## Reproducibility PI Manifesto, Transform to Open Science

Jupyter Community in MSK Imaging Research



## About me

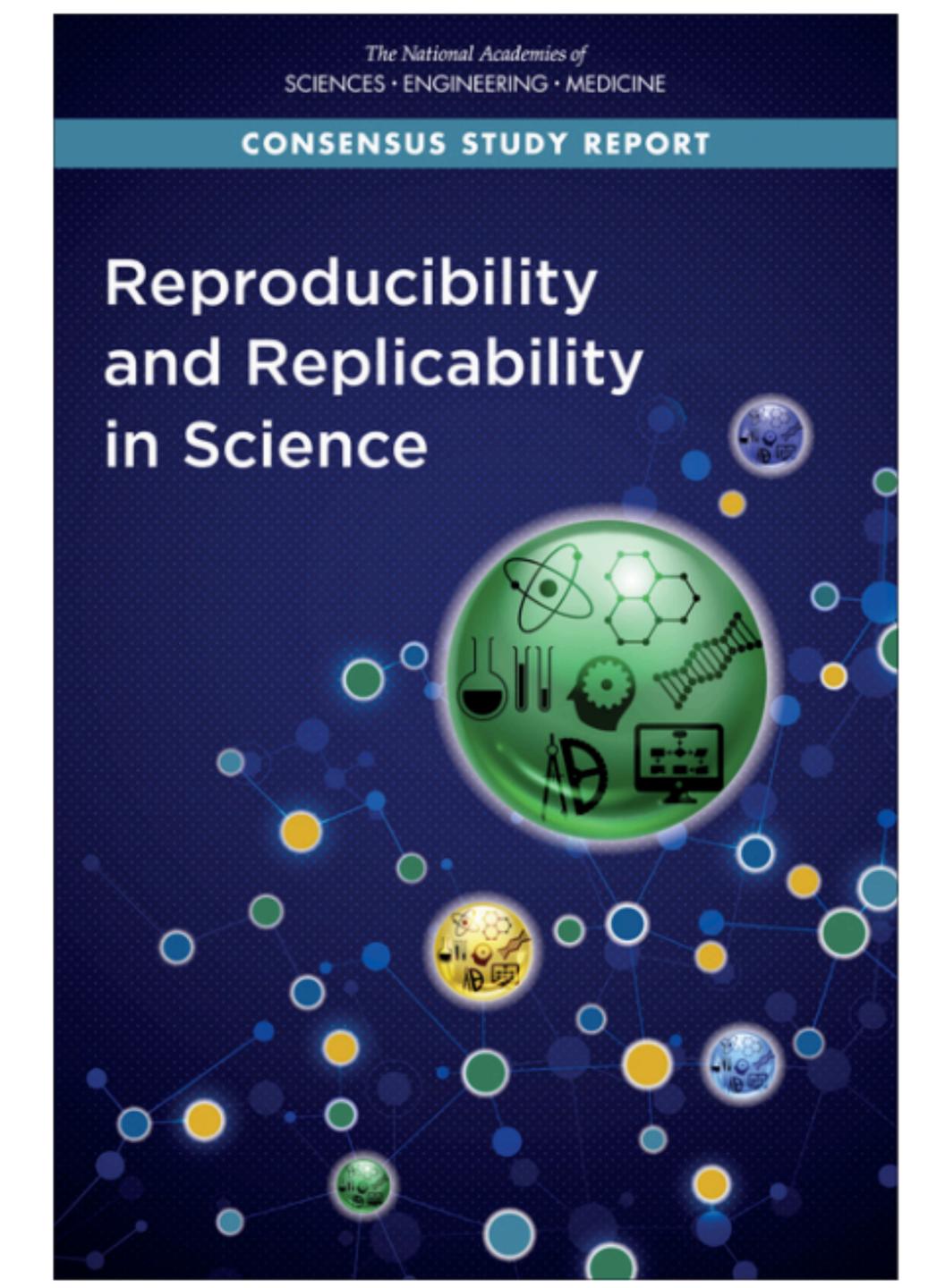
http://lorenabarba.com

- SC19 Reproducibility Chair; JupyterCon 2020 General Chair
- NASEM committee "Reproducibility and Replicability in Science"
- · NASEM committee "Open Source Software Policy Options for NASA"
- NumFOCUS Board of Directors, 2014–2021
- Founding editor and past AEiC of The Journal of Open Source Software
- Editor-in-Chief of IEEE Computing in Science and Engineering
- Author "Reproducibility PI Manifesto"

## Reproducibility PI Manifesto 2012

- I teach my graduate students about reproducibility
- All our research code (and writing) is under version control
- We always carry out verification & validation (and make them public)
- For main results, we share data, plotting script & figure under CC-BY
- We upload preprint to arXiv at the time of submission to a journal
- We release code at the time of submission of a paper to a journal
- We add a "Reproducibility" declaration at the end of each paper
- I develop a consistent open-science policy & keep an up-to-date web presence





- ► Study mandated by public law 114-329 (Jan. 2017)
- commissioned by NSF to The National Academies of Sciences, Engineering and Medicine (NASEM)
- ▶ 15 experts convened
- ▶ 18 months of in-person meetings, teleconferences, commissioned papers, deliberations, writing
- report released 7 May 2019

http://doi.org/c5jp



- ad-hoc committee to investigate and recommend best practices for NASA as it considers whether to establish an open code and open models policy, to complement its open data policy
- Convened Oct. 2017, report delivered Sep. 21, 2018
- "Accelerating the pace of science is a key goal."

https://numfocus.org

### LETTER FROM THE BOARD CO-CHAIRPERSON

NumFOCUS was founded in 2012 to provide a fiscal umbrella for many open-source software projects that have become essential for science and research. Our sponsored projects benefit from a range of services: fiscal, legal, operational, and more.

# NUMFCOCUS OPEN CODE = BETTER SCIENCE

ANNUAL REPORT

2019



LORENA BARBA
NumFOCUS Co-Chairperson

#### https://numfocus.org

84

Sponsored & **Affiliated Projects** 

\$5M

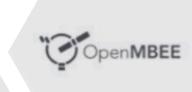
Total revenue in 2020

























ø









CONDA-FORGE















Arviz

Econ





ANALYSIS





mlpack





















**PYMC3** 



将軍

learn
machine learning in Python







## Top challenges of reproducible research

- creation, curation, usage and publication of research software
- acceptance, adoption and standardization of open-science practices;
- misalignment with academic incentive structures and institutional processes for career progression

## STAT

# When our world changed... January 8, 2020

**HEALTH** 

## WHO says mysterious illness in China likely being caused by new virus



By <u>Helen Branswell</u> y Jan. 8, 2020

Reprints





Nucleotide

Nucleotide

Advanced

GenBank -

Send to: -

## Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome

GenBank: MN908947.1



This sequence has been updated. See current version.

#### **FASTA** <u>Graphics</u>

**LOCUS** MN908947 30473 bp ss-RNA linear VRL 12-JAN-2020

Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete DEFINITION

genome.

ACCESSION MN908947 VERSION MN908947.1

**KEYWORDS** 

**SOURCE** Wuhan seafood market pneumonia virus ORGANISM <u>Wuhan seafood market pneumonia virus</u>

> Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; unclassified Betacoronavirus.

#### CORONAVIRUS

## **COVID-19 Datasets**

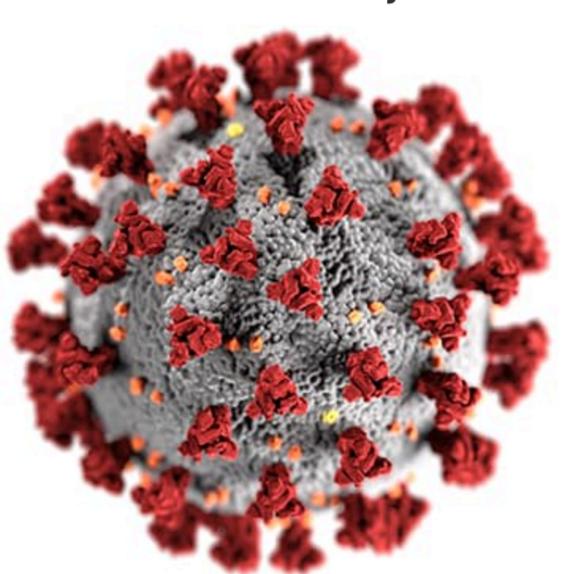
The Lens has assembled free and open datasets of patent documents, scholarly research works metadata and biological sequences from patents, and deposited them in a machine-readable and explorable form.

## **Human Coronaviruses Data Initiative**

https://about.lens.org/covid-19/

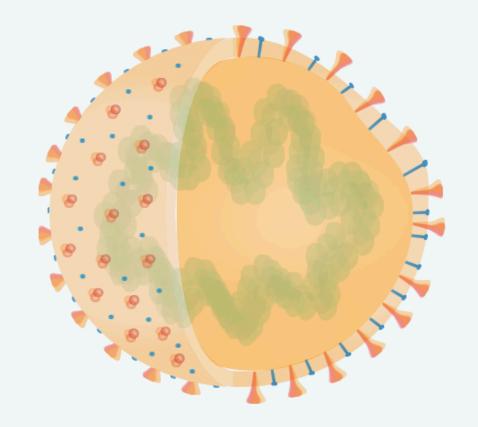
## COVID-19 Molecular Structure and Therapeutics Hub

Aggregating critical information to accelerate COVID-19 drug discovery for the molecular modeling and simulation community.





GENERAL MECHANISM TRANSMISSION DIAGNOSIS TREATMENT PREVENTION CASE REPORT FORECASTING



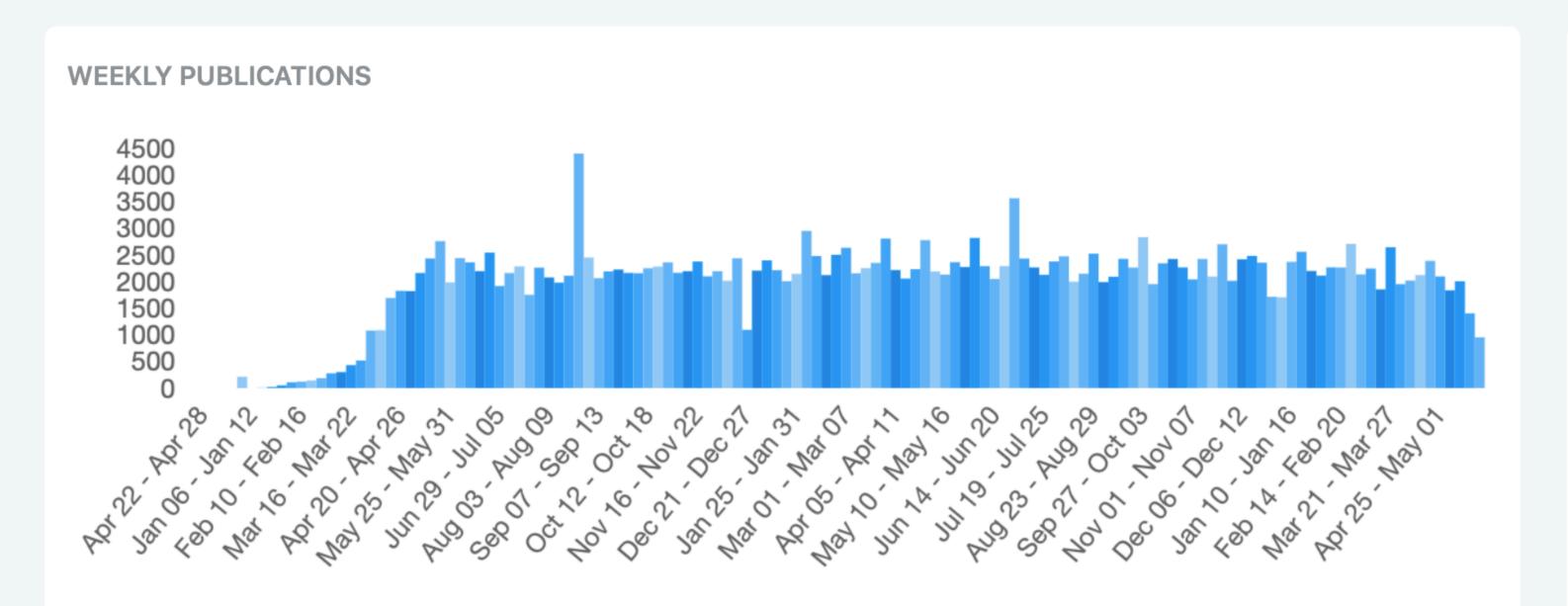
LitCovid is a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus. It is the most comprehensive resource on the subject, providing a central access to <u>254031</u> (and <u>growing</u>) relevant articles in PubMed. The articles are updated daily and are further categorized by different research topics (e.g. transmission) and geographic locations.











https://doi.org/gmh9wh

#### LATEST PUBLICATIONS

#### **TREATMENT**

Three doses of mRNA COVID-19 vaccine protects from SARS-CoV-2 infections in Japan.

Hotta, Katsuyuki et al. • J Intern Med

#### **MECHANISM**

Angiotensin-Converting Enzyme 2 and Furin Expression in the Appendix of Children with COVID-19.









HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS

medRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive, guide clinical practice/health-related behavior, or be reported in news media as established information.

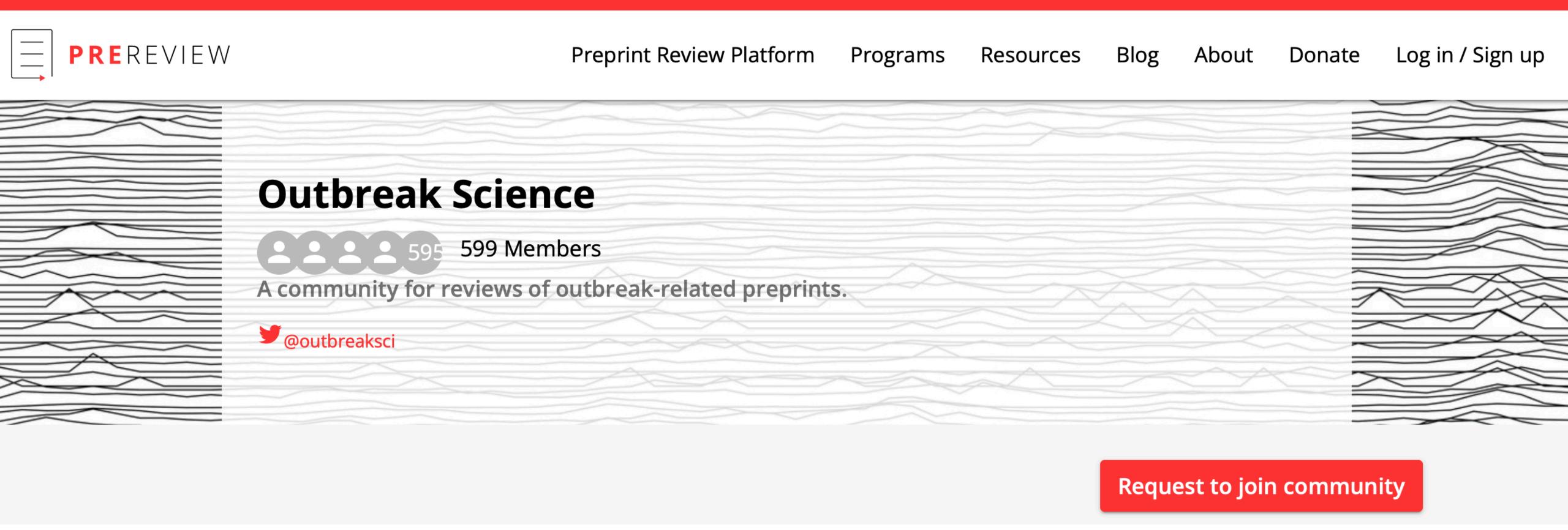
## COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

23,592 Articles (17,973 medRxiv, 5,619 bioRxiv)

Most recent first

Page I: Articles I-10 | Next 😜

#### Get involved with rapidly PREreviewing COVID-19 preprints and then view our COVID-19 Dashboard of PREreview activities and recommendations.



https://outbreaksci.prereview.org

"Our work is a direct response to the flawed way scientific research is evaluated. Behind closed doors, a handful of unpaid reviewers—selected opaquely and mainly through personal connections—use subjective criteria to decide the fate of a research article."

PREreview founders: Daniela Saderi, Samantha Hindle



Home > Newsletters > ACM Bulletins > Open Access To ACM Digital Library During Coronavirus Pandemic

## Open Access to ACM Digital Library During Coronavirus Pandemic

March 30, 2020

Dear ACM Members:

As the coronavirus/COVID-19 pandemic continues, we at ACM would like to do what we can to help support the computing community. Many computing researchers and practitioners are now working remotely. In addition, teaching and learning have also moved online as more and more campuses close.

We believe that ACM can help support research, discovery and learning during this time of crisis by opening the ACM Digital Library of to all. For the next three months, there will be no fees assessed for accessing or downloading work published by ACM. We hope this will help researchers, practitioners and students maintain access to our publications as well as increasing visibility and awareness of ACM's journals, proceedings and magazines. Please be sure to inform your colleagues that the ACM DL is now open, and will continue that way through June 30, 2020.

This global health crisis is a unique challenge that has impacted many ACM members. We would like to express our concern and support for all who are affected by this outbreak.

About Elsevier Products & Solutions Services Shop & Discover







Home > About > Press Releases

> Elsevier gives full access to its content on its COVID-19 Information Center for PubMed Central and other public health databases to accelerate ...

# Elsevier gives full access to its content on its COVID-19 Information Center for PubMed Central and other public health databases to accelerate fight against coronavirus

#### New York, March 13, 2020

From today, Elsevier, a global leader in research publishing and information analytics specializing in science and health, is making all its research and data content on its COVID-19 Information Center available to PubMed Central, the archive of biomedical and lifescience at the US. National Institutes of Health's National Library of Medicine, and other publicly funded repositories globally, such as the WHO COVID database, for as long as needed while the public health emergency is ongoing. This additional access allows researchers to use artificial intelligence to keep up with the rapidly growing body of literature and identify trends as countries around the world address this global health crisis.

**April 15, 2020** 







## Expanded access to JSTOR during COVID-19 crisis

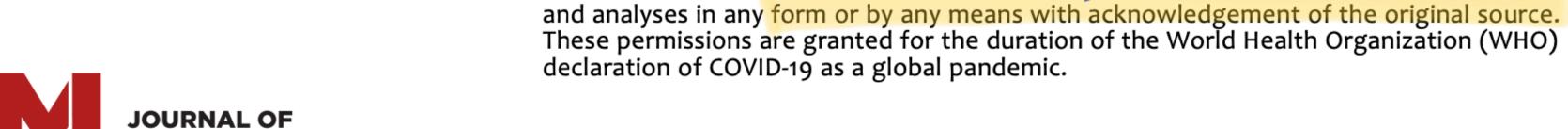
We are continuing to expand access to JSTOR for students, faculty, and institutions impacted by COVID-19. Access to all these resources will be available through **June 30, 2020**:

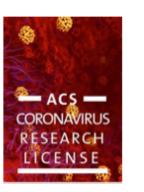
#### **Expanded JSTOR resources for institutions**

- Expanded access to journals and primary sources: For participating academic institutions that currently license some, but not all, JSTOR Archive and Primary Source collections, we will turn on access to all unlicensed collections at no cost.
- Expanded access to books: With support from dozens of publisher partners, more than 35,000 books are available at no charge for JSTOR participating academic institutions. The number of titles available through this effort is growing as more publishers opt in.

## During a pandemic crisis

## Another exponential spread: open science





pubs.acs.org/jcim Viewpoint

## A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19

Rommie E. Amaro\* and Adrian J. Mulholland\*





This article is made available via the ACS COVID-19 subset for unrestricted RESEARCH re-use

## During a pandemic crisis

Another exponential spread: open science





Our community letter in final form-hundreds of groups worldwide commit to open #COVID\_19 simulations, data, and methods to accelerate #SARSCoV2 science @AdrianMulholla1 @MolSSI\_NSF @XSEDEscience @PRACE\_RI @BioExcelCoE @hecbiosim



Richard Bryce @Bryce\_Group · May 2

A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19 pubs.acs.org/doi/10.1021/ac...

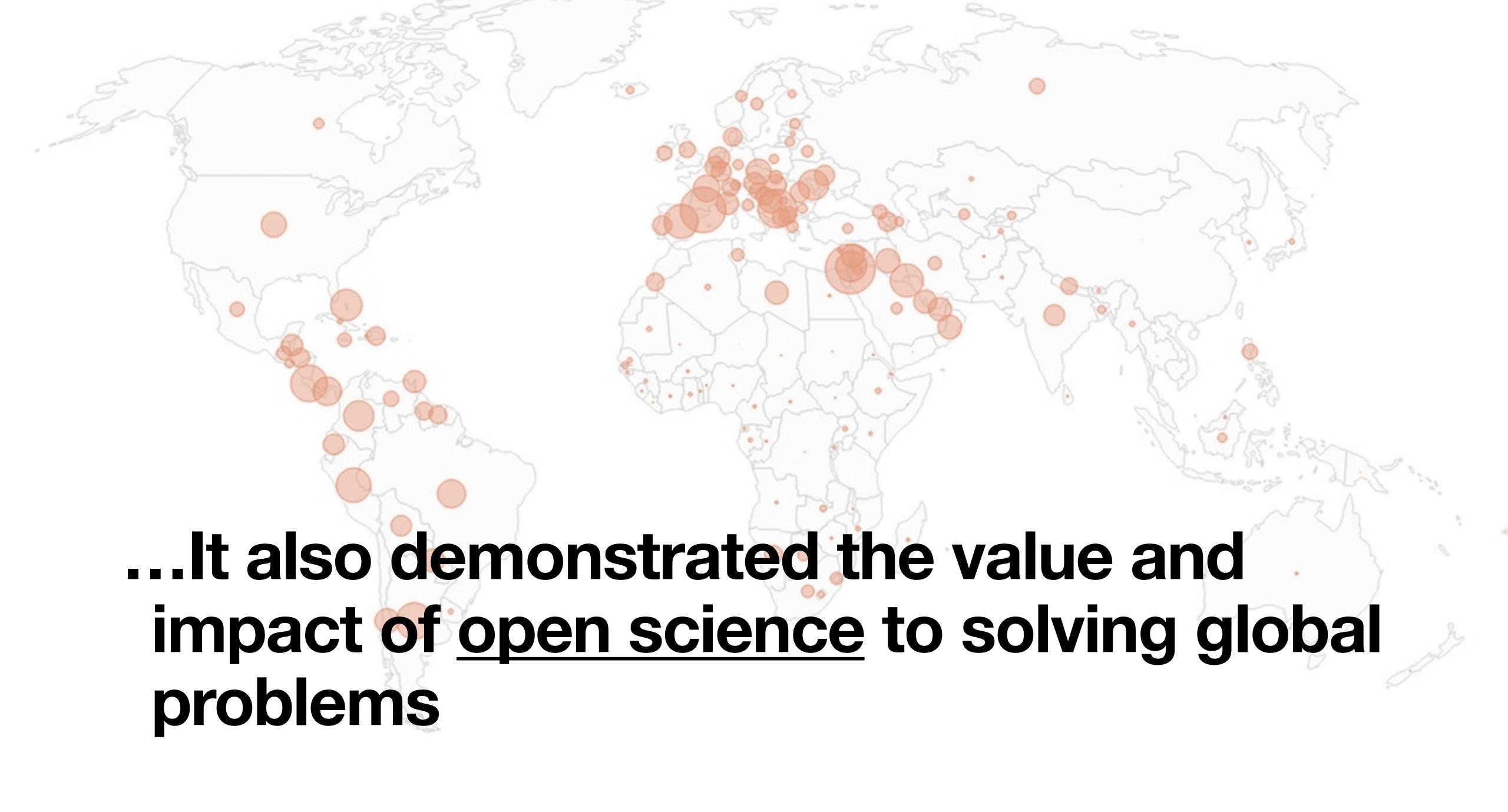
5:29 PM · May 2, 2020 · Twitter for iPhone

**29** Retweets **4** Quote Tweets **100** Likes

## **Sharing Biomolecular Simulation Data for COVID-19 Commitments**

- to making results available quickly via pre-prints: arXiv, bioRxiv, ChemRxiv...
- to make available input files, model-building and analysis scripts (e.g., Jupyter notebooks), and data necessary to reproduce the results;
- to use open data-sharing platforms to make available results as quickly as possible;
- to share algorithms and methods in order to accelerate reuse and innovation;
- to apply permissive open-source licensing strategies.

## It took a major global health crisis!



## Global problems

## **Existential threats**

- Global health
- Food security
- Planetary ecosystem sustainability
- Economic & social equity
- Energy & climate
- Water



CC BY 2.0 Stephen Morrison/Africa Practice for AusAID

# Open Science

What is it?

pen science "aims to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms that were used to generate those data"

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. Open Science by Design: Realizing a Vision for 21st Century Research. https://doi.org/gfxzc4

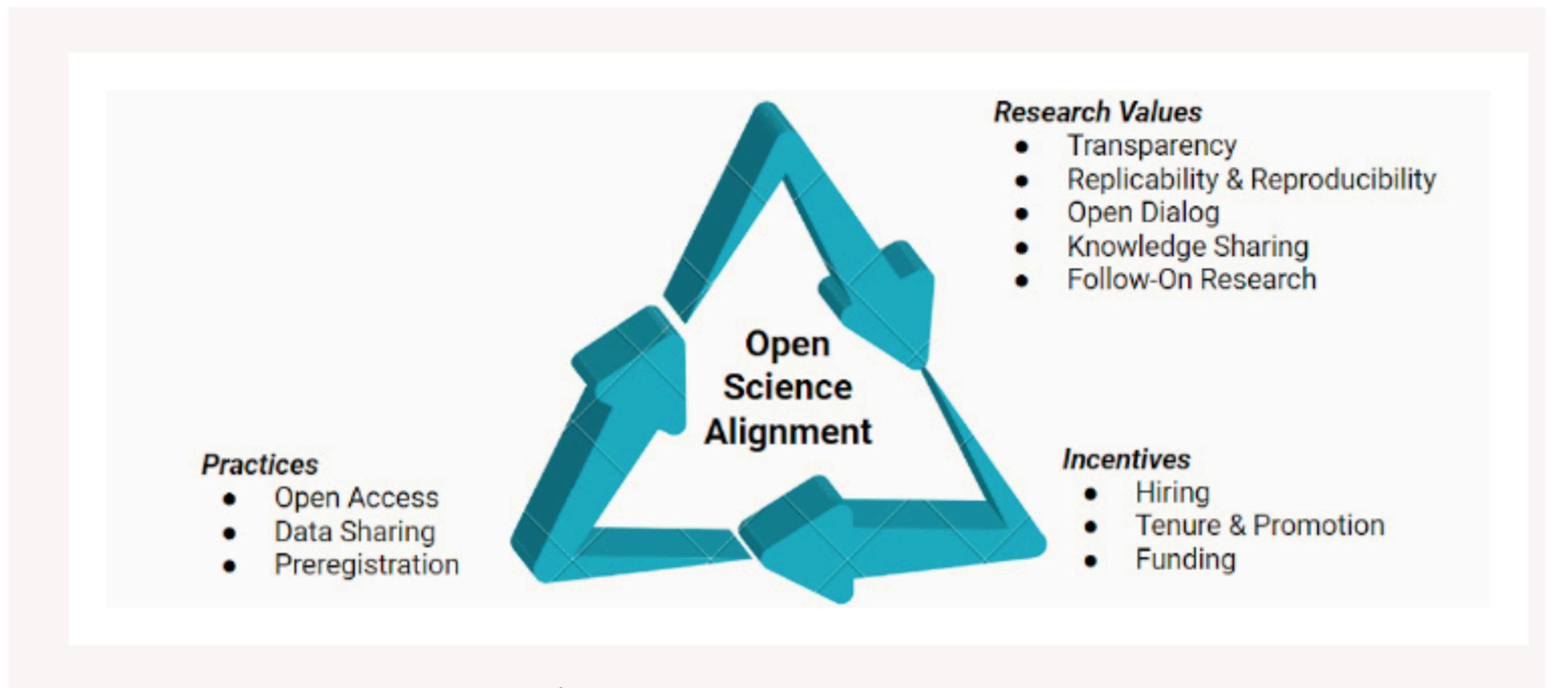


FIGURE 1-1 Open science alignment.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2021. Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop. https://doi.org/10.17226/26308

## Vision for EU 2016

"Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools."



https://doi.org/gk7tw3

## European Open Science Cloud, EOSC

A long-term initiative, launched 2018

- Ecosystem federating research data infrastructure, interlinking FAIR data and interoperable services to support open science
- Trusted digital space for hosting and processing research data
- offers distributed and cloud computing resources for researchers



| All Resources                      | 248 |
|------------------------------------|-----|
| CATEGORIES                         |     |
| Access physical & eInfrastructures | 63  |
| Aggregators & Integrators          | 22  |
| Processing & Analysis              | 115 |
| Security & Operations              | 14  |
| Sharing & Discovery                | 65  |
| Training & Support                 | 22  |

# "Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks"

Vicente-Saez, R. and Martinez-Fuentes, C., 2018. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88, pp.428-436. <a href="https://doi.org/gc5sjb">https://doi.org/gc5sjb</a>



#### **Computer Science > Computers and Society**

[Submitted on 26 Apr 2022 (v1), last revised 18 May 2022 (this version, v2)]

## Defining the role of open source software in research reproducibility

#### Lorena A. Barba

Reproducibility is inseparable from transparency, as sharing data, code and computational environment is a pre-requisite for being able to retrace the steps of producing the research results. Others have made the case that this artifact sharing should adopt appropriate licensing schemes that permit reuse, modification and redistribution. I make a new proposal for the role of open source software, stemming from the lessons it teaches about distributed collaboration and a commitment-based culture. Reviewing the defining features of open source software (licensing, development, communities), I look for explanation of the success from the defining features of open source software (licensing, development, communities), I look for explanation of the language-action framework of Winograd and Flores. I contend that reproducibility engenders trust, which we routinely build in community via conversations, and the practices of open source software help us to learn how to be more effective learning (discovering) together, contributing to the same goal.

Comments: 10 pages. Accepted for publication in IEEE Computer

Subjects: Computers and Society (cs.CY)

Cite as: arXiv:2204.12564 [cs.CY]

(or arXiv:2204.12564v2 [cs.CY] for this version) https://doi.org/10.48550/arXiv.2204.12564

NASA Transform to Open Science

Mission: 2022 to 2027

- 2023 is the Year of Open Science
- Deploy a learning platform and develop a core curriculum in open science
- Train 20,000 scientists from the earth and space science community
- \$130 million investment through FY2027

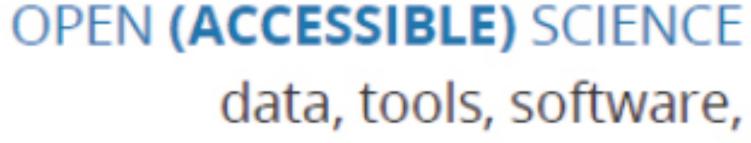


#### The Four Meanings of "Open" in Open-source Science



### OPEN (TRANSPARENT) SCIENCE

scientific process and results should be visible, accessible, and understandable



data, tools, software, documentation, and publications should be accessible to all (FAIR)



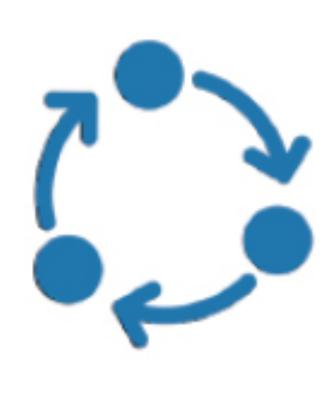


### OPEN (INCLUSIVE) SCIENCE

process and participants should welcome participation by and collaboration with diverse people and organizations

### OPEN (**REPRODUCIBLE**) SCIENCE

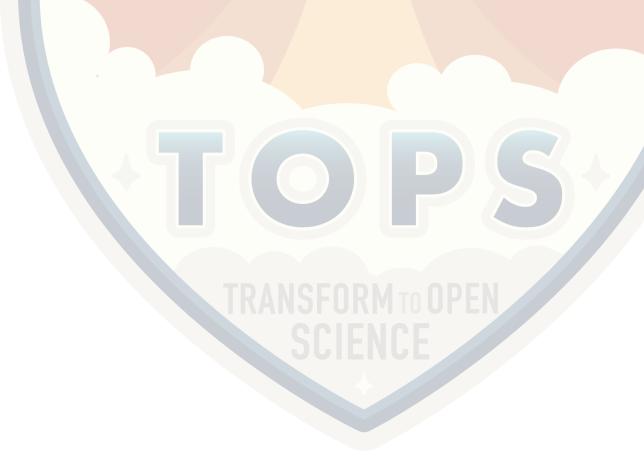
scientific process and results should be open such that they are reproducible by members of the community



National Academies of Sciences, Engineering, and Medicine. 2022. Open Scholarship Priorities and Next Steps: Proceedings of a Workshop—in Brief. <a href="https://doi.org/10.17226/26557">https://doi.org/10.17226/26557</a>

## "Scientific process and results should be visible, accessible, and understandable."

NASA TOPS open science definition



## Open Principles

## To guide your research

- Transparency
- Collaboration
- Quality
- Reproducibility
- Re-use
- Accountability

- Participation
- Sharing
- Fairness
- Inclusion
- Equity
- Connection

### How to start

### Prepare to share, be someone

- 1. Get ORCID (and use it everywhere)
- 2. Join GitHub, Figshare, Zenodo
- 3. Do you write any code? Familiarize yourself with OSI licenses.
- 4. Find preprint repository for you: arXiv, bioRxiv, chemRxiv, medRxiv...
- 5. Chose collaboration platform and internal data management
- 6. Think automation: turning protocols into code

## Open data and open science For reproducibility

reproducible computational research:
 "all details of the computations—code and data—are made conveniently available to others"

— Donoho et al., 2009





Our new study shows that data availability statements are not very useful; 1670 (93%) authors who indicated that data are available on request either did not respond or declined to share their data with us. Journal of Clinical Epidemiology: doi.org/10.1016/j.jcli...



#### Journal of Clinical Epidemiology

Available online 30 May 2022 In Press, Journal Pre-proof ?



Original Article

Many researchers were not compliant with their published data sharing statement: mixed-methods study

Mirko Gabelica <sup>1</sup> ⊠, Ružica Bojčić <sup>2</sup> ⊠, Livia Puljak <sup>3</sup> ⋈ ⊠

## Good data management FAIR Principles

 digital artifacts of research should be Findable, Accessible, Interoperable and Reusable for machines and for people

—Wilkinson et al., 2016, doi:10/bdd4



https://www.force11.org/group/fairgroup/fairprinciples

### Data repositories

### **Choose wisely**

- must provide a unique global identifier for your data (typically a digital object identifier, DOI)
- must offer long-term preservation guarantees (at least 10 years)













- general-purpose repository for all kinds of digital artifacts of research
- any file format, up to 5GB in size
- free and unlimited for public items

## ændo

- created by CERN and OpenAIRE
- free and non-commercial
- log in with your ORCID
- deposit large files: up to 50GB by default

https://zenodo.org/communities/barbagroup/

## Preprints

- A way of life
- In physics, math, CS... arXiv is a way of life
- Preprints growing by all metrics
- Explosion of 'Xiv sites









| Nature Publishing<br>Group                                     | Compatible | Communication between researchers includes not only conferences but also preprint servers. The ArXiv preprint server is the medium of choice for (mainly) physicists and astronomers who wish to share drafts of their papers with their colleagues, and with anyone else with sufficient time and knowledge to navigate it. [] If scientists wish to display drafts of their research papers  |           |
|--|------------|--|-----------|
| IOP Publishing   | Compatible | You m an IOF copyrig wordin IOP Pr it. The submit uploac the Au  Most journals accept article P's sillowing ournal] srived f wish to   | om [11] 🗗 |
| Oxford Journals  | Compatible | on the subjec This a   | [12]ਲ     |
| Elsevier   | Compatible | Submitted to preprint servers  Identifinote the Inform  Submitted to preprint servers  Submit |           |
| Springer, incl. SpringerOpen Journals and BioMed Central (BMC) | Compatible | Postin List of academic journals by preprint policy, Centra Wikipedia  Wikipedia   |           |
| Taylor & Francis   | Compatible | "This is your original manuscript (often called a "preprint"), and you can share this as much as you like. If you do decide to post it anywhere, including onto an academic networking site, we would recommend you use an amended version of the wording below to encourage usage and citation of your final, published article."   | [17] ₺    |

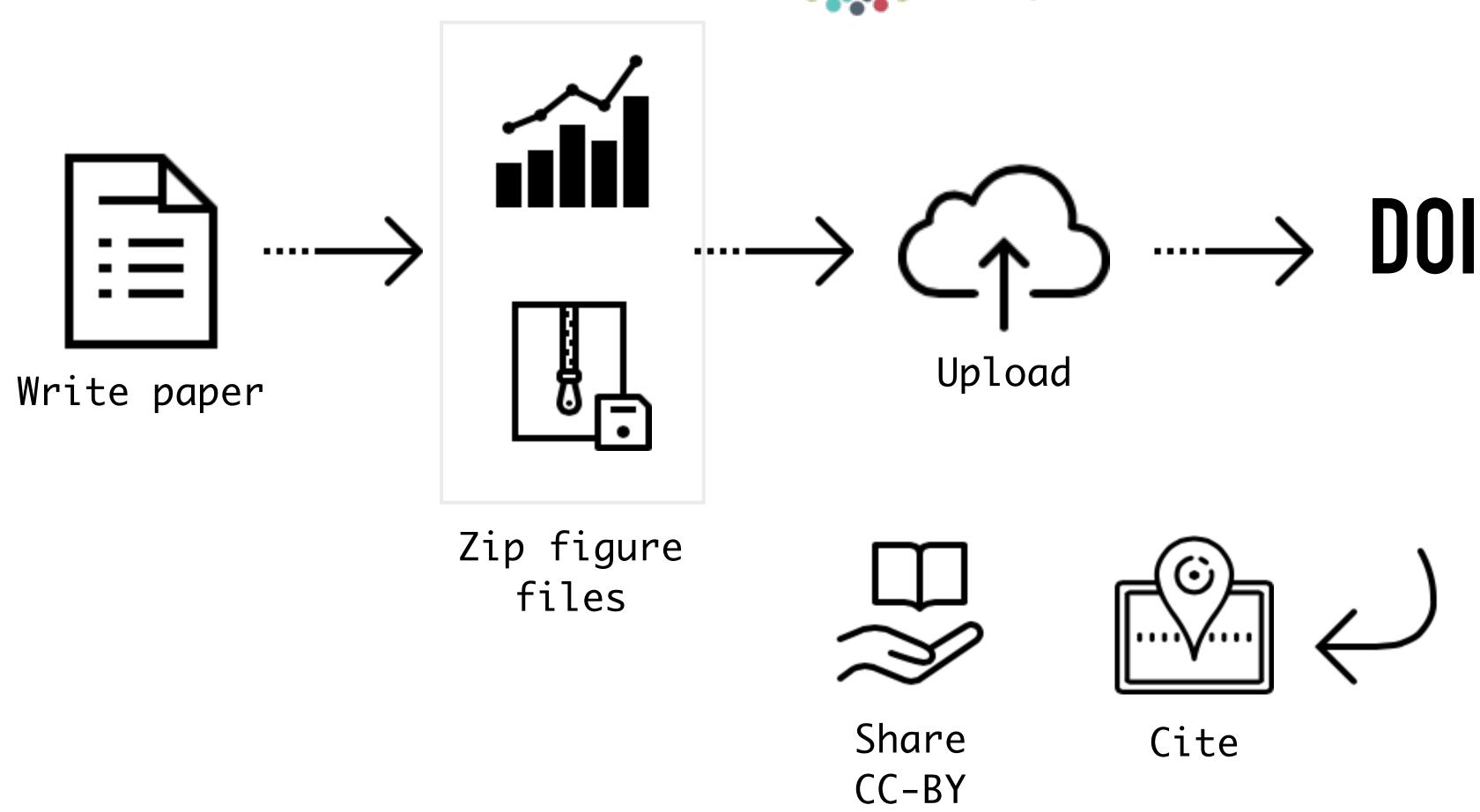
## ReproPacks

- For main results in a paper, we share data, plotting script & figure under CC-BY.
- Deposit the file bundle as a Figshare object and get a DOI
- We cite this DOI in the figure caption!



### Our workflow





Icons from Icons8.

## Publishing research software

To slash the hurdles for researchers to receive academic credit for all their output, including software and data.

Sign in

## The Journal of Open Source Software

A developer friendly journal for research software packages.

Learn more »

Accepted papers (391)

#### whimian / pyGeoPressure JOSS 10.21105/joss.00992

pyGeoPressure is a python package for geopressure (pore pressure) prediction with well log data and seismic velocity data in the field of hydrocarbon exploration. 10.5281/zenodo.1452001 [2

#### □ ropensci / piggyback JOSS 10.21105/joss.00971

An approach for managing larger data files on a github repository. See https://ropensci.github.io/piggyback/. This package was reviewed in the rOpenSci onboarding:... 10.5281/zenodo.1432527 [-]

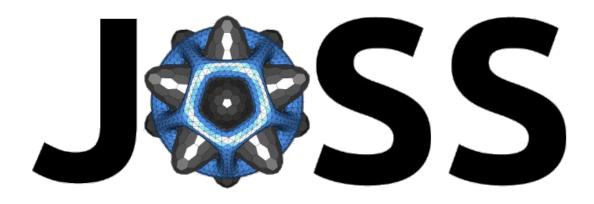
#### ejhigson / perfectns JOSS 10.21105/joss.00985

perfectns performs dynamic and standard nested sampling in spherically symmetric cases which allow the algorithms to be followed perfectly. This is extremely useful for statistical research into...

10.5281/zenodo.1443509 [2

### JOSS infrastructure

- GitHub—open-source software hosting & collaboration
- Zenodo—data repository by CERN
- ORCID—author identification
- CrossRef—DOI minting
- custom web app and Ruby bot





## The Journal of Open Source Software

A developer friendly journal for research software packages.





### Open peer review

### A cultural change

- peer review can suffer from implicit bias that disadvantages women, minorities and people from less-prestigious institutions
- educate reviewers about bias in the review process, and equip them with tools to interrupt this bias
- Double-blind peer review does not help build awareness of bias, is clunky, and easy to "guess" author identity
- Example of open review: The Journal of Open Source Software





tingyu66 commented on Apr 22, 2021 • edited -

On Ubuntu 18.04.5 LTS (GNU/Linux 4.15.0-118-generic x86\_64) the ./configure CXX=g++ stops with

```
checking for fftw3.h... yes checking for fftwf_malloc in user specified location... no checking for fftw_malloc in user specified location... no configure: error: cannot find fftwf library.
```

Exafmm-t requires the FFTW3 shared library in both single and double precision (libfftw3f.so and libfftw3.so), so we suggest to install libfftw3-dev package, as described in our documentation:

```
apt-get update
apt-get -y install libfftw3-dev
```

If you want to build FFTW3 library from source, you probably need to compile separately for each precision if I remember correctly. Then add the directory of the shared libraries to LIBRARY\_PATH and LD\_LIBRARY\_PATH environment variables.



pitsianis commented on May 16, 2021



I am done with review. I like the Jupyter notebook.



) ( 1

https://joss.theoj.org

Article Open Access Published: 18 January 2019

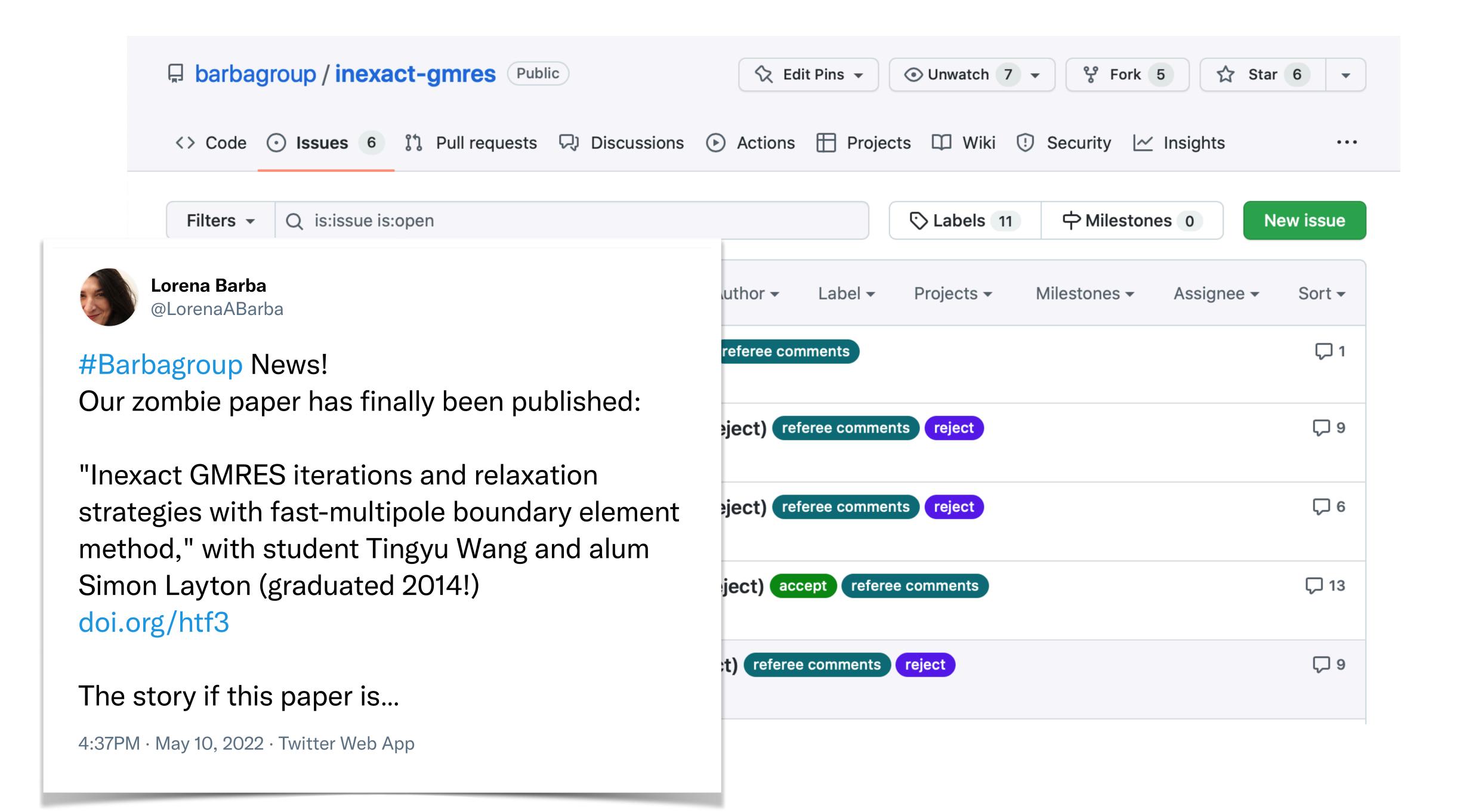
## The effect of publishing peer review reports on referee behavior in five scholarly journals

Giangiacomo Bravo, Francisco Grimaldo, Emilia López-Iñesta, Bahar Mehmani & Flaminio Squazzoni

Nature Communications 10, Article number: 322 (2019) Cite this article

18k Accesses 51 Citations 346 Altmetric Metrics

https://doi.org/gfvpwd

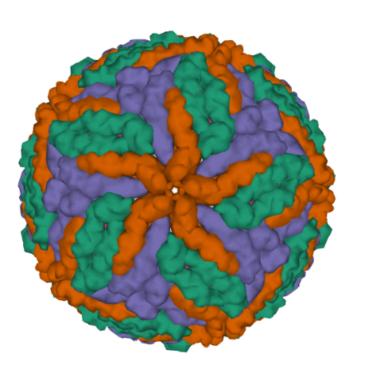


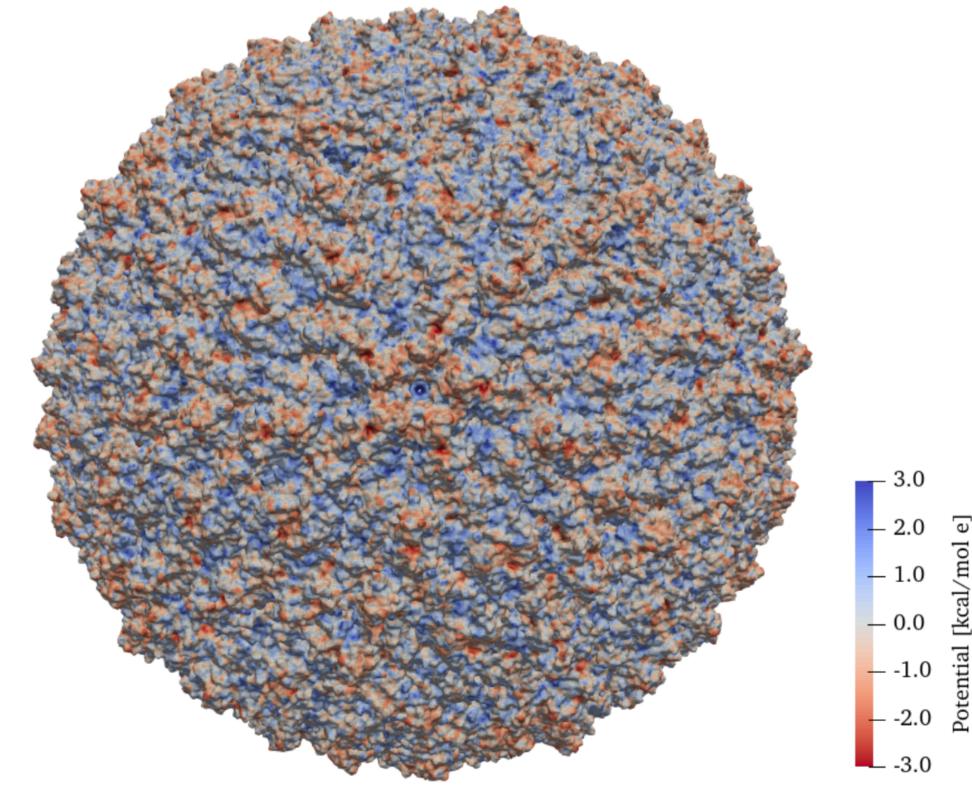
# High-productivity, high-performance workflow for virus-scale electrostatic simulations with Bempp-Exafmm

Tingyu Wang, Christopher D. Cooper, Timo Betcke, Lorena A. Barba.

Submitted for peer review

Preprint arXiv:2103.01048





https://github.com/barbagroup/bempp\_exafmm\_paper/

## "It is time to improve the quality, transparency, and accountability of the peer review system."

Haffar, S., Bazerbachi, F. and Murad, M.H., 2019, April. Peer review bias: a critical review. In Mayo Clinic Proceedings (Vol. 94, No. 4, pp. 670-676). https://doi.org/gnjhwk

## Reproducibility PI Manifesto, Transform to Open Science

Jupyter Community in MSK Imaging Research

