Bio-BERT** model on SQUAD 2.0 dataset.

- Initially, based on query, Top N candidate articles are selected from dataset based on similarity of query and metadata. Then, this articles are fed into Q&A tool to get probable answers for query in each document.

- BioBERT is Open source model trained on BioMedical corpus dataset so it can **recognise the Medical Ontology** like Gene, chemical and diseases.

- This BioBERT model is then **fine tuned on SQUAD 2.0** to get a Q&A model for Medical literature.

Q&A model Demo

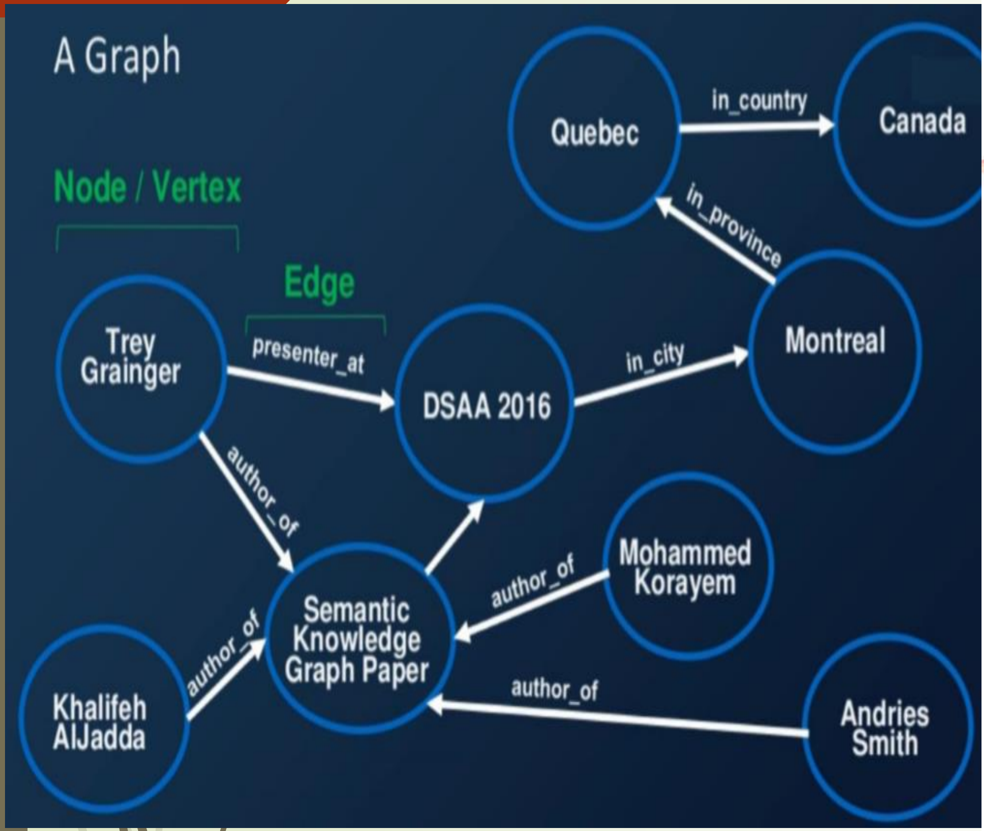


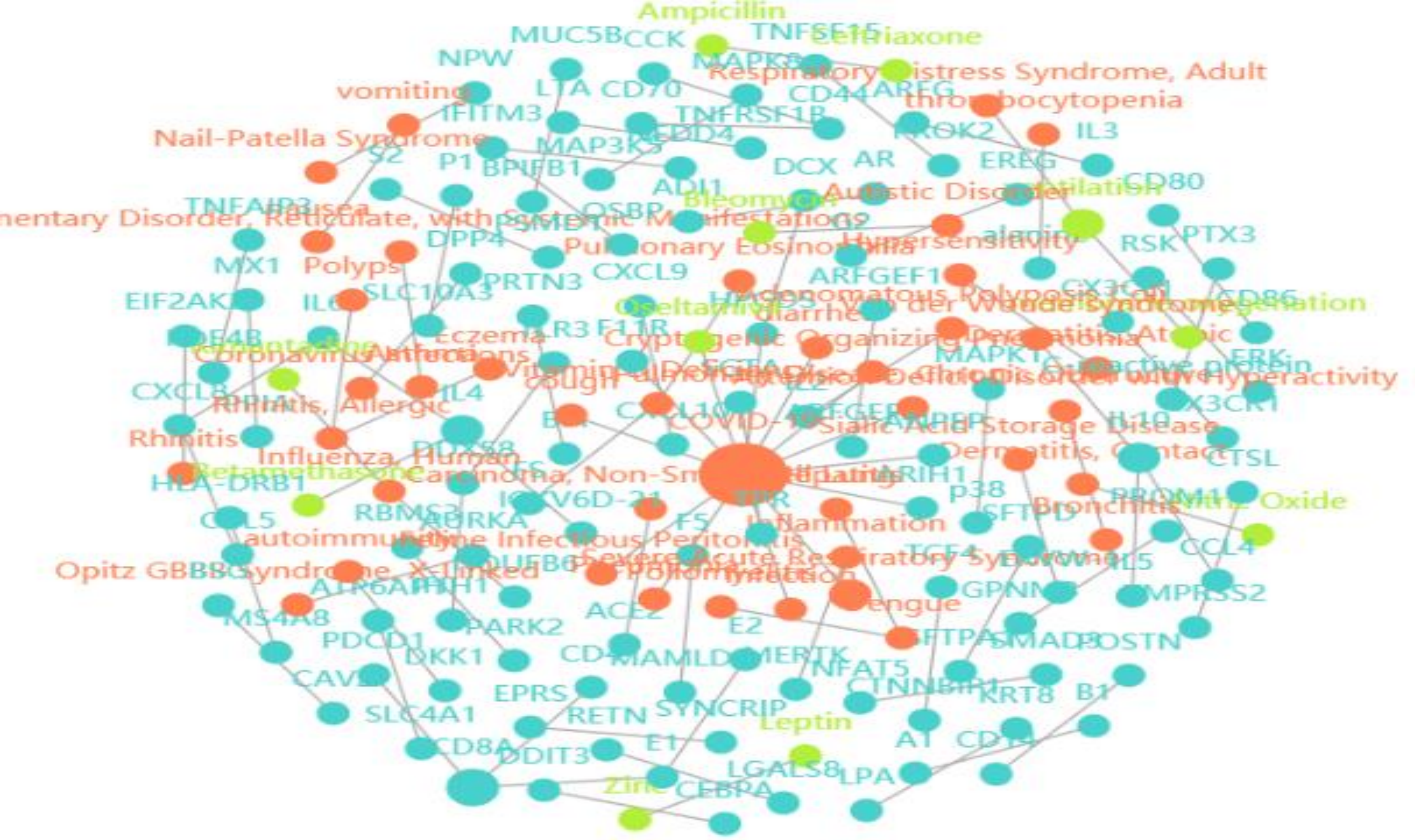
```
[9]: answers = qa.get_answers('What drugs are effective?')
     render_answers(answers)
```

answer	article	summ
Direct - acting antivirals are effective tools to control viral infections. score: 8.01	Unrevealing sequence and structural features of novel coronavirus using in silico approaches: The main protease as molecular target Ortega et al, 2020-03-17 EXCLI J	[{'summary_text': 'At present there is n...Name: body_text, dtype: object. 1488 \nINTRODUCTION\nAt present, there are n...'}]
The drugs Epclusa (velpatasvir / sofosbuvir) and Harvoni (ledipasvir / sofosbuvir) could be very effective owing to their dual inhibitory actions on two viral enzymes. score: 7.81	Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates Chen et al, 2020-02-21 F1000Res	[{'summary_text': 'On 7 January 2020, a ... will be the first day of the year in which the U.S. will have a female president. On 7 January 2021, the first female president will be inaugurated.'}]
Key findings The results suggest the effectiveness of Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir as potent drugs against SARS - CoV - 2 since they tightly bind to its RdRp. score: 7.53	Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study Elfiky et al, 2020-03-25 Life Sciences	[{'summary_text': 'In December 2019, a r...Name: body_text, dtype: object. 2829 \nIntroduction\nIn December 2015, a new book will be published. The book, The New York Review of Books, will be released in December 2015.'}]
Basic treatments were provided, such as antivirals, antibiotics, oxygen therapy, and glucocorticoids.	Comorbidities and multi-organ injuries in the treatment of COVID-19 Wang et al, 2020-03-27	[{'summary_text': 'Name: body_text, dtype: object. 2459 \n\nMore attention should be paid to ...'}]

The Relevant article, with answer highlighted, article link given and Summary also produced Based on Query

Knowledge graphs



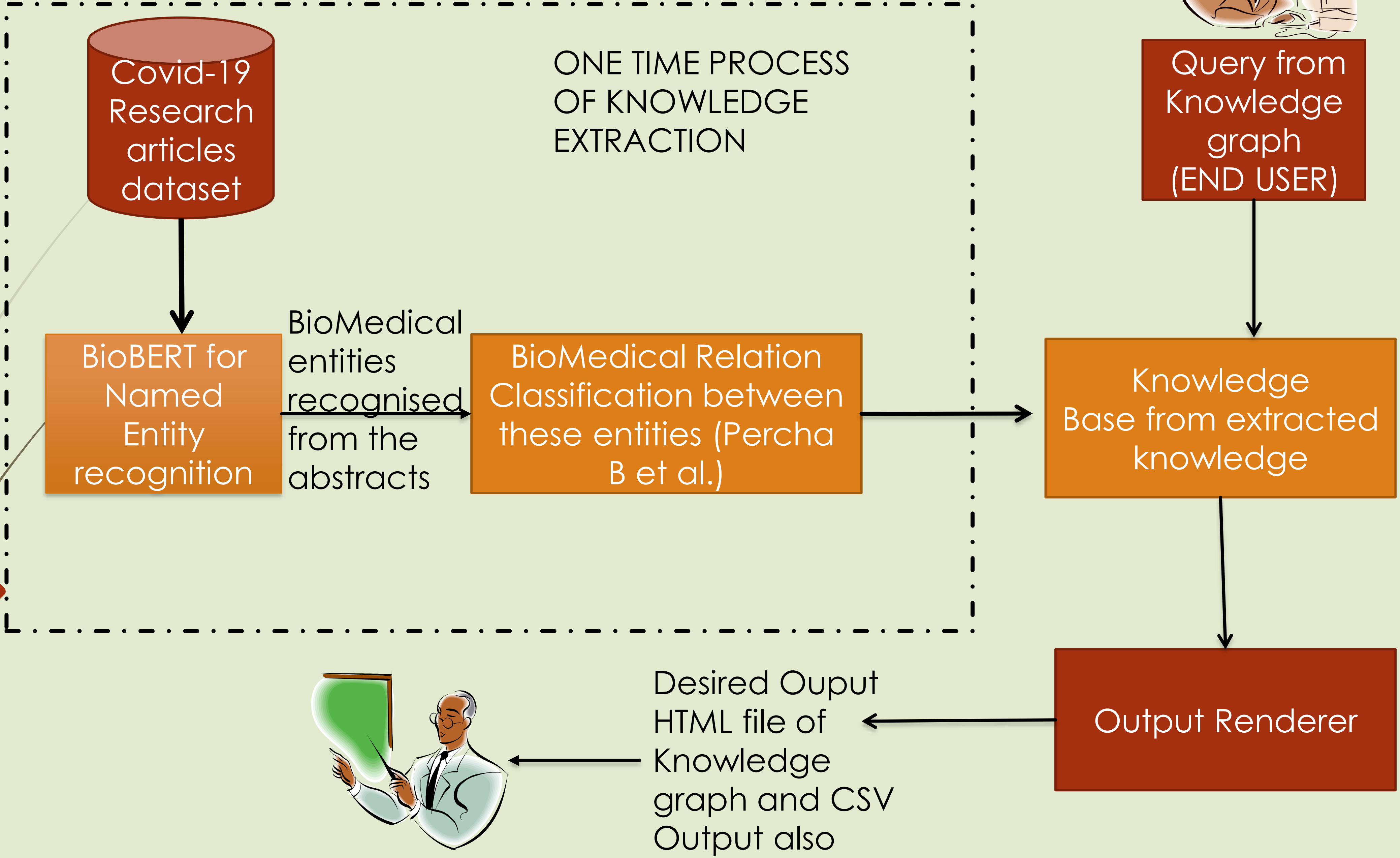


Other works : Knowledge graph of Meta data only like author details

Our work: Knowledge graph based on BioMedical relationships

- Knowledge graphs **resembles the way our human brain percieves** information from the text.
- Relations between chemicals, genes and disease offer **insights into the mechanisms behind higher order biochemical phenomena**, such as **drug-drug interactions, drug response** and **gene-disease** associations.

Knowledge graphs for Covid-19



ONE TIME PROCESS OF KNOWLEDGE EXTRACTION

Covid-19 Research articles dataset

BioBERT for Named Entity recognition

BioMedical entities recognised from the abstracts

BioMedical Relation Classification between these entities (Percha B et al.)

Query from Knowledge graph (END USER)

Knowledge Base from extracted knowledge

Output Renderer

Desired Ouput HTML file of Knowledge graph and CSV Output also

Details on Knowledge graphs

- Dataset used : **CORD-19 Dataset** of 36K research articles on N-Covid19.

- BioBert: Model trained for **Biomedical entity recogntion** : Chemical, Gene, Disease.The sentences containing more than two biomedical entities(Chemical,Gene,Disease) were extracted and marked. Trained on **NCBI-Disease dataset.**

- **Biomedical relation classifier** model : Marked sentences of above step given input here to Classify relation between detected entities to get the triples for knowledge graph creation.

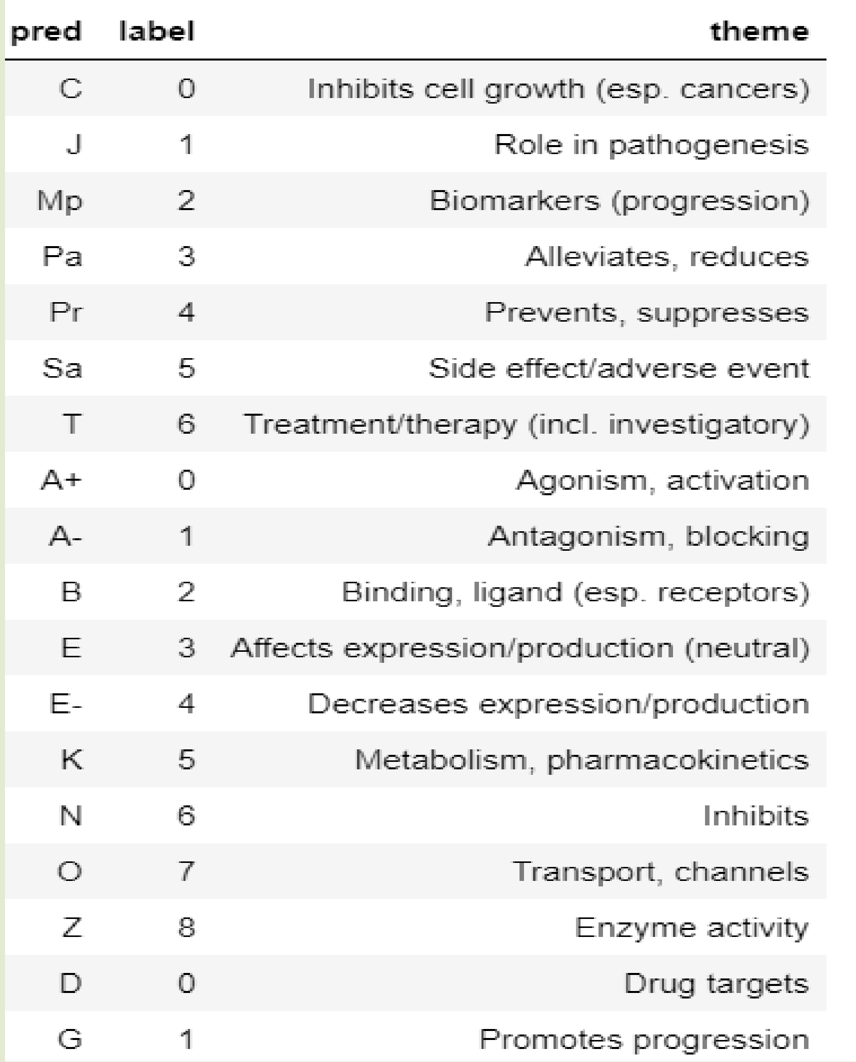
- Knowledge graph is a **collection of triples** like 1st entity -> Relation -> 2nd Entitty eg: **Hydroxycholoquine -> Treatment/therapy -> Covid-19**

- NER results : accuracy: 98.49%; precision: 86.67%; recall: 88.75%; F1: 87.70

- Relationship extraction model : f1 score:83.74% ; recall: 90.75% ;precision: 77.74%; specificity : 71.15%

BioMedical Relationships considered

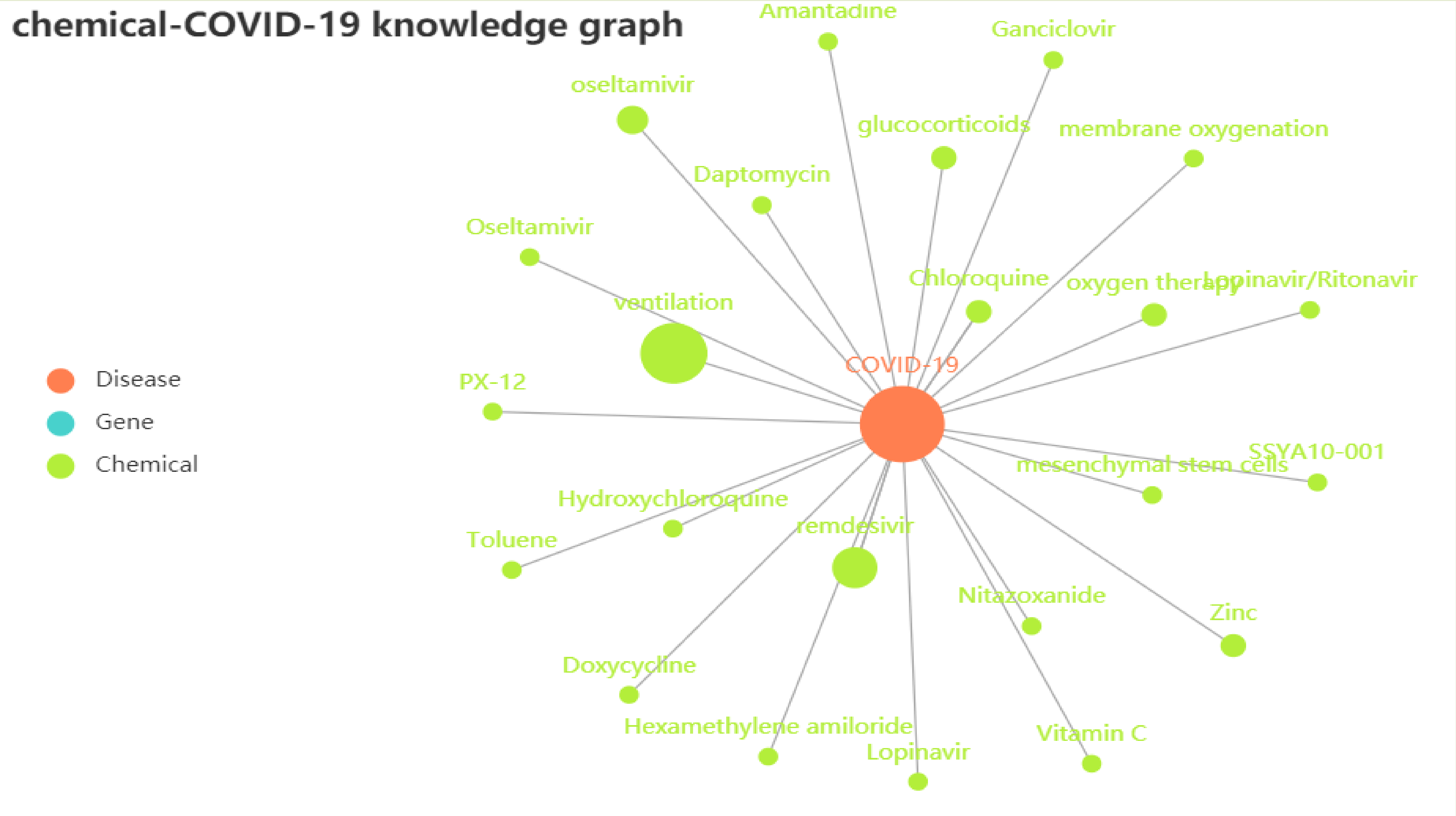
- Chemical-disease : ['T', 'C', 'Sa', 'Pr', 'Pa', 'J'],
- disease-chemical': ['Mp']
- chemical-gene: ['A+', 'A-', 'B', 'E+', 'E-', 'E', 'N'],
- gene-chemical': ['O', 'K', 'Z'],
- gene-disease': ['U', 'Ud', 'D', 'J', 'Te', 'Y', 'G'],
- disease-gene': ['Md', 'X', 'L'],
- gene-gene': ['B', 'W', 'V+', 'E+', 'E', 'I', 'H', 'Rg', 'Q']



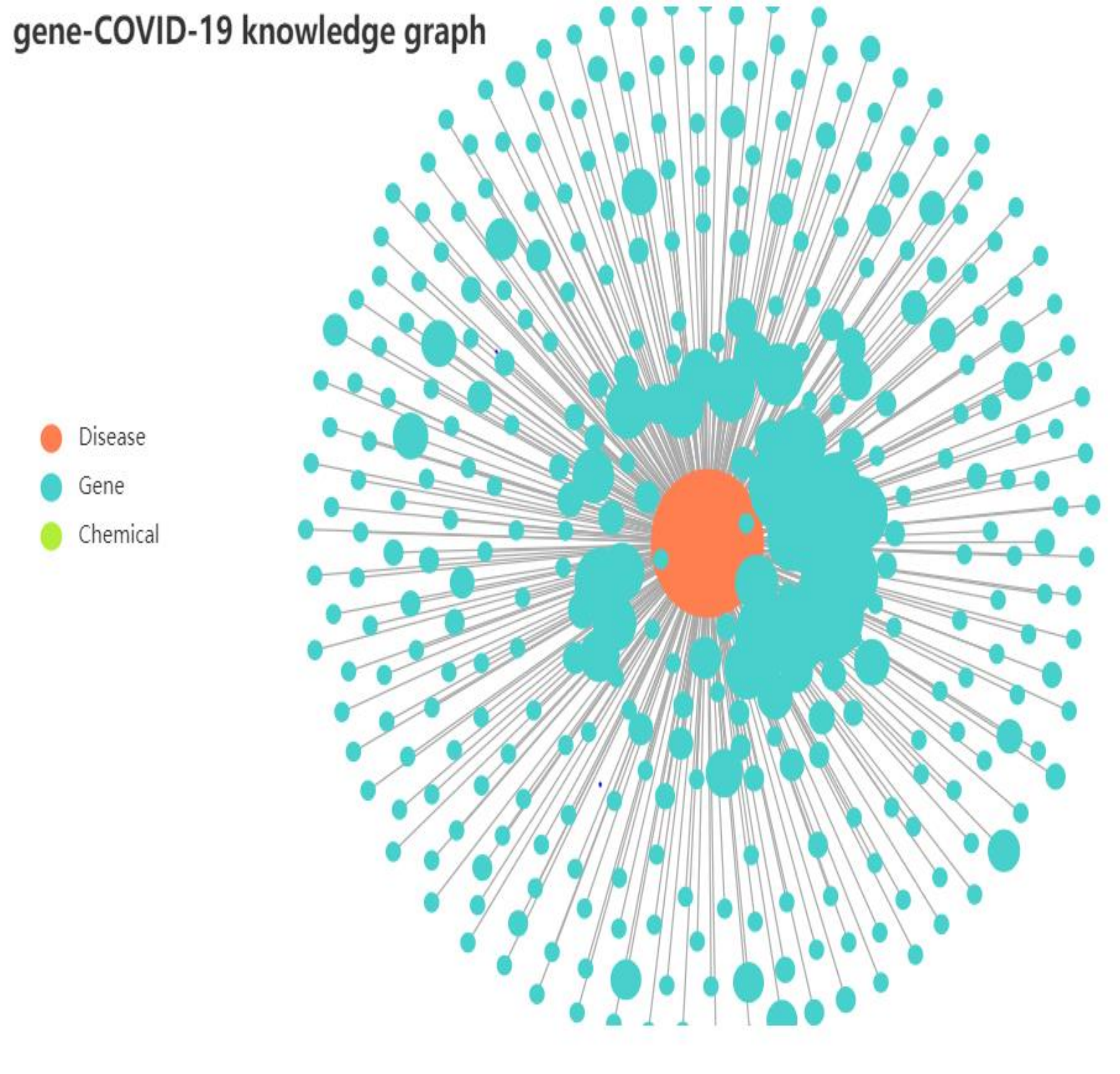
pred	label	theme
C	0	Inhibits cell growth (esp. cancers)
J	1	Role in pathogenesis
Mp	2	Biomarkers (progression)
Pa	3	Alleviates, reduces
Pr	4	Prevents, suppresses
Sa	5	Side effect/adverse event
T	6	Treatment/therapy (incl. investigatory)
A+	0	Agonism, activation
A-	1	Antagonism, blocking
B	2	Binding, ligand (esp. receptors)
E	3	Affects expression/production (neutral)
E-	4	Decreases expression/production
K	5	Metabolism, pharmacokinetics
N	6	Inhibits
O	7	Transport, channels
Z	8	Enzyme activity
D	0	Drug targets
G	1	Promotes progression

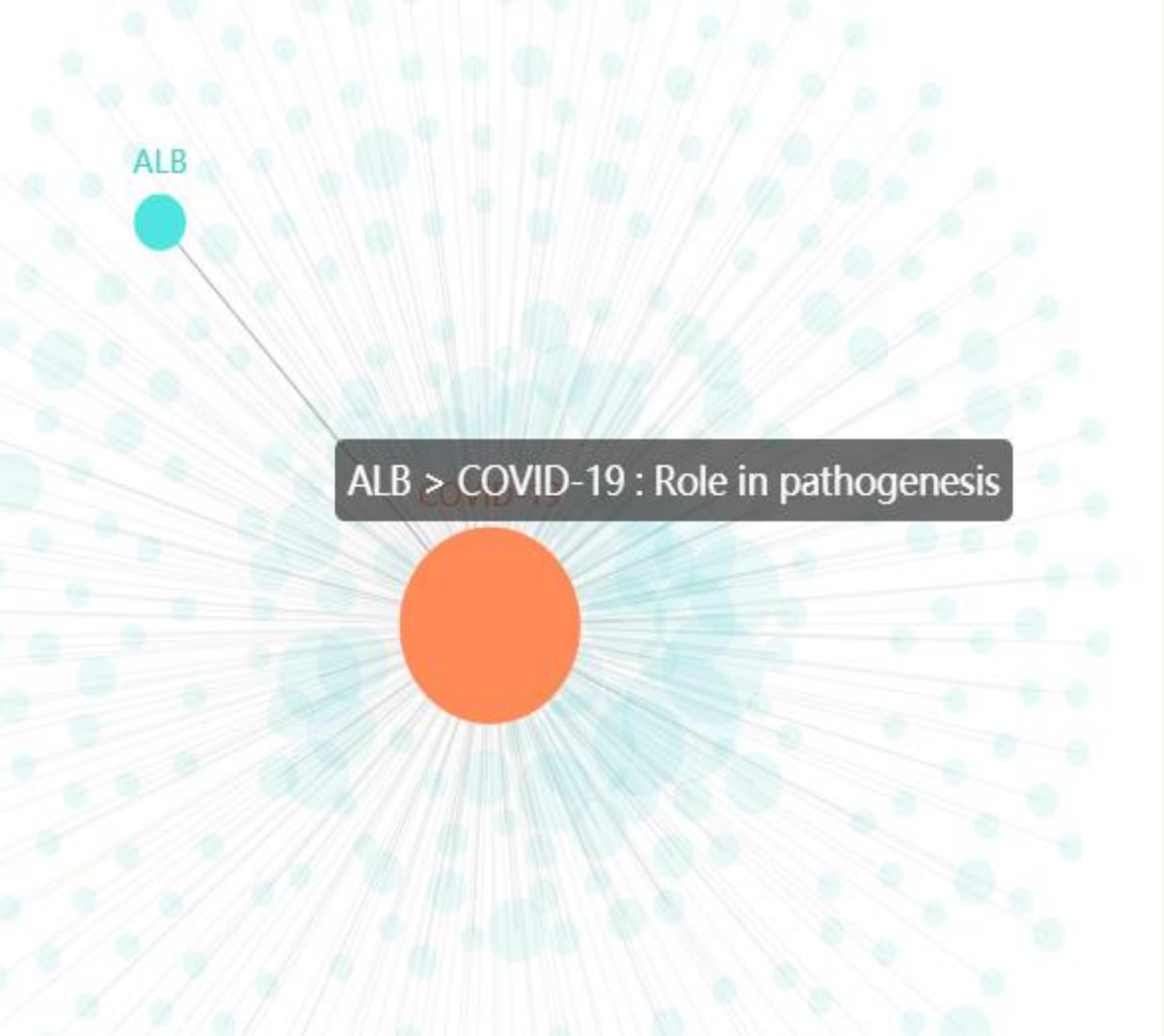
Table : Subset of Types of relationship detected

Chemical-covid 19 knowledge graph



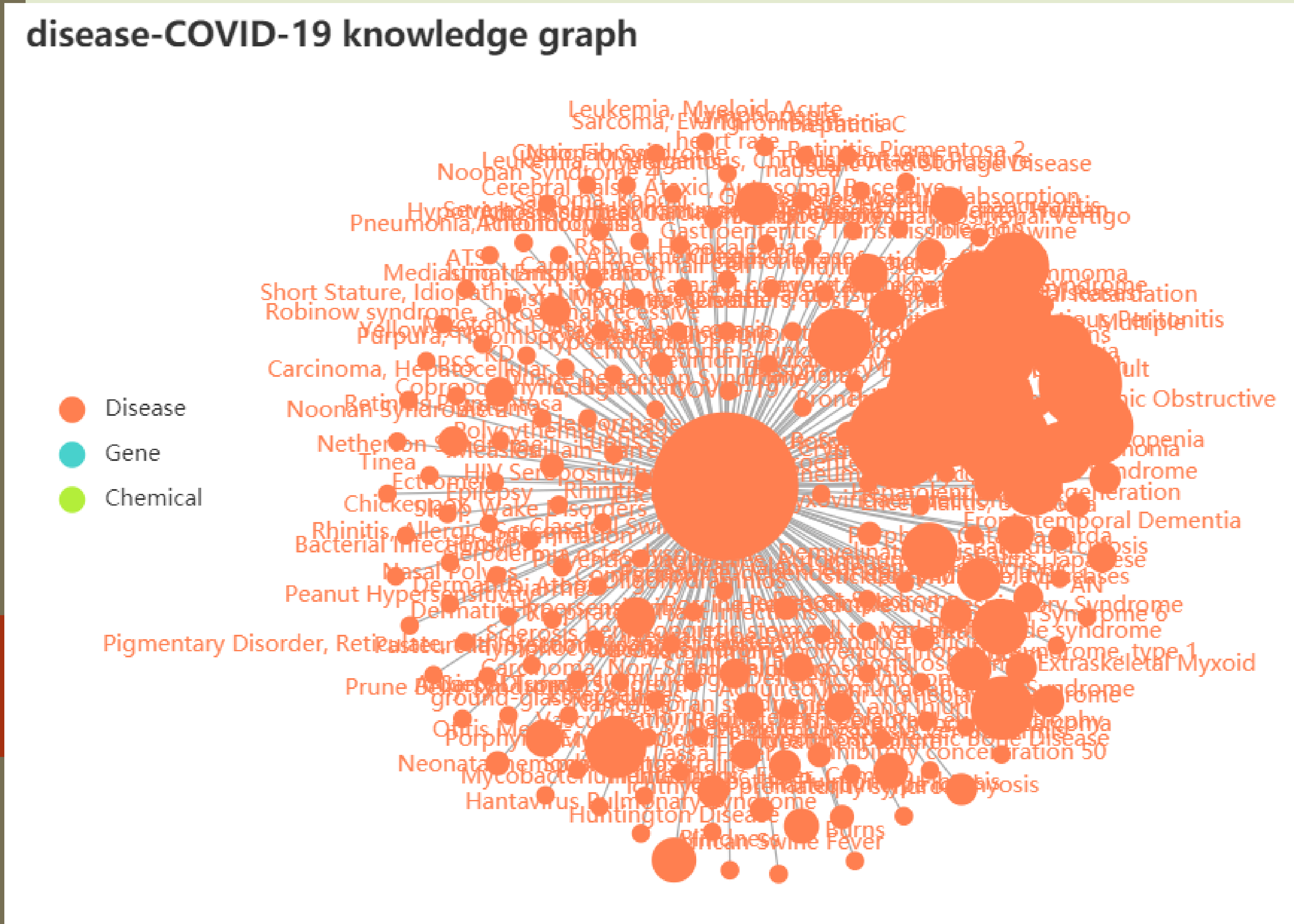
Gene-Covid 19 knowledge graph

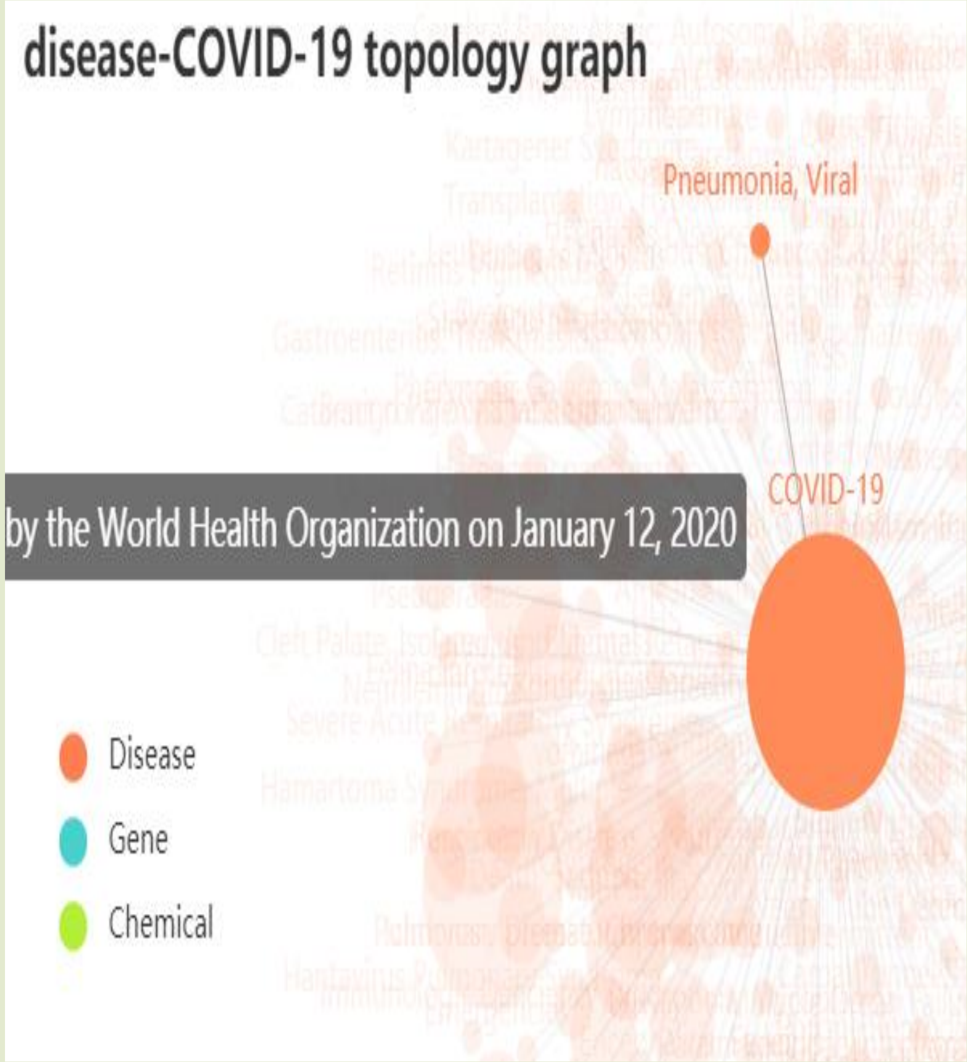




It shows that "ALB" gene "Role in pathogenesis" to "Covid19"

Disease-Covid 19 knowledge graph

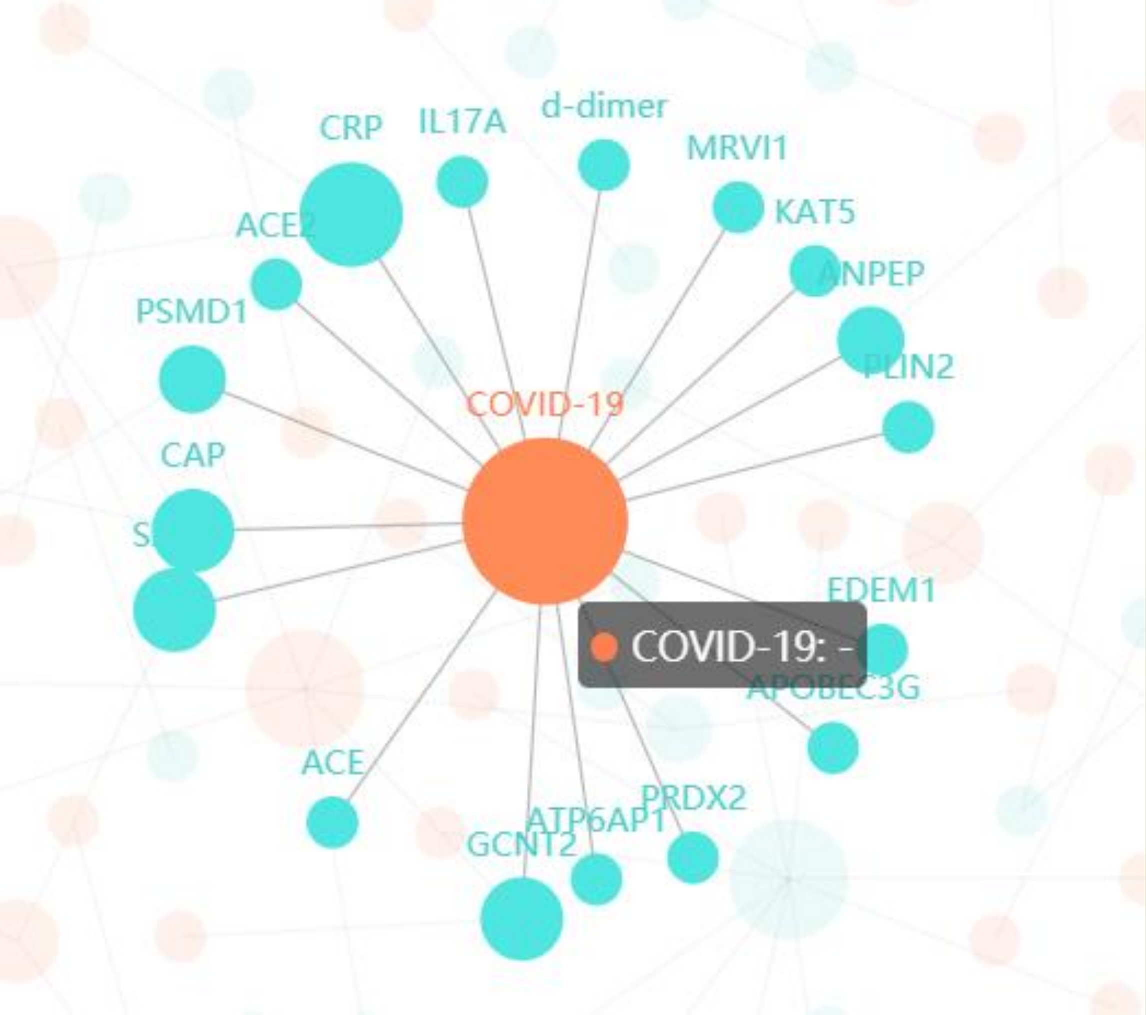




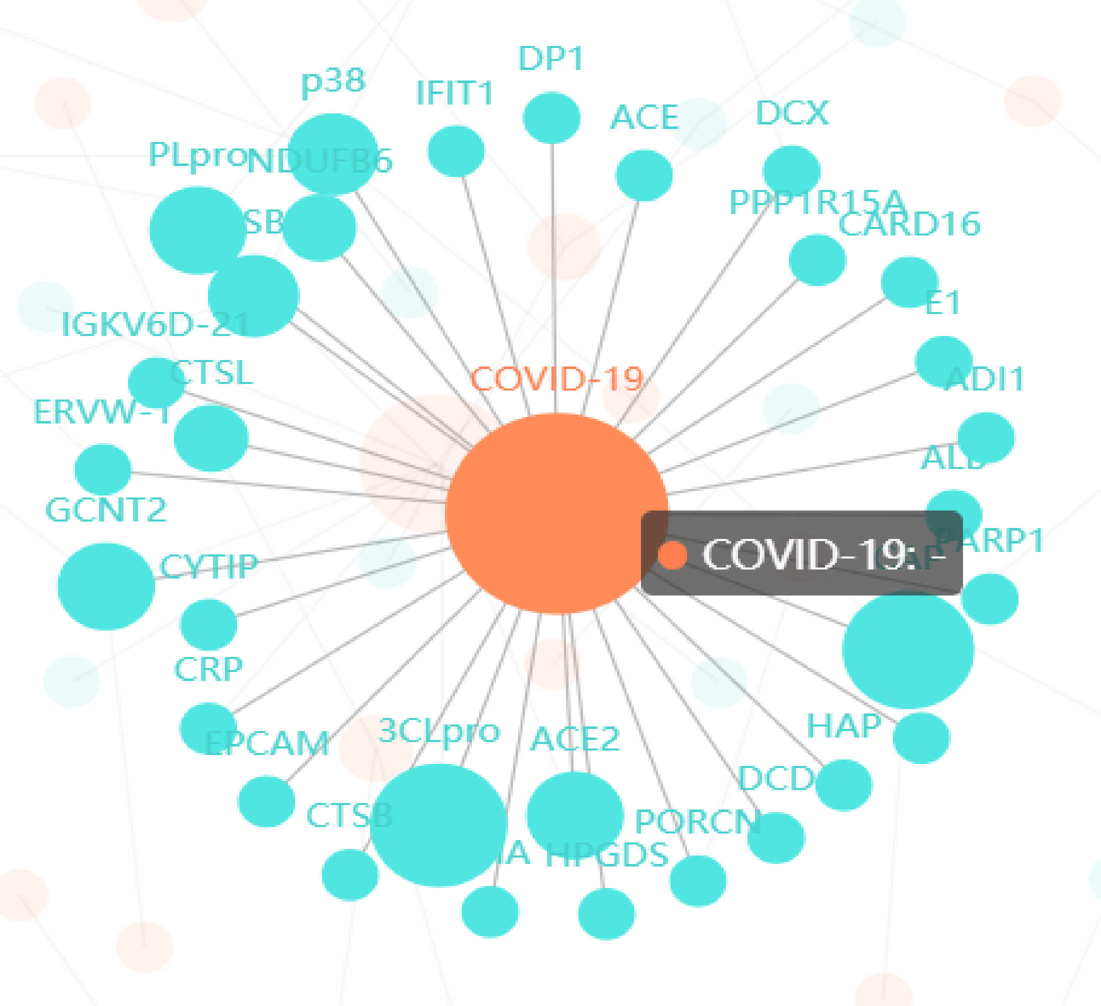
by the World Health Organization on January 12, 2020

It shows that "Pneumonia, viral" disease related as "WHO declaration on Jan 12,2020 " to "Covid19"

Custom query: want to know Drug targets and biomarkers(diagnostics) for covid 19

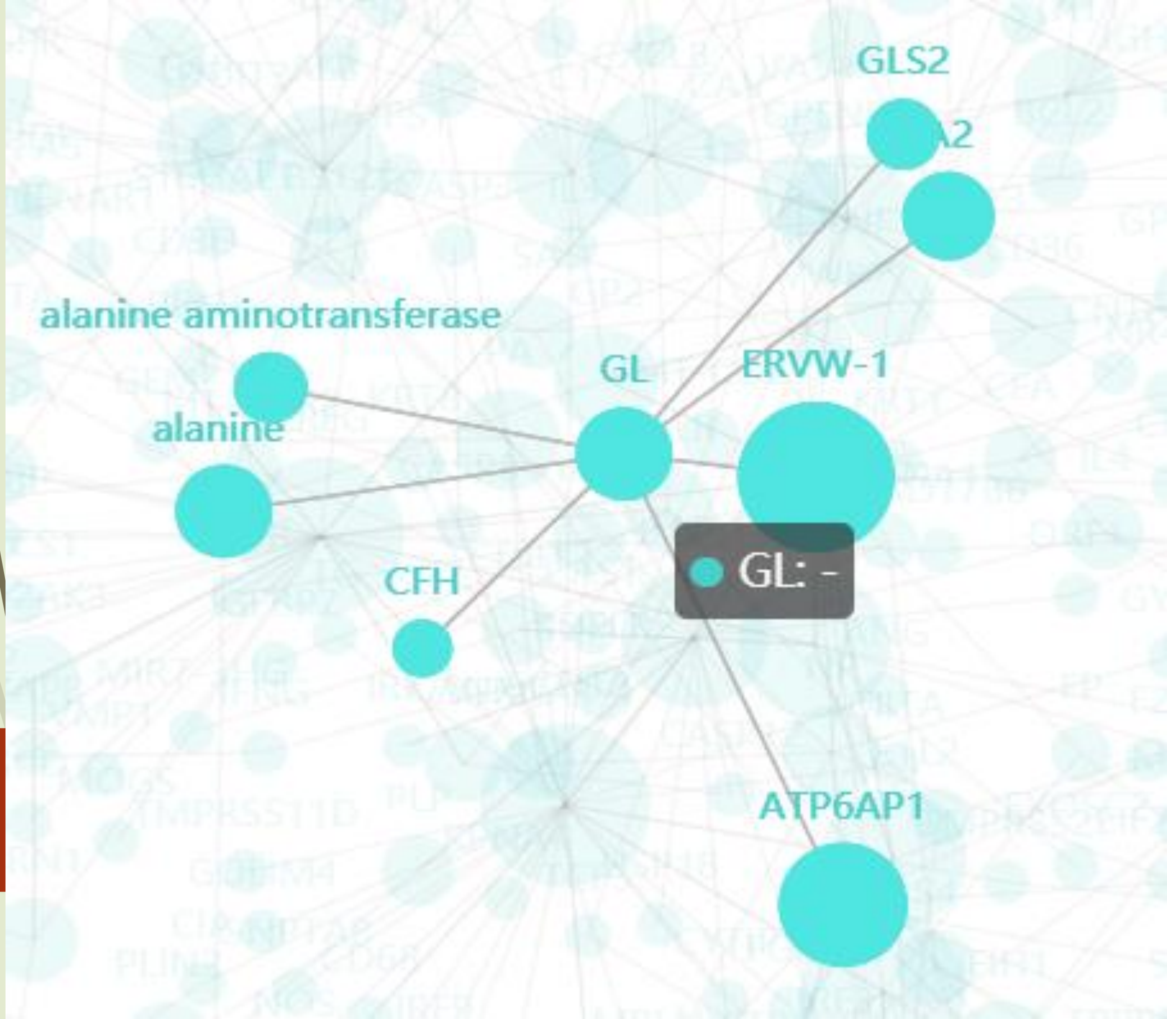


Select relation type as "Biomarkers(Diagnostic)" which is one of the supported relation and then Hover over Covid-19 in produced knowledge graph

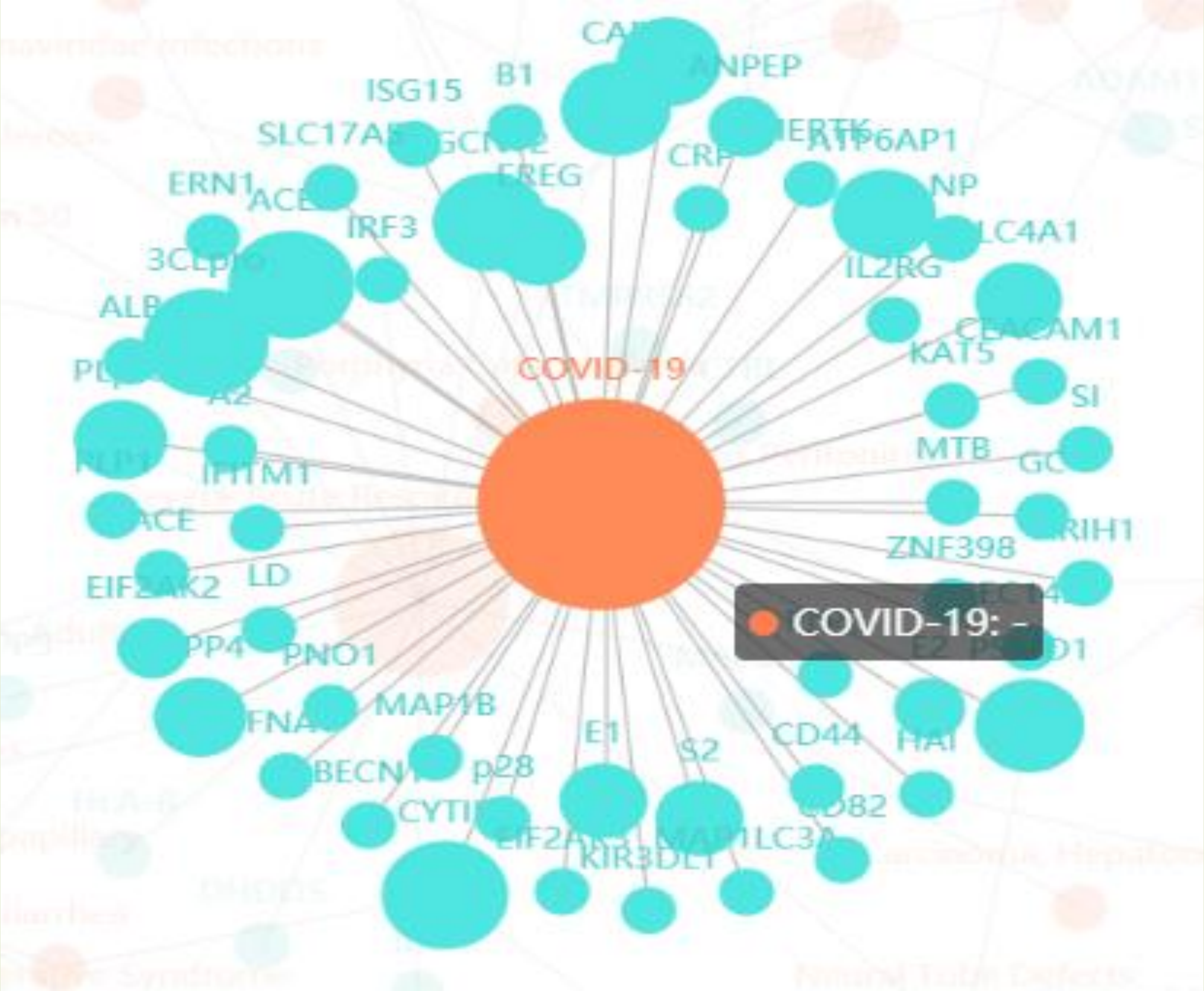


Select relation type as "Drug Targets" which is one of the supported relation and then Hover over Covid-19 in produced knowledge graph

Custom query: want to get relations like “Same protein or complex" and “Mutations affect disease course" for covid 19



Select relation type as “Same protein or complex" and hover over Gene of Interest



Select relation type as “Mutations affect disease course" and hover over Gene of Interest

Conclusion

- Biomedical literature access is fundamental for several types of users including biomedical researchers and clinicians. Particularly in situations like COVID 19 pandemic, where huge new information is being added in daily basis, tools like text mining can help to mine **<u>relevant information within minimal of time</u>**.

- Text mining **makes relationships between genes, diseases, bio markers and chemicals**, **computationally mappable** and thus helps to dig out **<u>“hidden”</u> relationships** which **<u>aren’t described distinctly</u>** in published literature.

- Knowledge graphs are more useful than conventional Q&A system as this solution gives **<u>to the point answer</u>** for most useful BioMedical queries.

Future directions

- Getting a **India specific knowledge graph** from articles published within India as they better represent the Information about Indians.
- Our vision is to develop open source tools for **Bio-Medical knowledge discovery** for Indian researchers.
- Developing an interface/mechanism so that **new papers can be added** to the knowledge base by any non-technical user.
- Integrating our model with apps like **Aarogya setu** etc.
- Knowledge graphs can be built on similar lines for other topics of Medical research like **Cancer, TB** etc.
- Q&A model **deployment on TensorRT engine** for faster Inference.

References

1. SQUAD 2.0 dataset https://rajpurkar.github.io/SQuAD-explorer/
2. BioBERT https://github.com/dmis-lab/biobert
3. BioSNAP http://snap.stanford.edu/biodata/
4. BioMedical relation classification https://github.com/jxzly/Biomedical-Relation-Classification
5. Medium.com
6. CORD-19 Dataset https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
7. Kernels at kaggle.com
8. NVIDIA TensorRT optimization https://developer.nvidia.com/tensorrt

Thank You !