

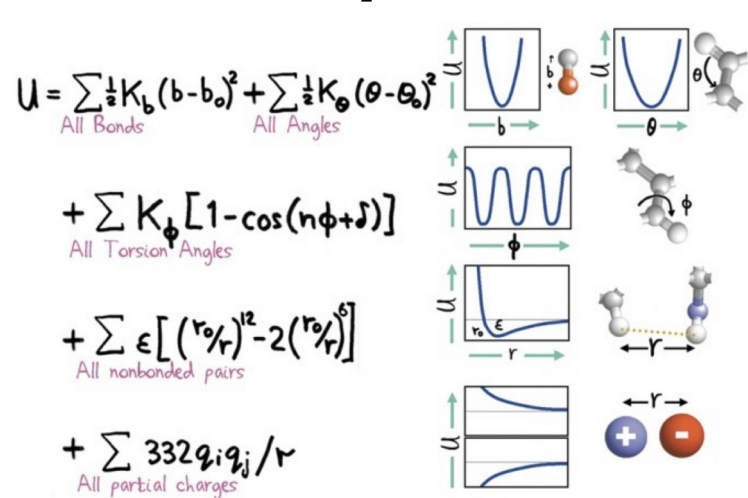


Shedding Light on the Dark Matter of Molecular Dynamics Simulations

Pierre Poulain

Molecular dynamics simulations require resources

expertise



M. Levitt

GROMACS
fast, flexible & free



computer power



Source: Penalva, Wikimedia, CC BY-SA

Sharing MD simulations files



- Requirements from funders, journals or institutions
- Open science
- Reproducibility

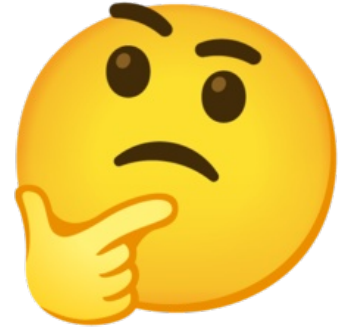
No consensus data repository for MD simulation files

Albeit many initiatives: MoDEL, GPCRmd, NMR-Lipids...

Use of generic **non-moderated** data repositories:
Zenodo, Figshare, OSF, Dryad...

Dark matter of MD simulations

Data that is **technically accessible**,
but neither **indexed**, **curated**,
or easily **searchable**.



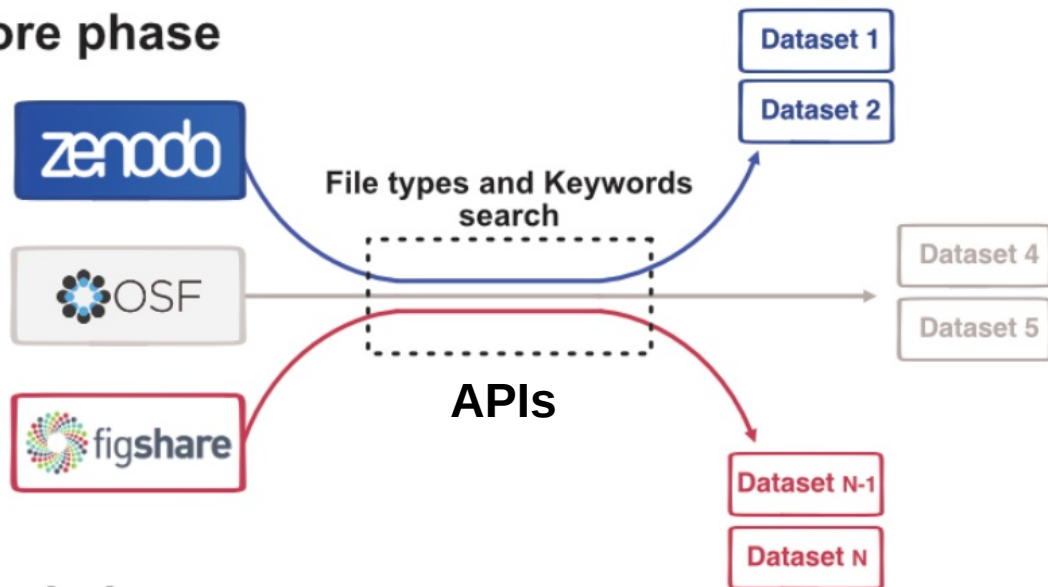
MDverse

1. Find and index **scattered** MD data
2. Extract, enhance and explore **metadata**
3. Assess **reusability** (R from FAIR principles)

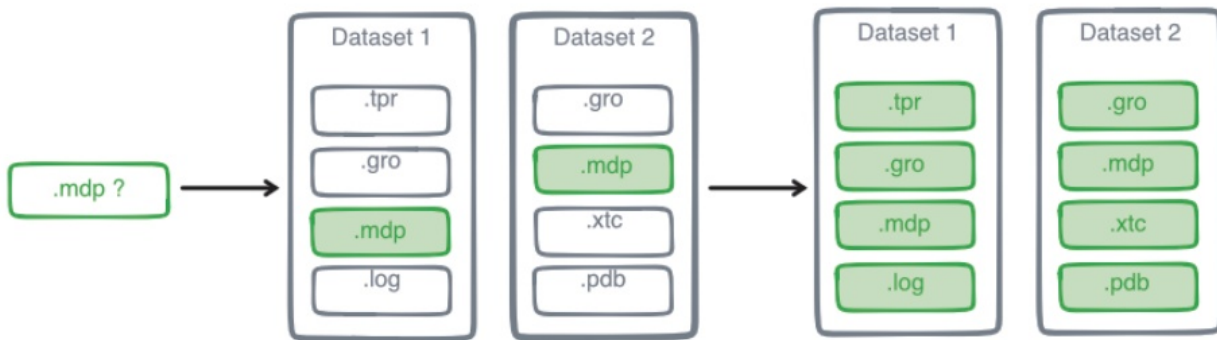
A universal search engine for MD open data
(Not yet another data repository)

Step 1: Find and index scattered MD data

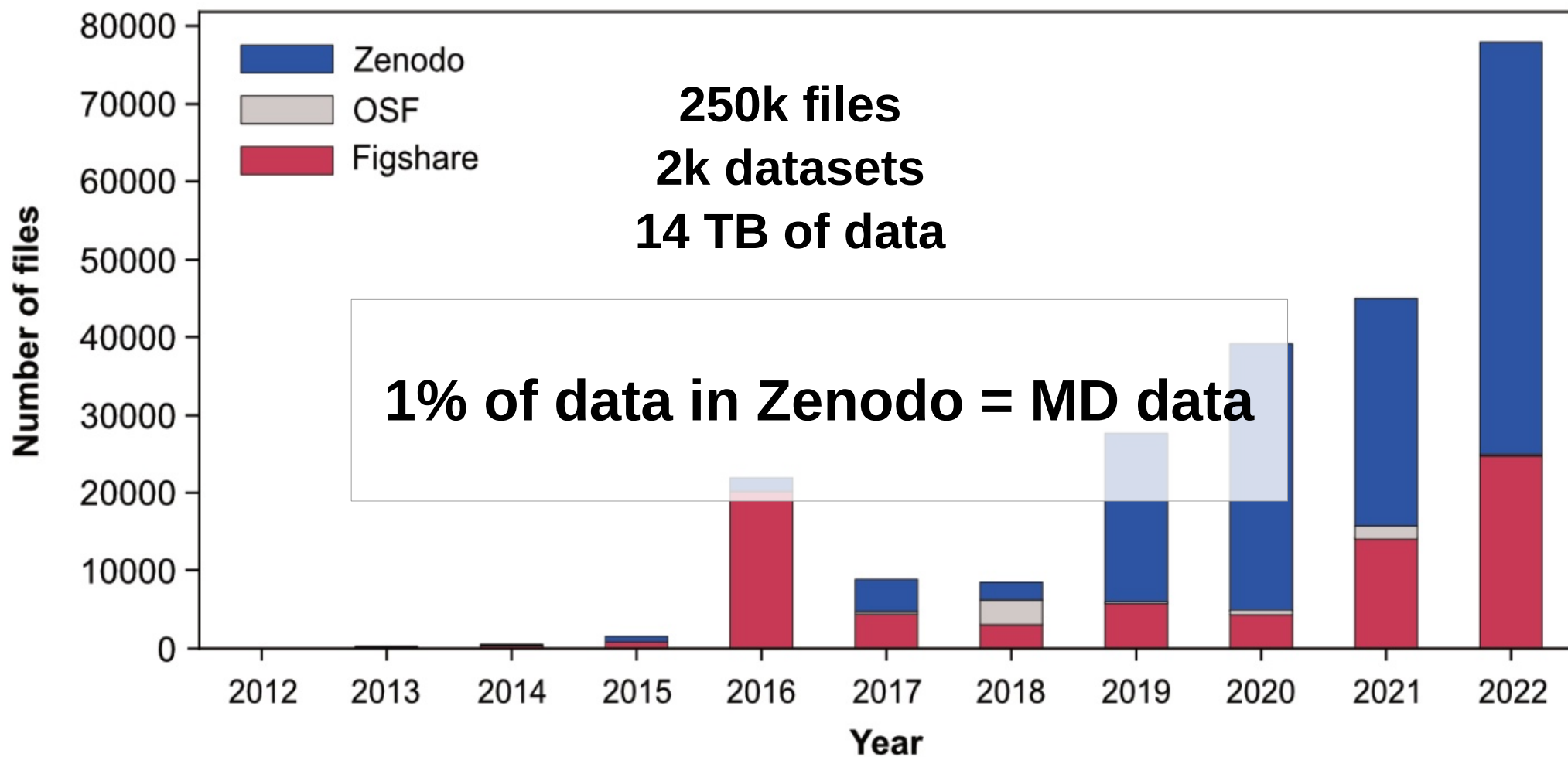
Explore phase



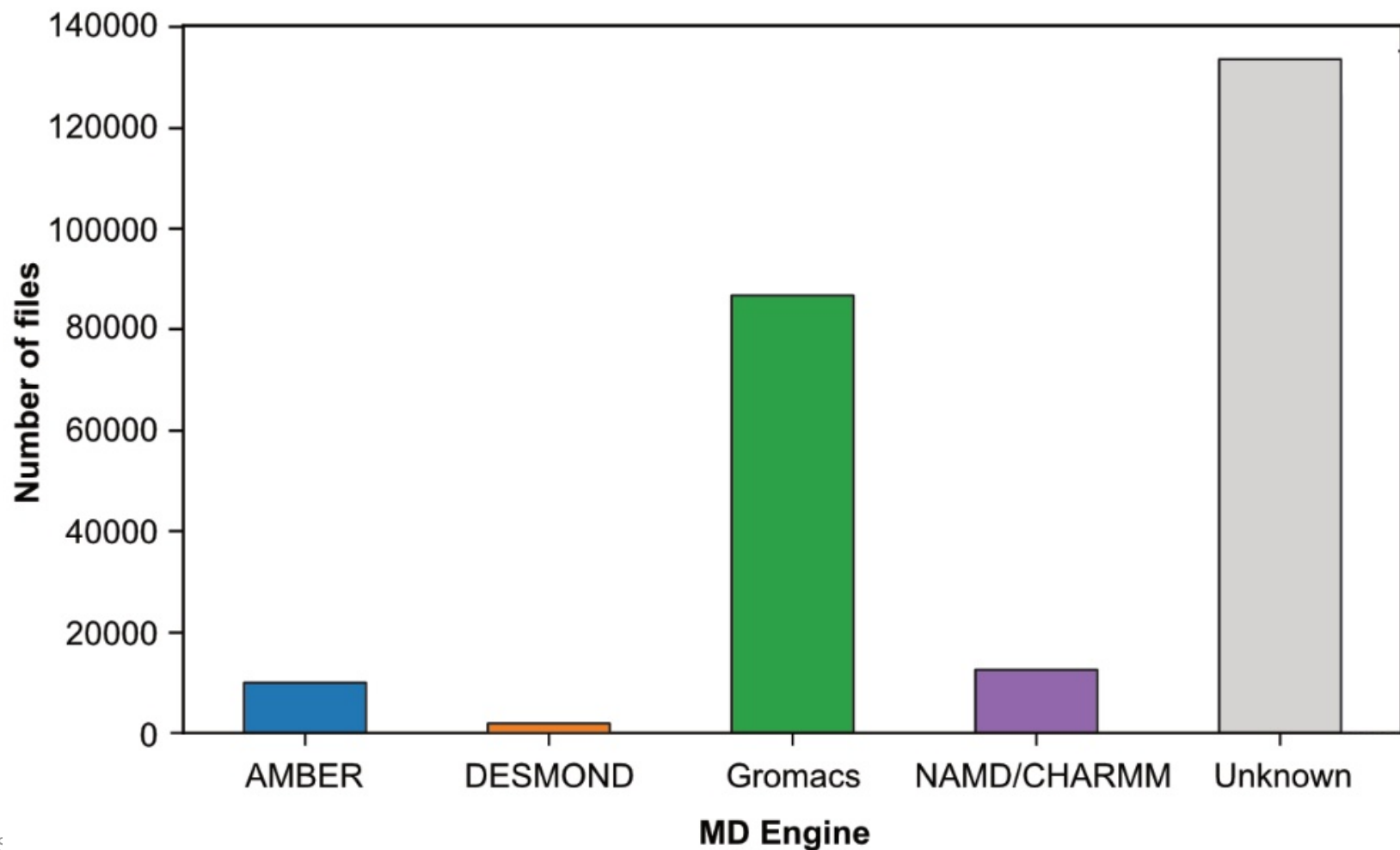
Expand phase



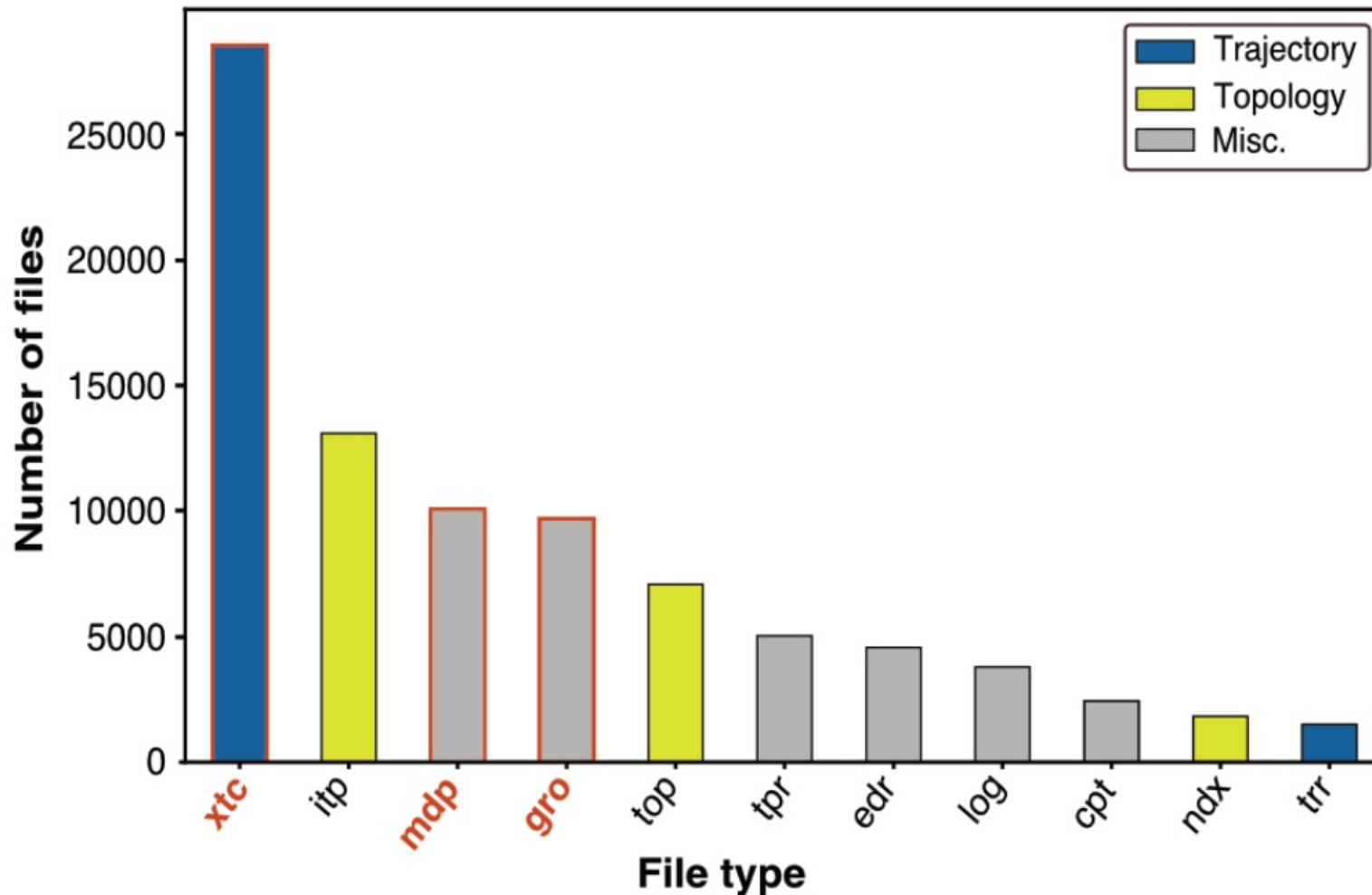
Step 1: Find and index scattered MD data



Step 1: Find and index scattered MD data



Step 1: Focus on Gromacs files



Step 2: Metadata = context



No metadata (bad)



Natural language
metadata (good)



</

Controlled vocabulary
metadata (better)



Metadata for MD simulation: identity of simulated molecules, temperature, length, force field, software...

Step 2: Source of MD metadata

February 27, 2019

Journal article

Open Access

SLIPID POPG-POPE 1:3 Bilayer Simulation (Last 100 ns, 150 mM NaCl, 310 K)

PEÓN, Antonio

Simulation of a POPG-POPE 1:3 bilayer of 500 lipids (126 L-POPG lipids and 374 POPE lipids, 250 per leaflet) is simulated for 500 ns using Gromacs v5.1.2 in water solution with Na⁺ counterions and 150 mM of NaCl. The SLIPID model is employed for lipids and TIP3P Water Model.

Trajectory (.xtc) is for the last 100 ns of a simulation of 500 ns with data saved every 10 ps. Additionally, the topology (.top), simulation parameter file (.mdp), index file (.ndx), portable binary run input file (.tpr) and the energy output file (.edr) are provided.

Files (5.2 GB)

Name

Size

index.ndx

5.5 MB

Download

md5:5a8c0d796996b9432ec7a15919544a17

m_400_500_PG_PE_1_3_slipid.xtc

5.2 GB

Download

md5:177e0b0c513b1910a9bbe6329ba7a814

md_02.tpr

3.9 MB

Download

md5:3b00b35301d3079a63f7055e09f4089a

md_400_500.edr

33.0 MB

Download

md5:9496927277c7b890c0a41974a3e6c746

topol.top

554 Bytes

Download

md5:1be0f066d982838ef60156ec2449d117

30

views

23

downloads

[See more details...](#)

Indexed in

OpenAIRE

Publication date:

February 27, 2019

DOI:

DOI 10.5281/zenodo.2579224

Keyword(s):

POPG-POPE 1:3, SLIPID

License (for files):

[Creative Commons Attribution 4.0 International](#)

Versions

Version 1

Feb 27, 2019

10.5281/zenodo.2579224

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.2579223. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Step 2: Uneven amount of metadata

February 27, 2019 Journal article Open Access

SLIPID POPG-POPE 1:3 Bilayer Simulation (Last 100 ns, 150 mM NaCl, 310 K)

PEÓN, Antonio

Simulation of a POPG-POPE 1:3 bilayer of 500 lipids (126 L-POPG lipids and 374 POPE lipids, 250 per leaflet) is simulated for 500 ns using Gromacs v5.1.2 in water solution with Na⁺ counterions and 150 mM of NaCl. The SLIPID model is employed for lipids and TIP3P Water Model .

Trajectory (.xtc) is for the last 100 ns of a simulation of 500 ns with data saved every 10 ps. Additionally, the topology (.top), simulation parameter file (.mdp), index file (.ndx), portable binary run input file (.tpr) and the energy output file (.edr) are provided.

Files (5.2 GB)

| Name | Size | |
|--|--------|--------------------------|
| index.ndx | 5.5 MB | Download |
| md5:5a8c0d796996b9432ec7a15919544a17 ? | | |
| m_400_500_PG_PE_1_3_slipid.xtc | 5.2 GB | Download |
| md5:177e0b0c513b1910a9bbe6329ba7a814 ? | | |

March 22, 2021 Dataset Open Access

Simulations of a pulmonary surfactant monolayer with additional compounds

To be added.

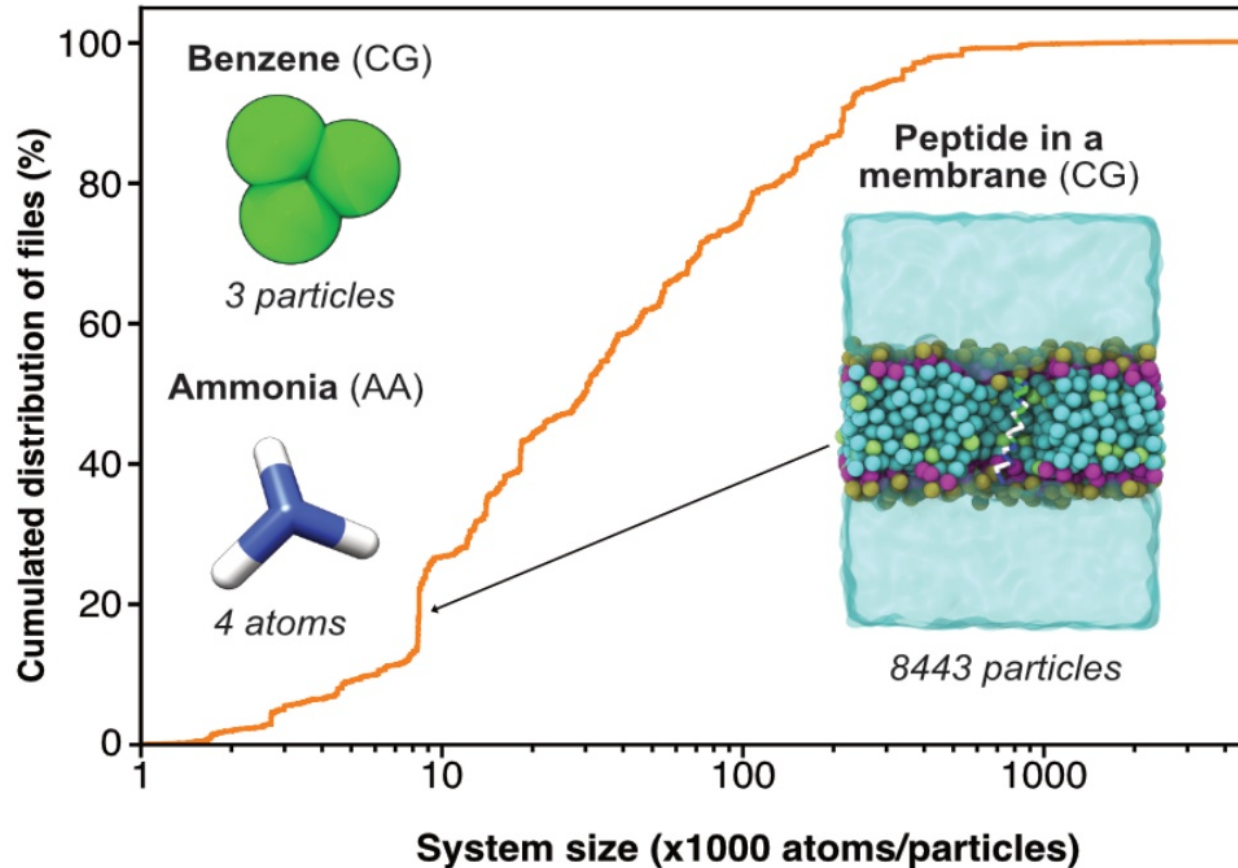
Files (47.1 GB)

| Name |
|--|
| benzaldehyde.ndx |
| md5:987f07a7e7fab150db9713558d7dd8ad ? |
| benzaldehyde.top |
| md5:e90dbc7f37ff3ea109ca4aba127cd3ca ? |

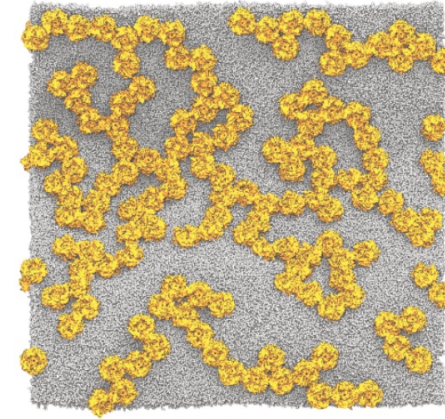


Open the can and taste the soup
Open the files and extract metadata

Step 2: Metadata from Gromacs (.gro) files

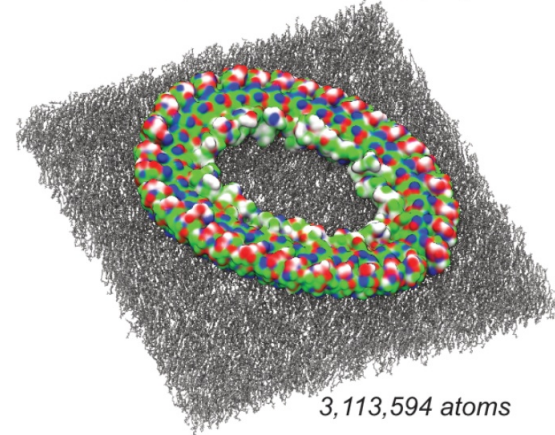


Kir Channels in a plasma membrane (CG)



3,522,816 particles

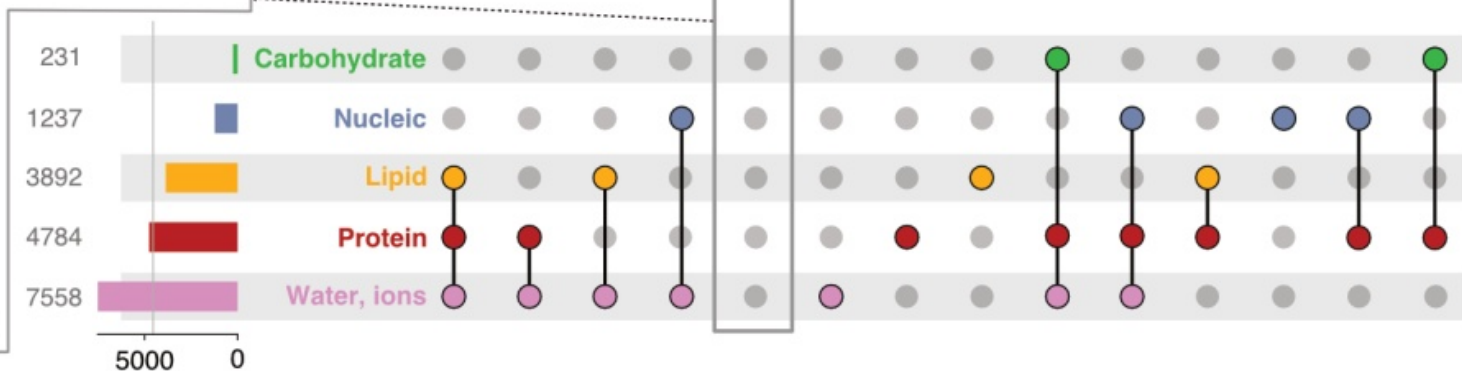
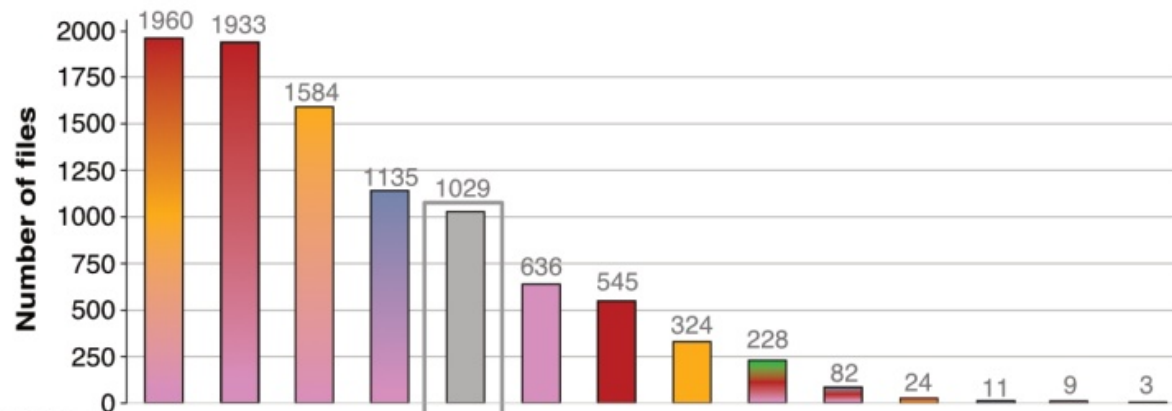
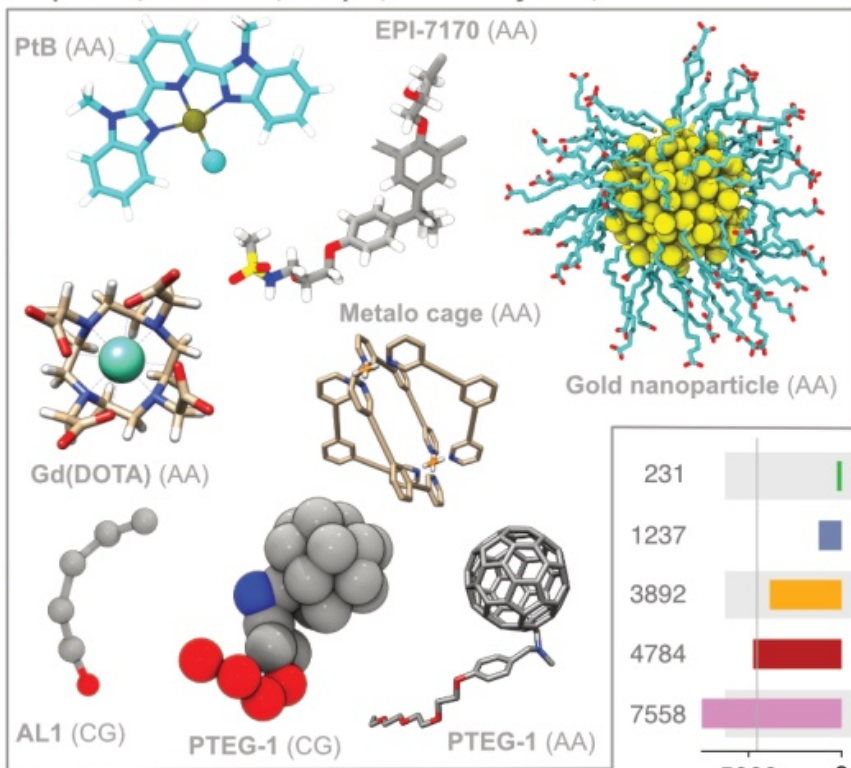
Gasdermin prepore in a plasma membrane (AA)



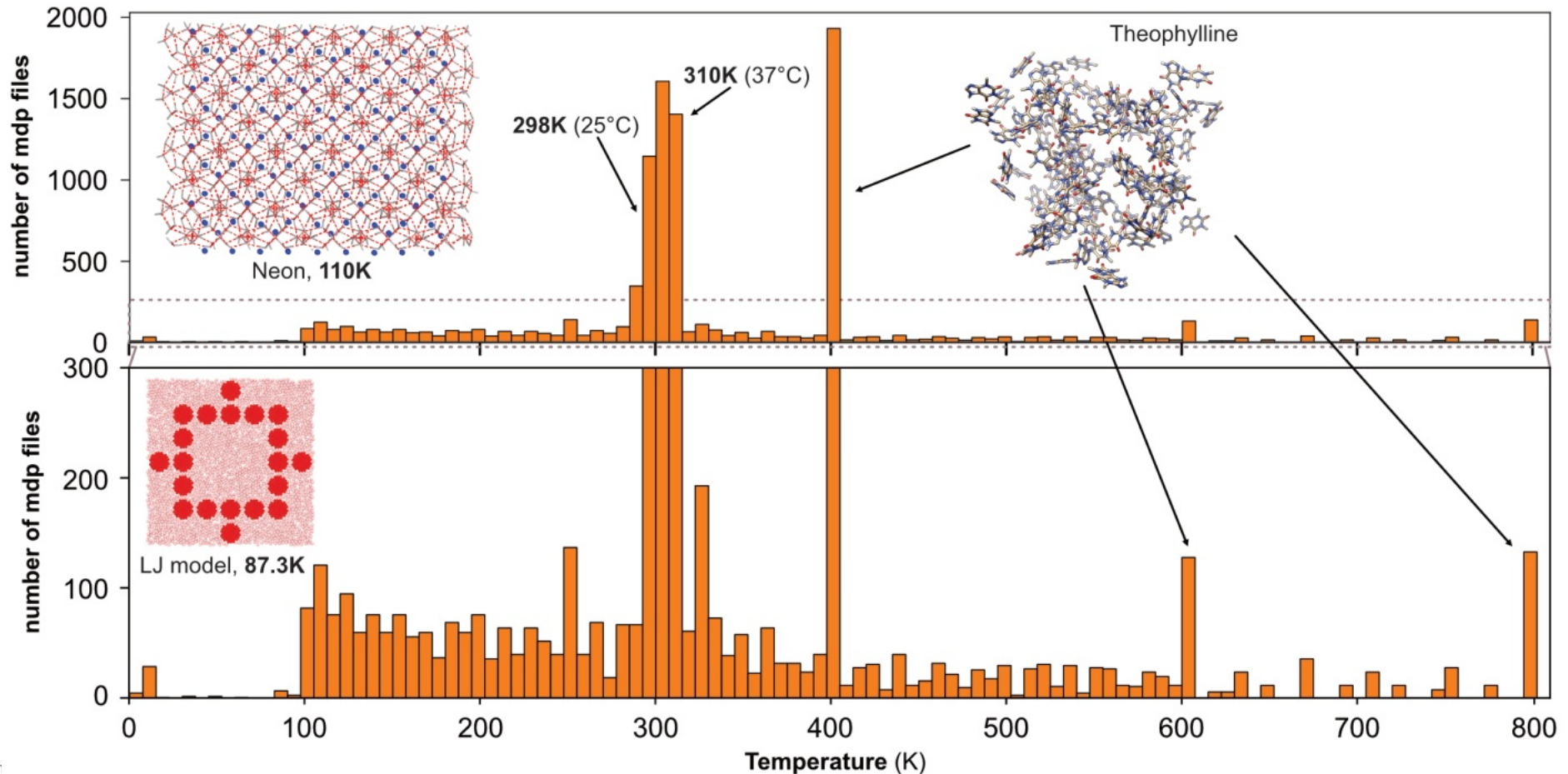
3,113,594 atoms

Step 2: Metadata from Gromacs (.gro) files

No protein, no nucleic, no lipid, no carbohydrate, no water or ions



Step 2: Metadata from Gromacs (.mdp) files



Step 2: Explore metadata



Datasets

GRO files

MDP files

Selected row:

3261

1

9780

Dataset: [zenodo 4943557](#)

Creation date: 2021-06-14

Author(s): Joseph, Thomas

Title: *Data from: Common internal allosteric network links anesthetic binding sites in a pentameric ligand-gated ion channel*

Description:
General anesthetics bind reversibly to ion channels, modifying their global conformational distributions, but the underlying atomic mechanisms are not completely known. We examine

MDverse data explorer

.gro files quick search

Enter search term. For instance: Covid, POPC, Gromacs, CHARMM36, 1402417

☐ Add filter

Export to tsv

9780 elements found

Show 20 entries

| index | Dataset | ID | Title | Creation date | Authors | Description | File name | Atom number | Protein | Lipid |
|-------|---------|-------------------------|--|---------------|------------|---|-------------|-------------|---------|-------|
| 1 | zenodo | 1468560 | C36 POPC simulation with 17 lipids per leaflet, 300K | 2018-10-22 | Hanne ... | POPC bilayer with 30 waters per lipid (17+17), at 300K, si... | whole17.gro | 7616 | false | true |
| 2 | zenodo | 6526243 | Amyloid-beta 16-22 peptide dimer simulation (150mM Na... | 2022-05-07 | Kav, Ba... | Amyloid-beta 16-22 peptide dimer simulation with the CH... | prod.gro | 32020 | true | false |
| 3 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 4 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 5 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 6 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 7 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 8 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 9 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |
| 10 | zenodo | 838641 | Large DPPC monolayer simulations with Charmm36+OPC ... | 2017-08-03 | Javanai... | DPPC monolayers simulated at a varying area per lipid in t... | DPPC-31... | 232488 | false | true |

<https://mdverse.streamlit.app/>

Step 3: Back to the FAIR principles

Findable

Accessible

Interoperable





Reusable

*“[...] the FAIR Principles put specific emphasis on enhancing the ability of **machines to automatically** find and **use the data**, in addition to supporting its reuse by individuals.”*

Wilkinson et al., Scientific Data, 2016.

Step 3: Assess accessibility (zip files)

All-atom molecular dynamics simulations of SARS-CoV-2 envelope protein E

 Kuzmin Alexander;  Orekhov Philipp;  Astashkin Roman;  Gordeliy Valentin;  Gushchin Ivan







The trajectories of all-atom (AA) MD simulations (NoPTM-1;2;3;4_POPC;Mix_CHARMM36m: 0.1x3 μ s) were obtained based on 4 starting representative conformations from the coarse-grained simulation (10.5281/zenodo.4740706). For each starting structure, there are six trajectories of the E protein: 3 with the protein embedded in the membrane containing POPC, and 3 with the membrane mimicking the natural ERGIC membrane (Mix: 50% POPC, 25% POPE, 10% POPI, 5% POPS, 10% cholesterol).

Simulations have been performed using the CHARMM36m (AA) force field, running with the GROMACS 2019.5 package on the supercomputer JURECA at Forschungszentrum Jülich under the conditions reported in bioRxiv 2021.03.10.434722.

<https://doi.org/10.1002/prot.26317>



[Preview](#)

Files (45.6 GB)

| Name | Size | |
|--|--------|--|
| NoPTM-1_Mix_CHARMM36m_0.1x3mks.zip | 6.0 GB |  Preview  Download |
| md5:3f3854a3de4a10489e9895b8ba5d368b  | | |
| NoPTM-1_POPC_CHARMM36m_0.1x3mks.zip | 5.1 GB |  Preview  Download |
| md5:e9ef288e489689c70956952091bba5bf  | | |

88 % of indexed files
were in zip files

 NoPTM-2_Mix_CHARMM36m_0.1x3mks.zip

| | |
|---|---------|
|  NoPTM-2-1_Mix_CHARMM36m_0.1mks.xtc | 1.9 GB |
|  NoPTM-2-2_Mix_CHARMM36m_0.1mks.xtc | 1.9 GB |
|  NoPTM-2-3_Mix_CHARMM36m_0.1mks.xtc | 1.9 GB |
|  NoPTM-2_Mix_CHARMM36m.pdb | 11.8 MB |
|  NoPTM-2_Mix_CHARMM36m.tpr | 4.1 MB |

Step 3: Assess reusability

A Gromacs trajectory file could be reused to:

- Analyse a simulation (.xtc + .pdb/.gro/.tpr)
- Continue a simulation (.gro/.trr/.cpt + .mdp/.tpr)

Do we have any proof we can actually reuse the data?

Conclusions

Depositing MD simulations data in a FAIR-enabled repository does **not** guarantee your data is actually FAIR:

- Provide metadata (context)
- Avoid zip files



Why is it important?

- Sharing and storing data cost money and energy
"Biomolecular Simulations for a Better World"
- Dynamic generative deep-learning models?
(Need for high quality, curated data)

What's next?

- Explore other “data” repositories: Dryad, GitHub (yep)...
- Extract structured metadata from raw text (text mining / Named Entity Recognition)
- Collectively define an ontology for (MD) simulations

Thanks 🙏

IJM, Paris, France

Lisa Bouarroudj
Mohamed Oussaren

IBPS, Toulouse, France

Magdalena Szczuka
Matthieu Chavent

LBT, Paris, France

Marc Baaden

Amsterdam, Netherlands

Steven Garcia

Univ Copenhagen, Denmark

Johanna K. S. Tiemann
Kresten Lindorff-Larsen

KTH Royal Inst. Tech. Stockholm, Sweden

Lucie Delemotte
Erik Lindahl

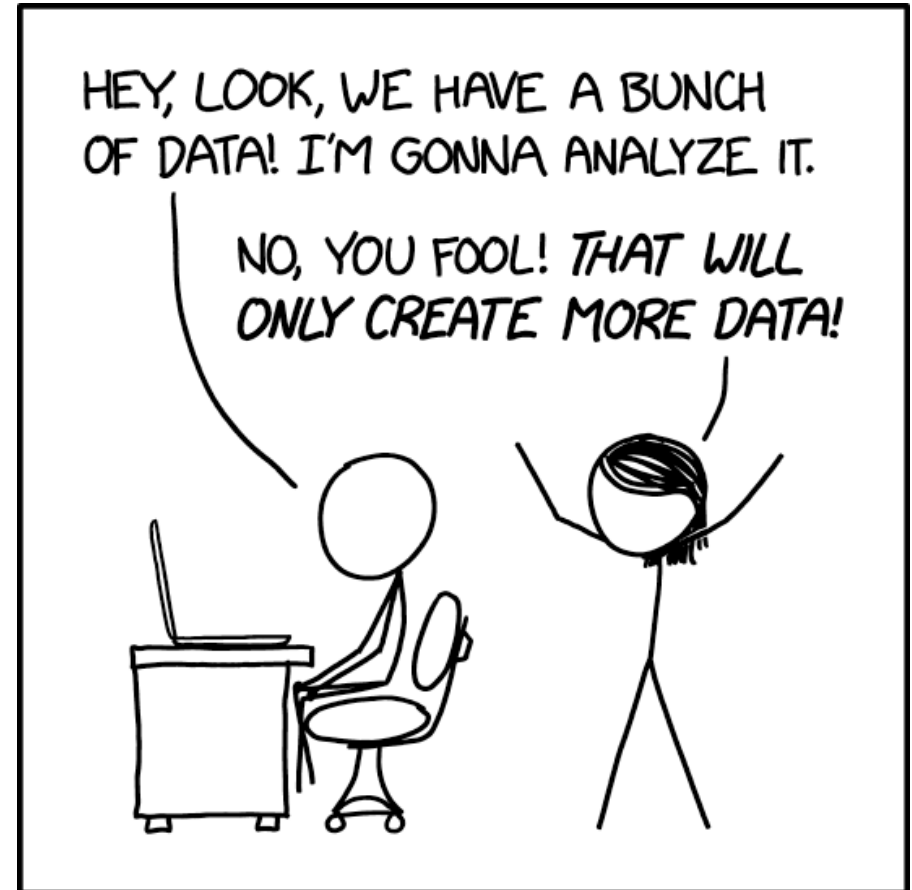
Stockholm Univ. , Sweden

Rebecca J. Howard

We ❤️ open science

Preprint [bioRxiv]
(data + source code)

This presentation [Zenodo]
DOI 10.5281/zenodo.8129492



Source: Data trap, XKCD, CC BY-NC