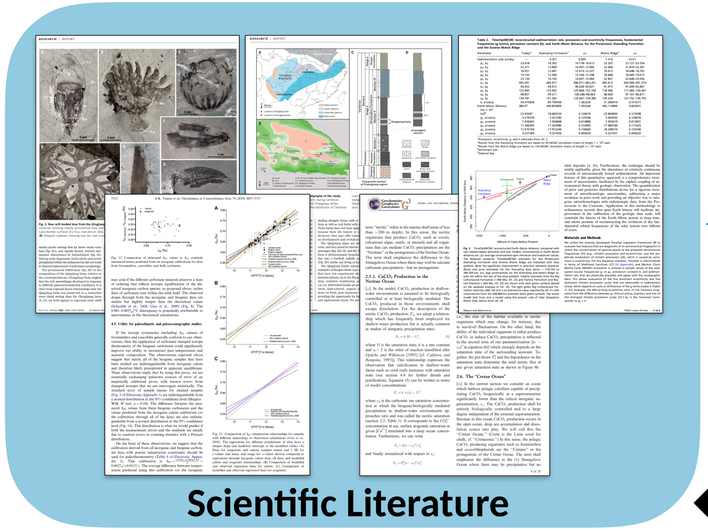
UW-COSMOS Final Demo:

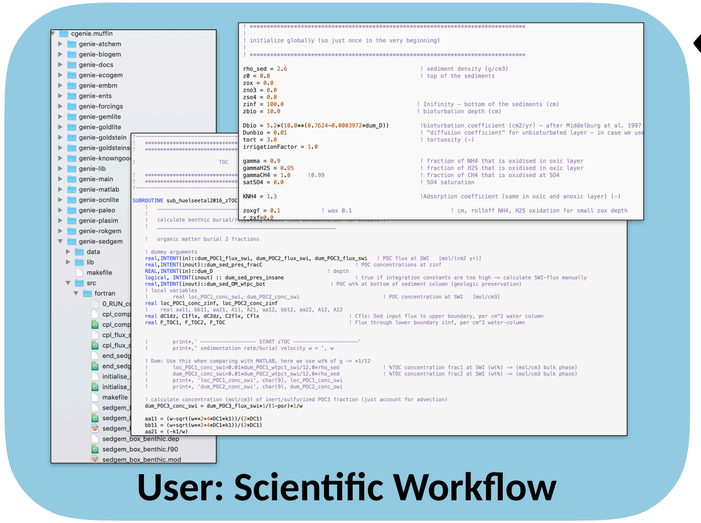
extracting semantic knowledge and observational data from scientific publications

Shanan Peters, Theo Rekatsinas, Miron Livny

Goal: Develop an AI technical assistant capable of assimilating information and knowledge from scientific publications



Scientific Literature



User: Scientific Workflow

These are our "raw data"

We want actionable data products

COSMOS

Universal Representation

- Retrieve-and-Read Q&A
- On-demand Knowledge Bases
- Basic Synonym Service
- Dataset Generation

- User-focused
- Answer questions, generate aggregate results
- Interactive evaluation and AI model training

- Entity relations across tables, figures, equations
- Discover knowledge in millions of papers

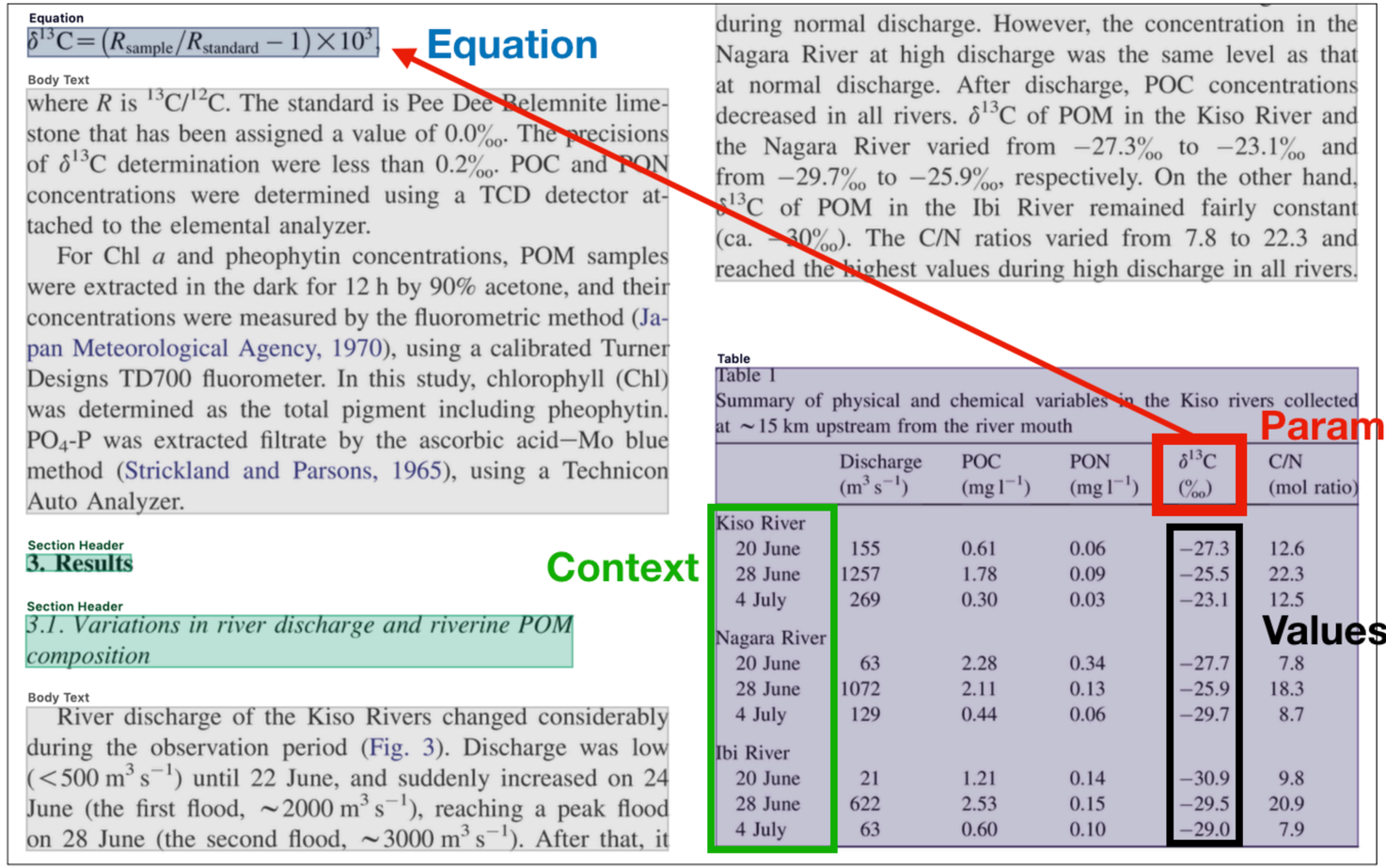
- Fine-tuned ML models over unified contextual representation to support diverse tasks
- Continual updating

: currently working on it

: alpha version completed

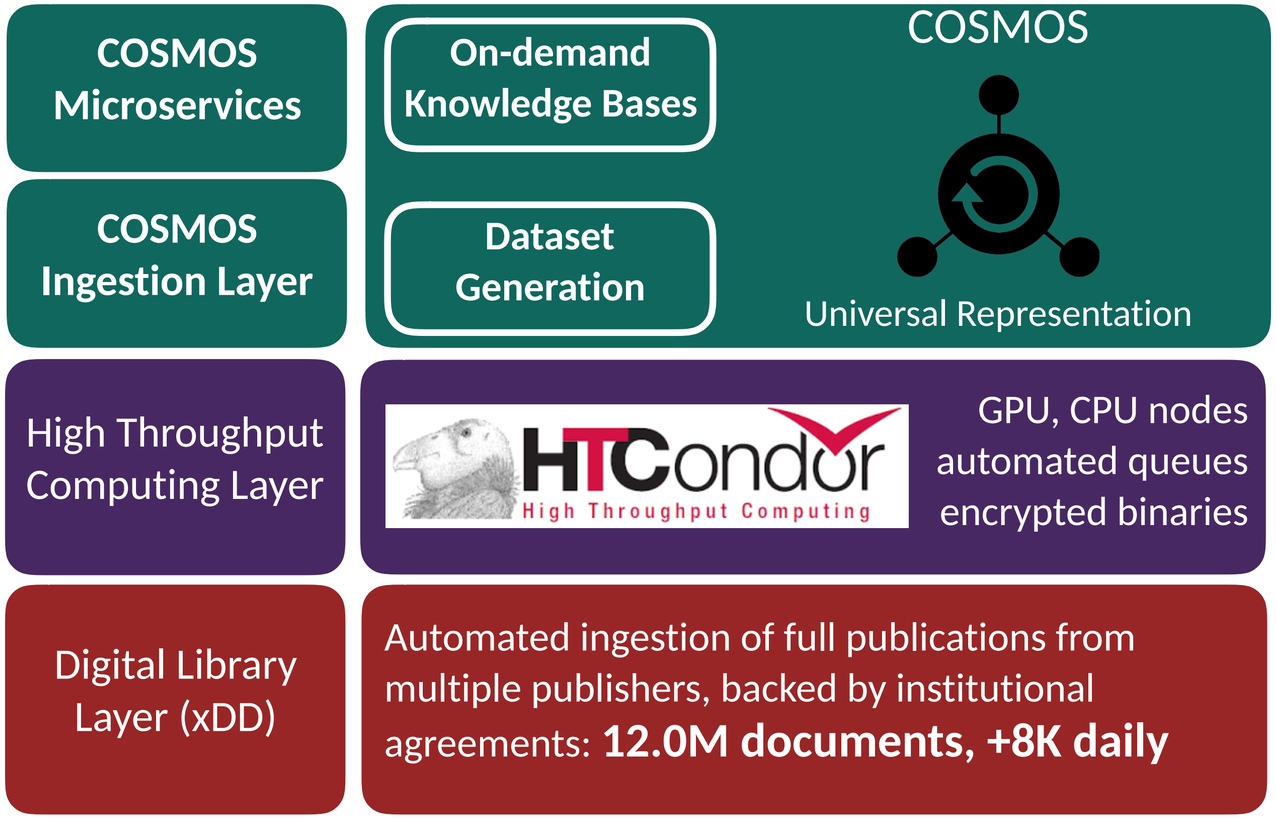
Semantic knowledge extraction

- **Goal:** create on-demand repositories with ***machine-readable enriched content*** that span all modalities in published documents and across domains of science
- **Our focus:** identify, extract, and collect semantic knowledge and data from the published literature
- **Semantic knowledge:** concepts, facts, relationships, entities, and observational data related to specific scientific phenomena and models
- Foundational for basic scientific research and for evaluating, revising and enhancing scientific models



Challenge: reasoning about semantic context

UW-COSMOS: an end-to-end stack for accelerating scientific research



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, API for simple queries
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA DEPARTMENT OF ENERGY UNITED STATES OF AMERICA

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)

Pertinent example: COSMOS COVID-19 response

Received: 30 January 2020 | Accepted: 4 February 2020
DOI: 10.1002/jmv.25707

REVIEW

JOURNAL OF MEDICAL VIROLOGY WILEY

Potential interventions for novel coronavirus in China: A systematic review

Lei Zhang | Yunhui Liu

Department of Neurosurgery, Shengjing Hospital of China Medical University, Shenyang, Liaoning, China

Correspondence
Yunhui Liu, Department of Neurosurgery, Shengjing Hospital, China Medical University, No. 36 Sanhao Street, Heping, Shenyang, 110004 Liaoning, China.
Email: liuyh@sj-hospital.org

Funding information
Project of Key Laboratory of Neurooncology in Liaoning Province, China,
Grant/Award Number: 112-2400017005

Abstract
An outbreak of a novel coro
significant threats to interr
treatment for this virus, th
control the spread of diseas
treatment options related to
fection and we have found th
and antiviral treatments shou
nutritional status of each inf
tration of general treatmen
cluding influenza vaccine sho

Emerging Microbes & Infections
2020, VOL. 9
https://doi.org/10.1080/22221751.2020.1745095

EMi Taylor & Francis Taylor & Francis Group

REVIEW

OPEN ACCESS Check for updates

Laboratory diagnosis of emerging human coronavirus infections – the state of the art

Michael J. Loeffelholz[a] and Yi-Wei Tang[b]

Platform, Shanghai, People's Republic of China

man coronavirus (HCoV) infections at the beginning of the twenty-
dily available, accurate and fast diagnostic testing methods. The
us infections have evolved substantially, with the development of
sts for emerging ones. Newer laboratory methods are fast, highly
the conventional gold standards. This presentation reviews the
onaviruses by focusing on the coronavirus disease 2019 (COVID-
ypically do not result in the production of purulent sputum. Thus,
nethod used to obtain a specimen for testing. Nasopharyngeal
cimen may need to be obtained by bronchoscopy. Alternatively,
e likelihood of the SARS-CoV-2 being present in the nasopharynx
f-care molecular devices are currently under development for fast
ese assays are simple, fast and safe and can be used in the local

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

INTERNATIONAL SOCIETY FOR INFECTIOUS DISEASES

A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action

Qianying Lin[a,1], Shi Zhao[b,c,1], Daozhou Gao[d], Yijun Lou[e], Shu Yang[f], Salihu S. Musa[e], Maggie H. Wang[b,c], Yongli Cai[g], Weiming Wang[g,*], Lin Yang[h,*], D

[a] Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI, USA
[b] JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong, China

Contents lists available at ScienceDirect

Diabetes & Metabolic Syndrome: Clinical Research & Reviews

journal homepage: www.elsevier.com/locate/dsx

Clinical considerations for patients with diabetes in times of COVID-19 epidemic

Ritesh Gupta [a], Amerta Ghosh [a], Awadhesh Kumar Singh [b], Anoop Misra [a, c, d, *]

[a] Fortis CDOC Hospital, Chirag Enclave, New Delhi, India
[b] GD Hospital and Diabetes Institute, Kolkata, India
[c] National Diabetes, Obesity and Cholesterol Foundation, New Delhi, India

16 March 2020 *Imperial College COVID-19 Response Team*

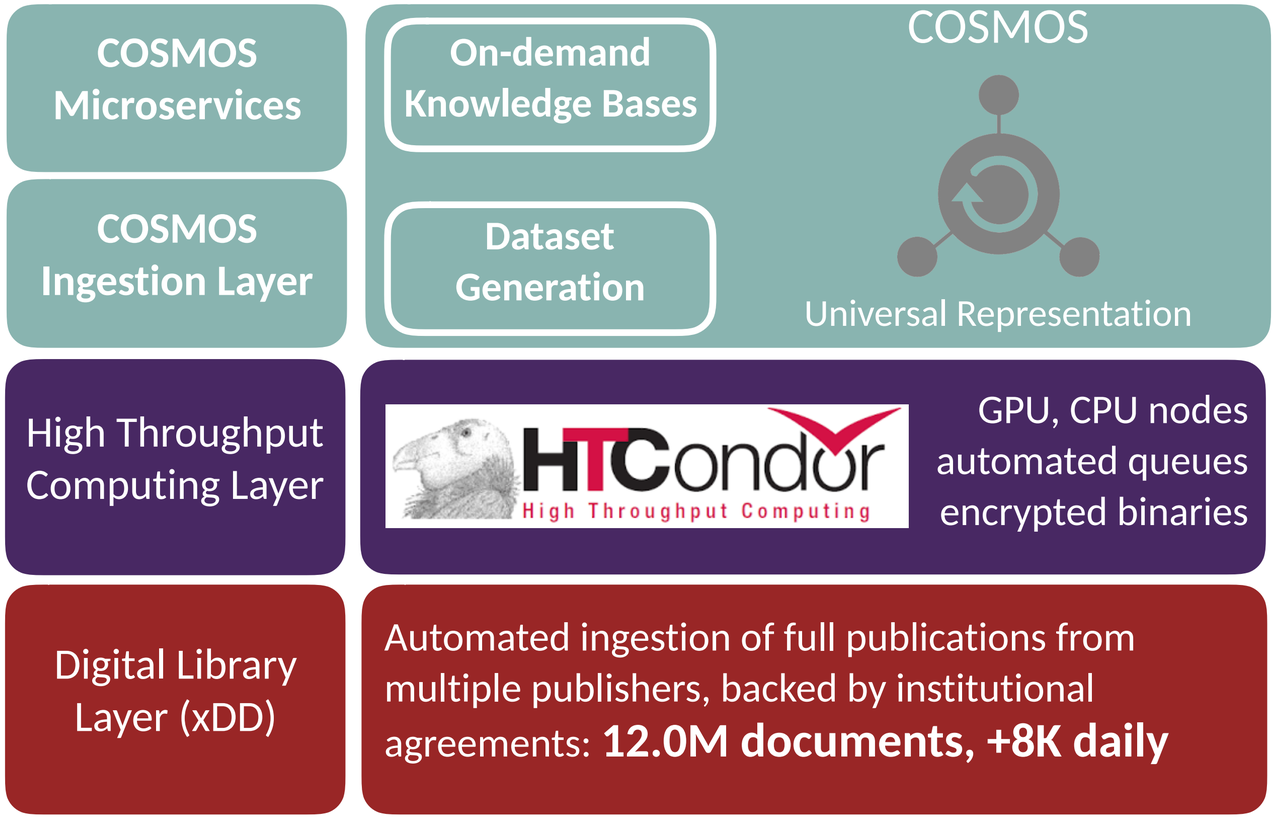
Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand

Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Lucy C Okell, Sabine van Elsland, Hayley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick GT Walker, Caroline Walters, Peter Winskill, Charles Whittaker, Christl A Donnelly, Steven Riley, Azra C Ghani.

On behalf of the Imperial College COVID-19 Response Team

WHO Collaborating Centre for Infectious Disease Modelling
MRC Centre for Global Infectious Disease Analysis

UW-COSMOS: an end-to-end stack for accelerating scientific research



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, API for simple queries
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA DEPARTMENT OF ENERGY UNITED STATES OF AMERICA

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)

COVID-19 response

COVID-19 Open Research Dataset (CORD-19)

Access this dataset to help with the fight against COVID-19

A Free, Open Resource for the Global Research Community

In response to the COVID-19 pandemic, the Allen Institute for AI has partnered with leading research groups to prepare and distribute the COVID-19 Open Research Dataset (CORD-19), a free resource of over 45,000 scholarly articles, including over 33,000 with full text, about COVID-19 and the coronavirus family of viruses for use by the global research community.

This dataset is intended to mobilize researchers to apply recent advances in natural language processing to generate new insights in support of the fight against this infectious disease. The corpus will be updated weekly as new research is published in peer-reviewed publications and archival services like bioRxiv, medRxiv, and others.

xDD: real-time corpus construction

STATUS: PRODUCTION READY

- Automated fetching and storage of documents and document metadata from multiple commercial and open-access publishers
- >12M full-content publications, domain agnostic, growing ~10K per day; new acquisition can be prioritized based on many criteria
- Mechanism for uses to define/supply hierarchical vocabularies for indexing across entire corpus (GitHub link below); simple REST-ful API to summarize terms across all documents
- COVID-19 set made in hours: currently >42K relevant full-content documents, growing ~2K/day


```
"name": "covid-19",
"base_classification": "virus",
"source": "https://raw.githubusercontent.com/UW-
"case_sensitive": true,
"last_updated": "2020-03-29T05:00:00.000Z",
"term_hits": {
  "angiotensin-converting enzyme 2": 3137,
  "MERS": 40696,
  "Spike Protein": 95,
  "N-Protein": 1847,
  "coronavirus": 76582,
  "COVID19": 65,
  "nucleocapsid protein": 13187,
  "hemagglutinin-esterase": 376,
  "TMPRSS2": 10036,
  "S Protein": 3748,
  "N Protein": 1847,
  "S-Protein": 3748,
  "ACE2": 33836,
  "COVID-19": 3492,
  "SARS-CoV-2": 1784,
  "2019-nCoV": 2704,
  "SARS": 152805,
  "betacoronavirus": 773,
  "MERS-CoV": 26365,
  "SARSr-CoVs": 69,
  "SARS-CoV": 43257
```


https://github.com/UW-Deepdive-Infrastructure/dictionary_example

https://geodeepdive.org/api/dictionaries?dict=covid-19&show_terms

xDD: corpus exploration

STATUS: PRODUCTION READY


```
{
  "pubname": "Journal of Medical Virology",
  "publisher": "Wiley",
  "_gddid": "5e73823d998e17af82650cff",
  "title": "Identification of coronavirus sequences in carp cDNA from Wuhan, China",
  "doi": "10.1002/jmv.25751",
  "coverDate": "",
  "URL": "https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25751",
  "authors": "Conway, Michael J.",
  "highlight": [
    "December 2019, and causes a respiratory illness called COVID-19, which can spread from",
    "respiratory illness called COVID-19, which can spread from person to person. As of"
  ]
},
```


- ElasticSearch index constructed over document full-texts, simple REST-ful API endpoints that expose mentions and surrounding context
- Retrieve intersections of domain vocabularies over document full texts (e.g., identify COVID-19 document, then ask “what genes does this document mention” and get answer immediately via separate dictionary)

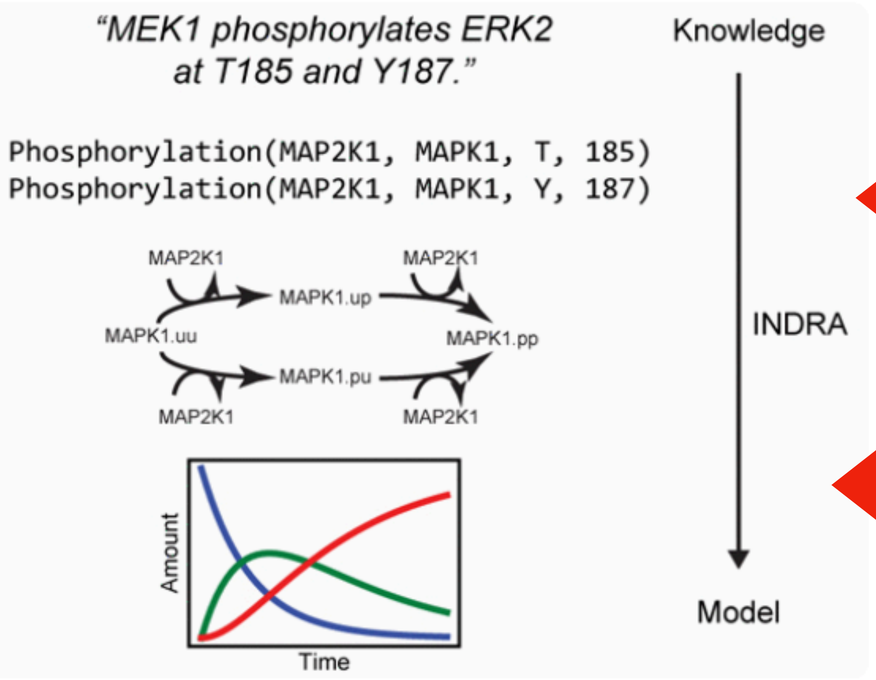
Example: https://geodeepdive.org/api/snippets?term=COVID-19&clean&full_results
Basic documentation: https://geodeepdive.org/api/snippets

xDD: code execution over corpus

INPUTS: large domain vocabulary, INDRA code container, xDD-supplied full texts

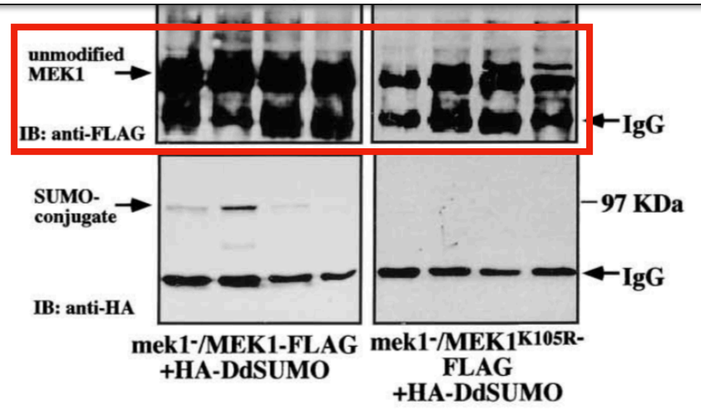
INDRA
Integrated Network and Dynamical Reasoning Assembler

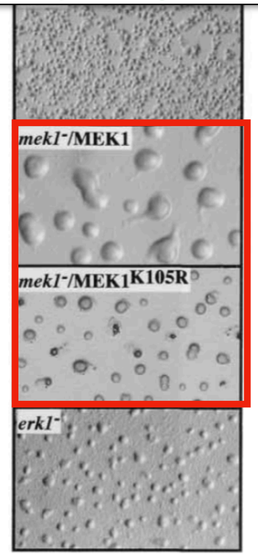

```
"title": "Regulated SUMOylation and Ubiquitination of DdMEK1 Is Required for Proper Chemotaxis",
"doi": "10.1016/S1534-5807(02)00186-7",
"coverDate": "June 2002",
"URL": "http://www.sciencedirect.com/science/article/pii/S1534580702001867",
"authors": "Sobko, Alex; Ma, Hui; Firtel, Richard A.",
"hits": 5,
"highlight": [
  ", San Diego 9500 Gilman Drive La Jolla, California 92093 Summary MEK1, which is required for aggregation",
  " to chemoattractant stimulation. SUMOylation is required for MEK1's function and its translocation from the
  " to the cytosol and cortex, including the leading edge of chemotaxing cells. MEK1 in which the site",
  " of SUMOylation is mutated is retained in the nucleus and does not complement the mek1 null phenotype",
  ". Constitutively active MEK1 is cytosolic and is constitutively SUMOylated, whereas the corresponding"
```

API-based access to full-text mentions

Pipeline to generate INDRA statements

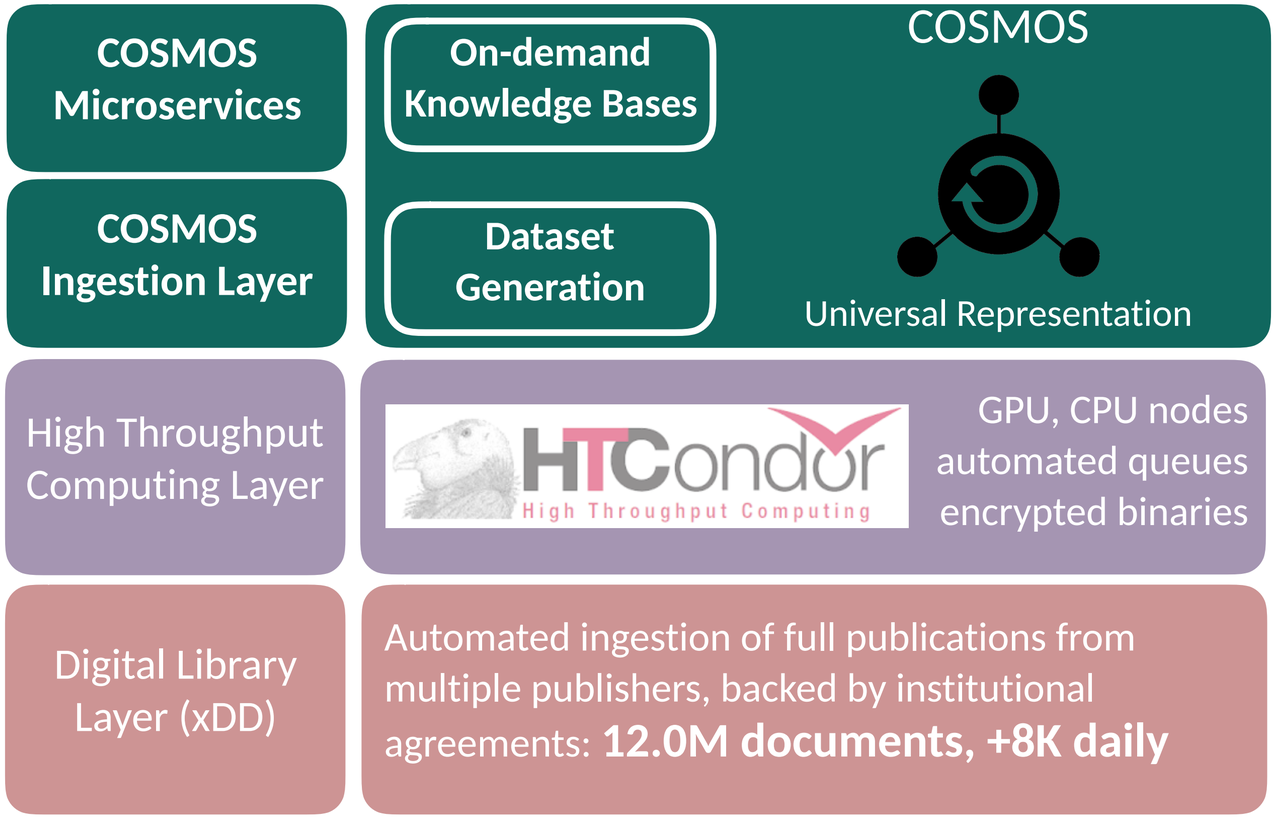




C.

Genotype (Strain)	Speed (μm/min)	Directional change (deg)	Roundness (%)	Directionality
Wild-type (KAx-3)	8.5±0.5	12.5±4.4	43.8±4.2	0.90±0.6
mek1-/MEK1 (ASF1)	8.1±0.4	12.4±0.8	41.3±4.8	0.89±0.05
mip1 null (HMF3)	5.7±0.3	41.1±4.5	51.9±0.7	0.59±0.06
mek1-/MEK1K105R	5.0±0.4	42.6±4.4	45.9±6.1	0.66±0.04

UW-COSMOS: an end-to-end stack for accelerating scientific research



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, API for simple queries
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA DEPARTMENT OF ENERGY UNITED STATES OF AMERICA

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)

COSMOS step 1: make sense of visual documents

STATUS: PRODUCTION READY

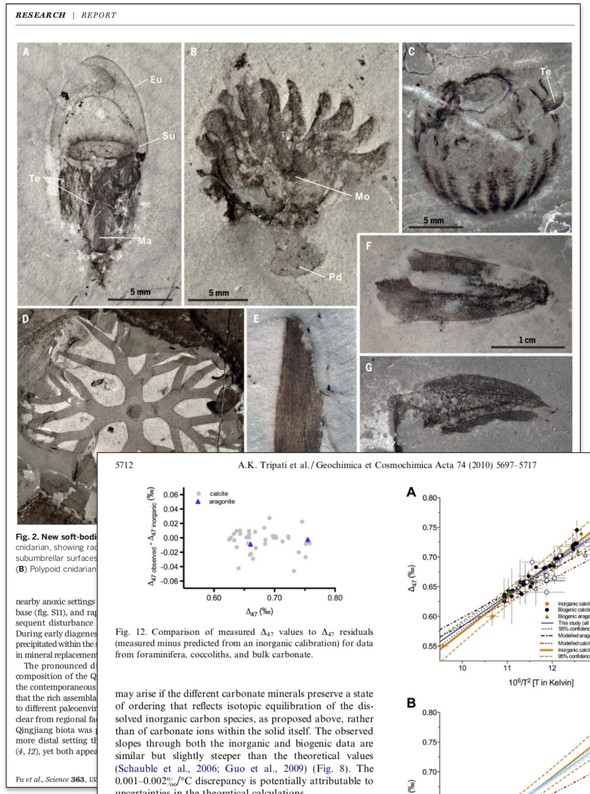


Fig. 2. New soft-bodi
cnidarian, showing rac
subumbrellar surfaces
(**B**) Polypoid cnidarian

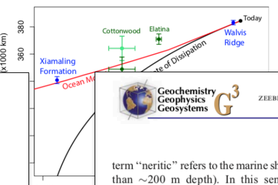
nearby anoxic settings
base (fig. S1I), and ru
sequent disturbance
During early diagenes
precipitated within the
in mineral replacemen
The pronounced d
composition of the Q
the contemporaneous
that the rich assembla
to different paleoenvir
clear from regional fa
Qingjiang biota was
more distal setting t
(*4*, *12*), yet both appea

Pu et al., Science **363**, 13

Table 2. TimeOptMCMC reconstructed sedimentation rate, precession and eccentricity frequencies, fundamental frequencies (*g* terms), precession constant (*k*), and Earth–Moon distance, for the Proterozoic Xiamaling Formation and the Eocene Walvis Ridge

Parameter	Today*	Xiamaling Formation†	±σ	Walvis Ridge‡	±σ
Sedimentation rate (cm/ky)	—	0.357	0.005	1.316	0.011
p_1, ky	23.678	14.392	14.178–14.613	23.335	23.121–23.554
p_2, ky	22.371	13.899	14.407–13.983	22.066	21.874–22.261
p_3, ky	18.951	12.497	12.673–12.327	18.613	18.448–18.781
p_4, ky	19.103	12.569	12.744–12.398	18.848	18.685–19.013
p_5, ky	23.120	14.192	14.407–13.983	22.827	22.608–23.050
e_1, ky	405.091	405.077	406.971–403.201	405.613	403.969–407.270
e_2, ky	94.932	94.912	96.028–93.821	91.975	91.099–92.867
e_3, ky	123.945	123.955	125.868–122.100	118.946	117.483–120.447
e_4, ky	98.857	99.211	100.386–98.064	98.009	97.161–98.871
e_5, ky	130.781	131.392	133.467–129.380	129.236	127.752–130.755
k, arcsec/y	50.475838	85.790450	1.362320	51.280910	0.515371
Earth–Moon distance, km × 10^3	384.4§	340.855800	1.293260	383.110800	0.822051
h/d¶	23.93447	18.684750	0.126679	23.804850	0.123398
g_1, arcsec/y	5.579378	5.531285	0.129246	5.494302	0.128076
g_2, arcsec/y	7.456665	7.456848	0.014886	7.452619	0.013001
g_3, arcsec/y	17.366595	17.320480	0.152805	17.480760	0.115325
g_4, arcsec/y	17.910194	17.912240	0.158683	18.348310	0.135346
g_5, arcsec/y	4.257564	4.257456	0.000020	4.257451	0.000020

*Precession, eccentricity, *g*, and *k* estimates from ref. 2.
†Results from the Xiamaling Formation are based on 50 MCMC simulation chains of length 1 × 10^6 each.
‡Results from the Walvis Ridge are based on 150 MCMC simulation chains of length 2 × 10^5 each.
§Semimajor axis.
¶Sidereal day.



tidal deposits (4, 16). Furthermore, the technique should be widely applicable, given the abundance of relatively continuous records of astronomically forced sedimentation. An important feature of this quantitative approach is a comprehensive treatment of uncertainties, facilitated by the explicit coupling of astronomical theory with geologic observation. The quantification of prior and posterior distributions allows for a rigorous treatment of astrochronologic uncertainties, addressing a major

5712 A.K. Tripati et al. / Geochimica et Cosmochimica Acta 74 (2010) 5697–5717

Fig. 12. Comparison of measured Δ_{47} values to Δ_{47} residuals (measured minus predicted from an inorganic calibration) for data from foraminifera, coccoliths, and bulk carbonate.

may arise if the different carbonate minerals preserve a state of ordering that reflects isotopic equilibration of the dissolved inorganic carbon species, as proposed above, rather than of carbonate ions within the solid itself. The observed slopes through both the inorganic and biogenic data are similar but slightly steeper than the theoretical values (Schauble et al., 2006; Guo et al., 2009) (Fig. 8). The 0.001–0.002‰/‰C discrepancy is potentially attributable to uncertainties in the theoretical calculations.

4.5. Utility for paleoclimatic and paleoceanographic studies

If the isotope systematics (including Δ_{47} values) of foraminifera and coccoliths generally conform to our observations, then the application of carbonate clumped isotope thermometry of the biogenic carbonates could significantly improve our ability to reconstruct past temperatures and seawater composition. The observations reported above suggest that nearly all of the biogenic samples that have been studied are indistinguishable from inorganic calcite and therefore likely precipitated in apparent equilibrium. These observations imply that by using this proxy, we are essentially exchanging unknown sources of error of an empirically calibrated proxy with known errors from clumped isotopes that we can interrogate statistically. The standard error of sample means for cleaned samples (Fig. 2 of Electronic Appendix 1) are indistinguishable from a normal distribution at the 95% confidence level (Shapiro–Wilk W test; $p = 0.54$). The difference between the measured Δ_{47} values from these biogenic carbonates and the values predicted from the inorganic calcite calibration (or the calibration through all of the data) are also indistinguishable from a normal distribution at the 95% confidence level (Fig. 14). This distribution is what we would predict if both the measurement errors and the residuals are simply due to random errors in counting statistics with a Poisson distribution.

On the basis of these observations, we suggest that the calibration derived from all inorganic and biogenic carbonate data with precise temperature constraints should be used for paleothermometry (Table 6 of Electronic Appendix 2). This calibration is $\Delta_{47} = \frac{0.0326_A(\pm 0.0006_A)\cdot 10^6}{T^2} + 0.0627_A(\pm 0.0115_A)$. The average difference between temperatures predicted using this calibration (or the inorganic

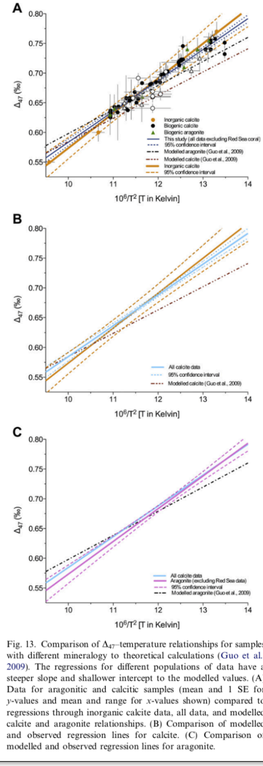


Fig. 13. Comparison of Δ_{47}–temperature relationships for samples with different mineralogy to theoretical calculations (Guo et al., 2009). The regressions for different populations of data have a steeper slope and shallower intercept to the modelled values. (**A**) Data for aragonitic and calcitic samples (mean and 1 SE for *y*-values and mean and range for *x*-values shown) compared to regressions through inorganic calcite data, all data, and modelled calcite and aragonite relationships. (**B**) Comparison of modelled and observed regression lines for calcite. (**C**) Comparison of modelled and observed regression lines for aragonite.

term "neritic" refers to the marine shelf areas of less than ~200 m depth). In this sense, the neritic organisms that produce $CaCO_3$ such as corals, calcareous algae, snails, or mussels and all organisms that can mediate $CaCO_3$ precipitation are the "Neritans" or the protagonists of the Neritan Ocean. The term shall emphasize the difference to the Strangelove Ocean where there may well be calcium carbonate precipitation—but no protagonists.

2.5.1. $CaCO_3$ Production in the Neritan Ocean

[22] In the model, $CaCO_3$ production in shallow-water environments is assumed to be biologically controlled or at least biologically mediated. The $CaCO_3$ produced in those environments shall escape dissolution. For the description of the neritic $CaCO_3$ production, F_n, we adopt a relationship which has frequently been employed for shallow-water production but is actually common in studies of inorganic precipitation rates:

$$F_n = k\ (\Omega - 1)^{\eta}, \tag{5}$$

where Ω is the saturation state, k is a rate constant and $\eta = 2$ is the order of reaction (modified after *Opdyke and Wilkinson* [1993] [cf. *Caldeira and Rampino*, 1993]). This relationship expresses the observation that calcification in shallow-water facies such as coral reefs increases with saturation state (see section 4.4 for further details and justification). Equation (5) can be written in terms of model concentrations:

$$F_n = k\ (c/c_n - 1)^2,$$

where c_n is the carbonate ion saturation concentration at which the biogenic/biologically mediated precipitation in shallow-water environments approaches zero and was called the neritic saturation (section 2.3, Table 1). It corresponds to the CO_3^{2-} concentration at say, surface aragonite saturation at given $[Ca^{2+}]$ translated into a deep ocean concentration. Furthermore, we can write

$$F_n = k(c - c_n)^2/c_n^2$$

and finally normalized with respect to c_s:

$$F_n = F_n^0(c - c_n)^n/c_s^2,$$

with $F_n^0 = k \times c_s^2/c_n^2$. The neritic, biogenic output flux then reads:

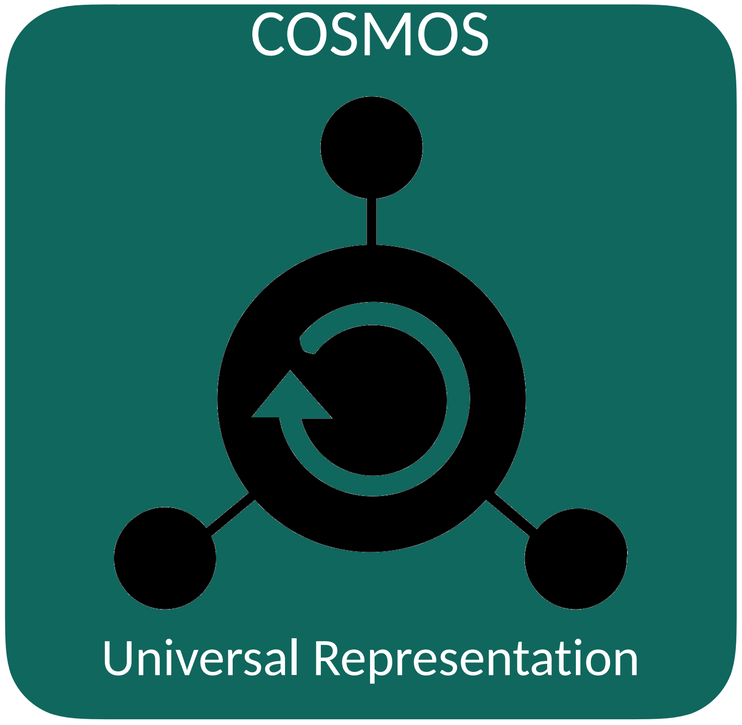
$$F_{out} = F_n = \begin{cases} 0 & c < c_n \\ F_n^0\ (c - c_n)^2/c_s^2 & c \geq c_n. \end{cases} \tag{6}$$

The pre-factor F_n^0 scales the neritic production and is just equal to it at $c = c_n + c_s$, i.e., at this point $F_n^0 = F_n$ (Figure 4b shows an example with $F_n^0 = 1.1$).

[23] The relation between the model parameterization of neritic $CaCO_3$ production and natural processes occurring in the ocean may be described as follows. The pre-factor F_n^0 is a measure of the $CaCO_3$ production of the total population of shallow-water calcifiers in the Neritan Ocean. This includes factors such as the ecological success of those groups but also the areal extent of shelves, i.e., the size of the habitat available to neritic organisms which may change, for instance, due to sea-level fluctuations. On the other hand, the ability of the individual organism to either produce $CaCO_3$ or induce $CaCO_3$ precipitation is reflected in the second term of our parameterization $[(c - c_n)^2$ in equation (6)] which strongly depends on the saturation state of the surrounding seawater. Together, the pre-factor F_n^0 and the dependence on the saturation state determine the total neritic flux at any given saturation state as shown in Figure 4b.

2.6. The "Cretan Ocean"

[24] In the current section we consider an ocean which harbors pelagic calcifiers capable of precipitating $CaCO_3$ biogenically at a supersaturation significantly lower than the critical inorganic supersaturation, c_c. The $CaCO_3$ production shall be entirely biologically controlled and to a large degree independent of the external supersaturation. Because in this ocean $CaCO_3$ production occurs in the open ocean, deep sea accumulation and dissolution comes into play. We will call this the "Cretan Ocean." (Creta is the Latin word for chalk, cf. "Cretaceous.") In this sense, the pelagic $CaCO_3$ producing organisms such as foraminifera and coccolithophorids are the "Cretans" or the protagonists of the Cretan Ocean. The term shall emphasize the difference to the (1) Strangelove Ocean where there may be precipitation but no



COSMOS step 1: train a model to visually parse

STATUS: PRODUCTION READY

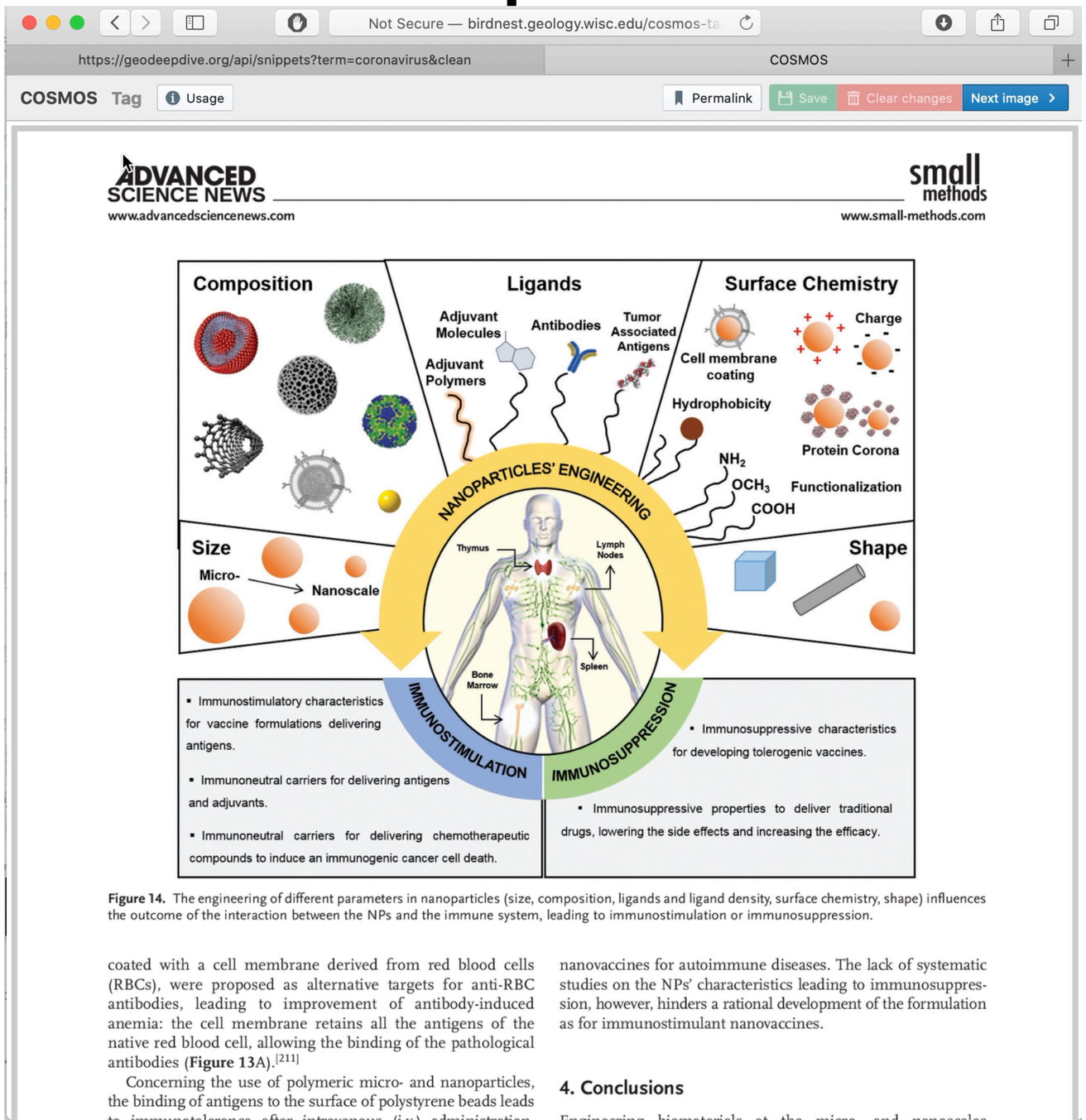


Figure 14. The engineering of different parameters in nanoparticles (size, composition, ligands and ligand density, surface chemistry, shape) influences the outcome of the interaction between the NPs and the immune system, leading to immunostimulation or immunosuppression.

coated with a cell membrane derived from red blood cells (RBCs), were proposed as alternative targets for anti-RBC antibodies, leading to improvement of antibody-induced anemia: the cell membrane retains all the antigens of the native red blood cell, allowing the binding of the pathological antibodies (**Figure 13**A).[211]

Concerning the use of polymeric micro- and nanoparticles, the binding of antigens to the surface of polystyrene beads leads

nanovaccines for autoimmune diseases. The lack of systematic studies on the NPs' characteristics leading to immunosuppression, however, hinders a rational development of the formulation as for immunostimulant nanovaccines.

4. Conclusions

- Custom web browser-based software to identify and classify objects in a visual reference space (deployable over any image for any purpose)
- Interface to make explicit links between tagged objects (e.g., variables to equations)
- Interface to view and assess annotator tags, share links to pages

COSMOS step 1: train a model to visually parse

STATUS: PRODUCTION READY

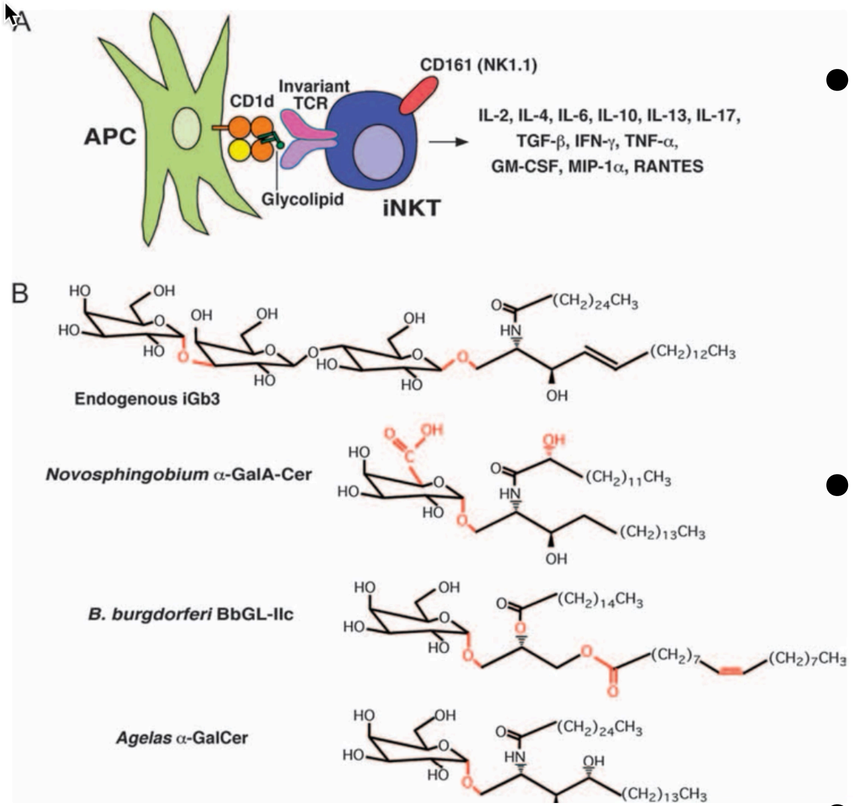
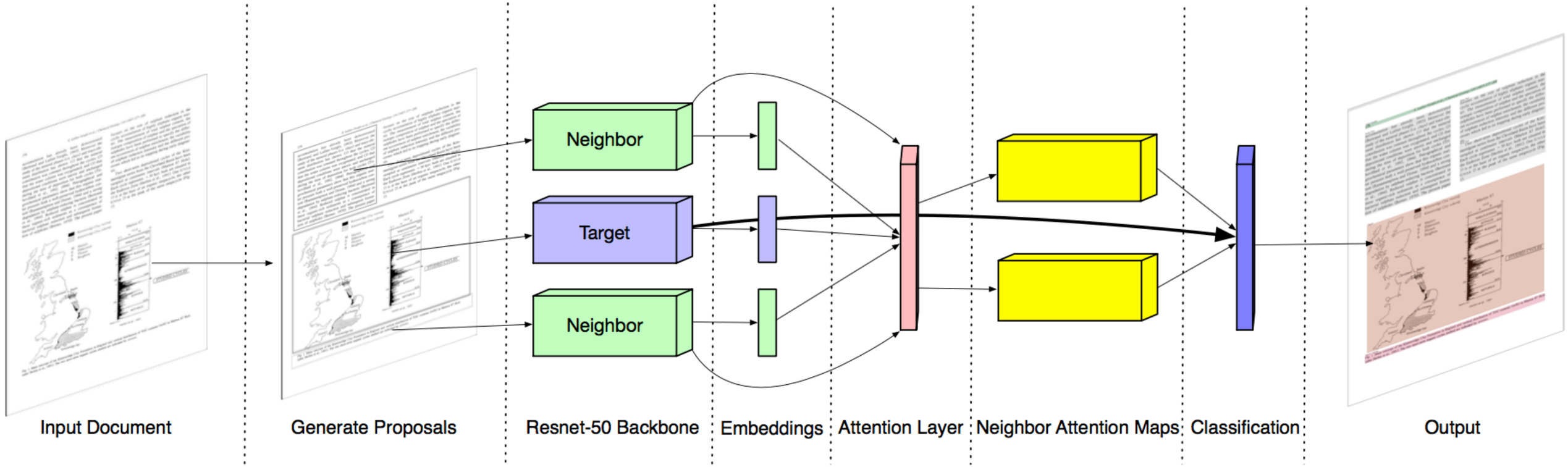


Figure 1 Specificity of iNKT cells. (A) iNKT cells express a semi-invariant T-cell receptor (TCR) that is specific for glycolipid or lipid antigens presented by CD1d on antigen-presenting cells (APCs). Most iNKT cells also express the natural killer cell marker CD161 (also called NK1.1 in mouse). Activated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

- Custom web browser-based software to identify and classify objects in a visual reference space (deployable over any image for any purpose)
- Interface to make explicit links between tagged objects (e.g., variables to equations)
- Interface to view and assess annotator tags, share links to pages

COSMOS step 1: train a model to visually parse

STATUS: PRODUCTION READY



cosmos.torch_model module

Torch model directory

SHOW SOURCE »

Sub-modules

cosmos.torch_model.inference

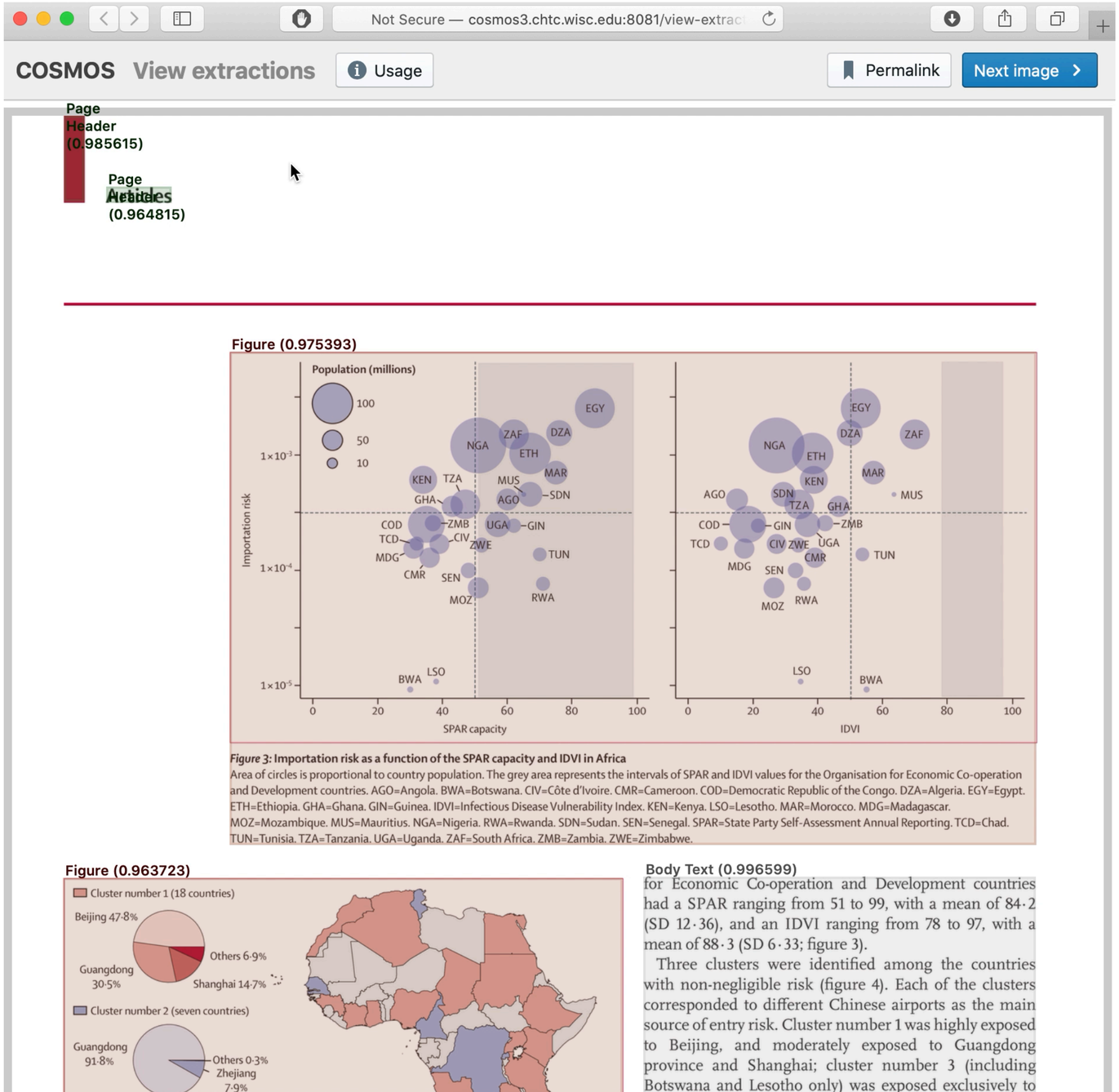
Inference module of model

cosmos.torch_model.model

Model specification dir

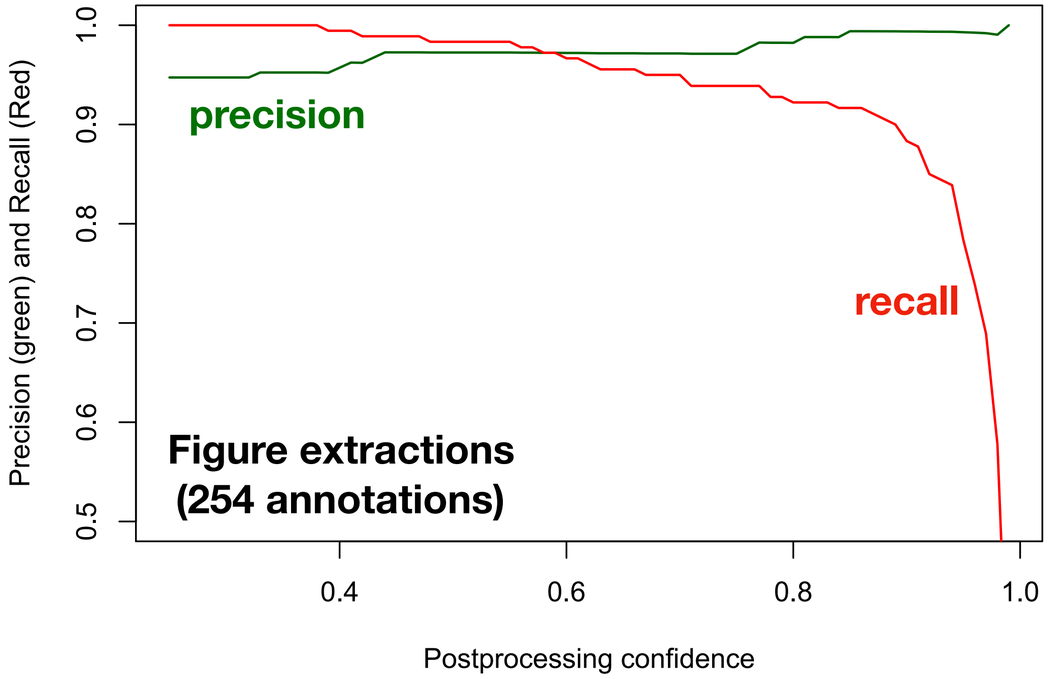
New distributed representation (in the visual space) for each element in the page.

COSMOS step 2: assess the visual model, improve



STATUS: PRODUCTION READY

- Custom web browser-based software to assess/annotate model output
- Precision > 0.85 for most elements, failures typically due to segmentation errors and out-of-distribution documents/objects (e.g., poorly typeset tables, widely spaced figure elements, extraneous objects)



COSMOS step 3: produce AI-constructed KB

STATUS: PRODUCTION READY

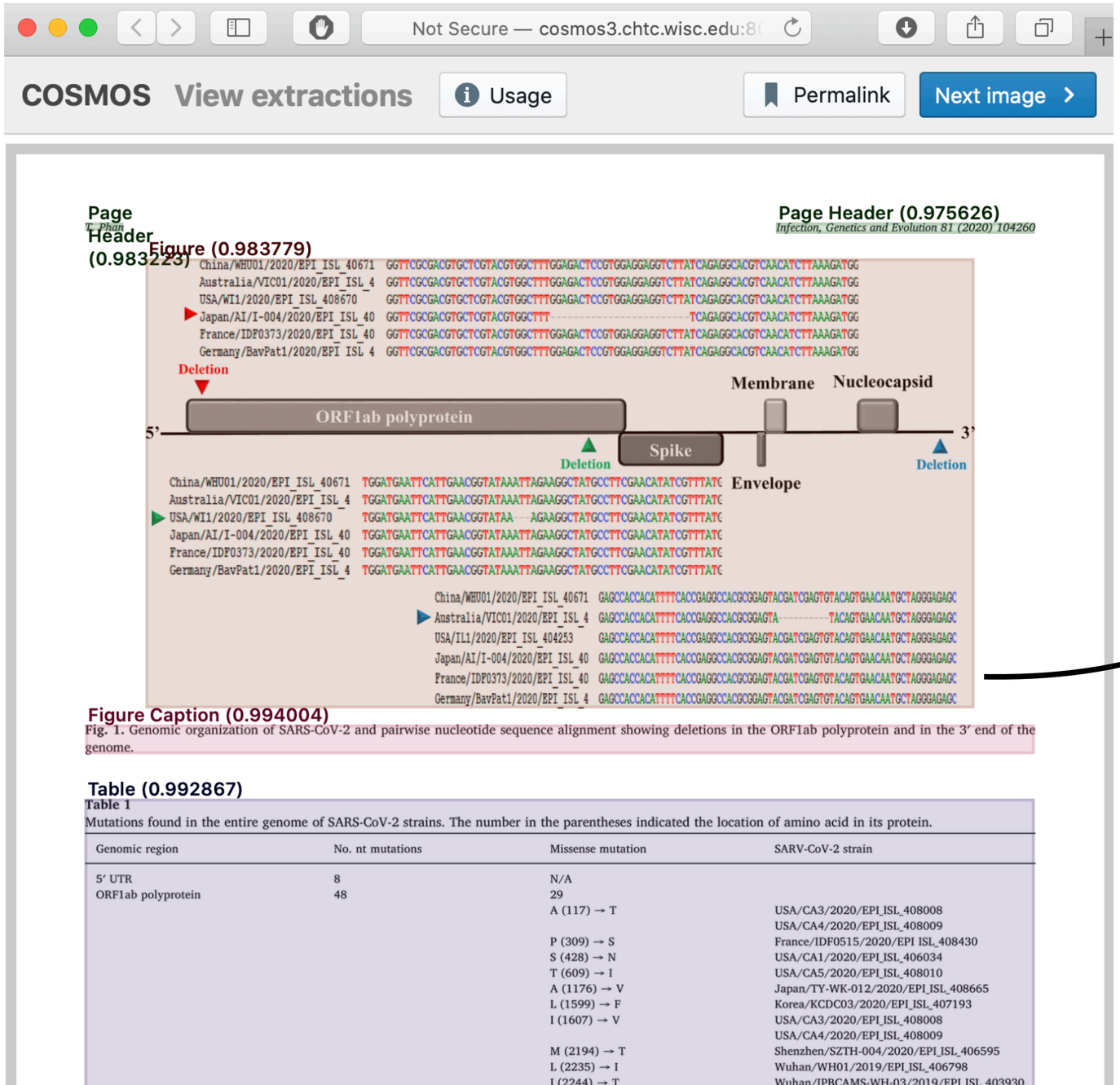


Table: page_objects

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

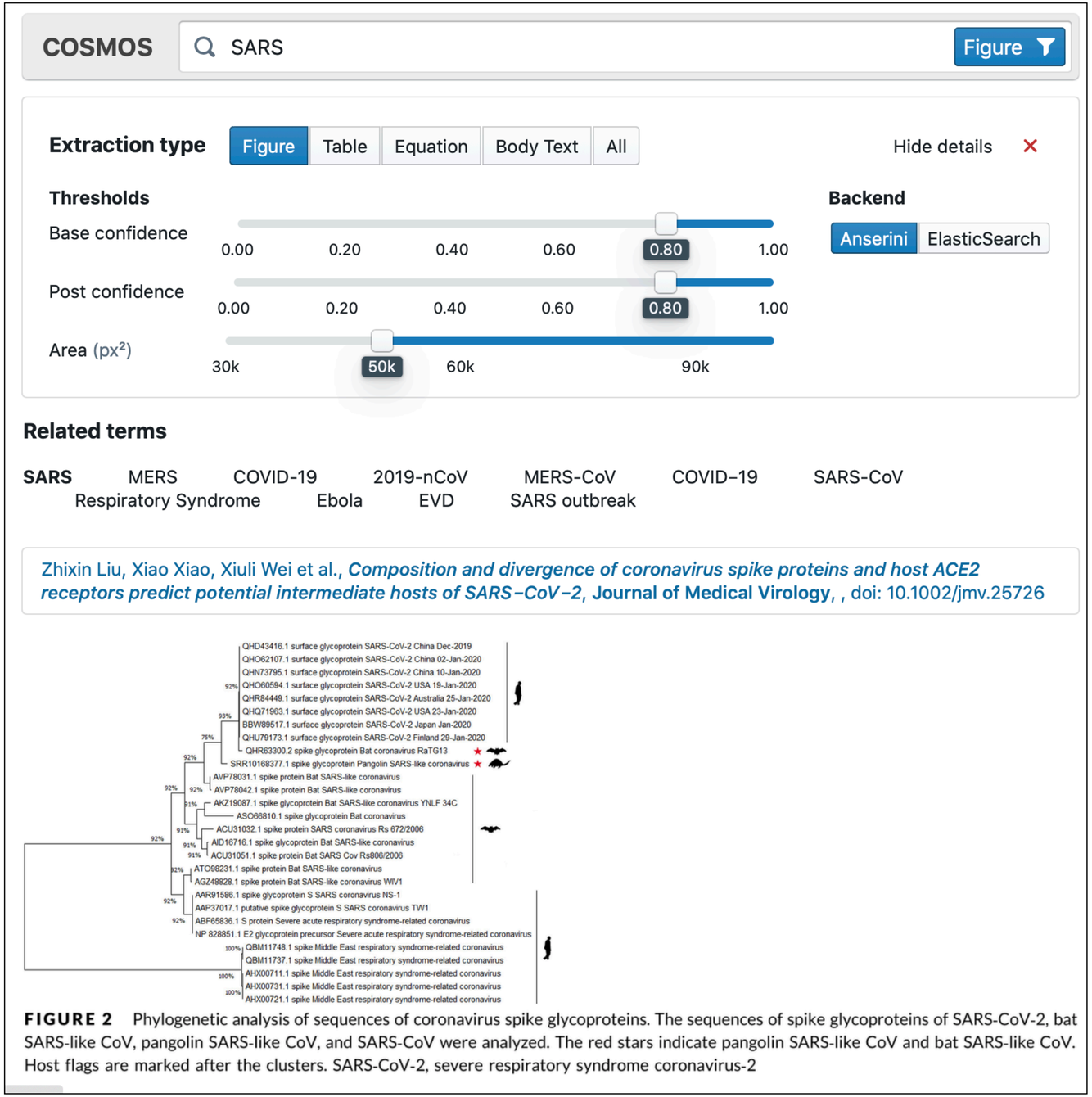
PRIMARY	*id*
INDEX	*pdf_id*
INDEX	*header_id*

Alter indexes

- Convert all modalities in heterogeneous PDFs into a structured representation
- Throughput of approximately one page every 2.5 seconds (GPU accelerated)
- Ability to deploy in parallel over distributed (CPU and/or GPU) nodes via HTCondor

COSMOS step 4: interface to search over KB

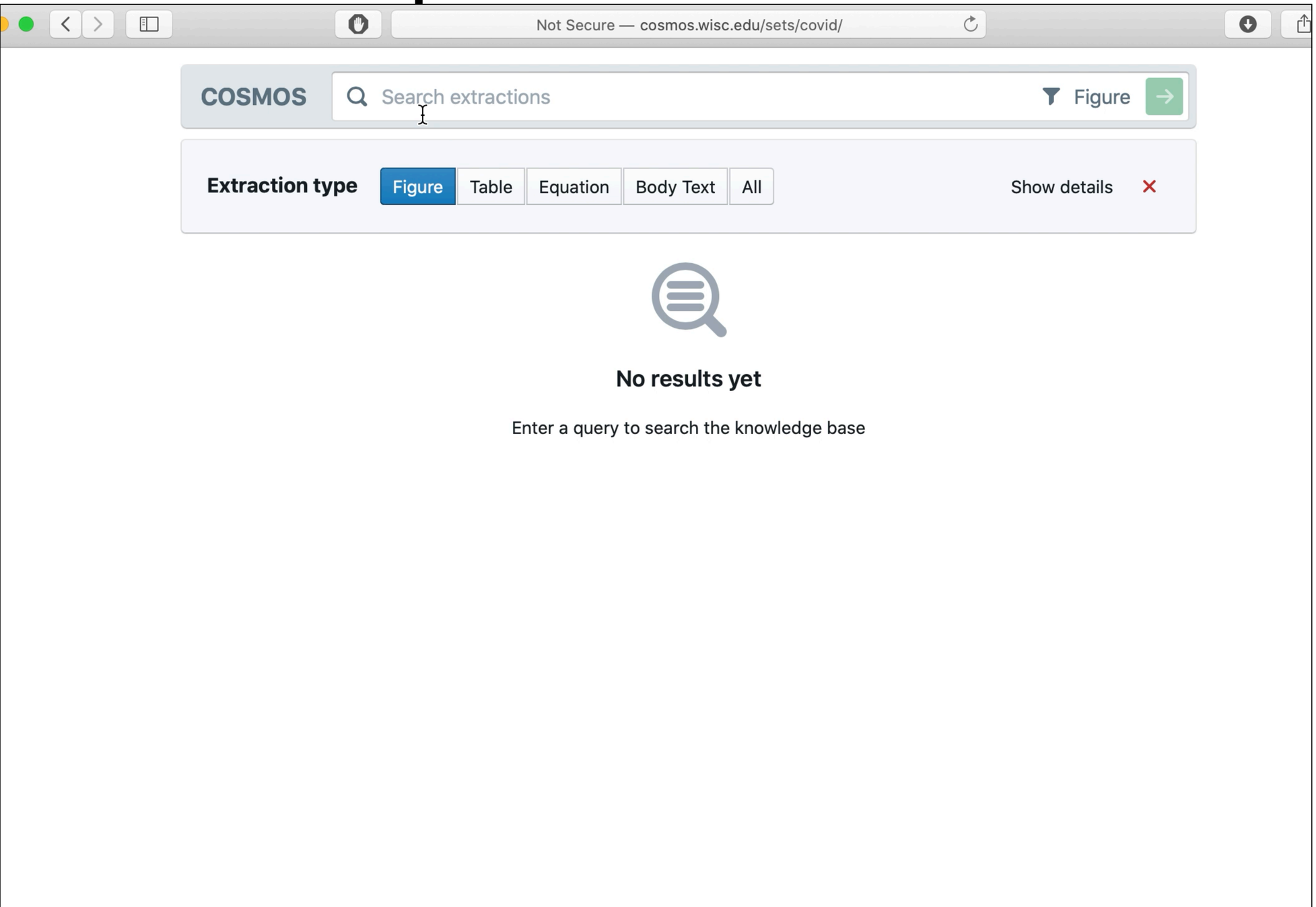
STATUS: PRODUCTION READY



- Backend REST-ful APIs to enable query over COSMOS-constructed KB
- Frontend web app surfacing COSMOS micro services, including
  - word embedding model trained over target corpus to facilitate vocabulary construction/search expansion
  - two modes of retrieval: ElasticSearch index over extracted objects, TF-IDF based retrieval
  - preliminary dataframe extraction for tables

COSMOS step 4: interface to search over KB

STATUS: PRODUCTION READY



COSMOS dataset generation

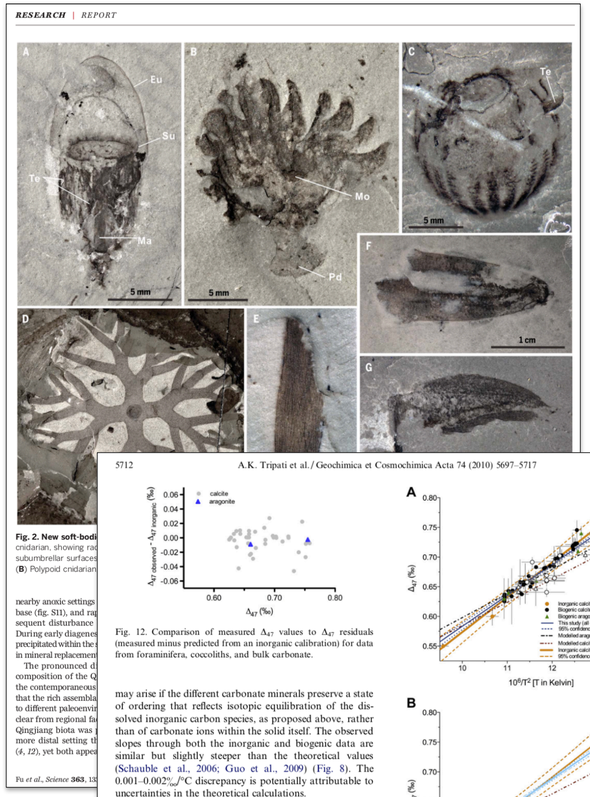


Fig. 2. New soft-bodi
cnidarian, showing ra
subumbrellar surfaces
(**B**) Polypoid cnidarian

nearby anoxic settings
base (fig. S11), and ru
sequent disturbance
During early diagenes
precipitated within the
in mineral replacemen
The pronounced d
composition of the Q
the contemporaneous
that the rich assembla
to different paleoenvir
clear from regional fa
Qingjiang biota was
more distal setting t
(*4*, *12*), yet both appea

Yu et al., *Science* **363**, 13

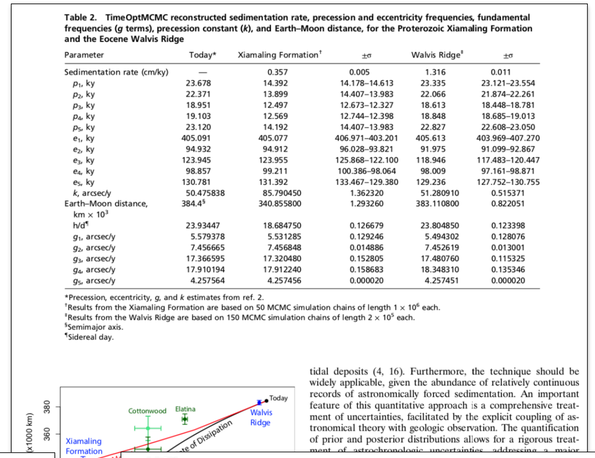
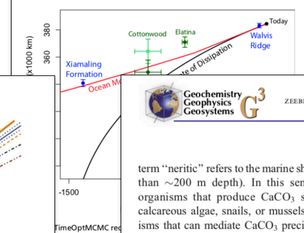


Table 2. TimeOptMCMC reconstructed sedimentation rate, precession and eccentricity frequencies, fundamental frequencies (*g* terms), precession constant (*k*), and Earth–Moon distance, for the Proterozoic Xiamaling Formation and the Eocene Walvis Ridge

Parameter	Today*	Xiamaling Formation†	±σ	Walvis Ridge‡	±σ
Sedimentation rate (cm/ky)	—	0.357	0.005	1.316	0.011
p_1, ky	23.678	14.392	14.178–14.613	23.335	23.121–23.554
p_2, ky	22.371	13.899	14.407–13.983	22.066	21.874–22.261
p_3, ky	18.951	12.497	12.673–12.327	18.613	18.448–18.781
p_4, ky	19.103	12.569	12.744–12.398	18.848	18.685–19.013
p_5, ky	23.120	14.192	14.407–13.983	22.827	22.608–23.050
e_1, ky	405.091	405.077	406.971–403.201	405.613	403.969–407.270
e_2, ky	94.932	94.912	96.028–93.821	91.975	91.099–92.867
e_3, ky	123.945	123.955	125.868–122.100	118.946	117.483–120.447
e_4, ky	98.857	99.211	100.386–98.064	98.009	97.161–98.871
e_5, ky	130.781	131.392	133.467–129.380	129.236	127.752–130.755
k, arcsec/y	50.475838	85.790450	1.362320	51.280910	0.515371
Earth–Moon distance, km × 10^3	384.4§	340.855800	1.293260	383.110800	0.822051
h/d¶	23.93447	18.684750	0.126679	23.804850	0.123398
g_1, arcsec/y	5.579378	5.531285	0.129246	5.494302	0.128076
g_2, arcsec/y	7.456665	7.456848	0.014886	7.452619	0.013001
g_3, arcsec/y	17.366595	17.320480	0.152805	17.480760	0.115325
g_4, arcsec/y	17.910194	17.912240	0.158683	18.348310	0.135346
g_5, arcsec/y	4.257564	4.257456	0.000020	4.257451	0.000020

*Precession, eccentricity, *g*, and *k* estimates from ref. 2.
†Results from the Xiamaling Formation are based on 50 MCMC simulation chains of length 1 × 10^6 each.
‡Results from the Walvis Ridge are based on 150 MCMC simulation chains of length 2 × 10^5 each.
§Semimajor axis.
¶Sidereal day.

tidal deposits (4, 16). Furthermore, the technique should be widely applicable, given the abundance of relatively continuous records of astronomically forced sedimentation. An important feature of this quantitative approach is a comprehensive treatment of uncertainties, facilitated by the explicit coupling of astronomical theory with geologic observation. The quantification of prior and posterior distributions allows for a rigorous treatment of stratigraphic uncertainties, addressing a major



TimeOptMCMC rec
lite-based estimat
are ±2σ and age u
esian posterior
g Formation and
Note the significa
nd prior estimate
km; 2σ). Age unce
within the size of
ood Formation (~
tone (~620 Ma; re
pdated analyses
stimate (364,192 k
nties from the 348
ed line), and a m
he), derive from re

nd Malinverno

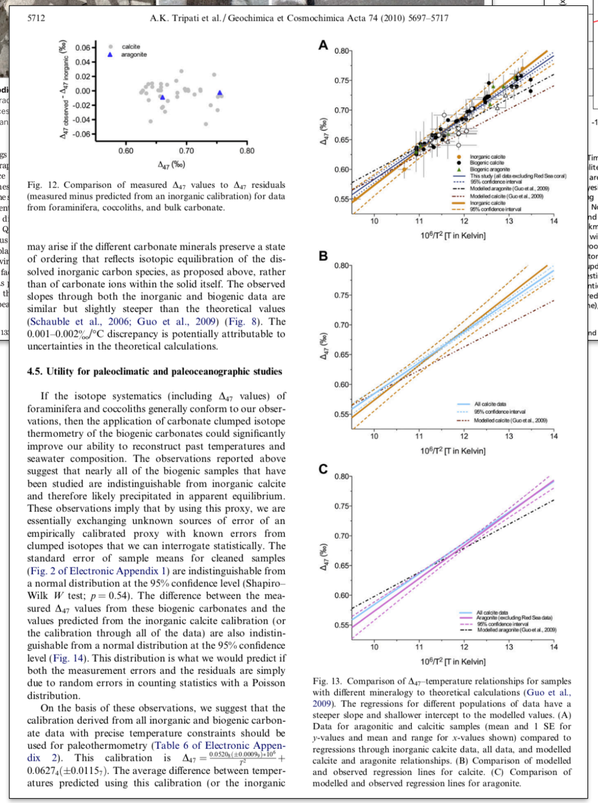


Fig. 12. Comparison of measured Δ_{47} values to Δ_{47} residuals (measured minus predicted from an inorganic calibration) for data from foraminifera, coccoliths, and bulk carbonate.

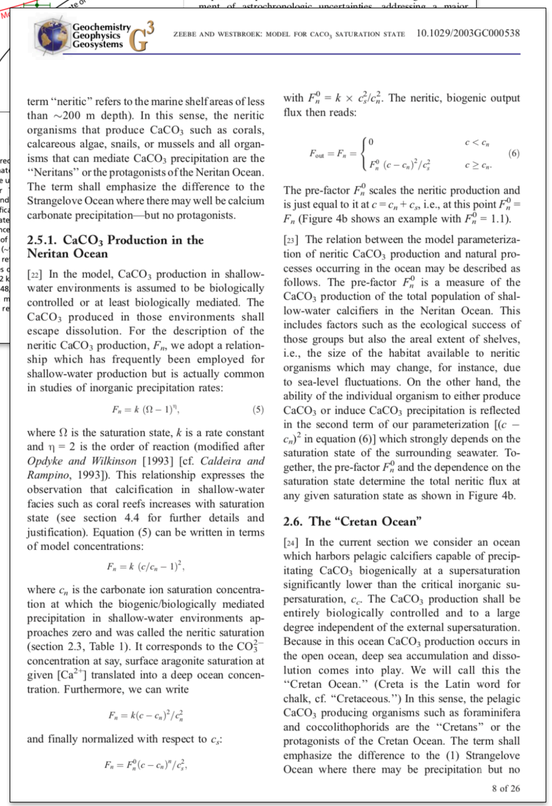
may arise if the different carbonate minerals preserve a state of ordering that reflects isotopic equilibration of the dissolved inorganic carbon species, as proposed above, rather than of carbonate ions within the solid itself. The observed slopes through both the inorganic and biogenic data are similar but slightly steeper than the theoretical values (Schauble et al., 2006; Guo et al., 2009) (Fig. 8). The 0.001–0.002‰ ^{47}/^{46}C discrepancy is potentially attributable to uncertainties in the theoretical calculations.

4.5. Utility for paleoclimatic and paleoceanographic studies

If the isotope systematics (including Δ_{47} values) of foraminifera and coccoliths generally conform to our observations, then the application of carbonate clumped isotope thermometry of the biogenic carbonates could significantly improve our ability to reconstruct past temperatures and seawater composition. The observations reported above suggest that nearly all of the biogenic samples that have been studied are indistinguishable from inorganic calcite and therefore likely precipitated in apparent equilibrium. These observations imply that by using this proxy, we are essentially exchanging unknown sources of error of an empirically calibrated proxy with known errors from clumped isotopes that we can interrogate statistically. The standard error of sample means for cleaned samples (Fig. 2 of Electronic Appendix 1) are indistinguishable from a normal distribution at the 95% confidence level (Shapiro–Wilk W test; $p = 0.54$). The difference between the measured Δ_{47} values from these biogenic carbonates and the values predicted from the inorganic calcite calibration (or the calibration through all of the data) are also indistinguishable from a normal distribution at the 95% confidence level (Fig. 14). This distribution is what we would predict if both the measurement errors and the residuals are simply due to random errors in counting statistics with a Poisson distribution.

On the basis of these observations, we suggest that the calibration derived from all inorganic and biogenic carbonate data with precise temperature constraints should be used for paleothermometry (Table 6 of Electronic Appendix 2). This calibration is $\Delta_{47} = \frac{0.0326_A(\pm 0.0006_A)\cdot 10^6}{T^2} + 0.0627_A(\pm 0.0115_A)$. The average difference between temperatures predicted using this calibration (or the inorganic

Fig. 13. Comparison of Δ_{47}–temperature relationships for samples with different mineralogy to theoretical calculations (Guo et al., 2009). The regressions for different populations of data have a steeper slope and shallower intercept to the modelled values. (**A**) Data for aragonitic and calcitic samples (mean and 1 SE for y-values and mean and range for x-values shown) compared to regressions through inorganic calcite data, all data, and modelled calcite and aragonite relationships. (**B**) Comparison of modelled and observed regression lines for calcite. (**C**) Comparison of modelled and observed regression lines for aragonite.



term "neritic" refers to the marine shelf areas of less than ~200 m depth). In this sense, the neritic organisms that produce $CaCO_3$ such as corals, calcareous algae, snails, or mussels and all organisms that can mediate $CaCO_3$ precipitation are the "Neritans" or the protagonists of the Neritan Ocean. The term shall emphasize the difference to the Strangelove Ocean where there may well be calcium carbonate precipitation—but no protagonists.

2.5.1. $CaCO_3$ Production in the Neritan Ocean

[22] In the model, $CaCO_3$ production in shallow-water environments is assumed to be biologically controlled or at least biologically mediated. The $CaCO_3$ produced in those environments shall escape dissolution. For the description of the neritic $CaCO_3$ production, F_n, we adopt a relationship which has frequently been employed for shallow-water production but is actually common in studies of inorganic precipitation rates:

$$F_n = k\ (\Omega - 1)^{\eta}, \tag{5}$$

where Ω is the saturation state, k is a rate constant and $\eta = 2$ is the order of reaction (modified after *Opdyke and Wilkinson* [1993] [cf. *Caldeira and Rampino*, 1993]). This relationship expresses the observation that calcification in shallow-water facies such as coral reefs increases with saturation state (see section 4.4 for further details and justification). Equation (5) can be written in terms of model concentrations:

$$F_n = k\ (c/c_n - 1)^2,$$

where c_n is the carbonate ion saturation concentration at which the biogenic/biologically mediated precipitation in shallow-water environments approaches zero and was called the neritic saturation (section 2.3, Table 1). It corresponds to the CO_3^{2-} concentration at say, surface aragonite saturation at given $[Ca^{2+}]$ translated into a deep ocean concentration. Furthermore, we can write

$$F_n = k(c - c_n)^2/c_n^2$$

and finally normalized with respect to c_r:

$$F_n = F_n^0(c - c_n)^n/c_r^2,$$

with $F_n^0 = k \times c_r^2/c_n^2$. The neritic, biogenic output flux then reads:

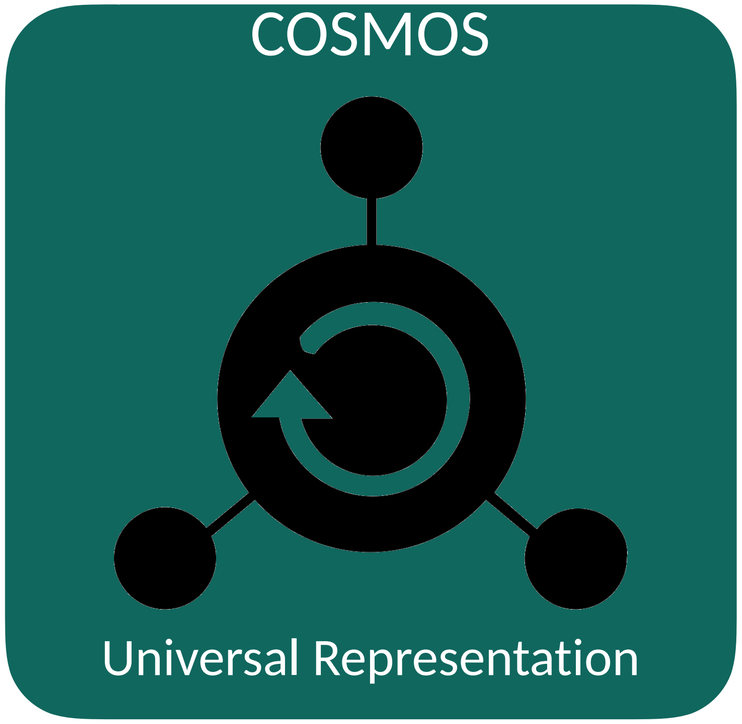
$$F_{\mathrm{out}} = F_n = \begin{cases} 0 & c < c_n \\ F_n^0\ (c - c_n)^2/c_r^2 & c \geq c_n. \end{cases} \tag{6}$$

The pre-factor F_n^0 scales the neritic production and is just equal to it at $c = c_n + c_r$, i.e., at this point $F_n^0 = F_n$ (Figure 4b shows an example with $F_n^0 = 1.1$).

[23] The relation between the model parameterization of neritic $CaCO_3$ production and natural processes occurring in the ocean may be described as follows. The pre-factor F_n^0 is a measure of the $CaCO_3$ production of the total population of shallow-water calcifiers in the Neritan Ocean. This includes factors such as the ecological success of those groups but also the areal extent of shelves, i.e., the size of the habitat available to neritic organisms which may change, for instance, due to sea-level fluctuations. On the other hand, the ability of the individual organism to either produce $CaCO_3$ or induce $CaCO_3$ precipitation is reflected in the second term of our parameterization $[(c - c_n)^2$ in equation (6)] which strongly depends on the saturation state of the surrounding seawater. Together, the pre-factor F_n^0 and the dependence on the saturation state determine the total neritic flux at any given saturation state as shown in Figure 4b.

2.6. The "Cretan Ocean"

[24] In the current section we consider an ocean which harbors pelagic calcifiers capable of precipitating $CaCO_3$ biogenically at a supersaturation significantly lower than the critical inorganic supersaturation, c_r. The $CaCO_3$ production shall be entirely biologically controlled and to a large degree independent of the external supersaturation. Because in this ocean $CaCO_3$ production occurs in the open ocean, deep sea accumulation and dissolution comes into play. We will call this the "Cretan Ocean." (Creta is the Latin word for chalk, cf. "Cretaceous.") In this sense, the pelagic $CaCO_3$ producing organisms such as foraminifera and coccolithophorids are the "Cretans" or the protagonists of the Cretan Ocean. The term shall emphasize the difference to the (1) Strangelove Ocean where there may be precipitation but no



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

COSMOS step 5: unsupervised structured knowledge extraction

STATUS: ALPHA READY

Re-Flex Retrieval

Input: Search query to COSMOS to get “relevant” parts of documents

Keyword query to COSMOS: “disease co-infection”

Retrieved relevant contexts

- “We observed that 21.2% (7/33) positive influenza A patients were admitted to an intensive care unit: five patients had H1N1 2009, one had IAV and one had coinfection with HRSV.”
- “It is likely that the clinical course of coinfections will differ; some combinations viruses are more or less deleterious than others. One combination that is believed to increase the severity of illness is RSV and hMPV coinfection.”
- “Two cases with multiple coinfecting viruses were detected (one NPA from 2007 and the other from 2009) and in both cases HBoV was detected in with RSV and PIV”
-

Input: Entities of interest and natural language queries representing relation

```
head_entities = ['influenza', 'hMPV', 'RSV', 'HBoV', 'hPyV', 'hCoV']
natural_language_queries = ['[Y] and [X] are frequently detected together',
                            '[X] coinfects with [Y]', '[X] was detected with [Y]']
```

Re-Flex: Unsupervised Relation Extraction from Language Models [Under submission, 2020]

Output: Structured relations

	head_entity	relation	tail_entity	context
0	influenza	coinfects_with	HRSV	We observed that 21.2% (7...
1	hMPV	coinfects_with	RSV	It is likely that the clin...
2	RSV	coinfects_with	PeV	Considering RSV A/B, doubl...
3	HBoV	coinfects_with	RSV	Two cases with multiple co...
4	hPyV	coinfects_with	coronavirusesOC43	Among the hBoV, hCoV, and ...
5	hCoV	coinfects_with	coronavirusesOC43	Among the hBoV, hCoV, and ...

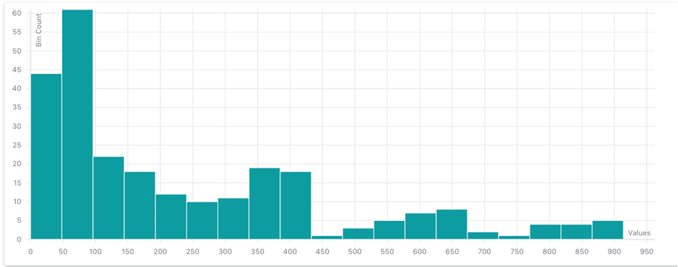
Next challenges:

traversing the gap to actionable science

Aggregate extractions

- Refined extractions (e.g., resolve synonyms, harmonize measurement units)
- Reason across multiple scales (e.g., from single element in table, to row and column, to many tables across documents)
- Aggregate extractions over multiple modes (e.g., combine information from body text and tables to identify and extract target data)

Extracted empirical distribution over TOC



Tabular data for TOC

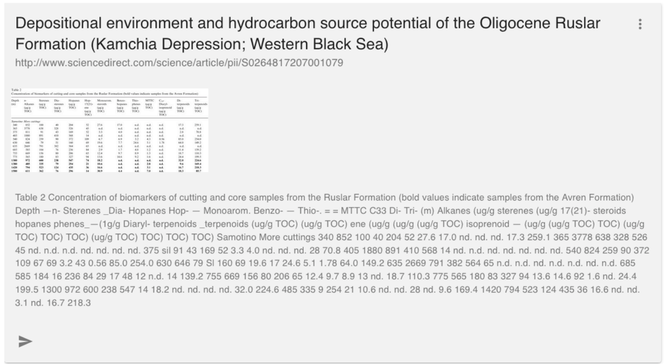
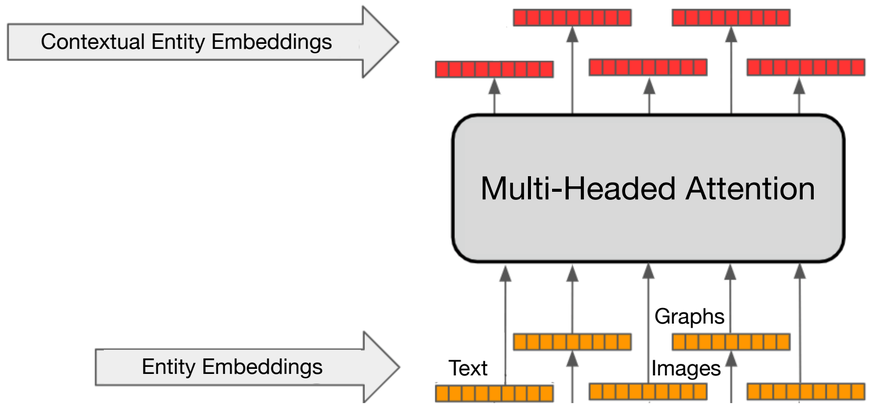
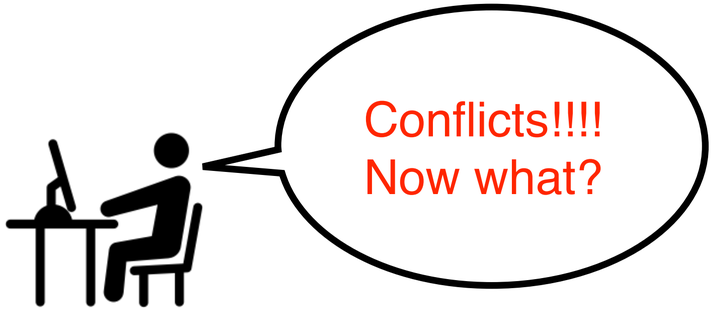

Depositional environment and hydrocarbon source potential of the Oligocene Ruslar Formation (Kamchia Depression; Western Black Sea)
http://www.sciencedirect.com/science/article/pii/S0264817207001079

Table 2 Concentration of biomarkers of cutting and core samples from the Ruslar Formation (bold values indicate samples from the Avren Formation) Depth —n- Sterenes _Dia- Hopanes Hop- — Monoarom. Benzo- — Thio-. = = MTTC C33 Di- Tri- (m) Alkanes (ug/g sterenes (ug/g 17(21)- steroids hopanes phenes_ —(1g/g Diaryl- terpenoids _terpenoids (ug/g TOC) (ug/g TOC) ene (ug/g (ug/g (ug/g TOC) isoprenoid — (ug/g (ug/g TOC) TOC) (ug/g TOC) TOC) TOC) (ug/g TOC) TOC) TOC) Samotino More cuttings 340 852 100 40 204 52 27.6 17.0 nd. nd. nd. 17.3 259.1 365 3778 638 328 526 45 nd. n.d. nd. nd. nd. nd. 375 sil 81 43 169 52 3.3 4.0 nd. nd. nd. 28 70.8 405 1880 891 410 568 14 nd. n.d. nd. nd. nd. nd. nd. 540 824 259 90 372 109 67 69 3.2 43 0.56 85.0 254.0 630 646 79 Sl 160 69 19.6 17 24.6 5.1 1.78 64.0 149.2 635 2669 791 382 564 65 n.d. n.d. nd. n.d. nd. n.d. 685 585 184 16 236 84 29 17 48 12 n.d. 14 139.2 755 669 156 80 206 65 12.4 9.7 8.9 13 nd. 18.7 110.3 775 565 180 83 327 94 13.6 14.6 92 1.6 nd. 24.4 199.5 1300 972 600 238 547 14 18.2 nd. nd. nd. nd. 32.0 224.6 485 335 9 254 21 10.6 nd. nd. 28 nd. 9.6 169.4 1420 794 523 124 435 36 16.6 nd. nd. 3.1 nd. 16.7 218.3



Expose and resolve contradictory information



Extractions

Source	Disease	Gene	CausedBy
OMIM	Li-Fraumeni Syndrome	CHEK2	**Yes**
Paper	Li-Fraumeni Syndrome	CHEK2	**No**

Genetic Heterogeneity of Li-Fraumeni Syndrome

A second form of Li-Fraumeni syndrome (LFS2; 609265) is caused by mutation in the CHEK2 gene (604373), and an LFS locus (LFS3; 609266) has been mapped to chromosome 1q23.

Source: OMIM

Increasing evidence that germline mutations in *CHEK2* do not cause Li-Fraumeni syndrome†

Nayanta Sodha, Richard S. Houlston, Sarah Bullock, Martin A. Yuille, Carol Chu, Gwen Turner, Rosalind A. Eeles

First published: 19 November 2002 Full publication history

- Identify and surface different conclusions/interpretations from related observations/ experiments
- Extract experimental setup and conditions, reasons for conclusions

Integration with scientific data collections

Abundance of data repositories

Repository details

CancerData.org

General | Institutions | Terms | Standards

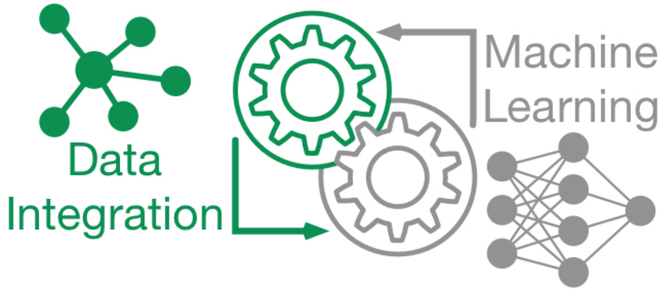
Name of repository	**CancerData.org**
Additional name(s)	Sharing data for cancer research
Repository URL	https://www.cancerdata.org/
Subject(s)	Basic Biological and Medical Research; Medicine; Biology; Life Sciences

Latest Datasets

- Mar 28 Rhodes2019 - Immune-Mediated theory of Metastasis
- Mar 27 Gene-expression changes in neurons active during contextu…
- Mar 27 RNAseq analysis in TNF treated HUVEC cells compared to …
- Mar 27 RNAseq of the protease transcriptome of primary murine alv…
- Mar 27 RNAseq of trachea, bronchi and lungs of mice to investigate…
- Mar 27 Transcriptomic analysis of immune response in healthy cont…
- Mar 26 RNAseq of 6 Different Rat Brain Areas after Treatment with …
- Mar 26 Microarray of liver from Red Junglefowl (Gallus gallus) raise…
- Mar 26 BCL-2 pathway counteracts aneuploidy in mammalian embr…
- Mar 26 Dunster2016 - Nondimensional Coagulation Model
- Mar 26 Smith2011 - Three Stage Innate Immune Response to a Pne…
- Mar 20 Histone H3 wild-type DIPG/DMG overexpressing EZHIP ext…

- Integration with data catalogues
- Disambiguation and linkage of dataset metadata to publications
- Zero-code data integration

Idea: use zero-code ML frameworks and contextual ML techniques for autonomous end-to-end data integration



Our vision: service-oriented knowledge extraction

Agent

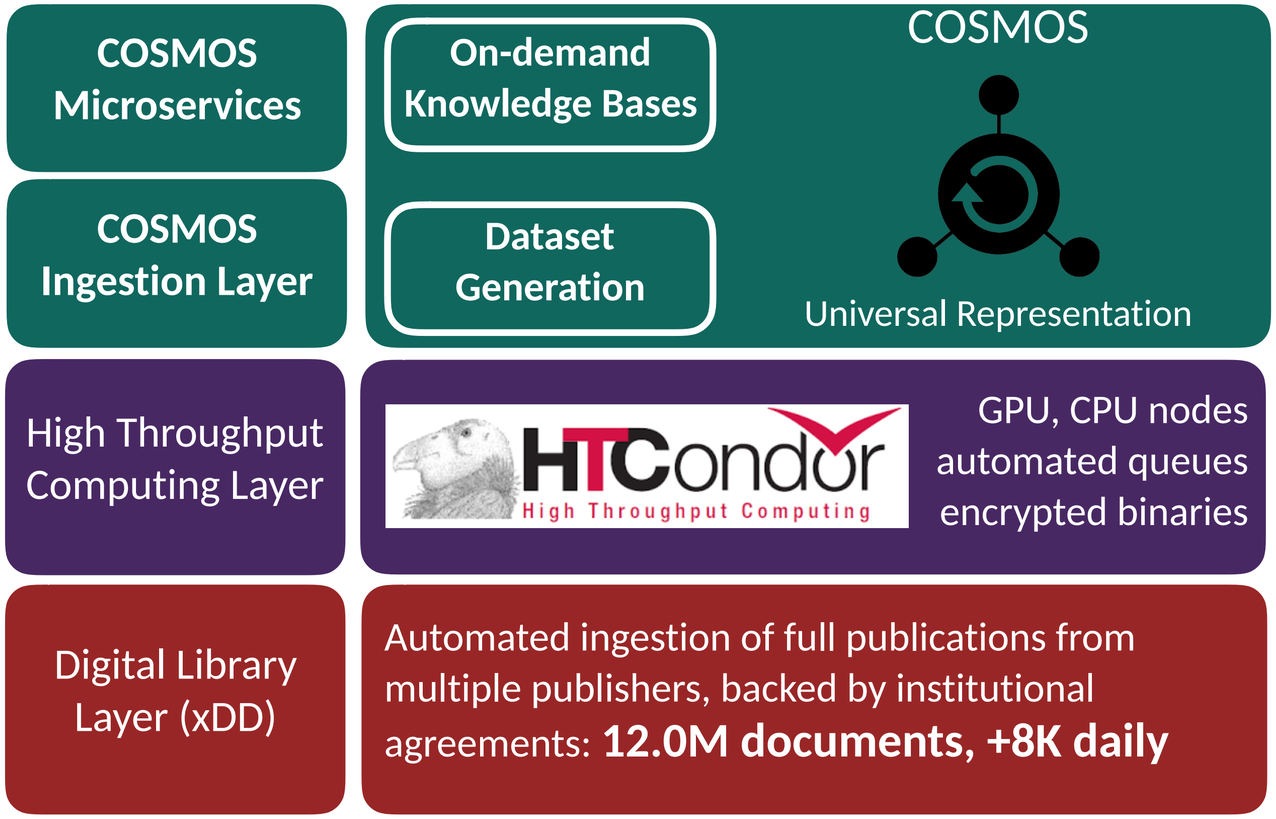
Applications

Services

Infrastructure and data

- An intelligent (AI-driven) virtual assistant that can consider and reason about semantics when linking/synthesizing artifacts from publications
- APIs that enable users to build custom applications over the assistant (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

UW-COSMOS: an end-to-end stack for accelerating scientific research



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, API for simple queries
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA DEPARTMENT OF ENERGY UNITED STATES OF AMERICA

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)