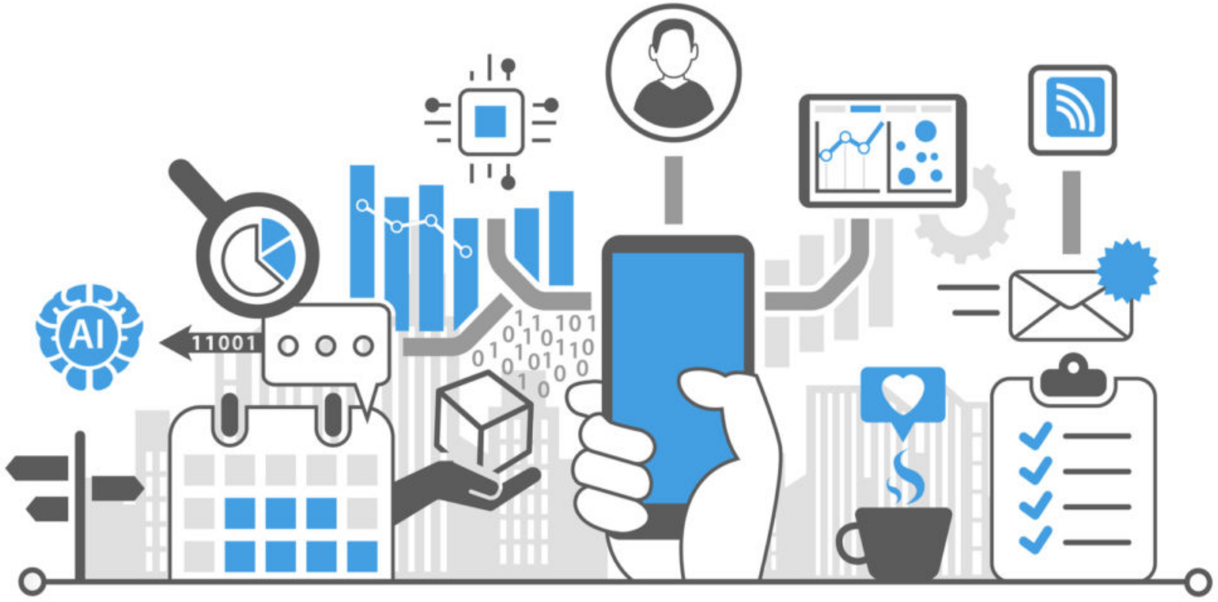
Scientific Text Mining and Knowledge Graphs

Meng Jiang and Jingbo Shang

University of Notre Dame University of California, San Diego

mjiang2@nd.edu jshang@ucsd.edu

Mining Knowledge from **Big Data**



Big Data

Structured Knowledge
& Insights

Massive Unstructured Text Data

News

Social Media

Business & Finance

Scientific Papers

Medical Records

......

Massive Unstructured Text Data

News

Social Media

Business & Finance

Scientific Papers

Medical Records

......

Goal: Texts → Knowledge & Insights

- **Traditional methods** rely on ***extensive annotations from domain experts***

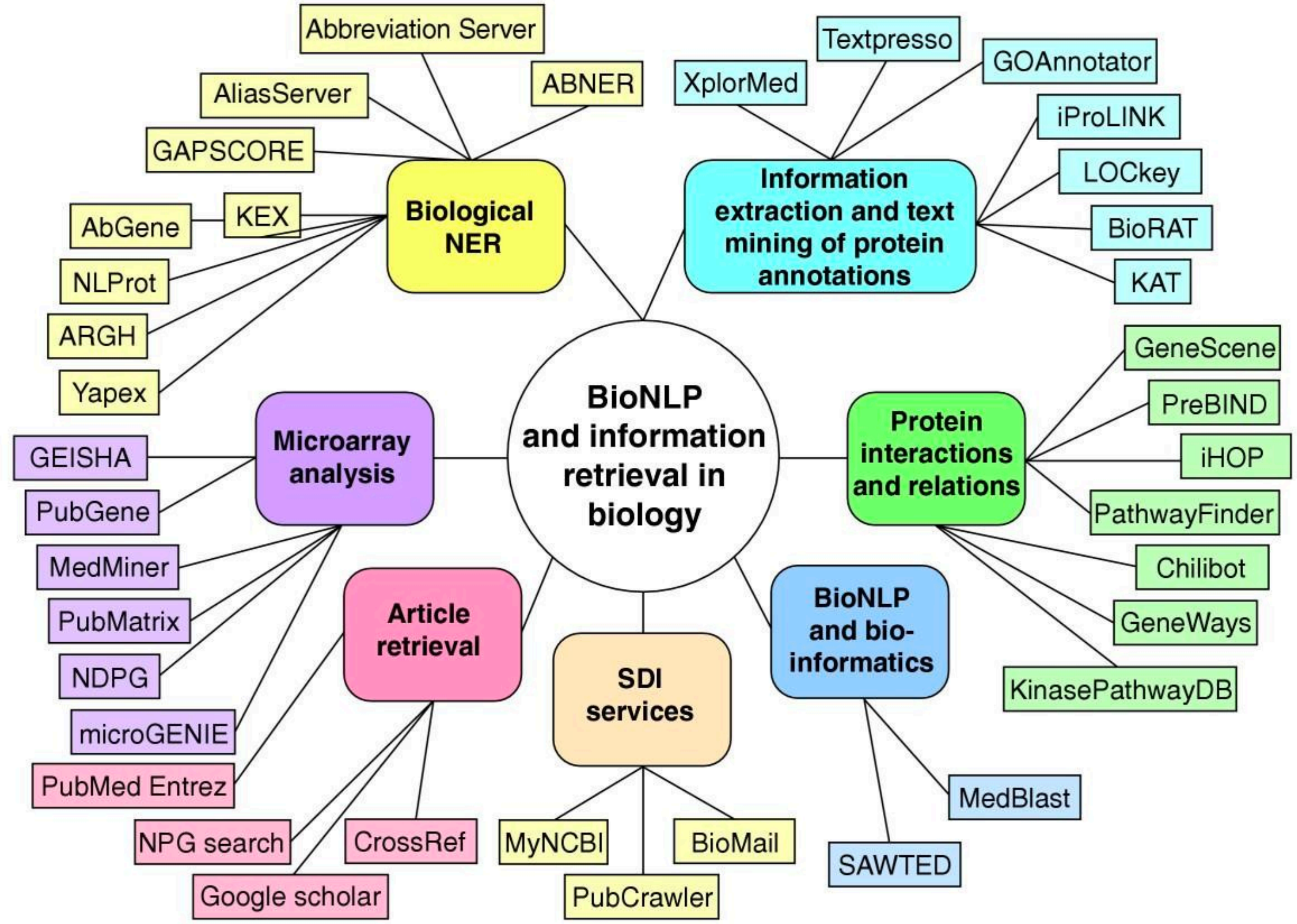
Unstructured Text Data

Automated Models

Structured Knowledge & Insights

- **This tutorial** focuses on an ***automatic*** way
- “Automatic” – using public knowledge bases only

An Example: BioNLP



BioNLP and information retrieval in biology

Biological NER
Abbreviation Server
AliasServer
ABNER
GAPSCORE
KEX
AbGene
NLProt
ARGH
Yapex

Information extraction and text mining of protein annotations
XplorMed
Textpresso
GOAnnotator
iProLINK
LOCkey
BioRAT
KAT

Protein interactions and relations
GeneScene
PreBIND
iHOP
PathwayFinder
Chilibot
GeneWays
KinasePathwayDB

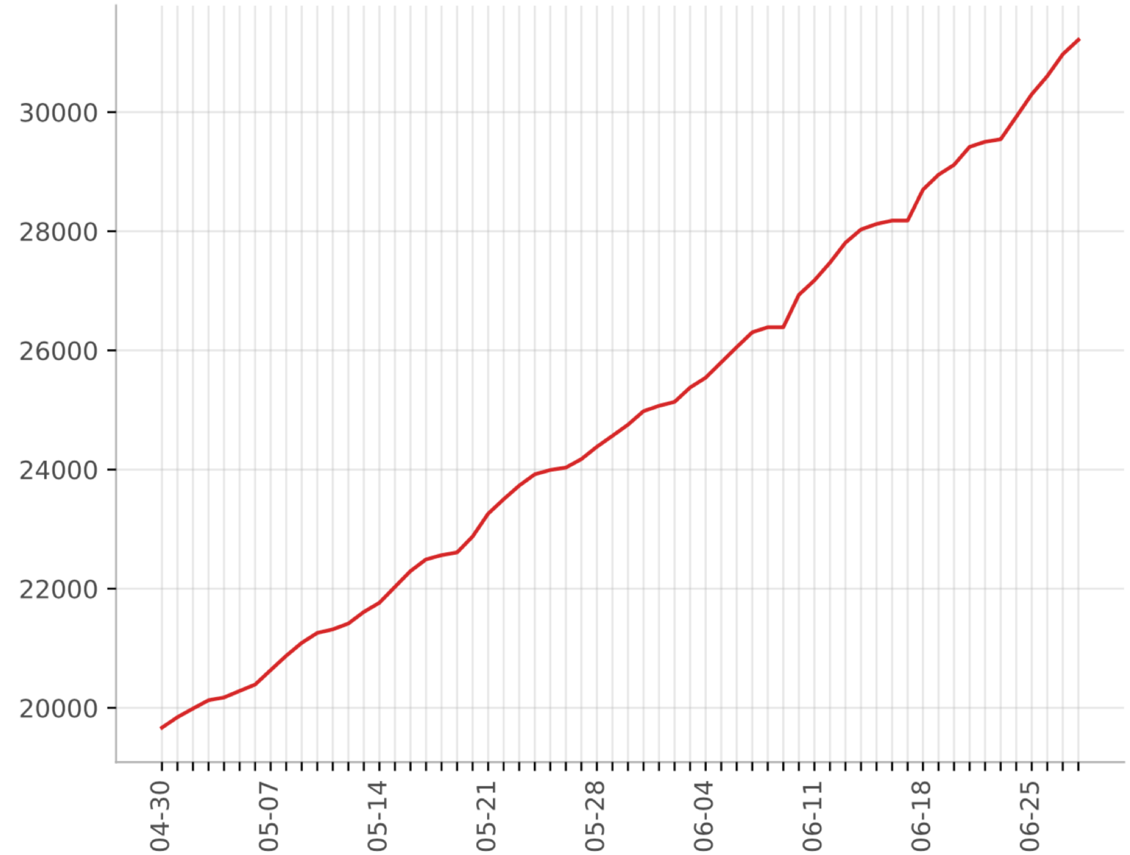
BioNLP and bio-informatics
MedBlast
SAWTED

SDI services
MyNCBI
PubCrawler
BioMail

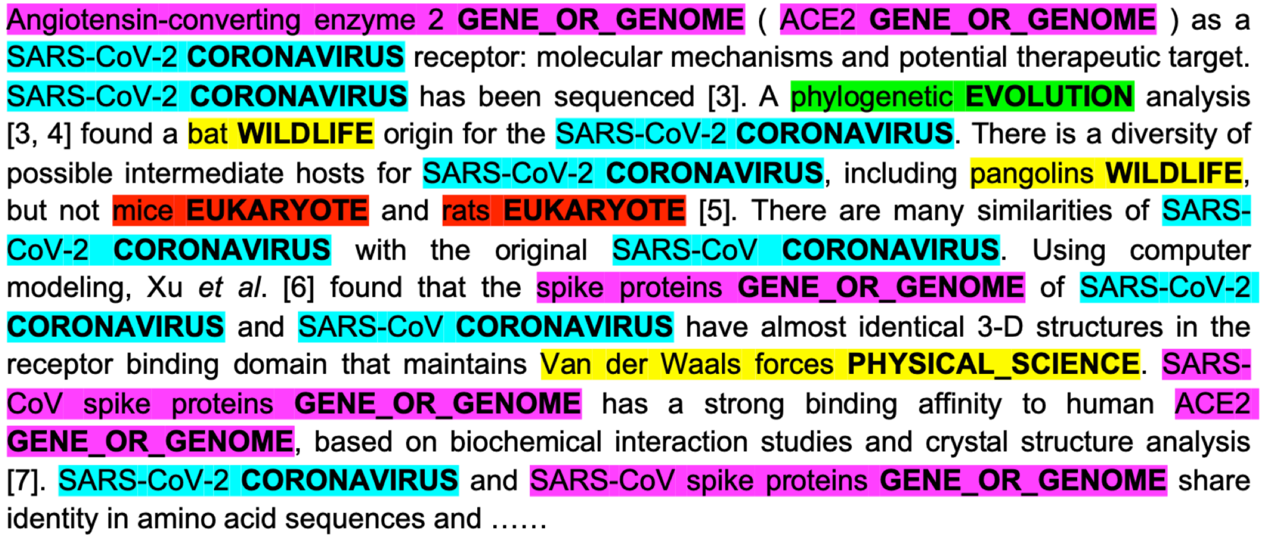
Article retrieval
PubMed Entrez
NPG search
Google scholar
CrossRef

Microarray analysis
GEISHA
PubGene
MedMiner
PubMatrix
NDPG
microGENIE

COVID-19 Research



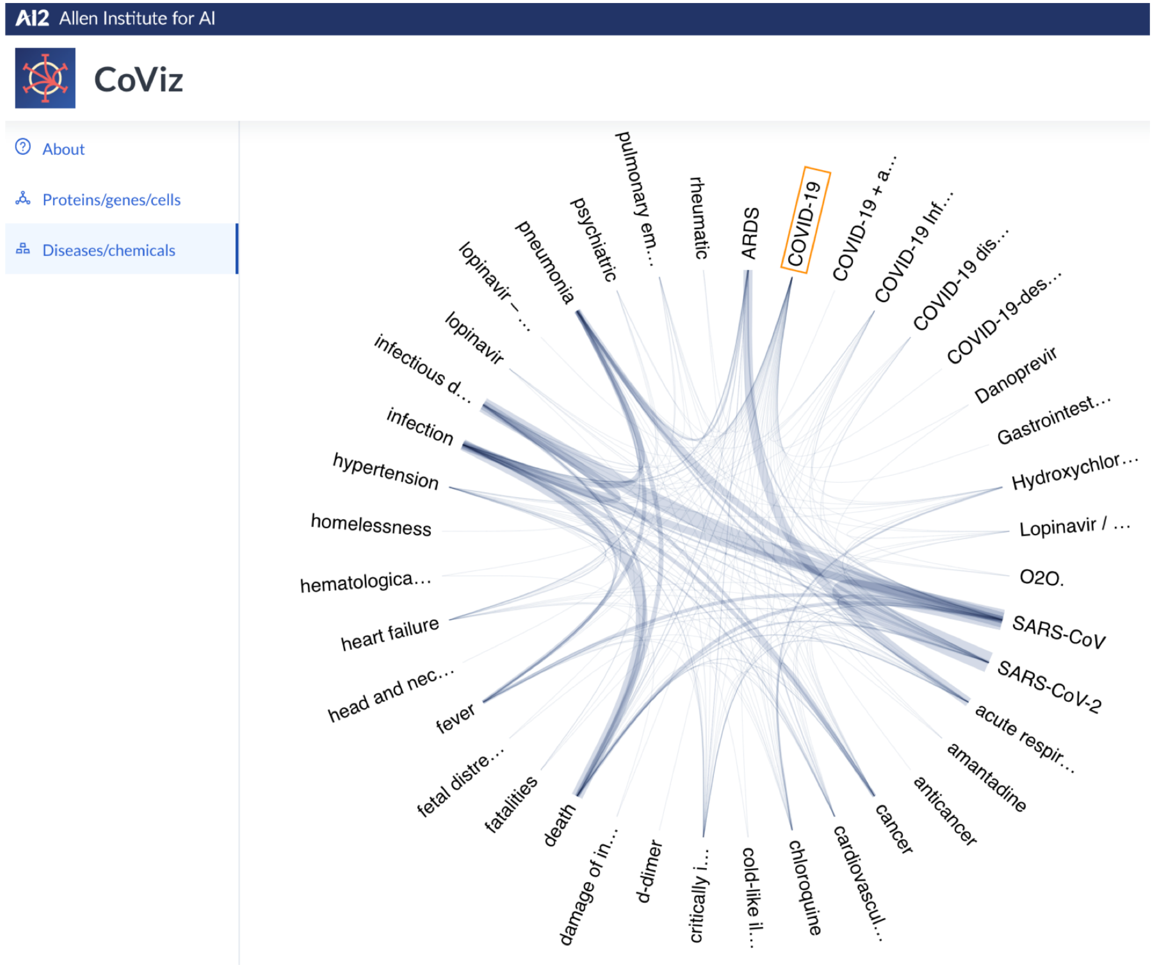
The Growing Number of COVID-19 Papers at PubMed

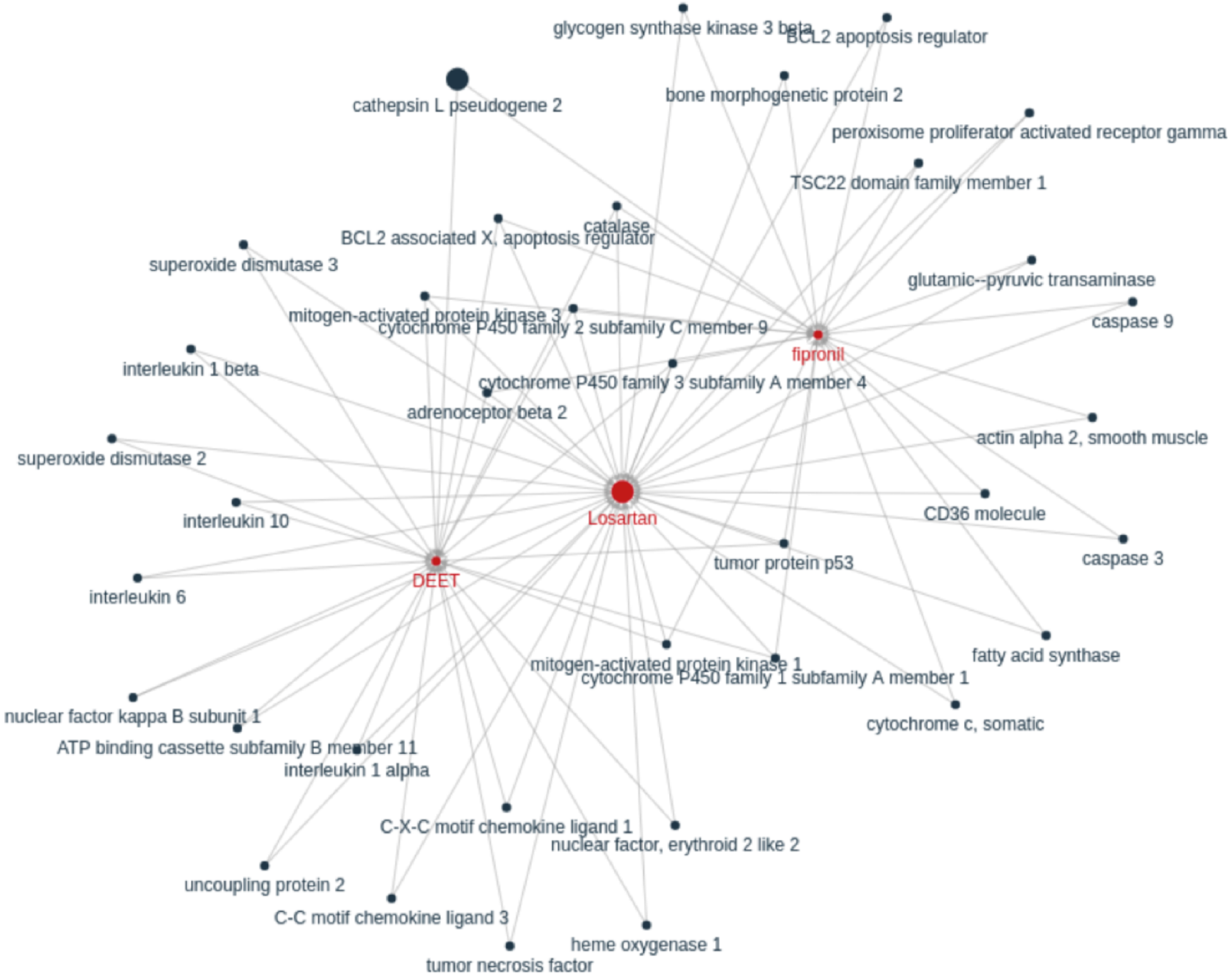

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3]. A phylogenetic **EVOLUTION** analysis [3, 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS**. There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS**, including pangolins **WILDLIFE**, but not mice **EUKARYOTE** and rats **EUKARYOTE** [5]. There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS**. Using computer modeling, Xu *et al*. [6] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE**. SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME**, based on biochemical interaction studies and crystal structure analysis [7]. SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and ……

Scientific Named Entity Recognition and Typing

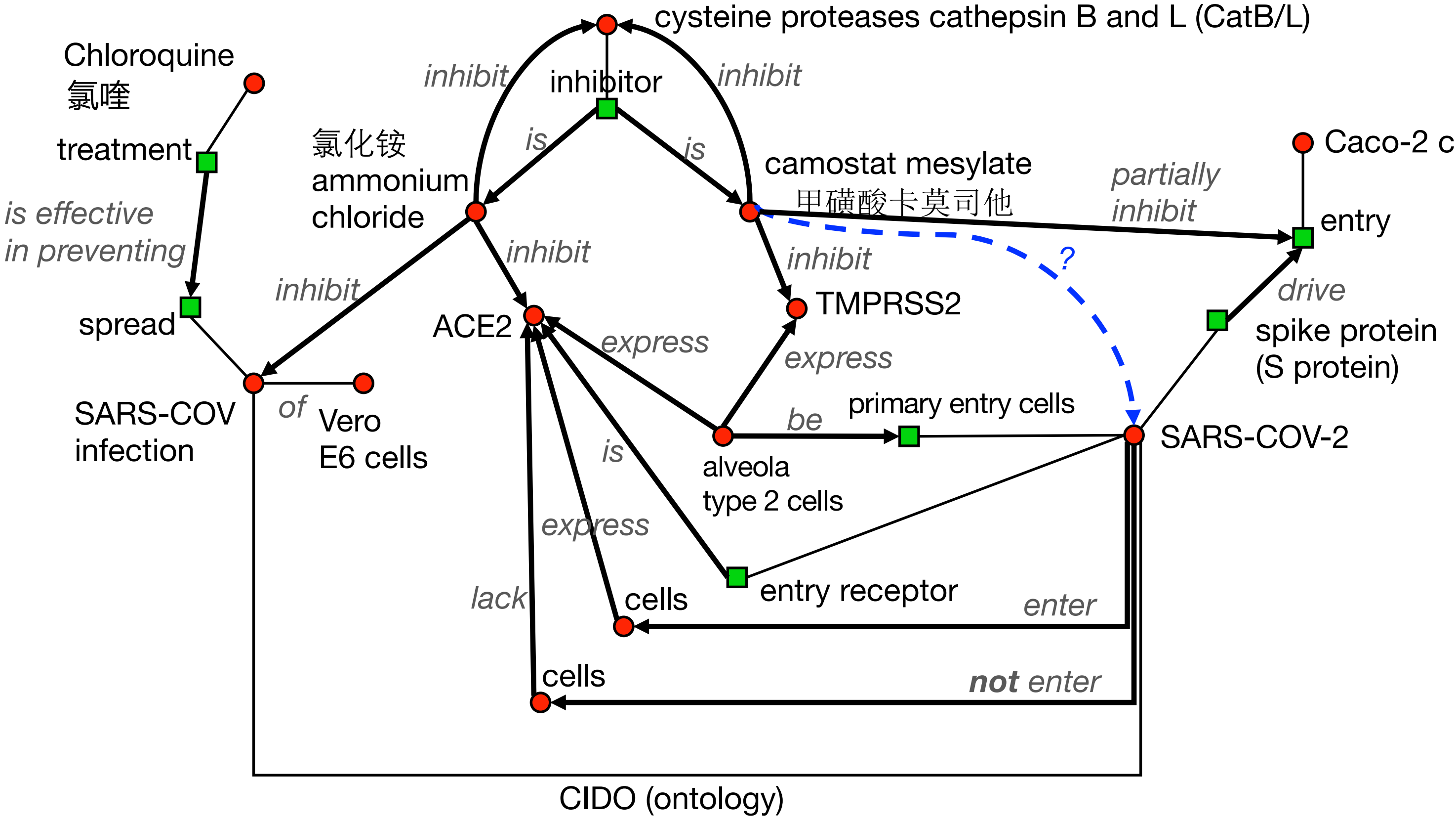
Wang et al. “COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation”

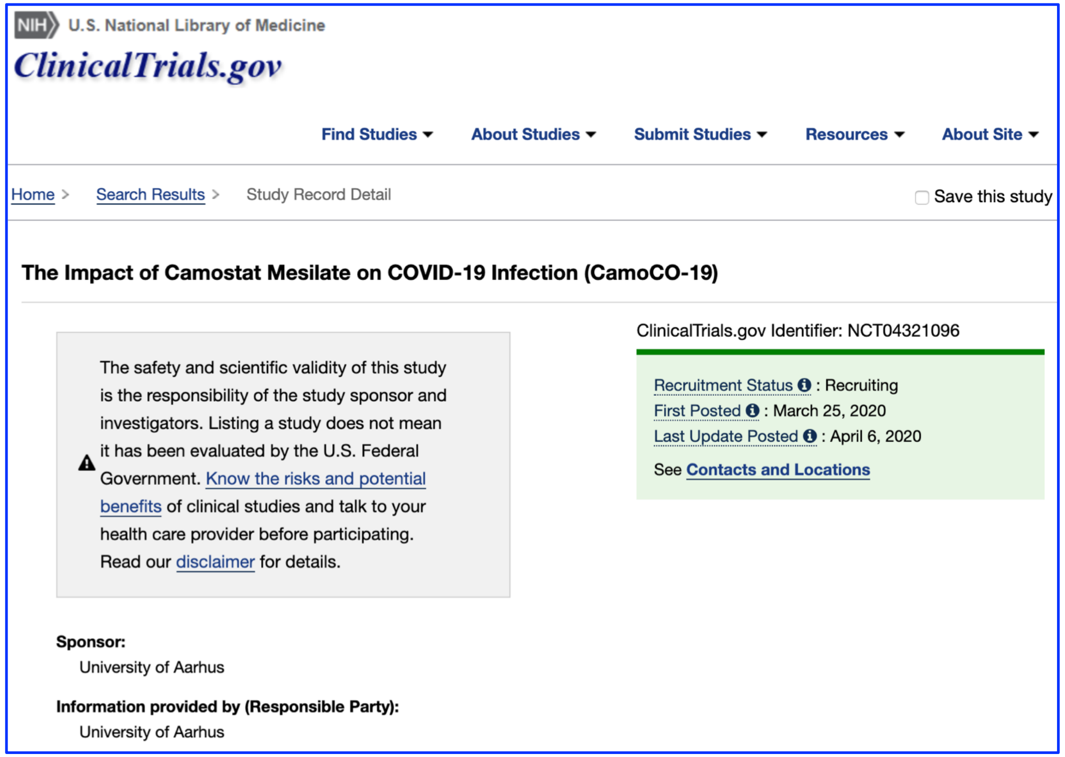
Knowledge Graphs in COVID-19 Research




glycogen synthase kinase 3 beta
BCL2 apoptosis regulator
cathepsin L pseudogene 2
bone morphogenetic protein 2
peroxisome proliferator activated receptor gamma
TSC22 domain family member 1
catalase
BCL2 associated X, apoptosis regulator
superoxide dismutase 3
glutamic--pyruvic transaminase
mitogen-activated protein kinase 3
cytochrome P450 family 2 subfamily C member 9
caspase 9
fipronil
interleukin 1 beta
cytochrome P450 family 3 subfamily A member 4
adrenoceptor beta 2
actin alpha 2, smooth muscle
superoxide dismutase 2
Losartan
CD36 molecule
interleukin 10
tumor protein p53
caspase 3
interleukin 6
DEET
fatty acid synthase
mitogen-activated protein kinase 1
cytochrome P450 family 1 subfamily A member 1
nuclear factor kappa B subunit 1
cytochrome c, somatic
ATP binding cassette subfamily B member 11
interleukin 1 alpha
C-X-C motif chemokine ligand 1
nuclear factor, erythroid 2 like 2
uncoupling protein 2
C-C motif chemokine ligand 3
heme oxygenase 1
tumor necrosis factor

Drug Repurposing




NIH U.S. National Library of Medicine

ClinicalTrials.gov

Home > Search Results > Study Record Detail

□ Save this study

The Impact of Camostat Mesilate on COVID-19 Infection (CamoCO-19)

⚠ The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Know the risks and potential benefits of clinical studies and talk to your health care provider before participating. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT04321096

Recruitment Status ℹ : Recruiting
First Posted ℹ : March 25, 2020
Last Update Posted ℹ : April 6, 2020

See Contacts and Locations

Sponsor:
University of Aarhus

Information provided by (Responsible Party):
University of Aarhus

Two Chapters

- Chapter 1: Mining Structures from Scientific Text
  - Phrase mining
  - Concept recognition (Named entity recognition)
  - Language models

Shang

  - Relation and attribute extraction
  - Conditional statement extraction
  - Experimental evidence extraction

Jiang

- Chapter 2: Constructing and Learning Scientific Knowledge Graphs
  - Taxonomy construction

Shang

  - Knowledge graph construction
  - Learning KG for literature search
  - Learning KG for scientific text generation

Jiang