COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation

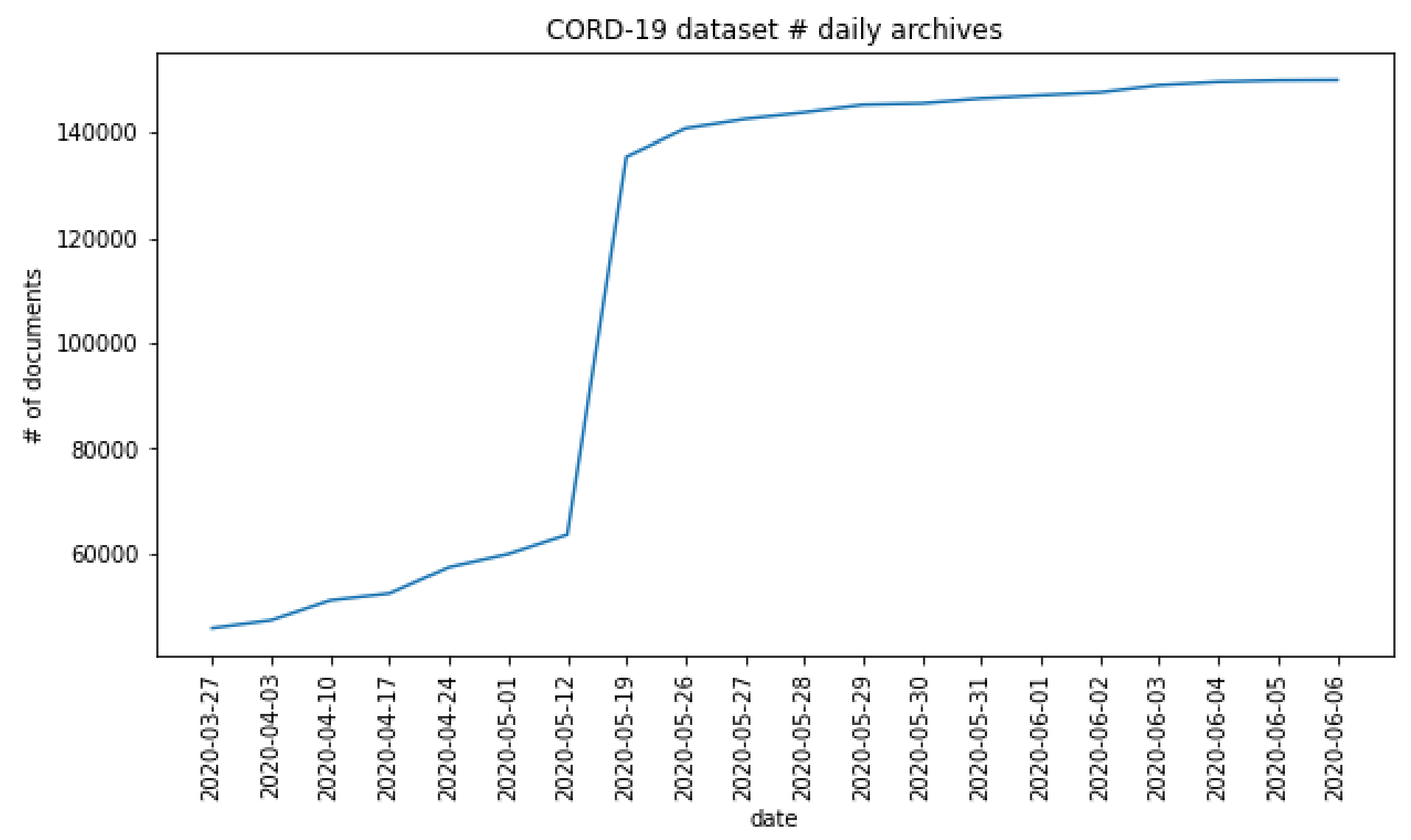
Qingyun Wang[1], Manling Li[1], Xuan Wang[1], Nikolaus Parulian[1], Guangxing Han[2], Jiawei Ma[2], Jingxuan Tu[3], Ying Lin[1], Haoran Zhang[1], Weili Liu[1], Aabhas Chauhan[1], Yingjun Guan[1], Bangzheng Li[1], Ruisong Li[1], Xiangchen Song[1], Yi R. Fung[1], Heng Ji[1], Jiawei Han[1], Shih-Fu Chang[2], James Pustejovsky[3], Jasmine Rah[4], David Liem[5], Ahmed Elsayed[6], Martha Palmer[6], Clare Voss[7], Cynthia Schneider[8], Boyan Onyshkevych[9]

[1]University of Illinois at Urbana-Champaign [2]Columbia University [3]Brandeis University [4]University of Washington [5]University of California, Los Angeles [6]Colorado University [7]Army Research Lab [8]QS2 [9]Defense Advanced Research Projects Agency

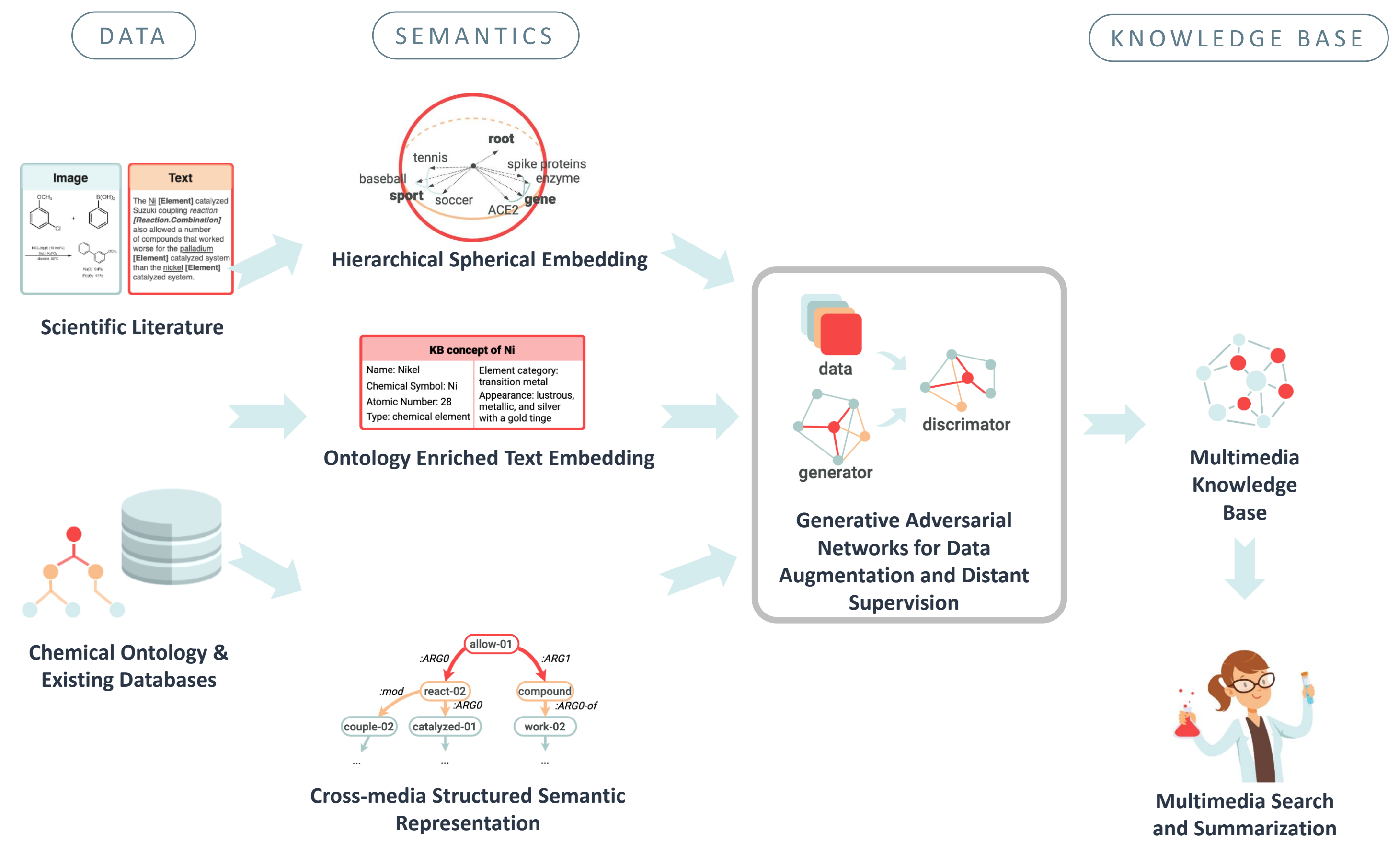
hengji@illinois.edu, hanj@illinois.edu, sc250@columbia.edu

➢ Motivation

- **Quantity**
  - More than **140K** paper are published about coronavirus by June 13, 2020.
- **Quality**
  - Many research results are **redundant**, **complementary** or even **conflicting** with each other



➢ Our Goals

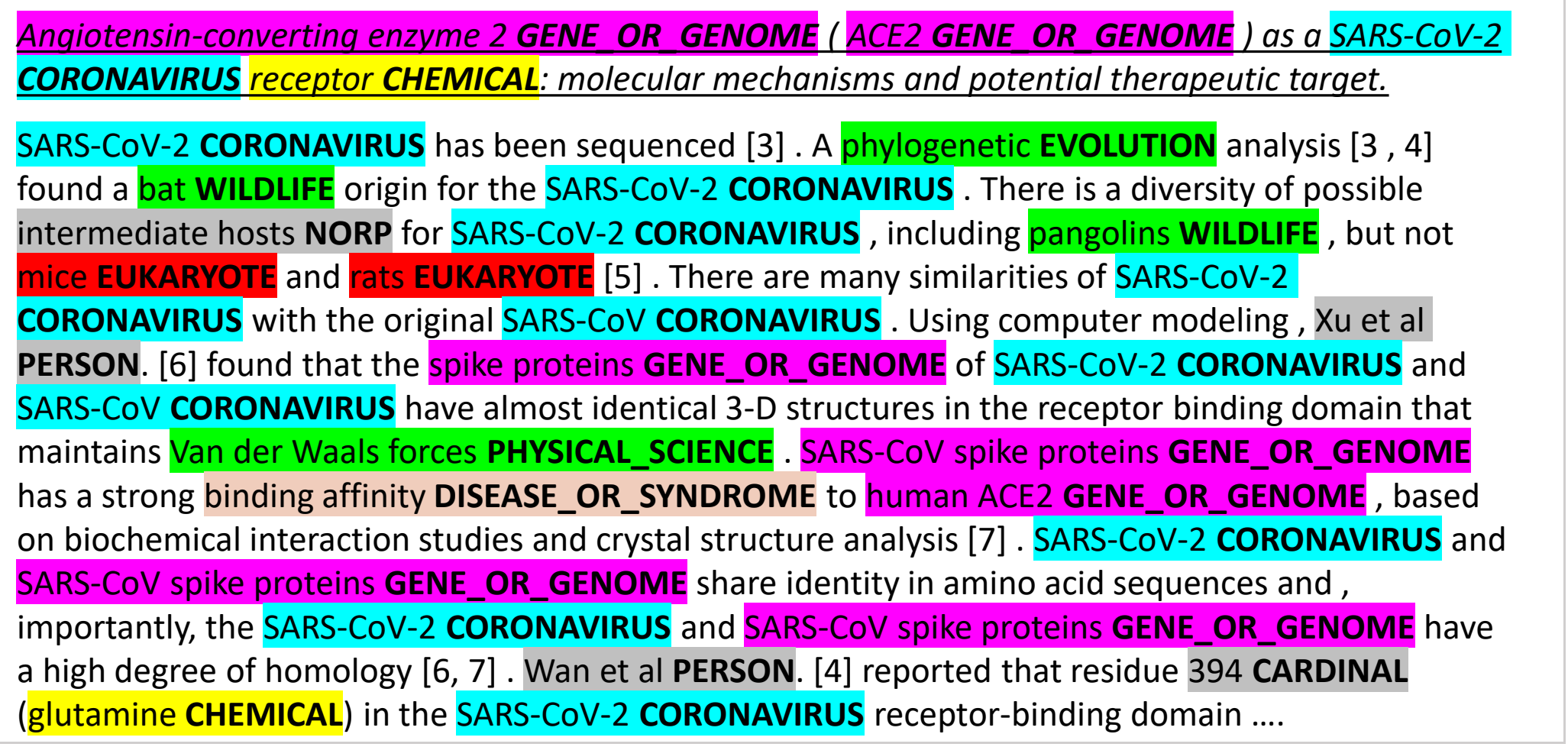

DATA
SEMANTICS
KNOWLEDGE BASE
Image
Text
The Ni [Element] catalyzed Suzuki coupling reaction [Reaction.Combination] also allowed a number of compounds that worked worse for the palladium [Element] catalyzed system than the nickel [Element] catalyzed system.
Scientific Literature
root
tennis
baseball
sport
soccer
spike proteins
enzyme
ACE2
gene
Hierarchical Spherical Embedding
KB concept of Ni
Name: Nikel
Chemical Symbol: Ni
Atomic Number: 28
Type: chemical element
Element category: transition metal
Appearance: lustrous, metallic, and silver with a gold tinge
Ontology Enriched Text Embedding
Chemical Ontology & Existing Databases
allow-01
:ARG0
:ARG1
:mod
react-02
compound
:ARG0
:ARG0-of
couple-02
catalyzed-01
work-02
Cross-media Structured Semantic Representation
data
generator
discriminator
Generative Adversarial Networks for Data Augmentation and Distant Supervision
Multimedia Knowledge Base
Multimedia Search and Summarization

➢ Coarse-grained Text Extraction

- **Entity Extraction + Entity Linking**
  - Extract entities from unstructured texts, link entity mentions to external biomedical ontologies including ***Comparative Toxicogenomics Database (CTD)*** and obtain ***Medical Subject Headings (MeSH) IDs***
- **Relation Extraction**
  - Extract **133** relation types including *Gene–Chemical–Interaction Relationships, Chemical–Disease Associations, Gene–Disease Associations, Chemical–GO Enrichment Associations* and *Chemical–Pathway Enrichment Associations*
- **Event Extraction**
  - Extract **13** Event types and the roles of entities involved in these events, including *Gene expression, Transcription, Localization, Protein catabolism , Binding, Protein modification, Phosphorylation , Ubiquitination, Acetylation, Deacetylation, Regulation, Positive regulation,* and *Negative regulation*

➢ Fine-grained Text Extraction

- **Fine-grained Knowledge Element**
  - Fine-grained entity extraction for ***75*** entity types (Xuan Wang and Jiawei Han, 2020), including many COVID-19 specific new entity types (e.g., *coronaviruses, viral proteins, evolution, materials, substrates, and immune responses*)
  - So we will be able to answer questions that include fine-grained knowledge elements such as “Which *amino acids* in glycoprotein (a spike protein of COVID-19) are most related to Glycan (CHEMICAL)?”

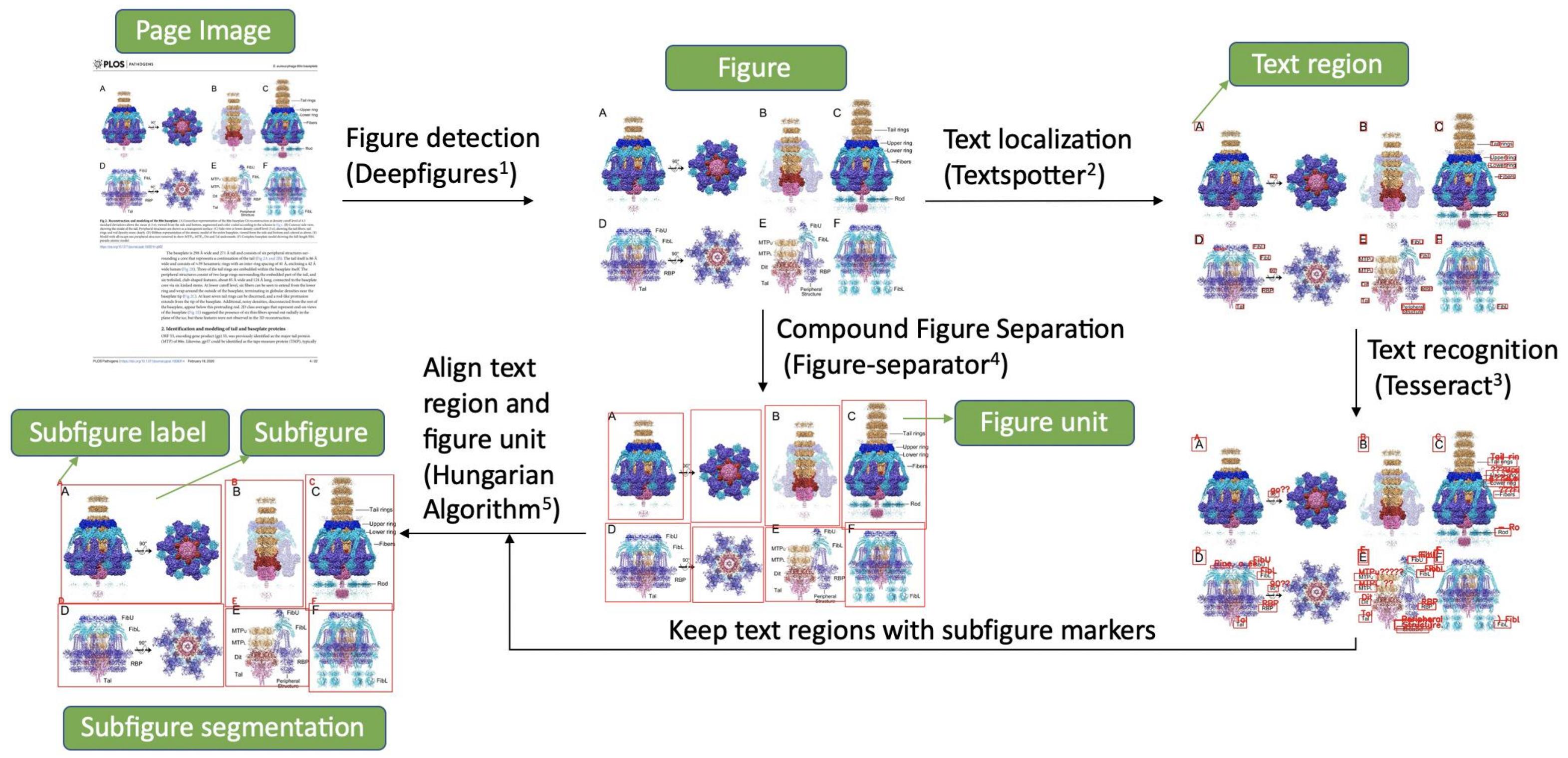


*Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor **CHEMICAL**: molecular mechanisms and potential therapeutic target.*

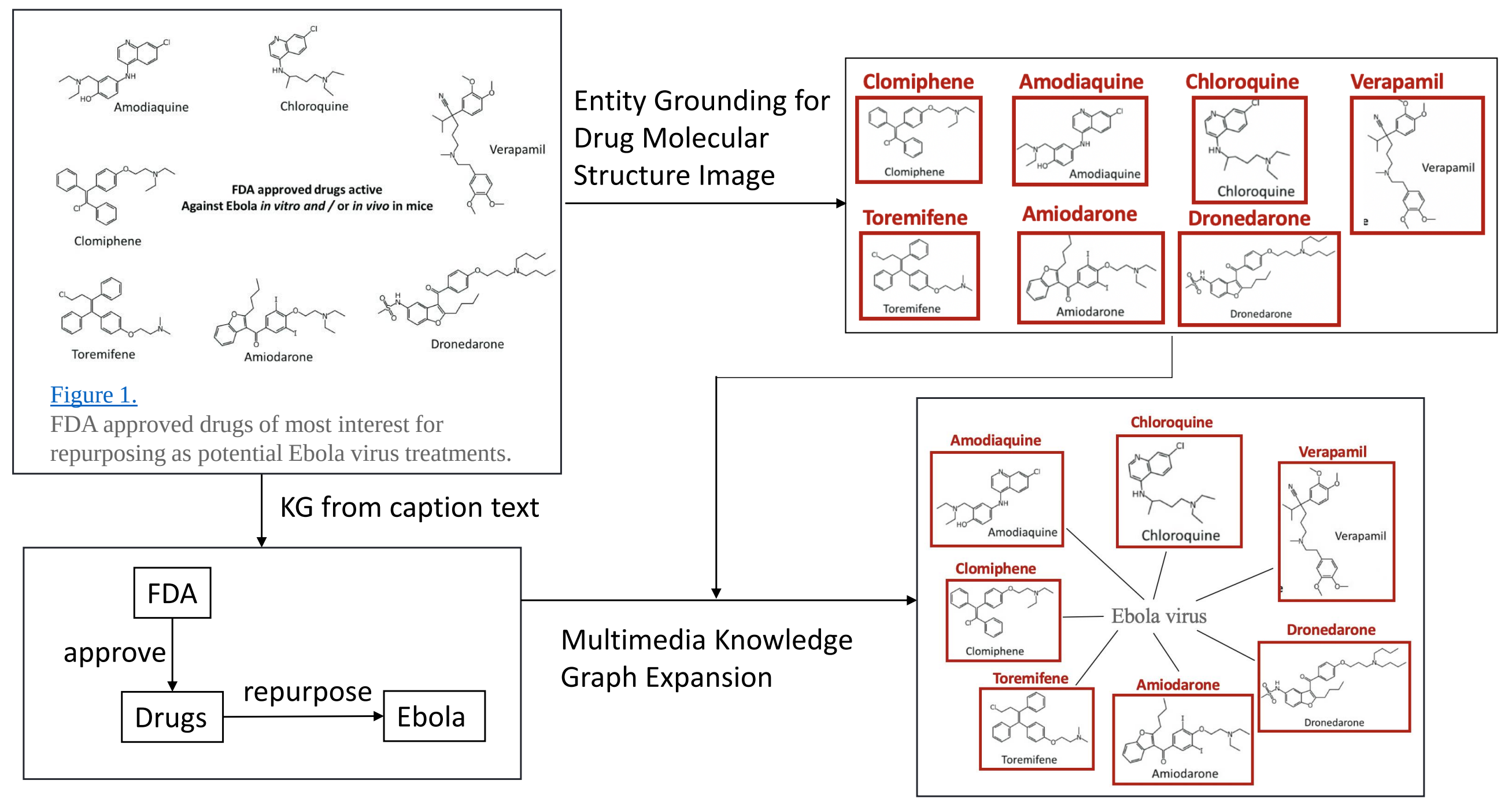
SARS-CoV-2 **CORONAVIRUS** has been sequenced [3] . A phylogenetic **EVOLUTION** analysis [3 , 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts **NORP** for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not 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 **PERSON**. [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 **DISEASE_OR_SYNDROME** 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 , importantly, the SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** have a high degree of homology [6, 7] . Wan et al **PERSON**. [4] reported that residue 394 **CARDINAL** (glutamine **CHEMICAL**) in the SARS-CoV-2 **CORONAVIRUS** receptor-binding domain ….

➢ Image Processing and Cross-media Entity Grounding

- **Automatic Figure Extraction and Subfigure Segmentation**
  - The figure shown here is from (Kizziah et al., 2020)

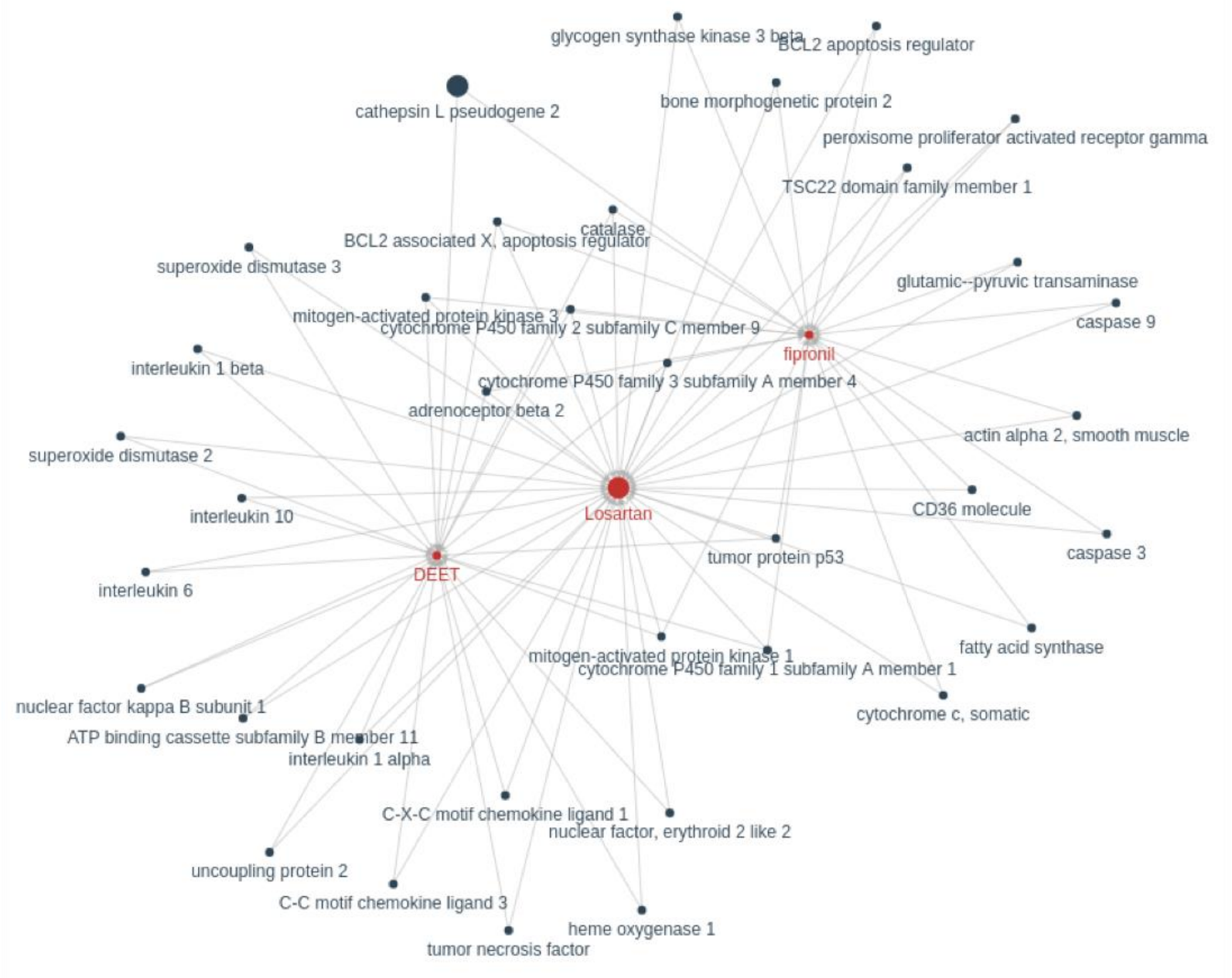

Page Image
Figure detection (Deepfigures[1])
Figure
Text localization (Textspotter[2])
Text region
Compound Figure Separation (Figure-separator[4])
Text recognition (Tesseract[3])
Figure unit
Align text region and figure unit (Hungarian Algorithm[5])
Subfigure label
Subfigure
Keep text regions with subfigure markers
Subfigure segmentation

- **Expanding KG through Subfigure Segmentation and Cross-modal Entity Grounding**
  - The figure shown here is from (Ekins and Coffee, 2015)

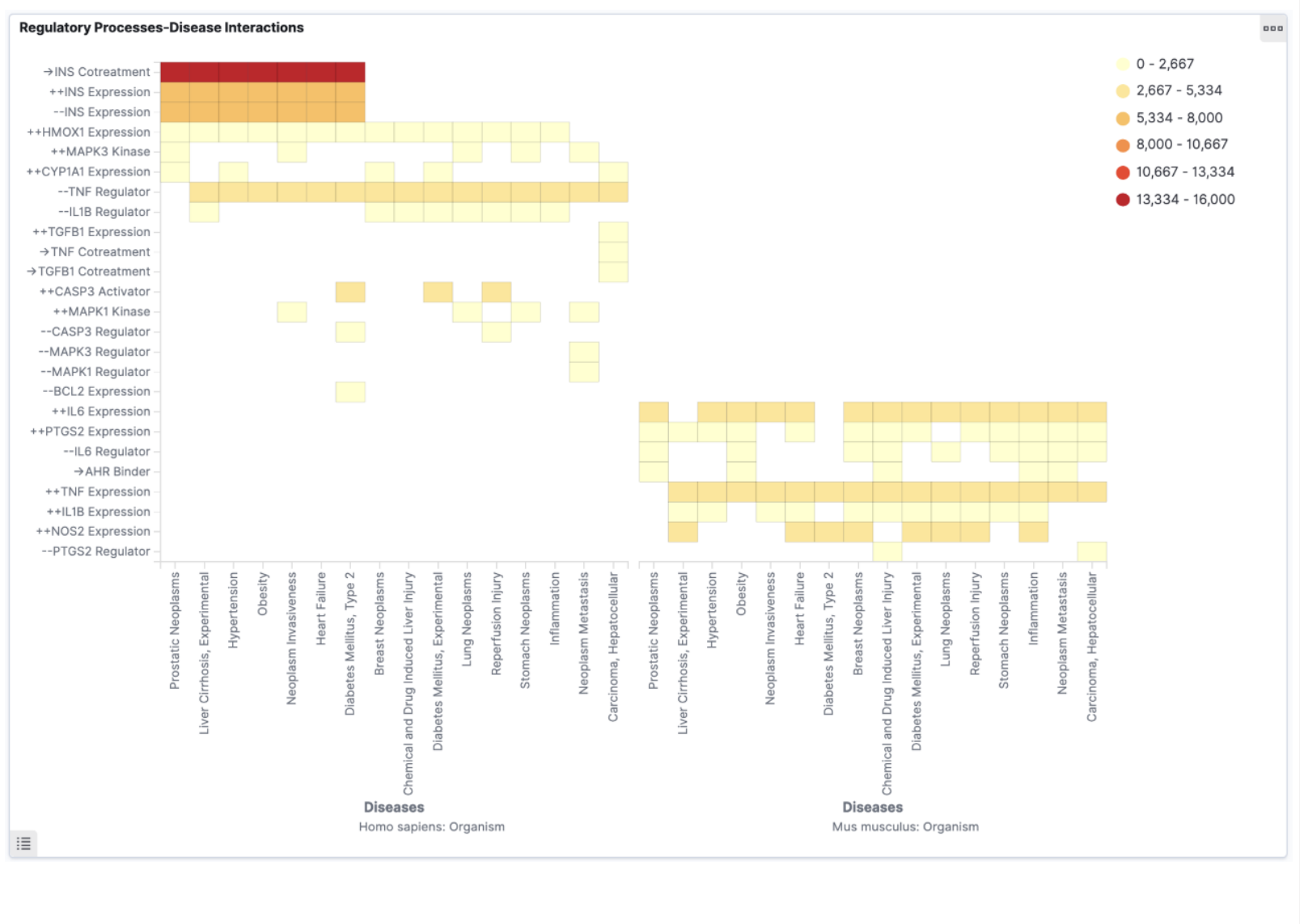

Amodiaquine
Chloroquine
Verapamil
Clomiphene
FDA approved drugs active Against Ebola in vitro and / or in vivo in mice
Toremifene
Amiodarone
Dronedarone
Figure 1.
FDA approved drugs of most interest for repurposing as potential Ebola virus treatments.
Entity Grounding for Drug Molecular Structure Image
KG from caption text
FDA
approve
Drugs
repurpose
Ebola
Multimedia Knowledge Graph Expansion
Ebola virus

➢ KG Visualization

- **Constructed KG Connecting Losartan and Cathepsin L pseudogene2**
  - Where ***red*** nodes represent *chemicals*, ***grey*** nodes represent *genes*, and edges represent gene-chemical relations

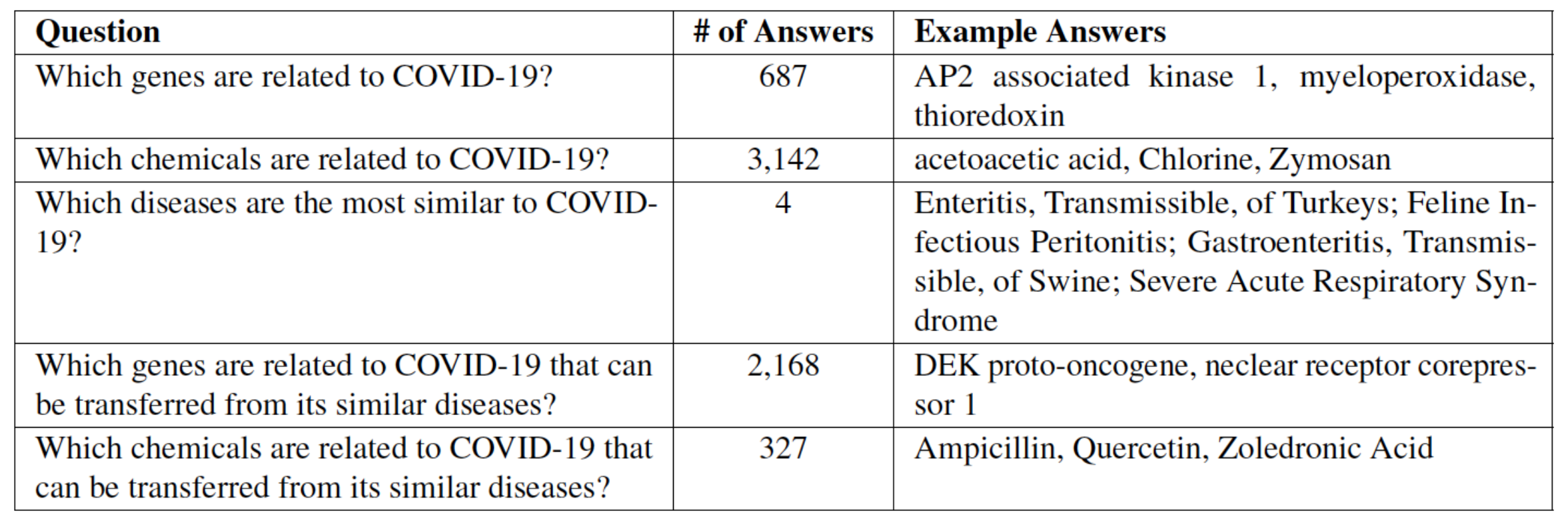

glycogen synthase kinase 3 beta
BCL2 apoptosis regulator
cathepsin L pseudogene 2
bone morphogenetic protein 2
peroxisome proliferator activated receptor gamma
TSC22 domain family member 1
catalase
BCL2 associated X, apoptosis regulator
superoxide dismutase 3
glutamic--pyruvic transaminase
mitogen-activated 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
caspase 3
tumor protein p53
DEET
interleukin 6
fatty acid synthase
mitogen-activated protein kinase 1
cytochrome P450 family 1 subfamily A member 1
cytochrome c, somatic
nuclear factor kappa B subunit 1
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

- **Regulatory Processes-Disease Interactions Heatmap**



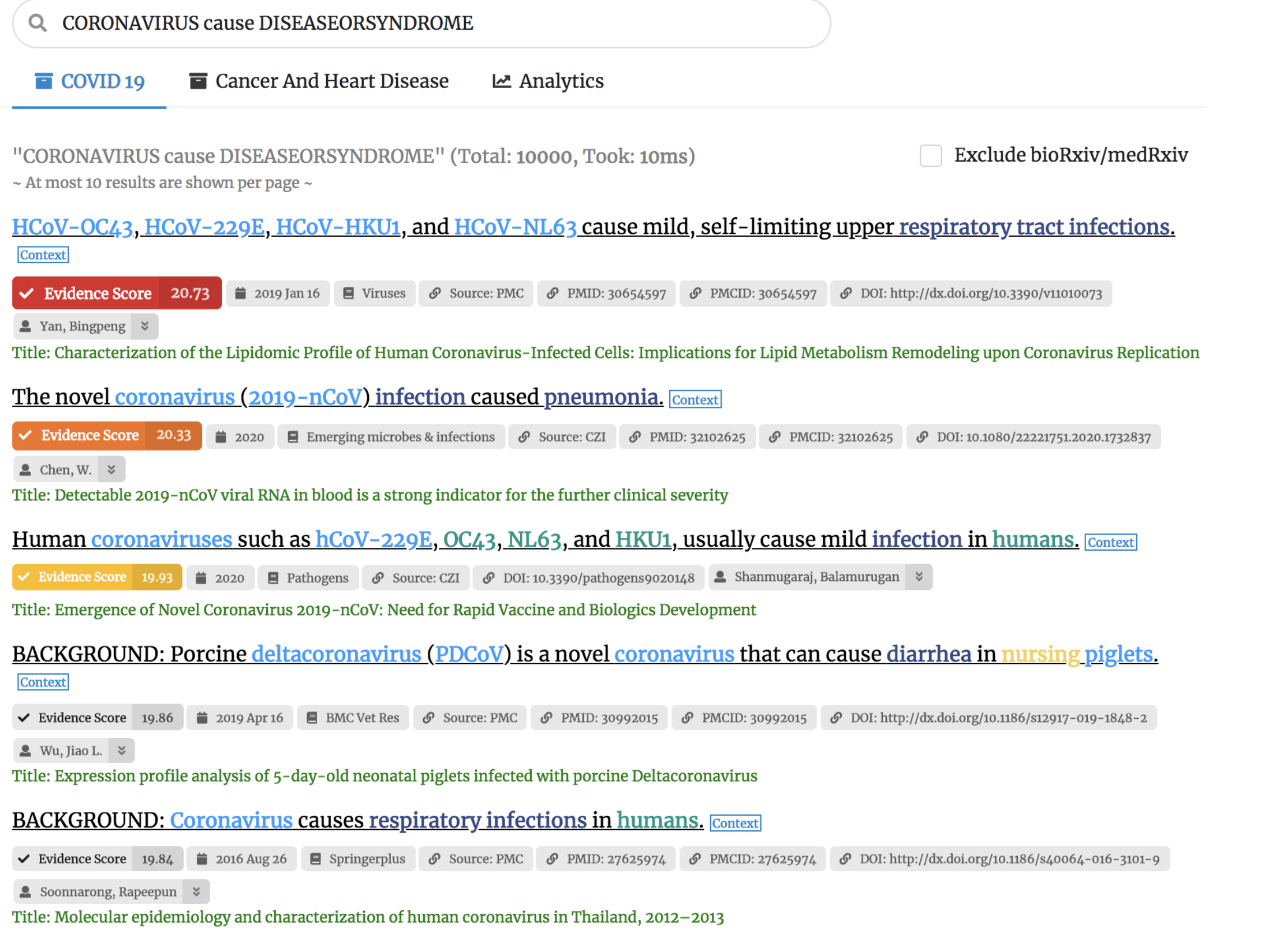
➢ Knowledge-driven Question Answering

- **Limitations of State-of-the-art Question Answering**
  - Fully rely on *word-level* or *sentence-level* semantic meaning matching
  - Questions are limited to non-experts (e.g., “Corona Virus Update?”) or too high-level (e.g., “What is known about transmission, incubation, and environmental stability?”)
- **What We Need**
  - Install a scientific brain (**KG**) for QA
  - Preliminary Results



Question	# of Answers	Example Answers
Which genes are related to COVID-19?	687	AP2 associated kinase 1, myeloperoxidase, thioredoxin
Which chemicals are related to COVID-19?	3,142	acetoacetic acid, Chlorine, Zymosan
Which diseases are the most similar to COVID-19?	4	Enteritis, Transmissible, of Turkeys; Feline Infectious Peritonitis; Gastroenteritis, Transmissible, of Swine; Severe Acute Respiratory Syndrome
Which genes are related to COVID-19 that can be transferred from its similar diseases?	2,168	DEK proto-oncogene, neclear receptor corepressor 1
Which chemicals are related to COVID-19 that can be transferred from its similar diseases?	327	Ampicillin, Quercetin, Zoledronic Acid

- EvidenceMiner with Query: “CORONAVIRUS cause DISEASEORSYNDROME”



"CORONAVIRUS cause DISEASEORSYNDROME" (Total: 10000, Took: 10ms)

~ At most 10 results are shown per page ~

HCoV-OC43, HCoV-229E, HCoV-HKU1, and HCoV-NL63 cause mild, self-limiting upper respiratory tract infections.

Evidence Score 20.73 · 2019 Jan 16 · Viruses · Source: PMC · PMID: 30654597 · PMCID: 30654597 · DOI: http://dx.doi.org/10.3390/v11010073 · Yan, Bingpeng

Title: Characterization of the Lipidomic Profile of Human Coronavirus-Infected Cells: Implications for Lipid Metabolism Remodeling upon Coronavirus Replication

The novel coronavirus (2019-nCoV) infection caused pneumonia.

Evidence Score 20.33 · 2020 · Emerging microbes & infections · Source: CZI · PMID: 32102625 · PMCID: 32102625 · DOI: 10.1080/22221751.2020.1732837 · Chen, W.

Title: Detectable 2019-nCoV viral RNA in blood is a strong indicator for the further clinical severity

Human coronaviruses such as hCoV-229E, OC43, NL63, and HKU1, usually cause mild infection in humans.

Evidence Score 19.93 · 2020 · Pathogens · Source: CZI · DOI: 10.3390/pathogens9020148 · Shanmugaraj, Balamurugan

Title: Emergence of Novel Coronavirus 2019-nCoV: Need for Rapid Vaccine and Biologics Development

BACKGROUND: Porcine deltacoronavirus (PDCoV) is a novel coronavirus that can cause diarrhea in nursing piglets.

Evidence Score 19.86 · 2019 Apr 16 · BMC Vet Res · Source: PMC · PMID: 30992015 · PMCID: 30992015 · DOI: http://dx.doi.org/10.1186/s12917-019-1848-2 · Wu, Jiao L.

Title: Expression profile analysis of 5-day-old neonatal piglets infected with porcine Deltacoronavirus

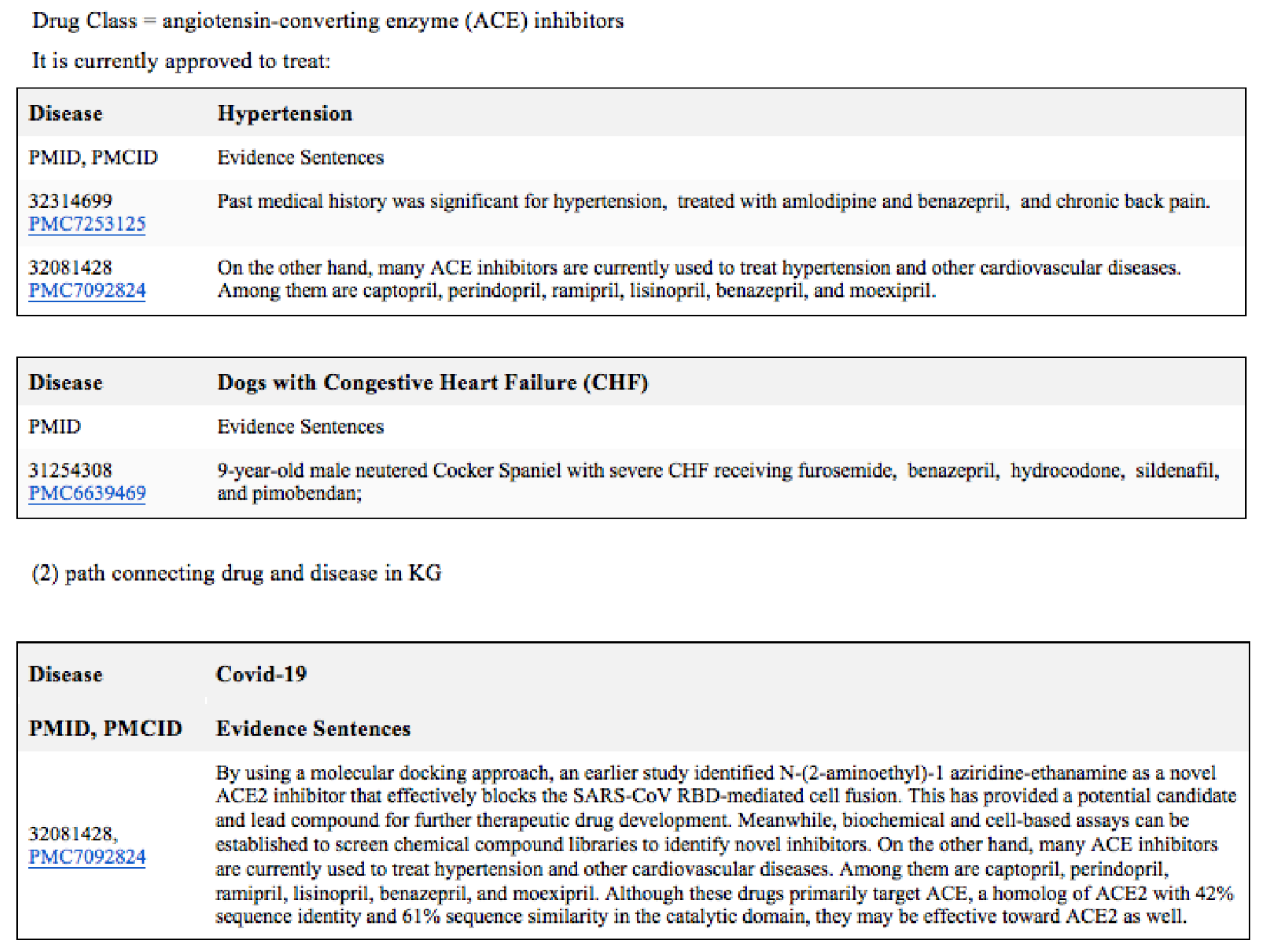
BACKGROUND: Coronavirus causes respiratory infections in humans.

Evidence Score 19.84 · 2016 Aug 26 · Springerplus · Source: PMC · PMID: 27625974 · PMCID: 27625974 · DOI: http://dx.doi.org/10.1186/s40064-016-3101-9 · Soonnarong, Rapeepun

Title: Molecular epidemiology and characterization of human coronavirus in Thailand, 2012–2013

➢ Case Study

- **KG Statistics**
  - ***50,864*** Gene nodes, ***7,230*** Disease nodes, ***9,123*** Chemical nodes, ***1,725,518*** chemical-gene links, ***5,556,670*** chemical-disease links, and ***7,7844,574*** gene-disease links
- **Sample Questions and Answers**
  - *Current indication: what is the drug class? What is it currently approved to treat?*
    - Results for Benazepril



Drug Class = angiotensin-converting enzyme (ACE) inhibitors

It is currently approved to treat:

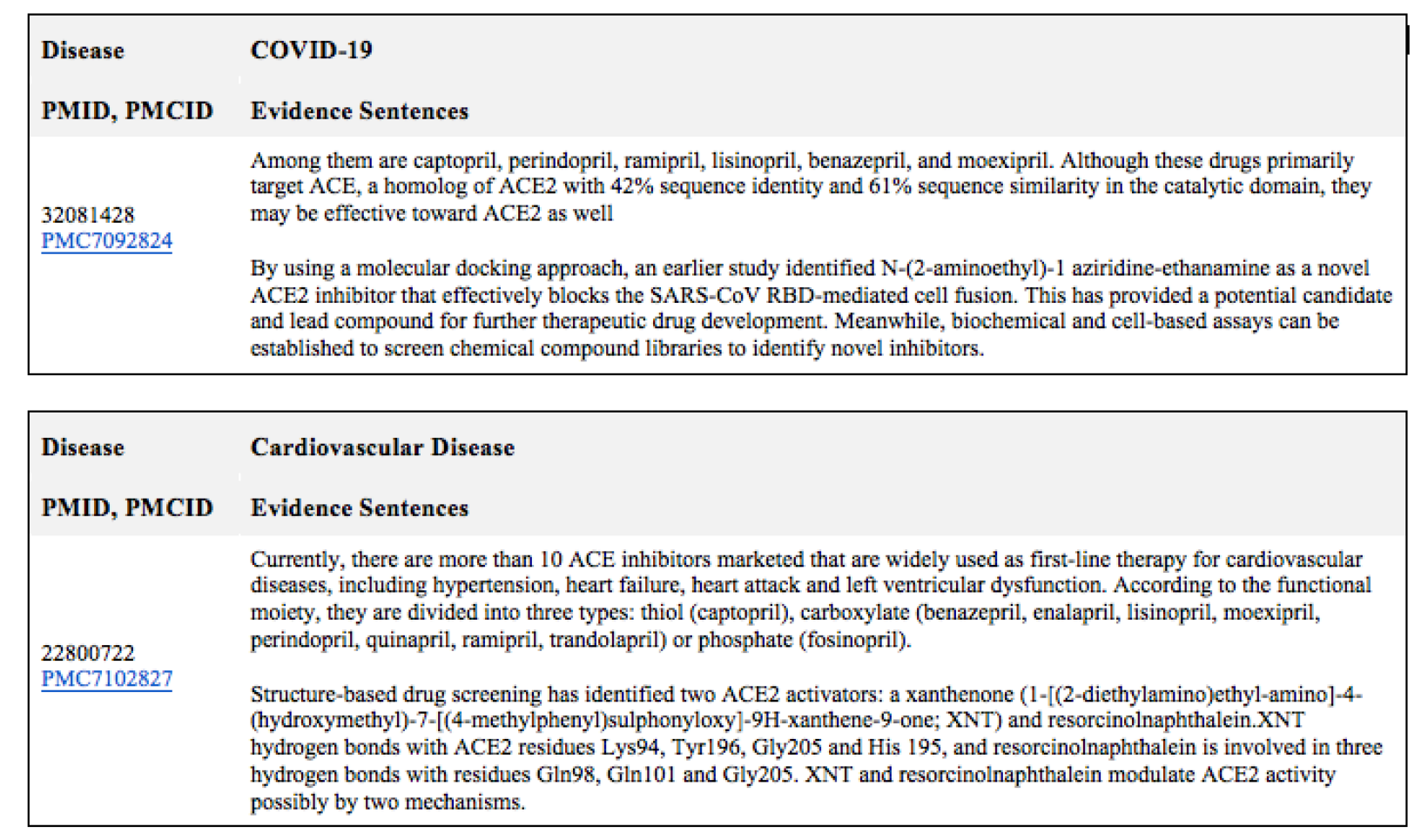
Disease	Hypertension
PMID, PMCID	Evidence Sentences
32314699 PMC7253125	Past medical history was significant for hypertension, treated with amlodipine and benazepril, and chronic back pain.
32081428 PMC7092824	On the other hand, many ACE inhibitors are currently used to treat hypertension and other cardiovascular diseases. Among them are captopril, perindopril, ramipril, lisinopril, benazepril, and moexipril.

Disease	Dogs with Congestive Heart Failure (CHF)
PMID	Evidence Sentences
31254308 PMC6639469	9-year-old male neutered Cocker Spaniel with severe CHF receiving furosemide, benazepril, hydrocodone, sildenafil, and pimobendan;

(2) path connecting drug and disease in KG

Disease	Covid-19
PMID, PMCID	Evidence Sentences
32081428, PMC7092824	By using a molecular docking approach, an earlier study identified N-(2-aminoethyl)-1 aziridine-ethanamine as a novel ACE2 inhibitor that effectively blocks the SARS-CoV RBD-mediated cell fusion. This has provided a potential candidate and lead compound for further therapeutic drug development. Meanwhile, biochemical and cell-based assays can be established to screen chemical compound libraries to identify novel inhibitors. On the other hand, many ACE inhibitors are currently used to treat hypertension and other cardiovascular diseases. Among them are captopril, perindopril, ramipril, lisinopril, benazepril, and moexipril. Although these drugs primarily target ACE, a homolog of ACE2 with 42% sequence identity and 61% sequence similarity in the catalytic domain, they may be effective toward ACE2 as well.

  - *Was the drug identified by manual or computation screen?*



Disease	COVID-19
PMID, PMCID	Evidence Sentences
32081428 PMC7092824	Among them are captopril, perindopril, ramipril, lisinopril, benazepril, and moexipril. Although these drugs primarily target ACE, a homolog of ACE2 with 42% sequence identity and 61% sequence similarity in the catalytic domain, they may be effective toward ACE2 as well By using a molecular docking approach, an earlier study identified N-(2-aminoethyl)-1 aziridine-ethanamine as a novel ACE2 inhibitor that effectively blocks the SARS-CoV RBD-mediated cell fusion. This has provided a potential candidate and lead compound for further therapeutic drug development. Meanwhile, biochemical and cell-based assays can be established to screen chemical compound libraries to identify novel inhibitors.

Disease	Cardiovascular Disease
PMID, PMCID	Evidence Sentences
22800722 PMC7102827	Currently, there are more than 10 ACE inhibitors marketed that are widely used as first-line therapy for cardiovascular diseases, including hypertension, heart failure, heart attack and left ventricular dysfunction. According to the functional moiety, they are divided into three types: thiol (captopril), carboxylate (benazepril, enalapril, lisinopril, moexipril, perindopril, quinapril, ramipril, trandolapril) or phosphate (fosinopril). Structure-based drug screening has identified two ACE2 activators: a xanthenone (1-[(2-diethylamino)ethyl-amino]-4-(hydroxymethyl)-7-[(4-methylphenyl)sulphonyloxy]-9H-xanthene-9-one; XNT) and resorcinolnaphthalein.XNT hydrogen bonds with ACE2 residues Lys94, Tyr196, Gly205 and His 195, and resorcinolnaphthalein is involved in three hydrogen bonds with residues Gln98, Gln101 and Gly205. XNT and resorcinolnaphthalein modulate ACE2 activity possibly by two mechanisms.

Demo video: http://159.89.180.81/demo/covid/Covid-KG_DemoVideo.mp4, Project website: http://blender.cs.illinois.edu/covid19, Drug Repurposing Report: http://blender.cs.illinois.edu/covid19/DrugRe-purposingReport_V2.0.docx