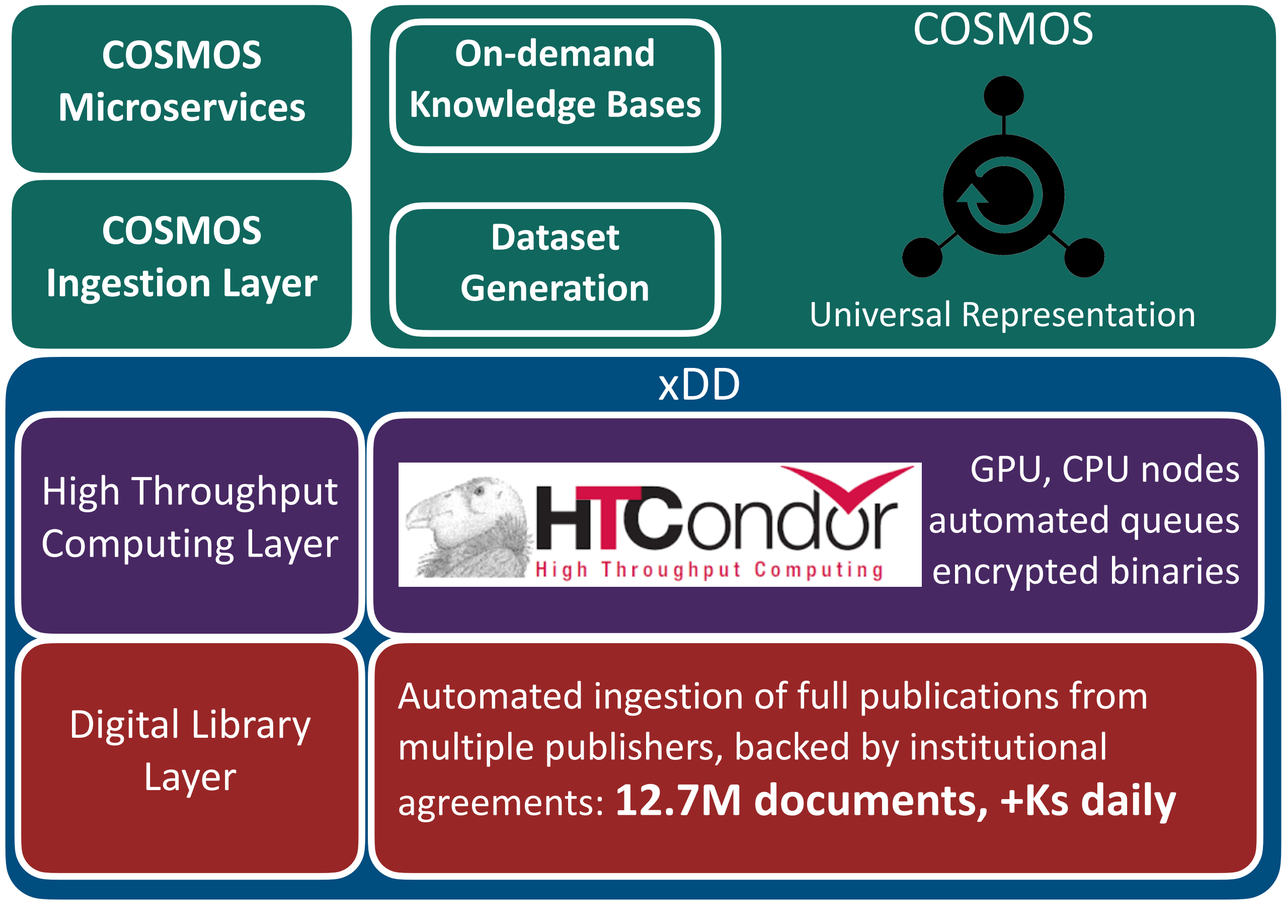
UW-COSMOS Phase 3:

A platform for extracting semantic knowledge and data from scientific publications

Shanan Peters, Theo Rekatsinas, Miron Livny

xDD & UW-COSMOS: an end-to-end stack to accelerate scientific research



- AI-powered “technical assistant” providing ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- HTC infrastructure to parse and analyze documents; API for full-text, fine-grained retrieval
- Automated fetching/storage of new and archival publications spanning many publishers and covering all domains of science

NSF USGS science for a changing world DARPA

support 2014-2018 by NSF-ICER 1343760, DARPA ASKE
partial current support from USGS, DOE

xDD API:
https://geodeepdive.org/api
Code available at:
https://github.com/UW-COSMOS

Correspondance:
Shanan Peters (peters@geology.wisc.edu)
Theodoros Rekatsinas (thodrek@cs.wisc.edu)
Miron Livny (miron@cs.wisc.edu)

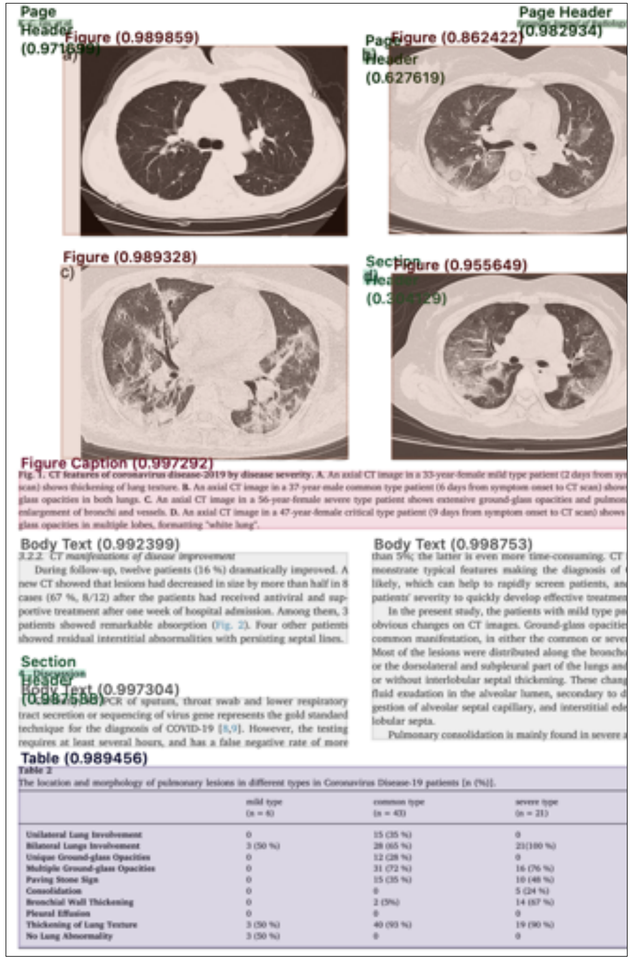
UW-COSMOS Phase 3 Concept:

provide an AI-powered technical assistant and cloud infrastructure for leveraging the scientific literature in diverse, distributed workflows

Part 1:
Publication acquisition

WILEY Publishers Since 1807
bioRχiv
medRχiv
CDC
ELSEVIER
Taylor & Francis
Taylor & Francis Group
PLOS ONE

Part 2:
xDD/COSMOS parsing, indexing



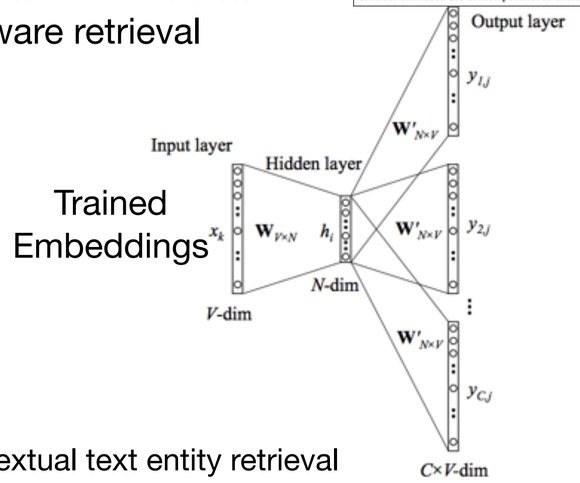
Part 3:
COSMOS-derived data products

Recombinant VACV	Ethacrynic acid	
	Luc	Titer
	IC50 (μM) ± SD	IC50 (μM)
(A)		
HeLa G	2.6 ± 0.6	1.6 ± 0.1
BSC-40	35.3 ± 7.0	11.9 ± 0.5
L929	26.5 ± 1.9	
MΦ + IFNγ	24.2 ± 2.8	
MΦ + IFNγ	18.3 ± 3.7	
Jurkat	–	

$$N(t) \le N_0 e^{-\mu t} + \frac{\Lambda}{\mu}\left(1 - e^{-\mu t}\right)$$

Fig. 2. CXR (left) with patchy peripheral left mid to lower lung opacities contrast-enhanced the contemporaneous chest CT (right).

Content- and context-aware retrieval



Contextual text entity retrieval

virus binds to a ACE2 on human lung…

Deliverable:
AI-powered API platform

Ecosystem of REST-ful API data services providing ASKE teams and collaborators with real-time, fine-grained and aggregate data and knowledge extractions from scientific publications.

Phase 3 Tasks

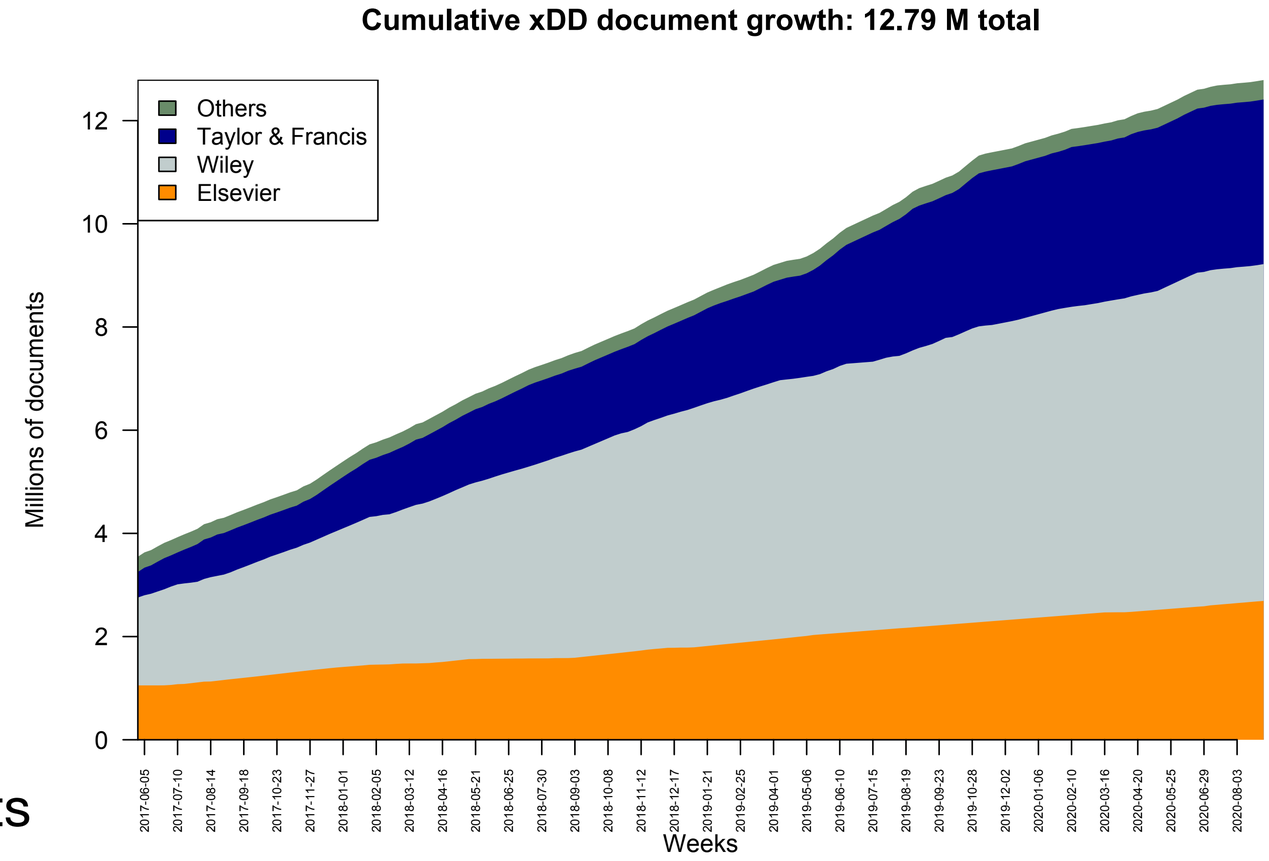
- **Task 1: Scale and enhance xDD infrastructure and APIs**
  A. Expand corpus, build new partnerships with publishers
  B. Develop and release template for collaborator code execution
  C. Incrementally annotate corpus with collaborator output
  D. Improve and scale xDD APIs

- **Task 2: Scale and enhance COSMOS retrieval, KBC and APIs**
  A. Improve and refine visual segmentation
  B. Incorporate full-document context into text/table/figure/eqn retrieval
  C. Automate knowledge base construction tasks
  D. Release public API over COVID-19 COSMOS document set

Task 1A: Expand xDD Corpus

continue to grow the document set

- Ongoing work to bring in additional publishers and enhance our current agreements, extend into AKSE space
- Implement xDD “fetching as a service”
  - From a list of user-provided DOIs, journals, etc.
    - Immediately identify what's in xDD
    - Prioritize acquisition of documents already within xDD's publisher pipeline
    - Recognize and more broadly acquire open-access full-text documents
    - Influence/motivate future publisher agreements
  - Have some DOIs and/or journal runs that you need? **Let us know!**



Taylor & Francis
Taylor & Francis Group

WILEY
Publishers Since 1807

PLOS ONE

ELSEVIER

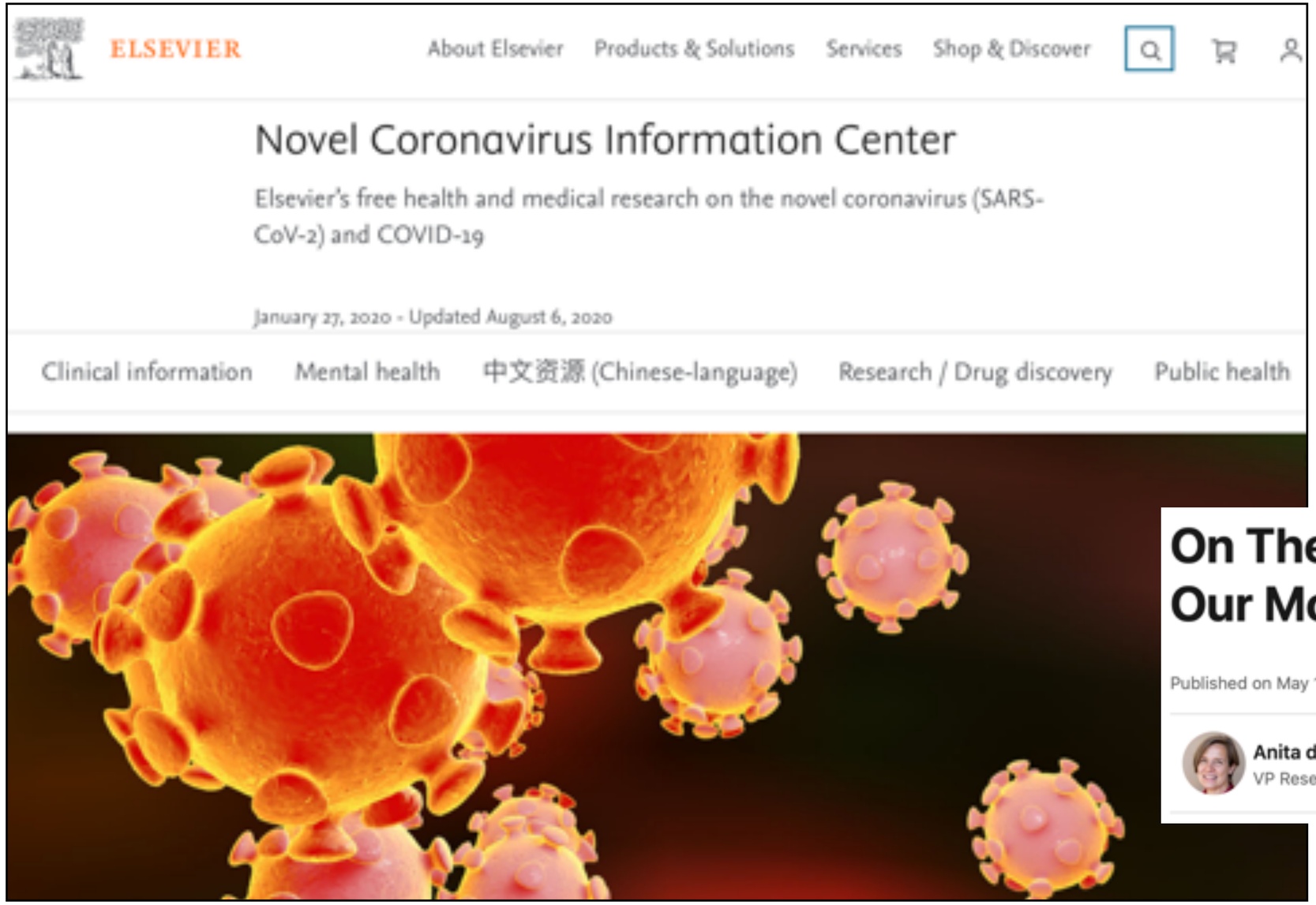
bioRχiv

medRχiv

CDC

Task 1A: Expand xDD Corpus

continue to grow the document set



Welcome to Elsevier's Novel Coronavirus Information Center. Here you will find expert, curated information for the research and health community on SARS-CoV-2 (the novel coronavirus) and COVID-19 (the disease).

All resources are free to access and include guidelines for clinicians and patients.

See our directory of all Elsevier's COVID-19 resources >

On The Power of Networks, Or: Is This Our Moonshot?

Published on May 14, 2020

Anita de Waard
VP Research Collaborations at Elsevier

1 article

+ Fo

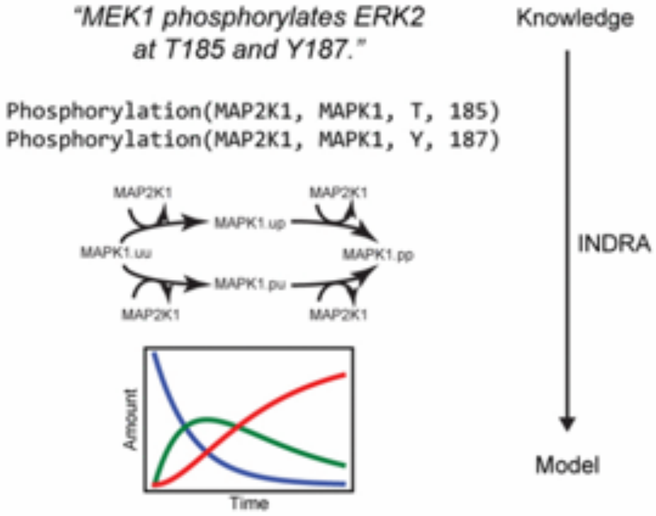
Task 1B: Container Template for Collaborator Code

Release recipes for deploying code over xDD

- Develop an easy-to-use container template that allows ASKE collaborator code to be executed over xDD
- Collaboration with HMS provides initial use case:
  1. Define document corpus (DOIs, journals, keywords, etc.)
  2. HMS-provided docker image with logic to read input from (+ write output to) prescribed location in container
  3. Output handling: write to S3 bucket
  4. xDD infrastructure executes over 100sK of full-text documents, returns results regularly (daily?) as new documents are acquired
- Development is underway for better container support within the xDD CHTC infrastructure
- Plan to have initial template ready this month
  - Interested? **Let us know!**

INDRA

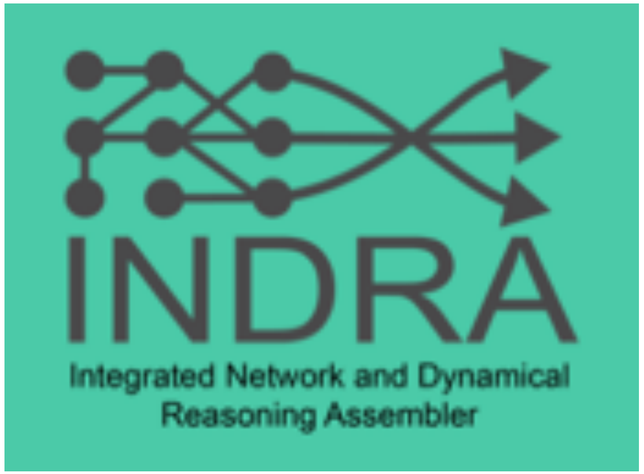
Integrated Network and Dynamical Reasoning Assembler

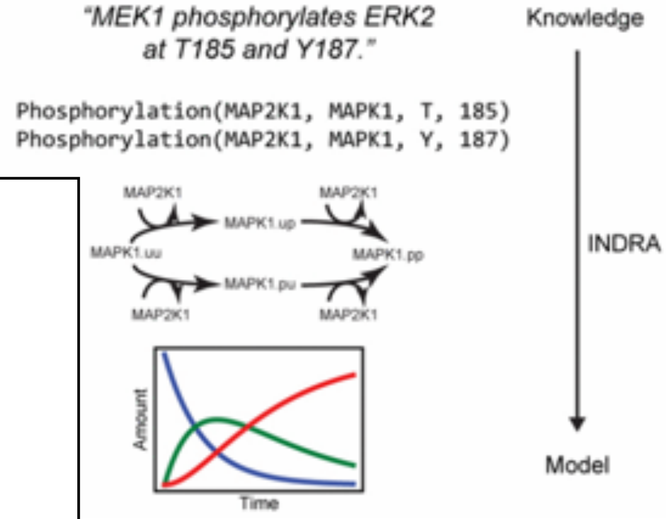


Task 1C: Augment xDD corpus with container outputs

Incrementally annotate corpus with knowledge products

- Develop data model for and mechanisms to incorporate container code output back into xDD infrastructure.
- Expose annotations and extractions via the xDD API, even when performing arbitrary searches.
- End goal: incrementally augment knowledge structure around xDD corpus in order to automatically enrich document- and mention-level API output





Mechanisms of Sensitivity and Resistance to CDK4/6 Inhibition
Published: 13 April 2020. Cancer Cell

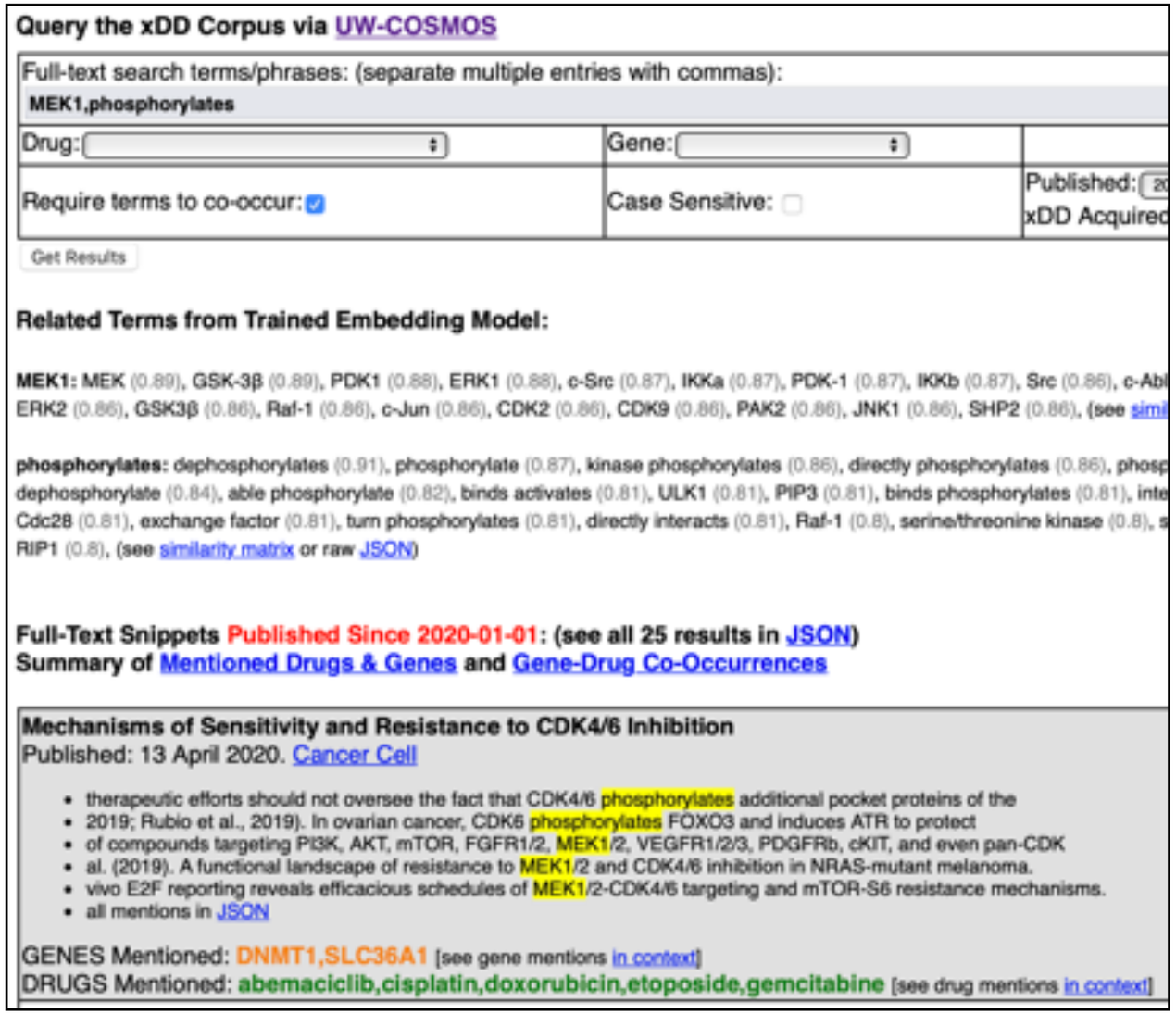
- of compounds targeting PI3K, AKT, mTOR, FGFR1/2, MEK1/2, VEGFR1/2/3, PDGFRb, cKIT, and even pan-CDK
- al. (2019). A functional landscape of resistance to MEK1/2 and CDK4/6 inhibition in NRAS-mutant melanoma.
- vivo E2F reporting reveals efficacious schedules of MEK1/2-CDK4/6 targeting and mTOR-S6 resistance mechanisms.
- all mentions in JSON

GENES Mentioned: DNMT1,SLC36A1 [see gene mentions in context]
DRUGS Mentioned: abemaciclib,cisplatin,doxorubicin,etoposide,gemcitabine [see drug mentions in context]

Task 1D: Improve and scale xDD API

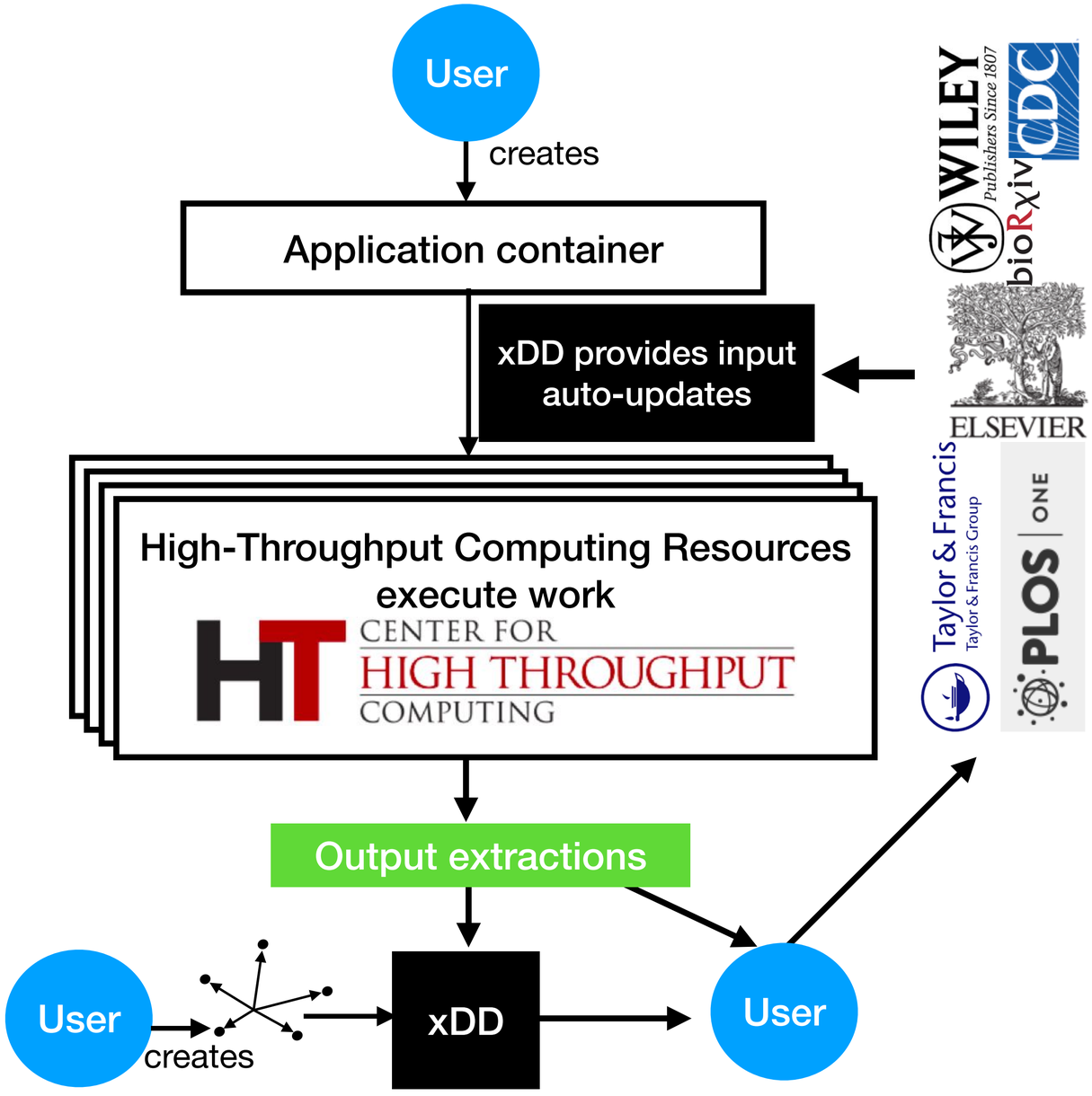
Expand power of contextual searches, ensure responsiveness

- Additional hardware and architecture updates to enhance performance and scale
- Improve documentation for users and add powerful new retrieval/app-building features including:
  - Universal sortable scan and scroll
  - Improved query logic support (e.g., regex-like and proximity-based retrieval, such as “<term X> appearing within <30> tokens of <term Y>” and restricting results to sections within documents)
  - Incorporation of collaborator knowledge products into document- and mention-level API responses.
- Have a need? **Let us know!**



Task 1: Increasing efficiency and working together to grow a knowledge structure around the scientific literature

- Example:
  - User develops an application using container template that extracts entity relationships from xDD documents ("x *causes* y").
  - Output extractions are sent to the user *and* fed back into xDD to provide additional context and knowledge at document/mention levels, allowing, e.g.:
    - *"Give me all tables mentioning <x> or <y> from from the literature which provide evidence of <x causes y>."*
    - *"Give me all extracted <y> caused by <x> within the context of <covid-19>."*
    - Links back to full document- and mention-level context are never lost! Entire end-to-end updated daily.

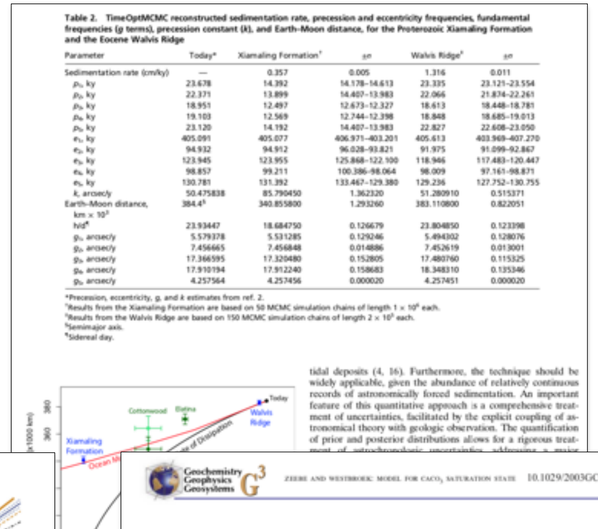


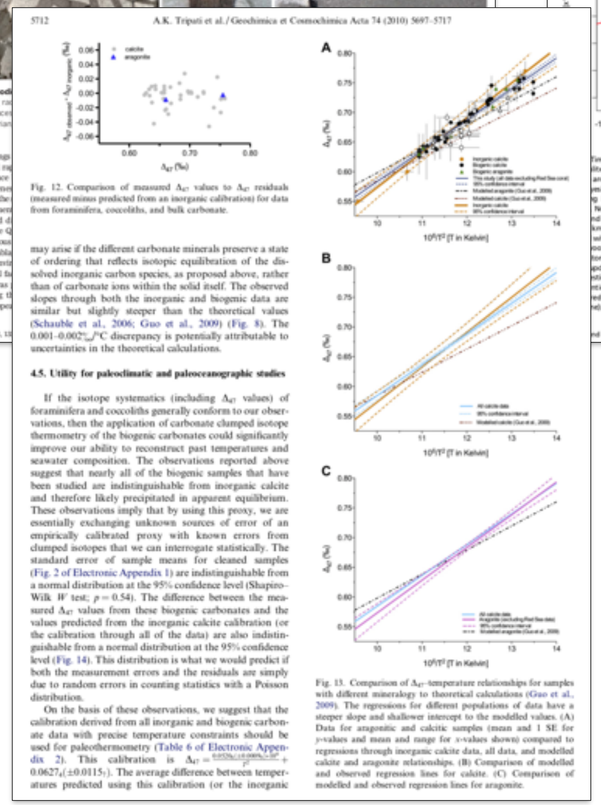
Phase 3 Tasks

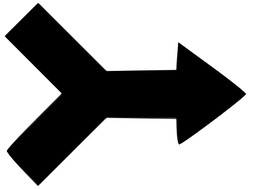
- **Task 1: Scale and enhance xDD infrastructure and APIs**
  A. Expand corpus, build new partnerships with publishers
  B. Develop and release template for collaborator code execution
  C. Incrementally annotate corpus with collaborator output
  D. Improve and scale xDD APIs

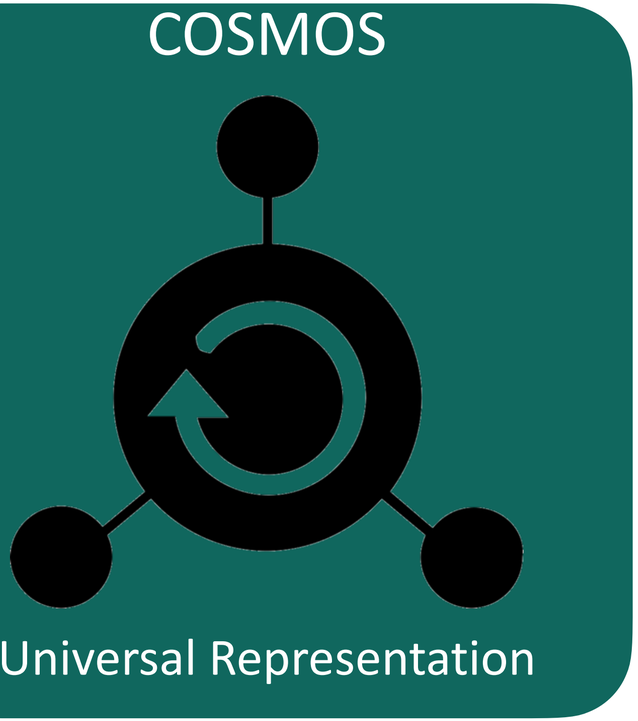
- **Task 2: Scale and enhance COSMOS retrieval, KBC and APIs**
  A. Improve and refine visual segmentation
  B. Incorporate full-document context into text/table/figure/eqn retrieval
  C. Automate knowledge base construction tasks
  D. Release public API over COVID-19 COSMOS document set

COSMOS dataset generation



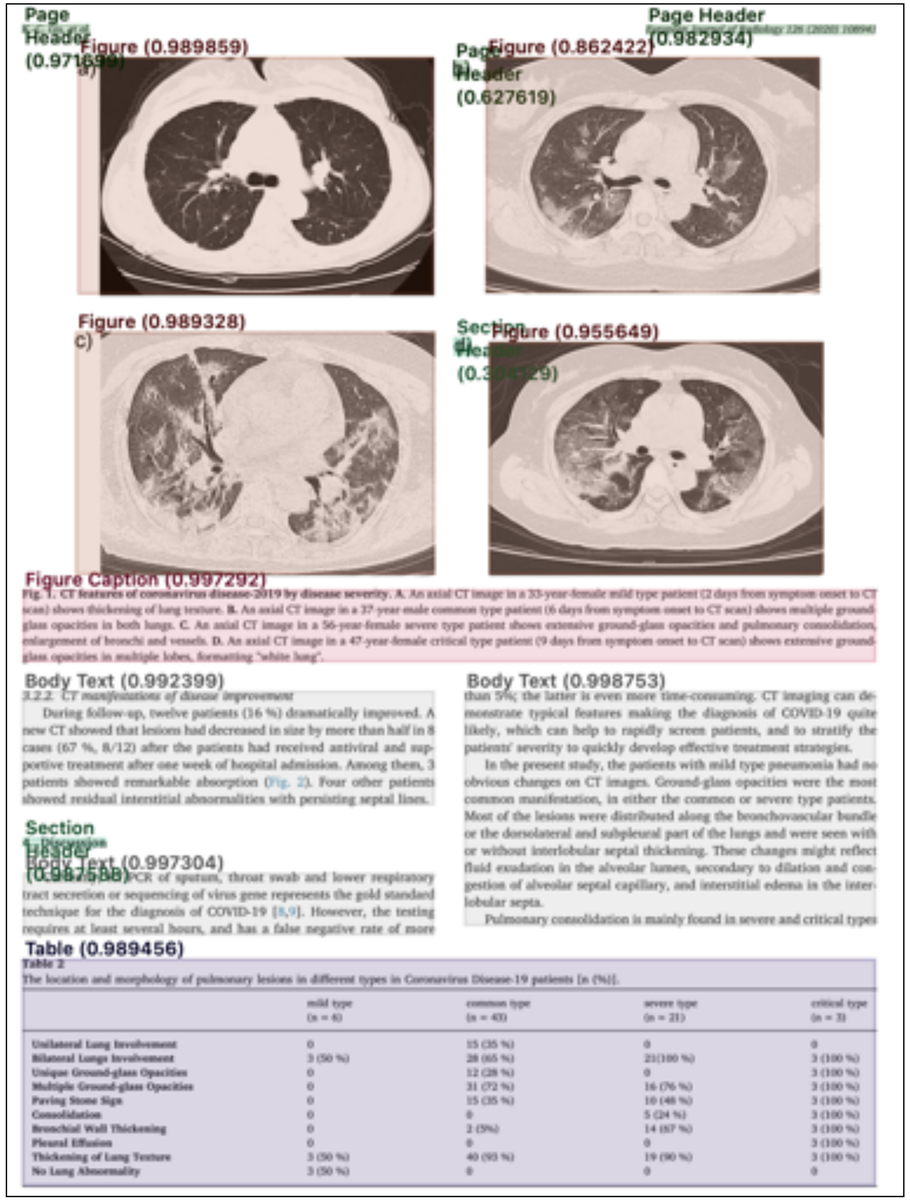


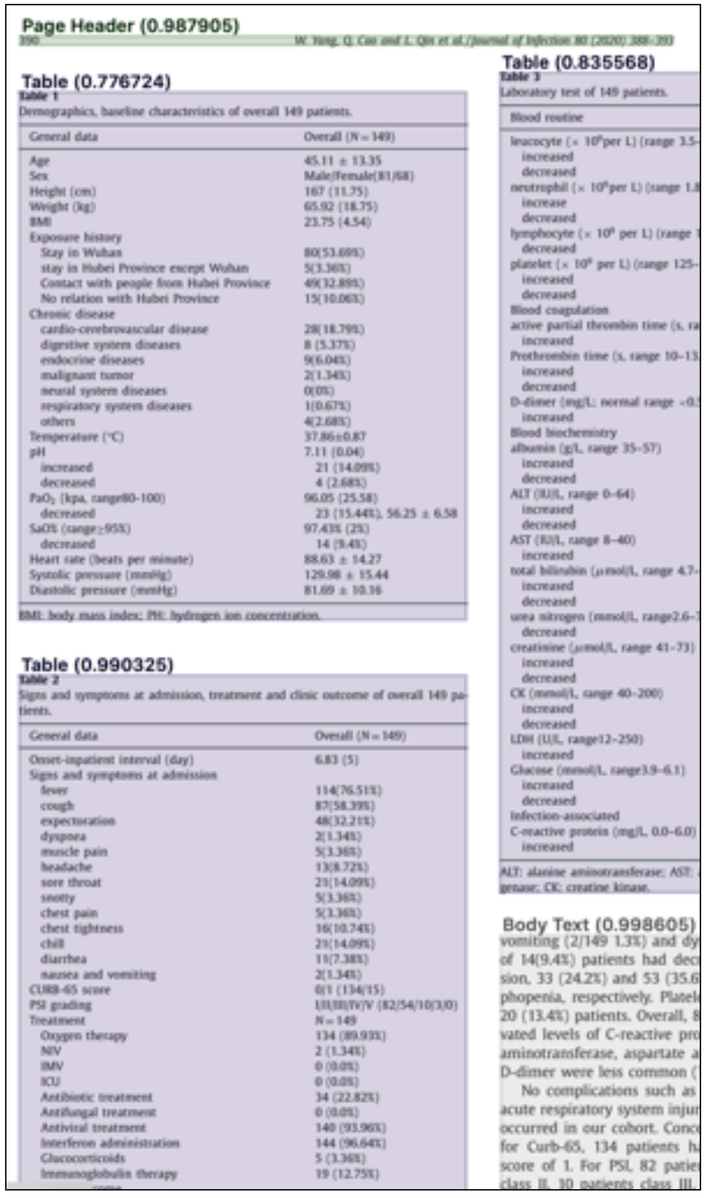


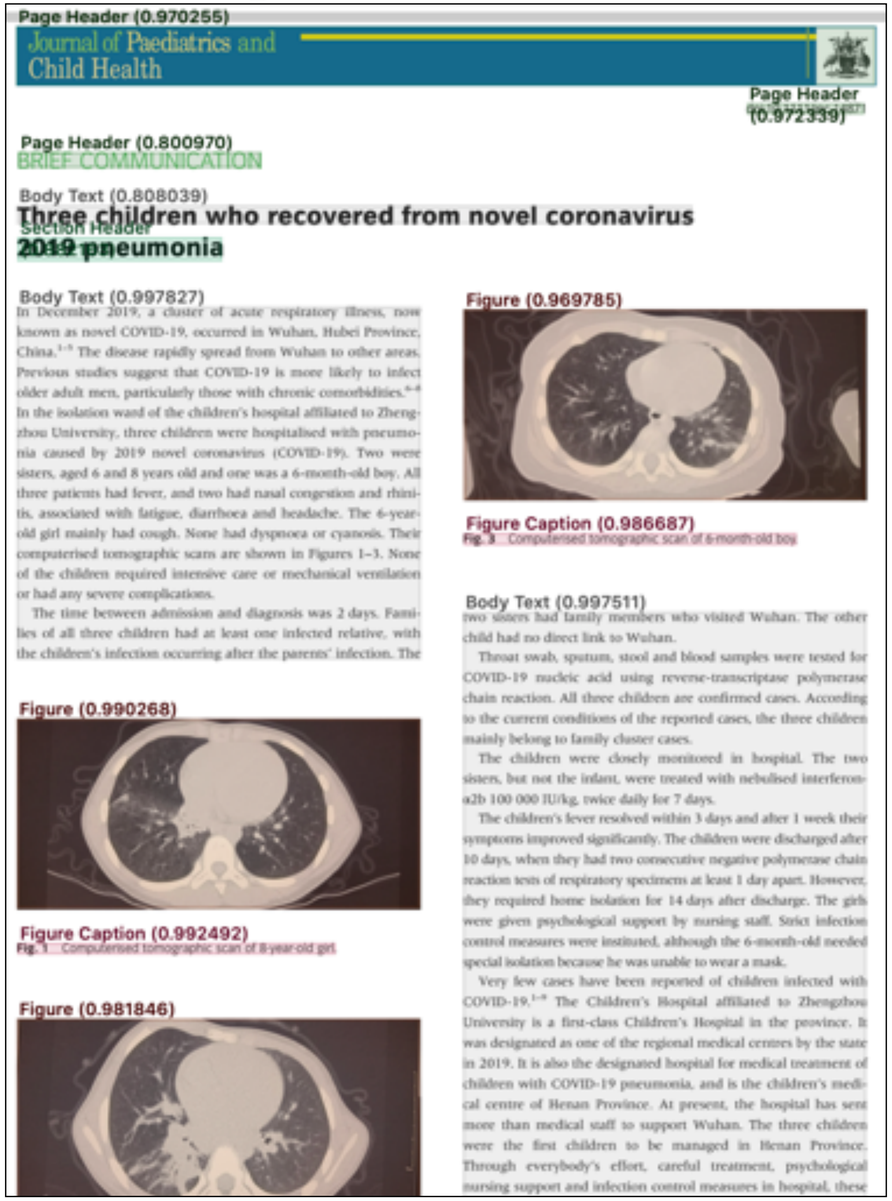


Fine-grained retrieval, word/document embeddings, extraction-ready tables, equations, figures

UW-COSMOS Visual Document Segmentation







COSMOS AI-powered results

COSMOS AI-constructed KB

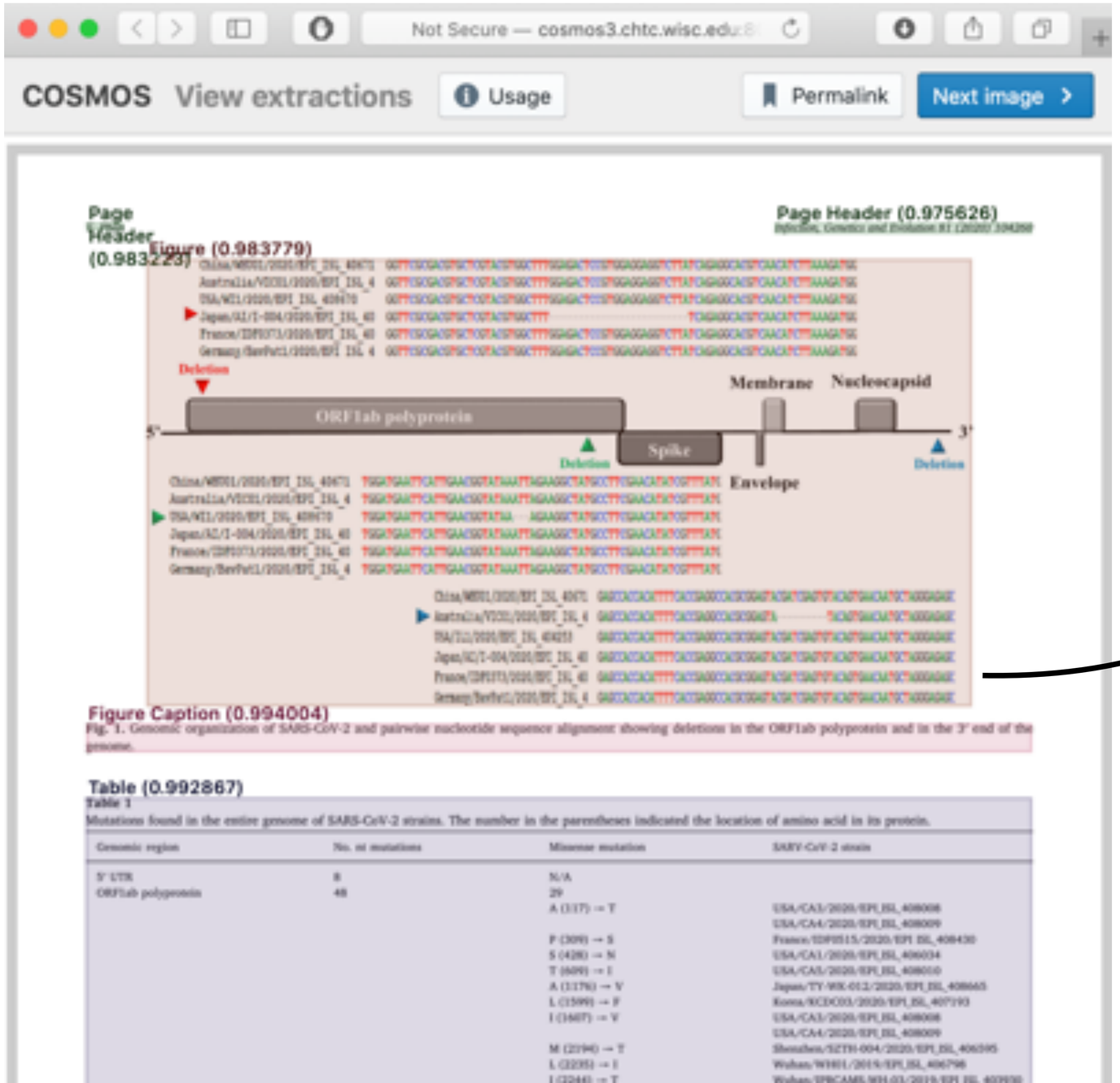


Table: page_objects

Select data **Show structure** Alter table New item

Column	Type	Comment
id	int(11) *Auto Increment*	
page_id	int(11) *NULL*	
context_id	int(11) *NULL*	
bytes	longblob *NULL*	
content	varchar(10000) *NULL*	
bounding_box	json *NULL*	
init_cls_confidences	json *NULL*	
cls	varchar(200) *NULL*	
pp_rule_cls	varchar(200) *NULL*	
annotated_cls	varchar(200) *NULL*	
confidence	decimal(9,6) *NULL*	
classification_success	tinyint(1) *NULL*	
proposal_success	tinyint(1) *NULL*	

Table: object_contexts

Select data **Show structure** Alter table

Column	Type	Comment
id	int(11) *Auto Increment*	
pdf_id	int(11) *NULL*	
cls	varchar(200) *NULL*	
header_id	int(11) *NULL*	
header_content	text *NULL*	
content	longtext *NULL*	

Indexes

PRIMARY	*id*
INDEX	*pdf_id*
INDEX	*header_id*

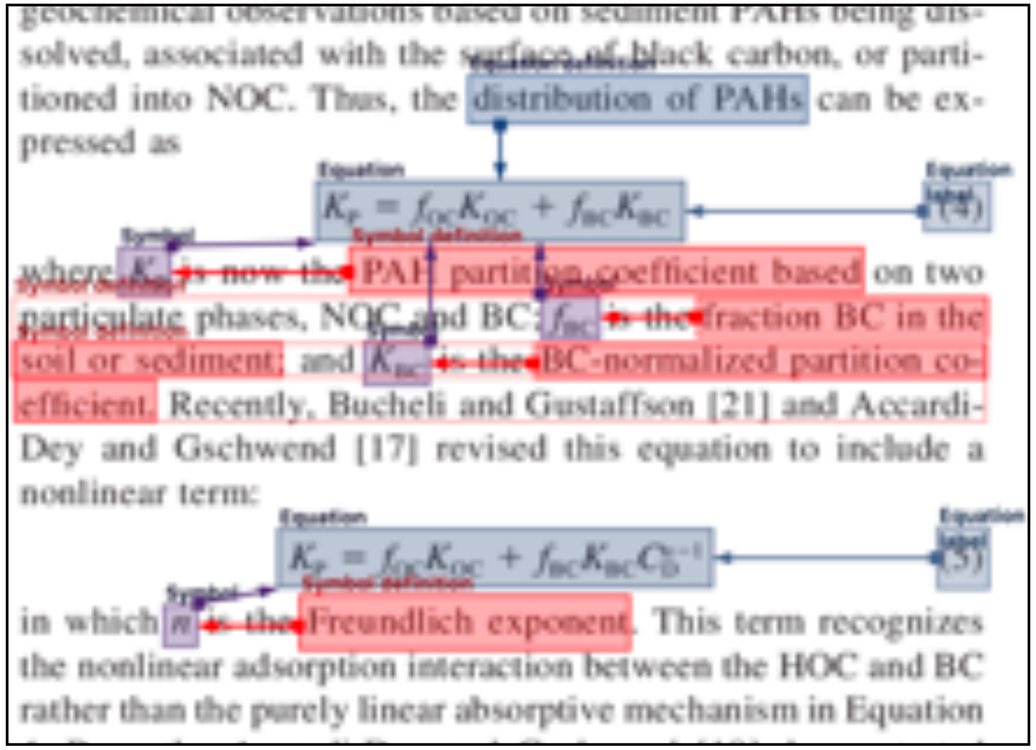
Alter indexes

- Convert all modalities in heterogeneous PDFs into a structured representation
- Throughput of approximately one page every 2 seconds (GPU accelerated)
- Ability to deploy in parallel over distributed (CPU and/or GPU) nodes via HTCondor

Task 2A: Improve COSMOS visual segmentation

Incorporate user annotations into continual model refinement

- Improve and deploy document-tagging functionality for model training and evaluation
- Leverage our support for tagging complex data types and relations (e.g. linked information hierarchies)



Task 2A: Improve COSMOS visual segmentation

Incorporate user annotations into continual model refinement

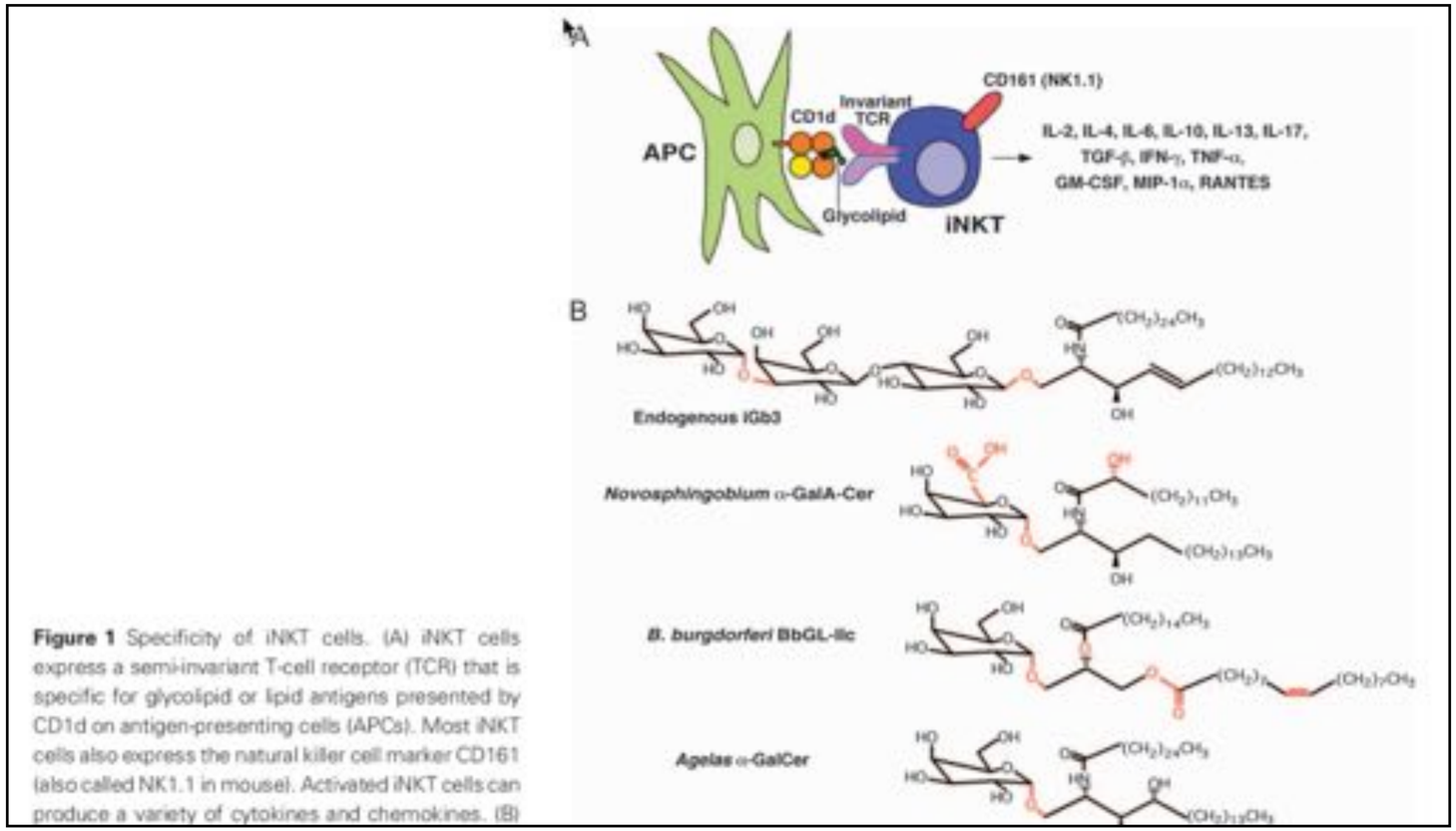


Figure 1 Specificity of iNKT cells. (A) iNKT cells express a semi-invariant T-cell receptor (TCR) that is specific for glycolipid or lipid antigens presented by CD1d on antigen-presenting cells (APCs). Most iNKT cells also express the natural killer cell marker CD161 (also called NK1.1 in mouse). Activated iNKT cells can produce a variety of cytokines and chemokines. (B)

- COSMOS web browser-based software for manual tagging of objects in a visual reference space (deployable over any image for any purpose); assessing AI proposals
- Interface to make explicit links between tagged objects (e.g., variables to equations)
- Need to leverage this in order to improve visual model and add modalities

Task 2B: Add document-level contextual retrieval

Incorporate body text descriptions into table and figure retrieval

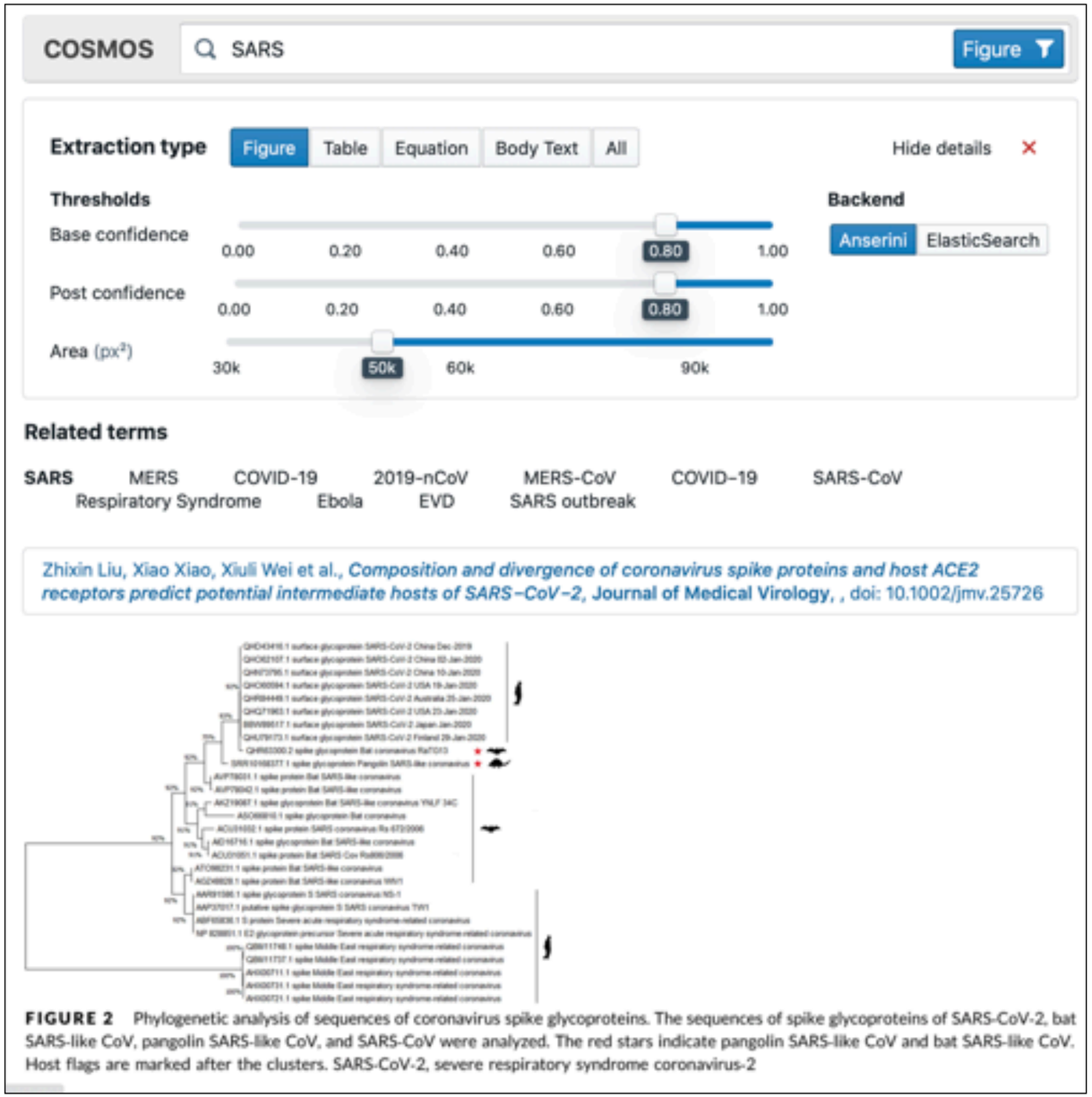


FIGURE 2 Phylogenetic analysis of sequences of coronavirus spike glycoproteins. The sequences of spike glycoproteins of SARS-CoV-2, bat SARS-like CoV, pangolin SARS-like CoV, and SARS-CoV were analyzed. The red stars indicate pangolin SARS-like CoV and bat SARS-like CoV. Host flags are marked after the clusters. SARS-CoV-2, severe respiratory syndrome coronavirus-2

- Frontend web app surfacing COSMOS KB and micro services, including
  - word embedding model trained over target corpus
  - two modes for retrieval: ElasticSearch index over extracted objects, TF-IDF based retrieval
  - preliminary data-frame extraction for tables

https://cosmos.wisc.edu/sets/covid

Task 2B: Add document-level contextual retrieval

Incorporate body text descriptions into table and figure retrieval

Table Reference **Table Caption**

TABLE. Hospitalization, intensive care unit (ICU) admission, and case–fatality percentages for reported COVID–19 cases, by age group — United States, February 12–March 16, 2020

Age group (yrs) (no. of cases)	%*		
	Hospitalization	ICU admission	Case-fatality
0–19 (123)	1.6–2.5	0	0
20–44 (705)	14.3–20.8	2.0–4.2	0.1–0.2
45–54 (429)	21.2–28.3	5.4–10.4	0.5–0.8
55–64 (429)	20.5–30.1	4.7–11.2	1.4–2.6
65–74 (409)	28.6–43.5	8.1–18.8	2.7–4.9
75–84 (210)	30.5–58.7	10.5–31.0	4.3–10.5
≥85 (144)	31.3–70.3	6.3–29.0	10.4–27.3
Total (2,449)	**20.7–31.4**	**4.9–11.5**	**1.8–3.4**

Table

* Lower bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group; upper bound of range = number of persons hospitalized, admitted to ICU, or who died among total in age group with known hospitalization status, ICU admission status, or death.

Table Note

Ref: CDC COVID-19 Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. MMWR Morb Mortal Wkly Rep 2020;69:343–6. 10.15585/mmwr.mm6912e2

Table Reference Context

aged ≤19 years. Percentages of ICU admissions were lowest among adults aged 20–44 years (2%–4%) and highest among adults aged 75–84 years (11%–31%) (Table).

Table Reference

Among 44 cases with known outcome, 15 (34%) deaths were reported among adults aged ≥85 years, 20 (46%) among adults aged 65–84 years, and nine (20%) among adults aged 20–64 years. Case-fatality percentages increased with increasing age, from no deaths reported among persons aged ≤19 years to highest percentages (10%–27%) among adults aged ≥85 years (Table) (Figure 2).

Table Reference Context

Table Reference

Among 508 (12%) patients known to have been hospitalized, 9% were aged ≥85 years, 36% were aged 65–84 years, 17% were aged 55–64 years, 18% were 45–54 years, and 20% were aged 20–44 years. Less than 1% of hospitalizations were among persons aged ≤19 years (Figure 2). The percentage of persons hospitalized increased with age, from 2%–3% among persons aged ≤9 years, to ≥31% among adults aged ≥85 years. (Table).

Table Reference Context **Table Reference**

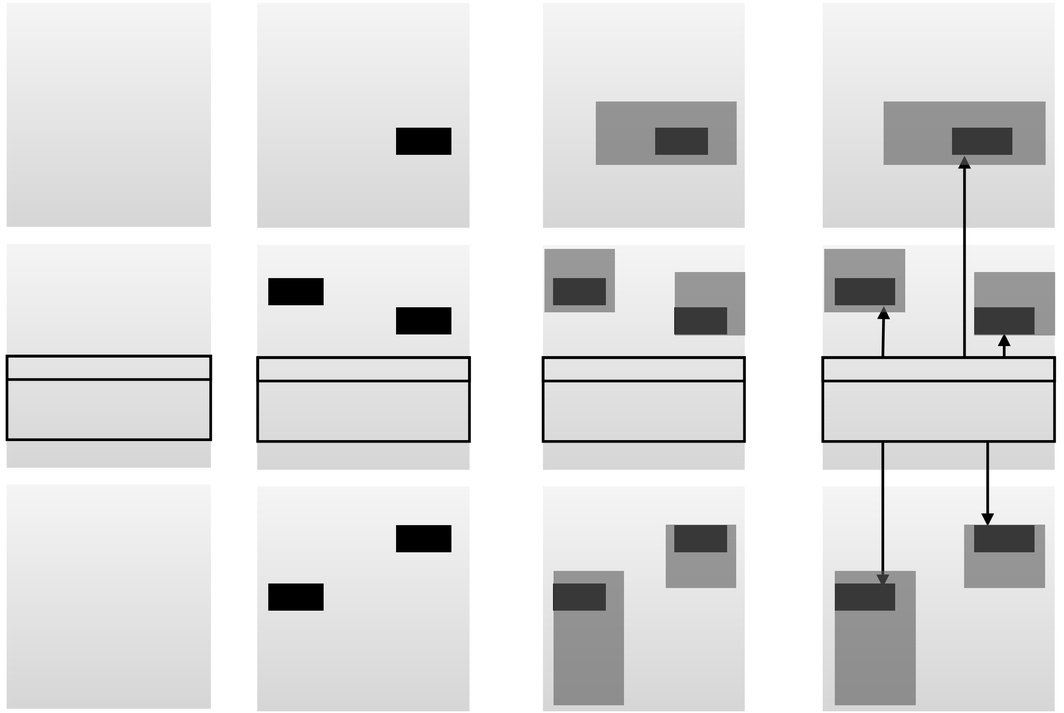
COVID-19 Admissions

Returns the original table, and the natural language descriptions:

- Percentage hospitalized increases with age
- Fatalities increase with age
- ICU admissions lowest among adults aged 20-44
- ICU admissions highest among adults aged 75-84

Task 2B: Document-level contextual retrieval

Incorporate body text descriptions into table and figure retrieval

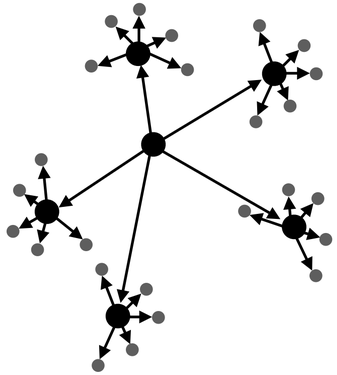


document with table

coreferences defined

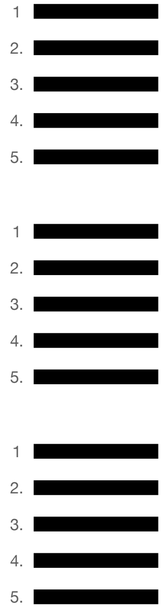
coreference context captured

context linked back to table

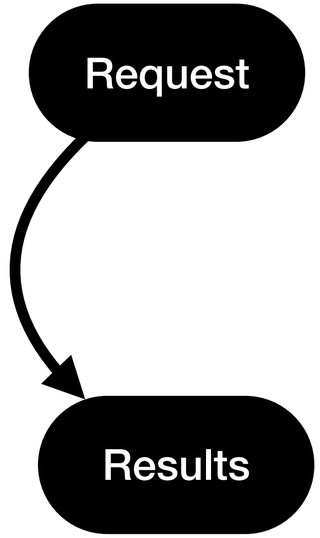


body-text content captured in graph structure

index the graph structures in a single document



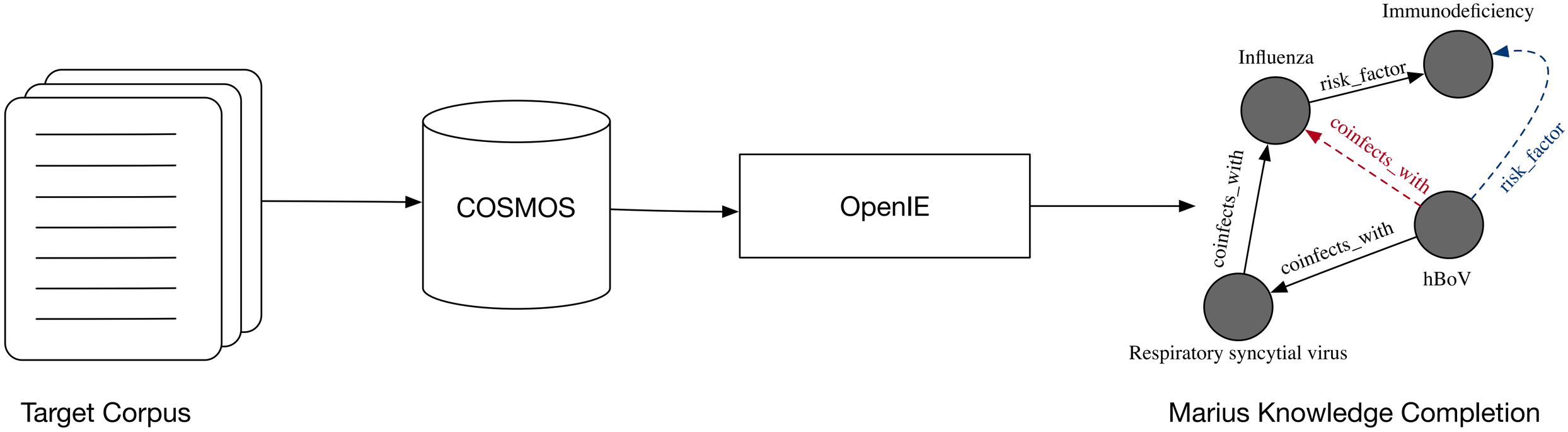
index the graph structures across all documents in corpus



Retrieve object plus body text content with query

Task 2C: Automated KBC with CosmosIE and Marius

- Use state of the art natural language models to extract knowledge graphs from target corpora, at scale
- Deploy state of the art scalable graph embedding framework, Marius, for knowledge base completion



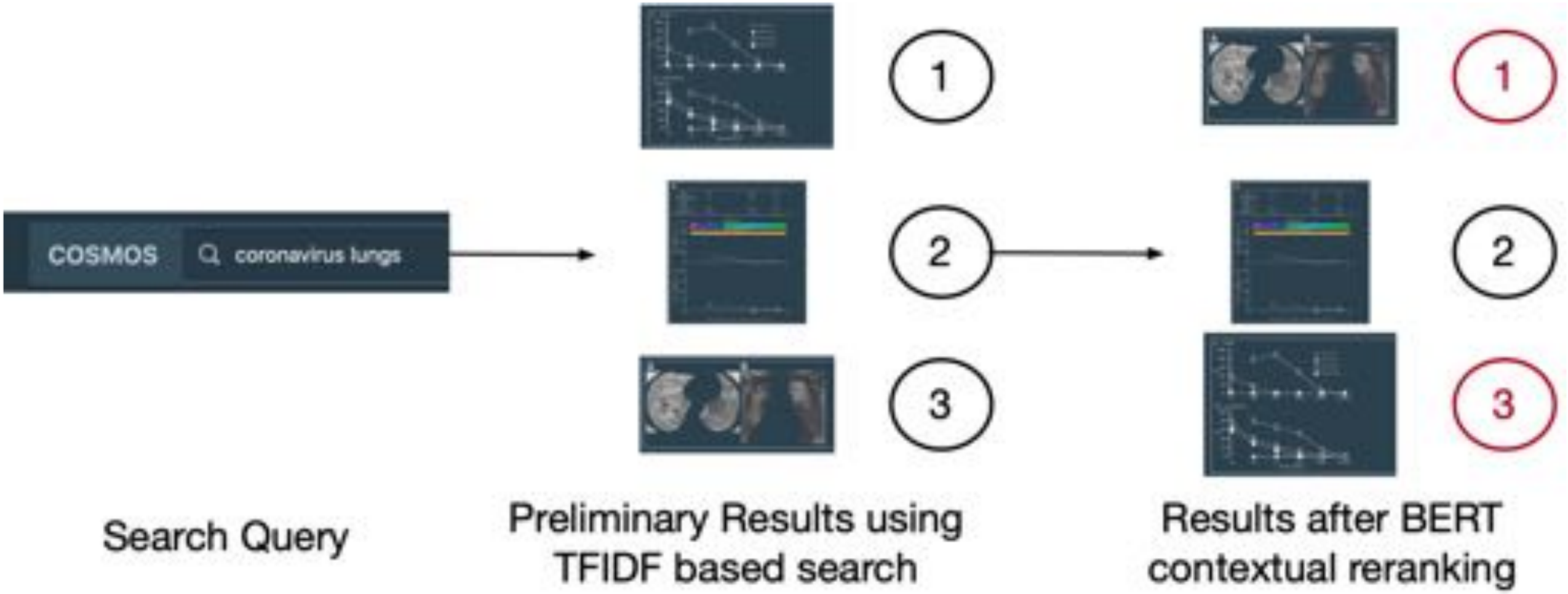
Target Corpus

Marius Knowledge Completion

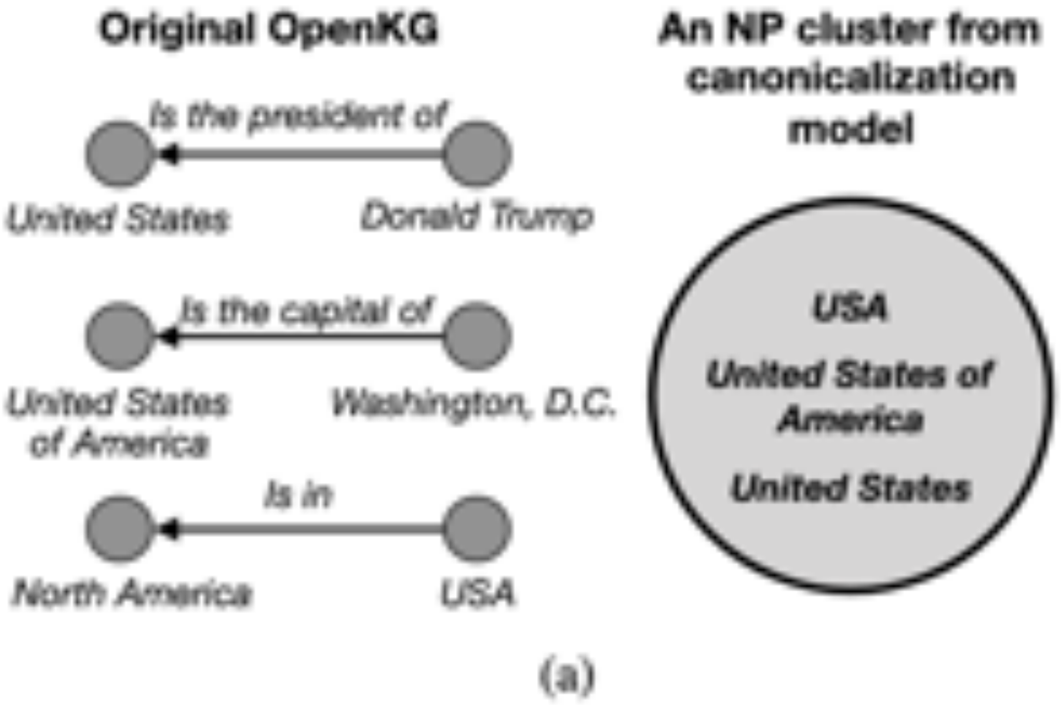
Task 2C: Automated KBC with CosmosIE and Marius

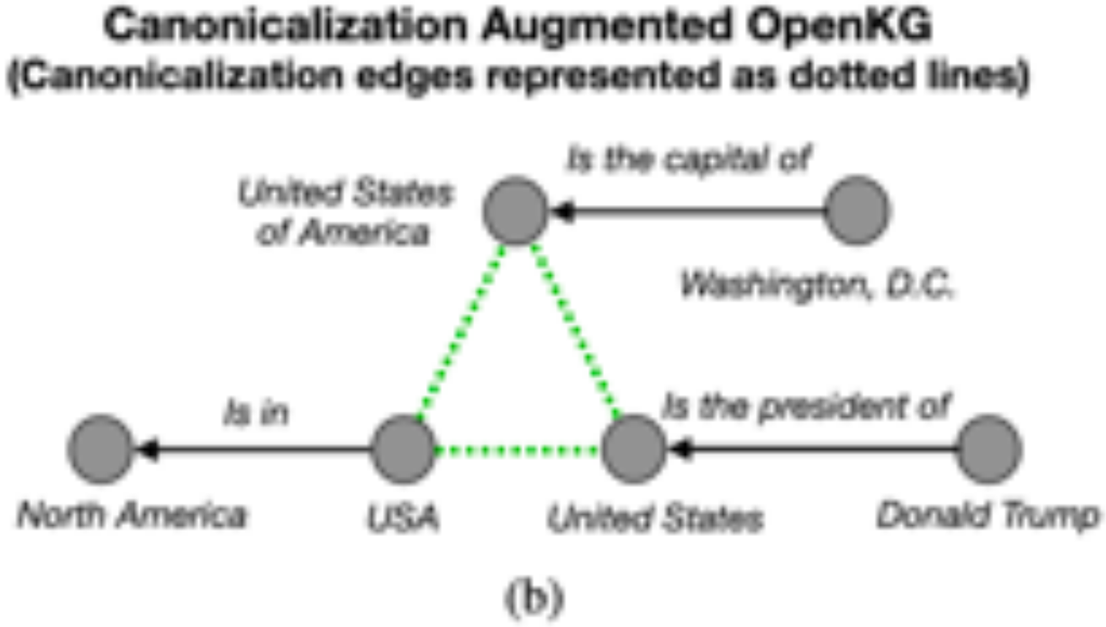
Improving COSMOS Search using Contextual Reranking

- Use deep learning to predict the most relevant search results using natural language understanding, improving over basic string-based retrieval.



Task 2C: Automated KBC with CosmosIE and Marius





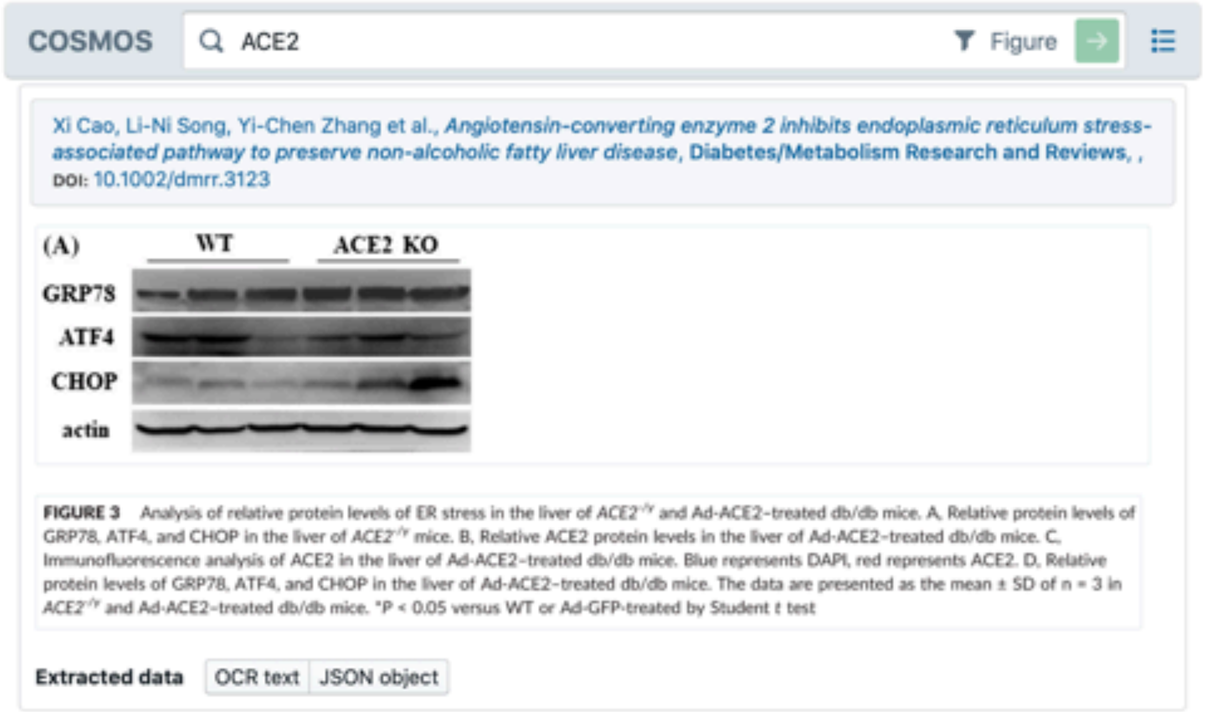
[1] Gupta et. al: https://www.aclweb.org/anthology/D19-1036/

- Construct an Open Knowledge Base from COSMOS, then use Canonicalized-infused Representations [1] to canonicalize extracted entities
- Enhance COSMOS with question-answering, KB completion capabilities

Task 2D: Release public COSMOS API over COVID-19 set

Allow third-party search/retrieval of table, figure, eqns

- Current interface sits on the COSMOS REST-ful API service
- Expand capabilities to allow team interfacing with COSMOS extractions, including:
  - Direct search of visual objects and contexts, multiple retrieval/rank methods
  - Stable URLs, with versioned access to past, current outputs:
    - table, figure, equations
    - custom trained word embedding with normalization variants (“raw” vs. lower cased, ligature-cleaned)



Our vision: service-oriented knowledge extraction

Agent

Applications

Services

Infrastructure and data

- COSMOS: An intelligent (AI-driven) virtual assistant that can consider and reason about semantics when linking/synthesizing artifacts from xDD-automated publication stream
- REST-ful APIs that enable users to build custom web and analytical applications over the scientific literature (simple to complex)
- Services that bring fine-grained knowledge extraction from publications within reach of every scientist—without requiring programming expertise (interactions based on natural language and point-and-click interfaces) or need to find and aggregate documents
- Compute infrastructure and document acquisition pipeline to support continual updating over diverse domains of science and engineering