Sharing & reusing molecular dynamics data: what did we miss?

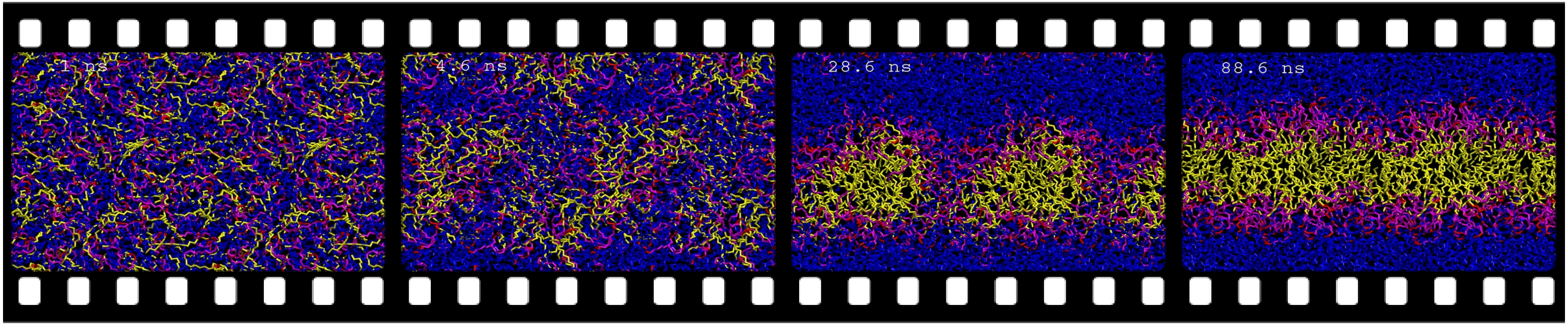
Pierre Poulain

pierre.poulain@u-paris.fr

Laboratoire de Biochimie Théorique, Université Paris Cité

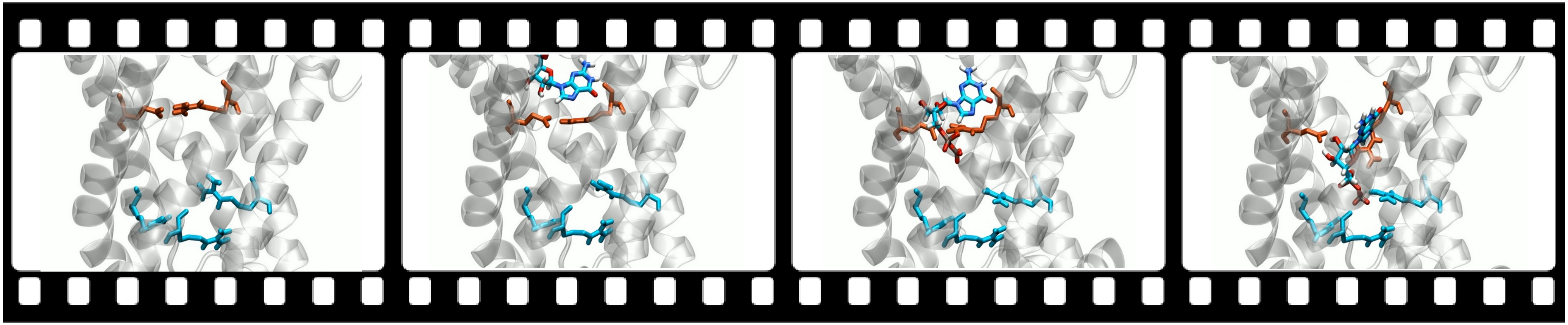
Open Science Days @UGA, Grenoble, 2024

What is molecular dynamics (MD)?



water + detergent

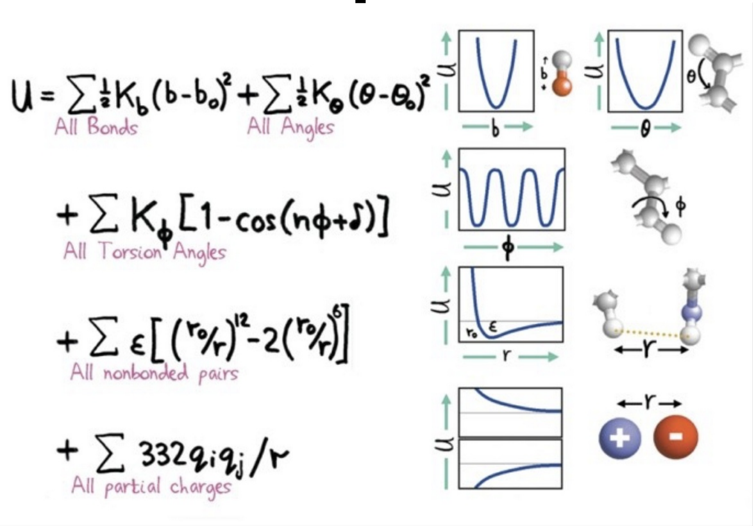
Senac et al, Langmuir, 2017. Movie by Patrick Fuchs



Gagelin et al, Nature communications, 2023.

Molecular dynamics simulations require resources

expertise


$U = \sum \frac{1}{2}K_b(b-b_0)^2 + \sum \frac{1}{2}K_\theta(\theta-\theta_0)^2$
All Bonds
All Angles
$+ \sum K_\phi[1-\cos(n\phi+\delta)]$
All Torsion Angles
$+ \sum \epsilon[(\frac{r_0}{r})^{12} - 2(\frac{r_0}{r})^6]$
All nonbonded pairs
$+ \sum 332 q_i q_j / r$
All partial charges
U
b
θ
φ
r
r_0
ε
+
-

M. Levitt

GROMACS

fast, flexible & free

computer power

Source ; Photothèque CNRS/Cyril Frésillon (droits réservés)

MD simulations require large computational resources → high cost

Source ; Photothèque CNRS/Cyril Frésillon (droits réservés)

GENCI

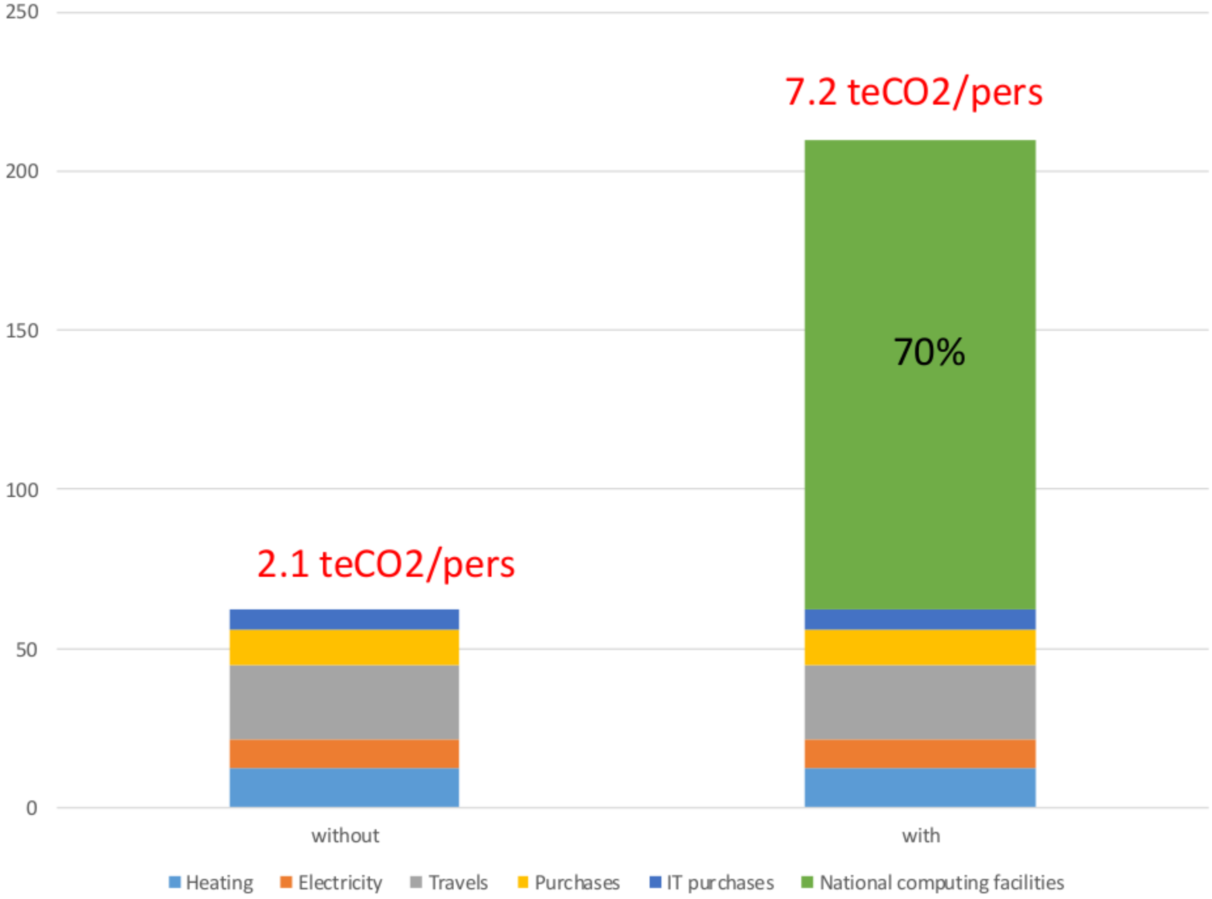
HPC at the service of knowledge

In 2023:

- GENCI, CT 7 (simulation in biology)
- 160 Mh CPU
- 10 Mh GPU
- **Total cost: 8.5 M€**

Patrick Fuchs, 2024.

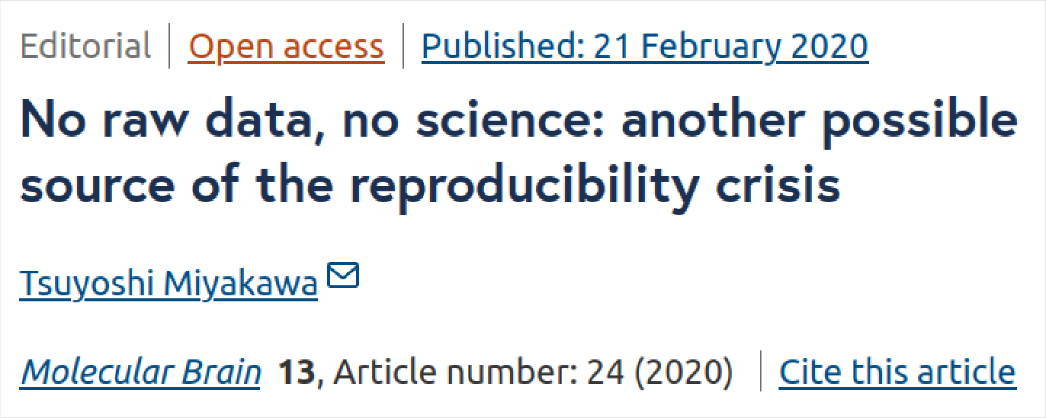
MD simulations require large computational resources → high environmental cost



Élise Duboué-Dijon & Antoine Taly, LBT, 2024.

Sharing research data is important

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

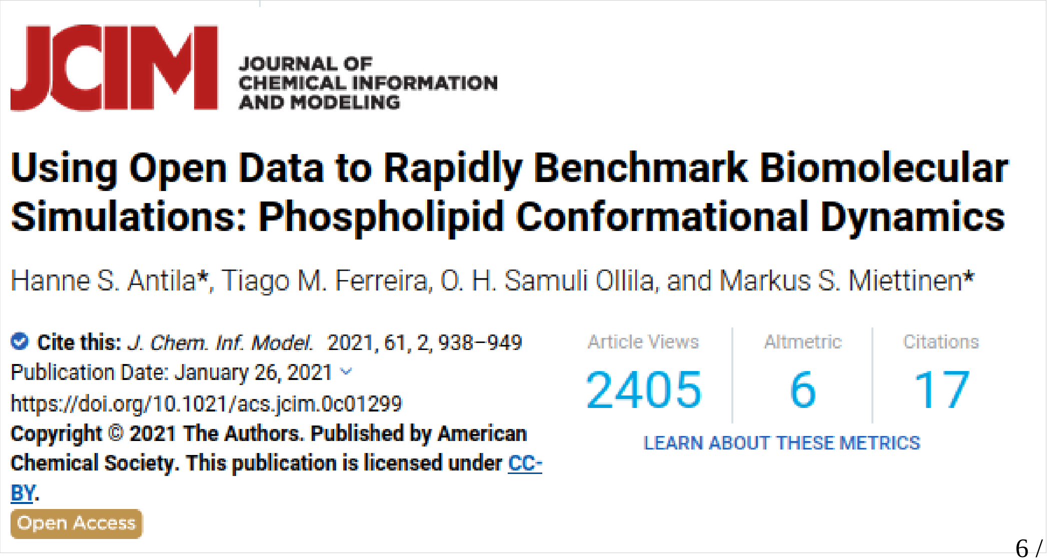


Editorial | Open access | Published: 21 February 2020

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa

Molecular Brain **13**, Article number: 24 (2020) | Cite this article



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Using Open Data to Rapidly Benchmark Biomolecular Simulations: Phospholipid Conformational Dynamics

Hanne S. Antila*, Tiago M. Ferreira, O. H. Samuli Ollila, and Markus S. Miettinen*

Cite this: *J. Chem. Inf. Model.* 2021, 61, 2, 938–949
Publication Date: January 26, 2021
https://doi.org/10.1021/acs.jcim.0c01299
Copyright © 2021 The Authors. Published by American Chemical Society. This publication is licensed under CC-BY.
Open Access

Article Views	Altmetric	Citations
2405	6	17

LEARN ABOUT THESE METRICS

Sharing research data in biology 🎉

Community-approved data repositories for experimental (omics) data:

- SRA, GEO, ENA: genomics & transcriptomics
- PRIDE, MassIVE: proteomics & metabolomics

Sharing MD simulation files

😢

Sharing MD simulation files

- **Many initiatives**: MoDEL, GPCRmd, NMRLipids…

- **No consensus data repository** for MD simulation files (yet)

Sharing MD simulation files

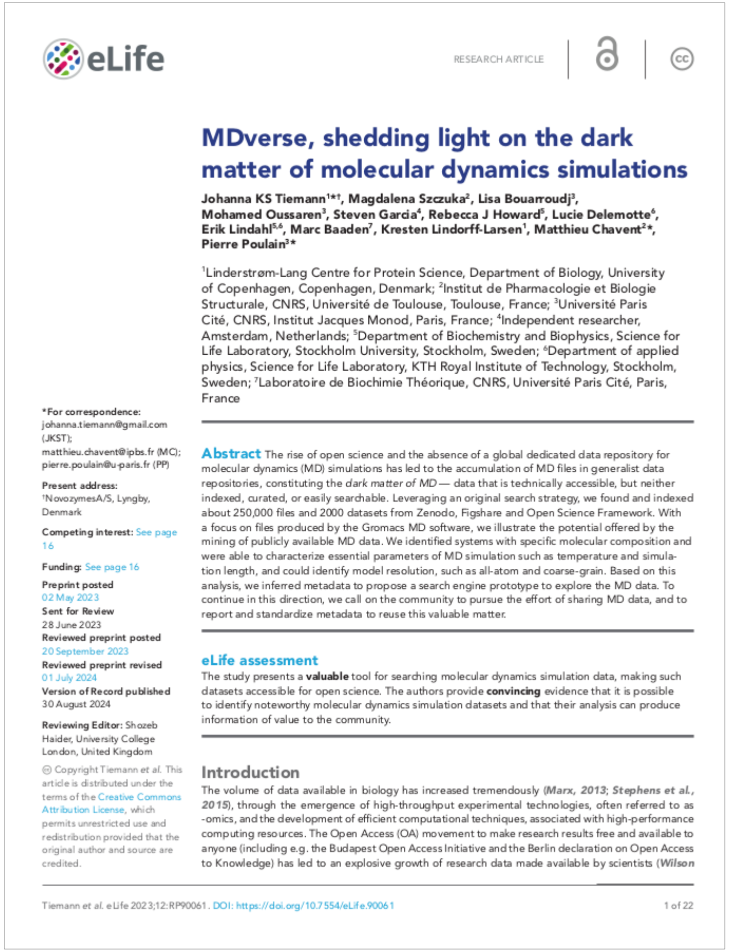
- **Many initiatives**: MoDEL, GPCRmd, NMRLipids…
- **No consensus data repository** for MD simulation files (yet)
- Use of generic **non-moderated** data repositories: Zenodo, Figshare, OSF, Dryad…

Welcome to the Dark Matter of MD

Data that is **technically accessible**,

but neither **indexed**, **curated**,

or easily **searchable**.



eLife RESEARCH ARTICLE

MDverse, shedding light on the dark matter of molecular dynamics simulations

Johanna KS Tiemann[1*†], Magdalena Szczuka[2], Lisa Bouarroudj[3], Mohamed Oussaren[3], Steven Garcia[4], Rebecca J Howard[5], Lucie Delemotte[6], Erik Lindahl[5,6], Marc Baaden[7], Kresten Lindorff-Larsen[1], Matthieu Chavent[2]*, Pierre Poulain[3]*

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [2]Institut de Pharmacologie et Biologie Structurale, CNRS, Université de Toulouse, Toulouse, France; [3]Université Paris Cité, CNRS, Institut Jacques Monod, Paris, France; [4]Independent researcher, Amsterdam, Netherlands; [5]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden; [6]Department of applied physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden; [7]Laboratoire de Biochimie Théorique, CNRS, Université Paris Cité, Paris, France

***For correspondence:** johanna.tiemann@gmail.com (JKST); matthieu.chavent@ipbs.fr (MC); pierre.poulain@u-paris.fr (PP)

Present address: †NovozymesA/S, Lyngby, Denmark

Competing interest: See page 16

Funding: See page 16

Preprint posted
02 May 2023
Sent for Review
28 June 2023
Reviewed preprint posted
20 September 2023
Reviewed preprint revised
01 July 2024
Version of Record published
30 August 2024

Reviewing Editor: Shozeb Haider, University College London, United Kingdom

Abstract The rise of open science and the absence of a global dedicated data repository for molecular dynamics (MD) simulations has led to the accumulation of MD files in generalist data repositories, constituting the *dark matter of MD* — data that is technically accessible, but neither indexed, curated, or easily searchable. Leveraging an original search strategy, we found and indexed about 250,000 files and 2000 datasets from Zenodo, Figshare and Open Science Framework. With a focus on files produced by the Gromacs MD software, we illustrate the potential offered by the mining of publicly available MD data. We identified systems with specific molecular composition and were able to characterize essential parameters of MD simulation such as temperature and simulation length, and could identify model resolution, such as all-atom and coarse-grain. Based on this analysis, we inferred metadata to propose a search engine prototype to explore the MD data. To continue in this direction, we call on the community to pursue the effort of sharing MD data, and to report and standardize metadata to reuse this valuable matter.

eLife assessment

The study presents a **valuable** tool for searching molecular dynamics simulation data, making such datasets accessible for open science. The authors provide **convincing** evidence that it is possible to identify noteworthy molecular dynamics simulation datasets and that their analysis can produce information of value to the community.

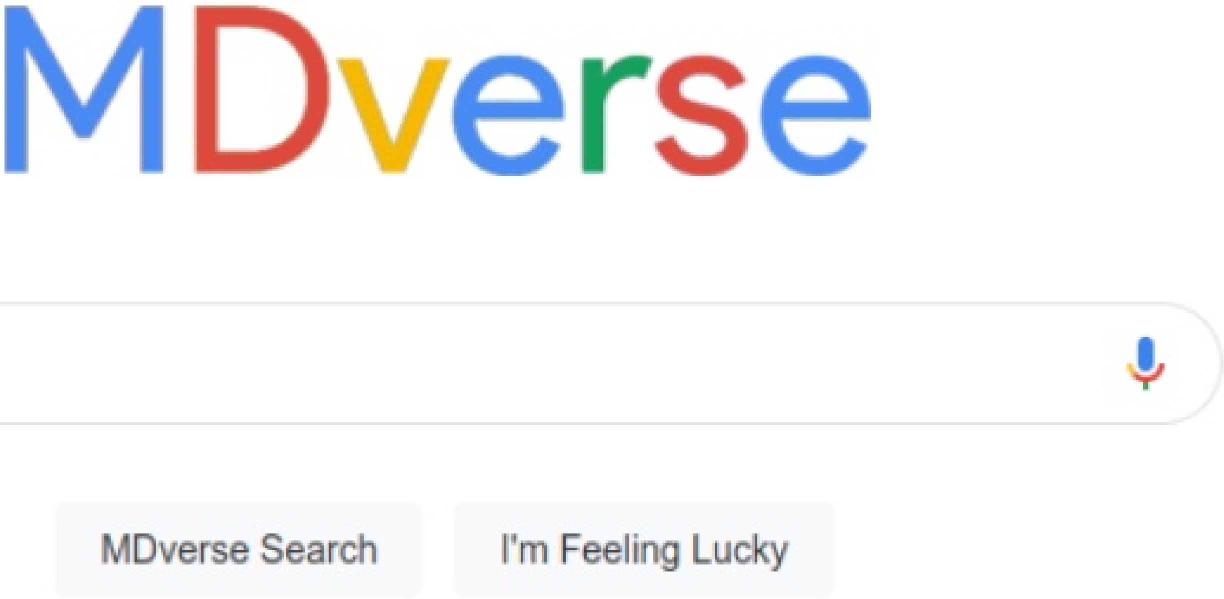
Introduction

The volume of data available in biology has increased tremendously (***Marx, 2013***; ***Stephens et al., 2015***), through the emergence of high-throughput experimental technologies, often referred to as -omics, and the development of efficient computational techniques, associated with high-performance computing resources. The Open Access (OA) movement to make research results free and available to anyone (including e.g. the Budapest Open Access Initiative and the Berlin declaration on Open Access to Knowledge) has led to an explosive growth of research data made available by scientists (***Wilson***

Tiemann *et al.* eLife 2023;12:RP90061. DOI: https://doi.org/10.7554/eLife.90061 1 of 22

DOI 10.7554/eLife.90061.3

A universal search engine for MD open data



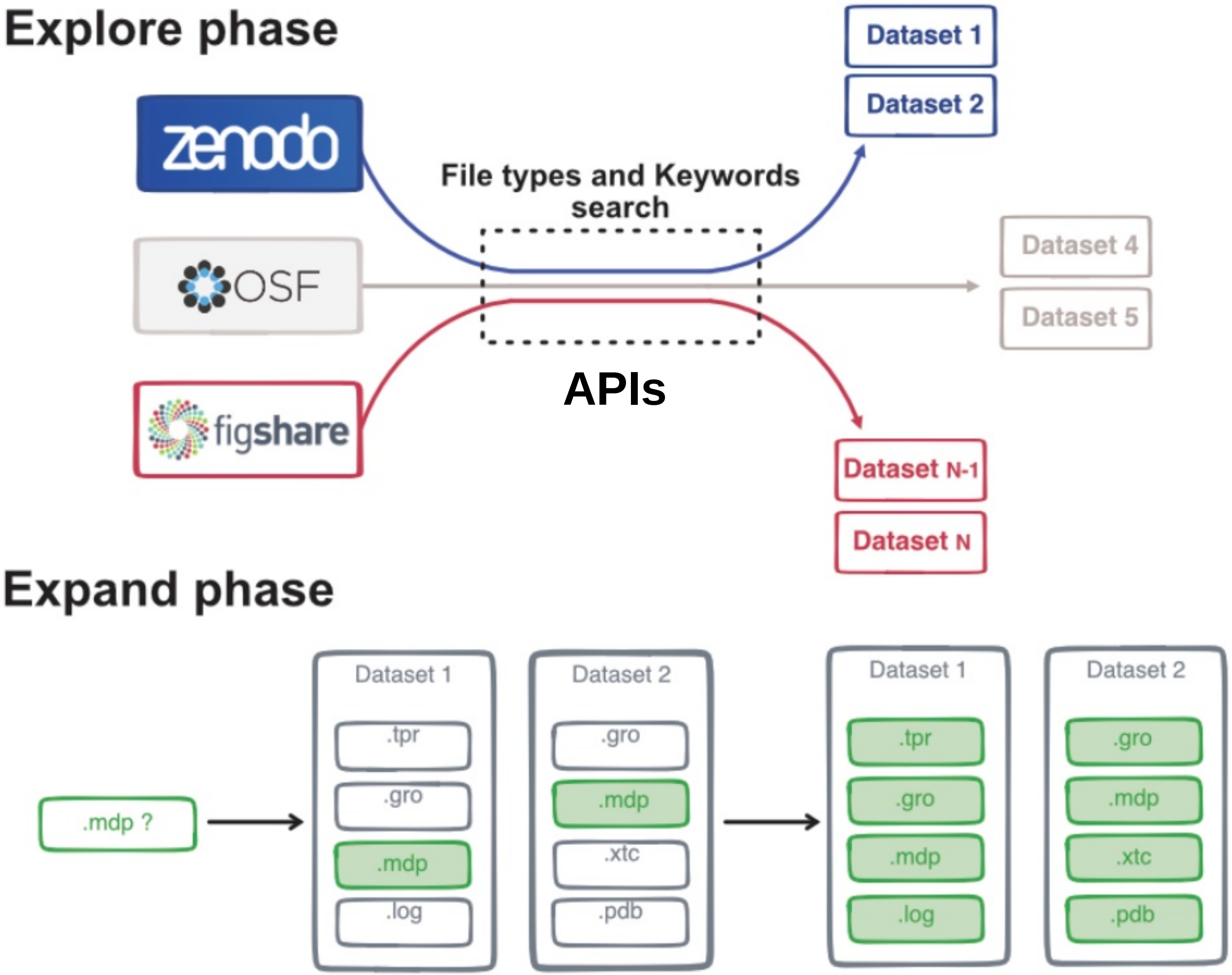
MDverse

1. Find and index **scattered** MD data

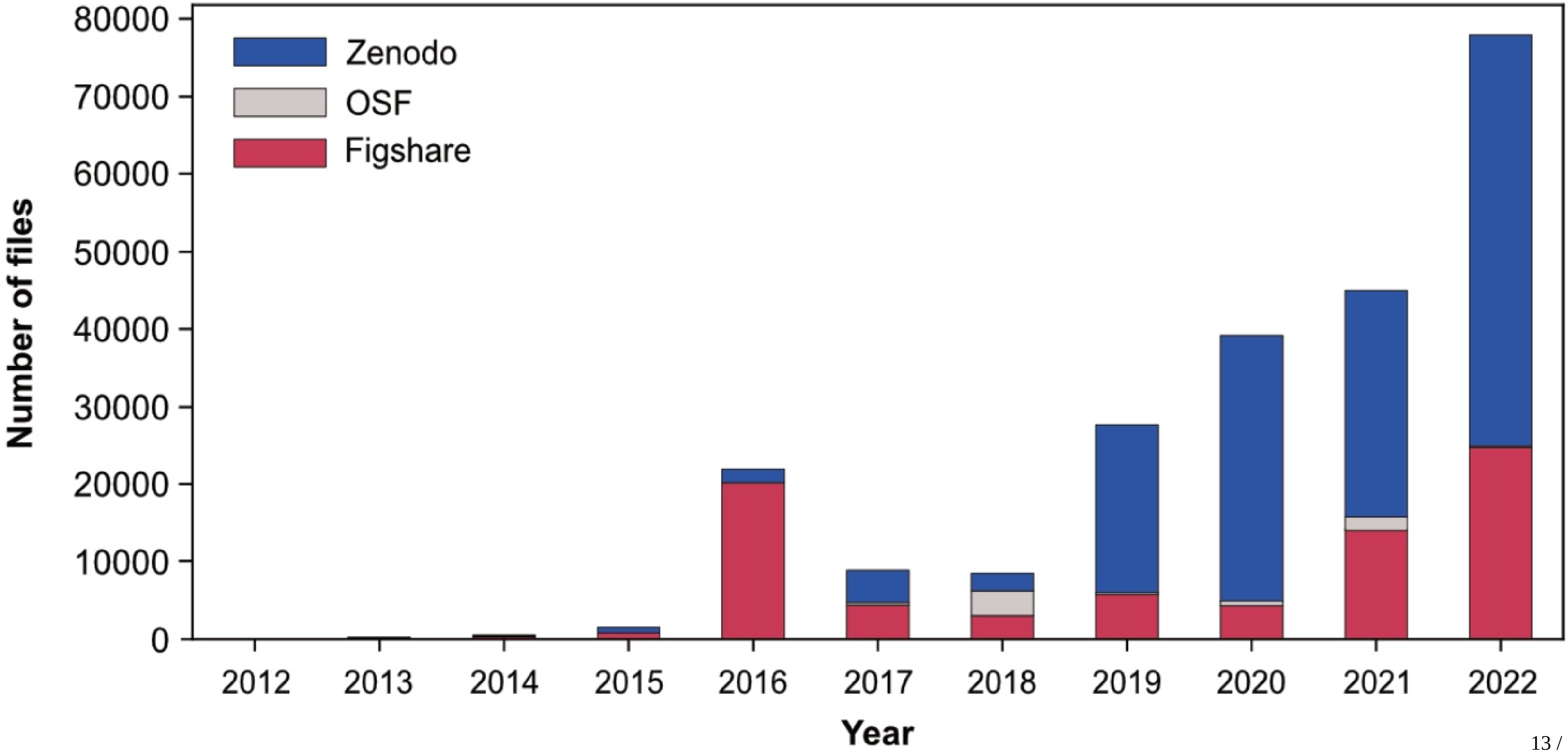
2. Extract, enhance and explore **metadata**

3. Assess **accessibility** and **reusability**
   (from FAIR principles)

Step 1: Find and index scattered MD data

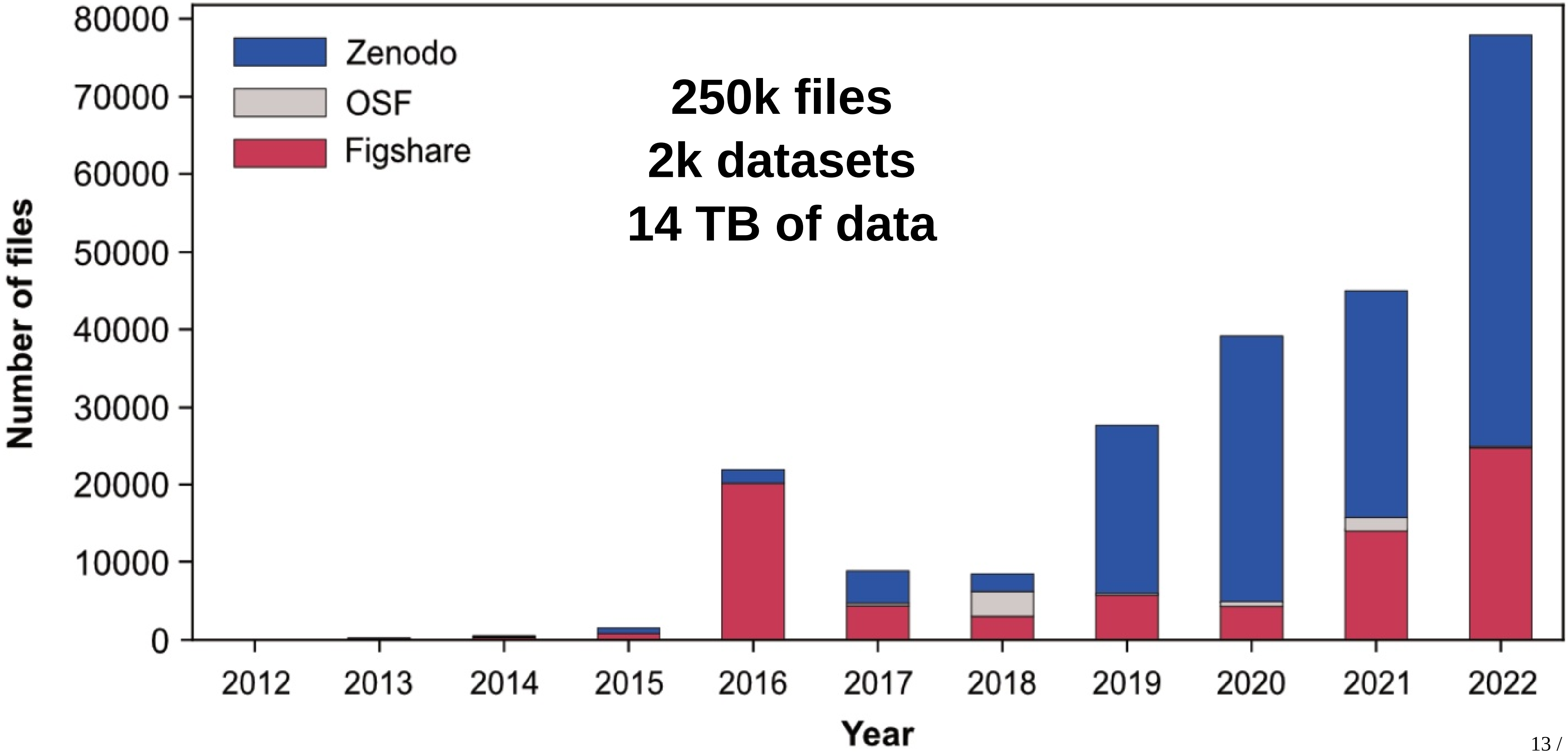


Step 1: Find and index scattered MD data



P.

Step 1: Find and index scattered MD data

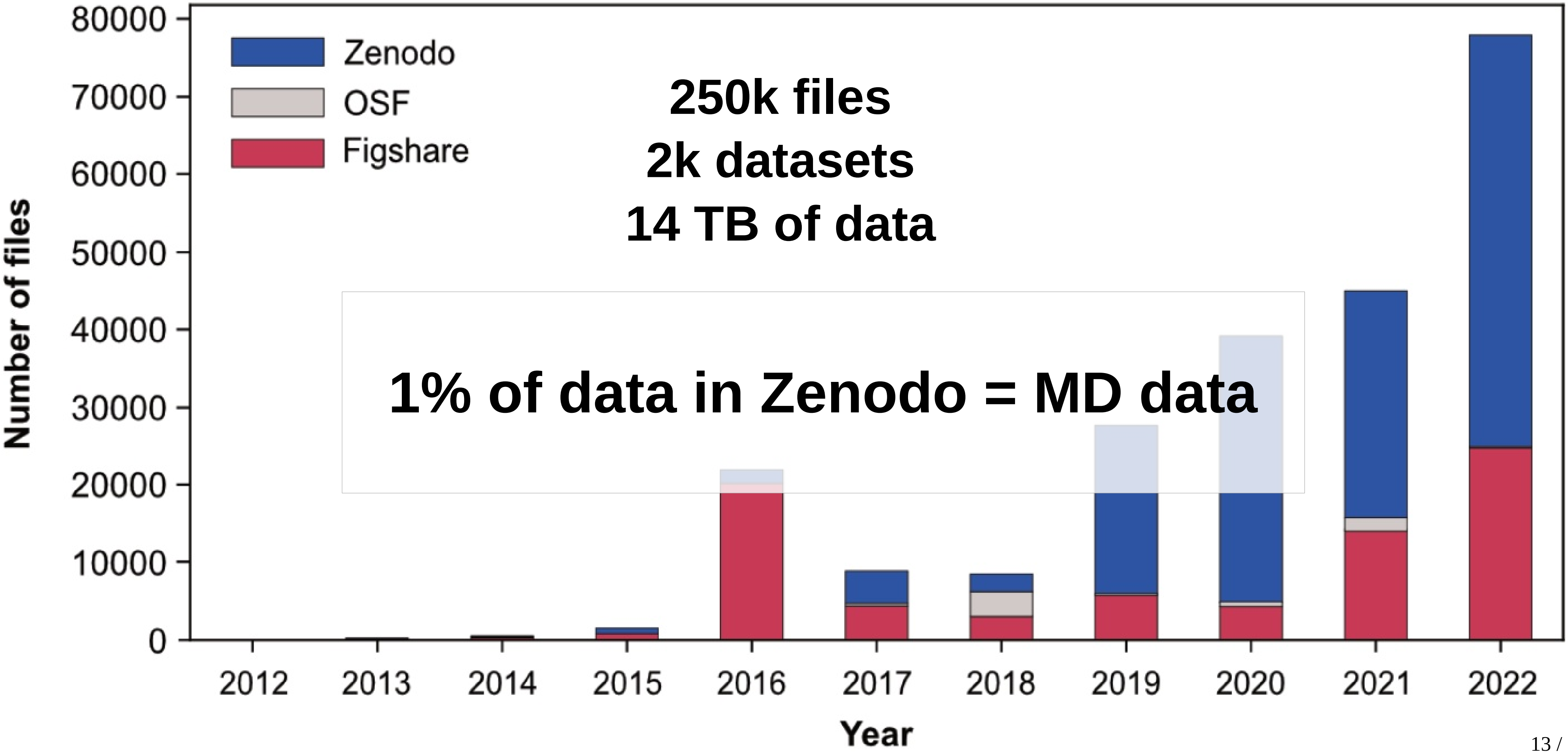


250k files

2k datasets

14 TB of data

Step 1: Find and index scattered MD data



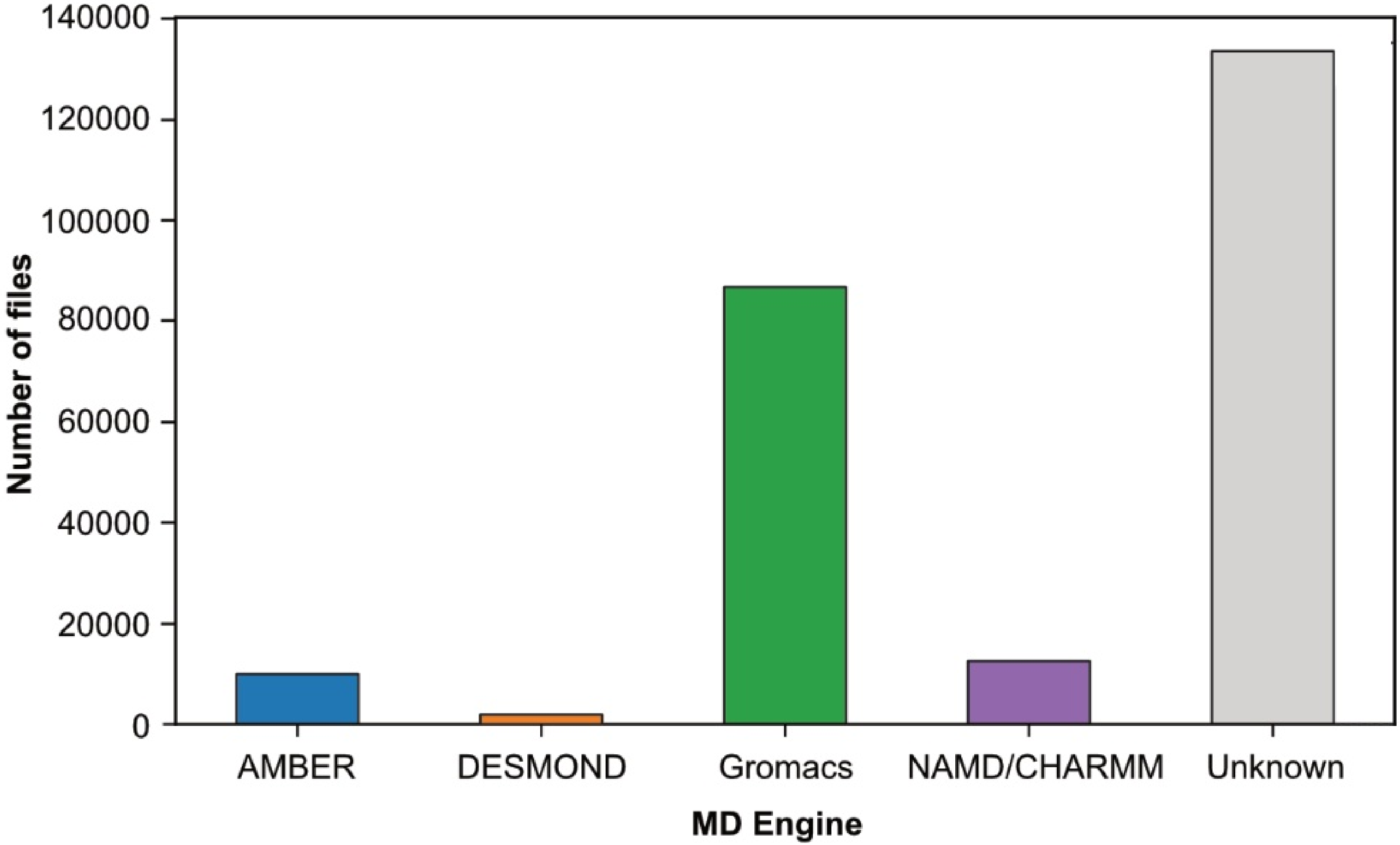
250k files

2k datasets

14 TB of data

1% of data in Zenodo = MD data

Step 1: Find and index scattered MD data

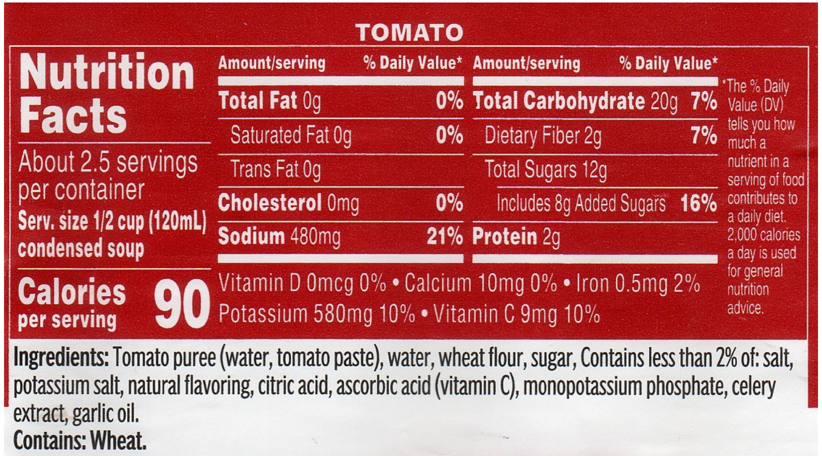


Step 2: Metadata = context

Step 2: Metadata = context

No metadata
(bad)

Natural language metadata (good) 🧑‍🔬



TOMATO

Nutrition Facts

About 2.5 servings per container
Serv. size 1/2 cup (120mL) condensed soup

Calories per serving 90

Amount/serving	% Daily Value*	Amount/serving	% Daily Value*
Total Fat 0g	**0%**	**Total Carbohydrate** 20g	**7%**
Saturated Fat 0g	**0%**	Dietary Fiber 2g	**7%**
Trans Fat 0g		Total Sugars 12g	
Cholesterol 0mg	**0%**	Includes 8g Added Sugars	**16%**
Sodium 480mg	**21%**	**Protein** 2g	

Vitamin D 0mcg 0% • Calcium 10mg 0% • Iron 0.5mg 2%
Potassium 580mg 10% • Vitamin C 9mg 10%

*The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Ingredients: Tomato puree (water, tomato paste), water, wheat flour, sugar, Contains less than 2% of: salt, potassium salt, natural flavoring, citric acid, ascorbic acid (vitamin C), monopotassium phosphate, celery extract, garlic oil.
Contains: Wheat.

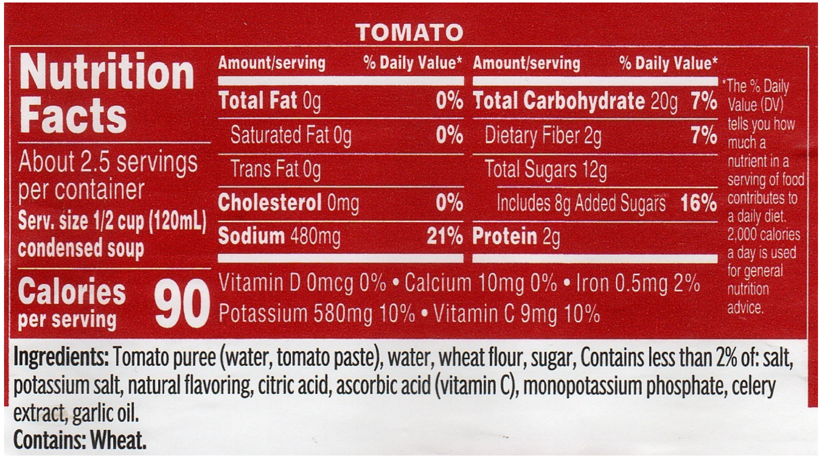
Controlled vocabulary metadata (better) 🤖

Soup can idea from "Metadata for Effective Research Data Management", Chiu et al., 2019.

Step 2: Metadata = context

No metadata
(bad)

Natural language
metadata (good) 🧑‍🔬



TOMATO

Nutrition Facts

About 2.5 servings per container

Serv. size 1/2 cup (120mL) condensed soup

Calories per serving 90

Amount/serving	% Daily Value*	Amount/serving	% Daily Value*
Total Fat 0g	**0%**	**Total Carbohydrate** 20g	**7%**
Saturated Fat 0g	**0%**	Dietary Fiber 2g	**7%**
Trans Fat 0g		Total Sugars 12g	
Cholesterol 0mg	**0%**	Includes 8g Added Sugars	**16%**
Sodium 480mg	**21%**	**Protein** 2g	

Vitamin D 0mcg 0% • Calcium 10mg 0% • Iron 0.5mg 2%
Potassium 580mg 10% • Vitamin C 9mg 10%

*The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

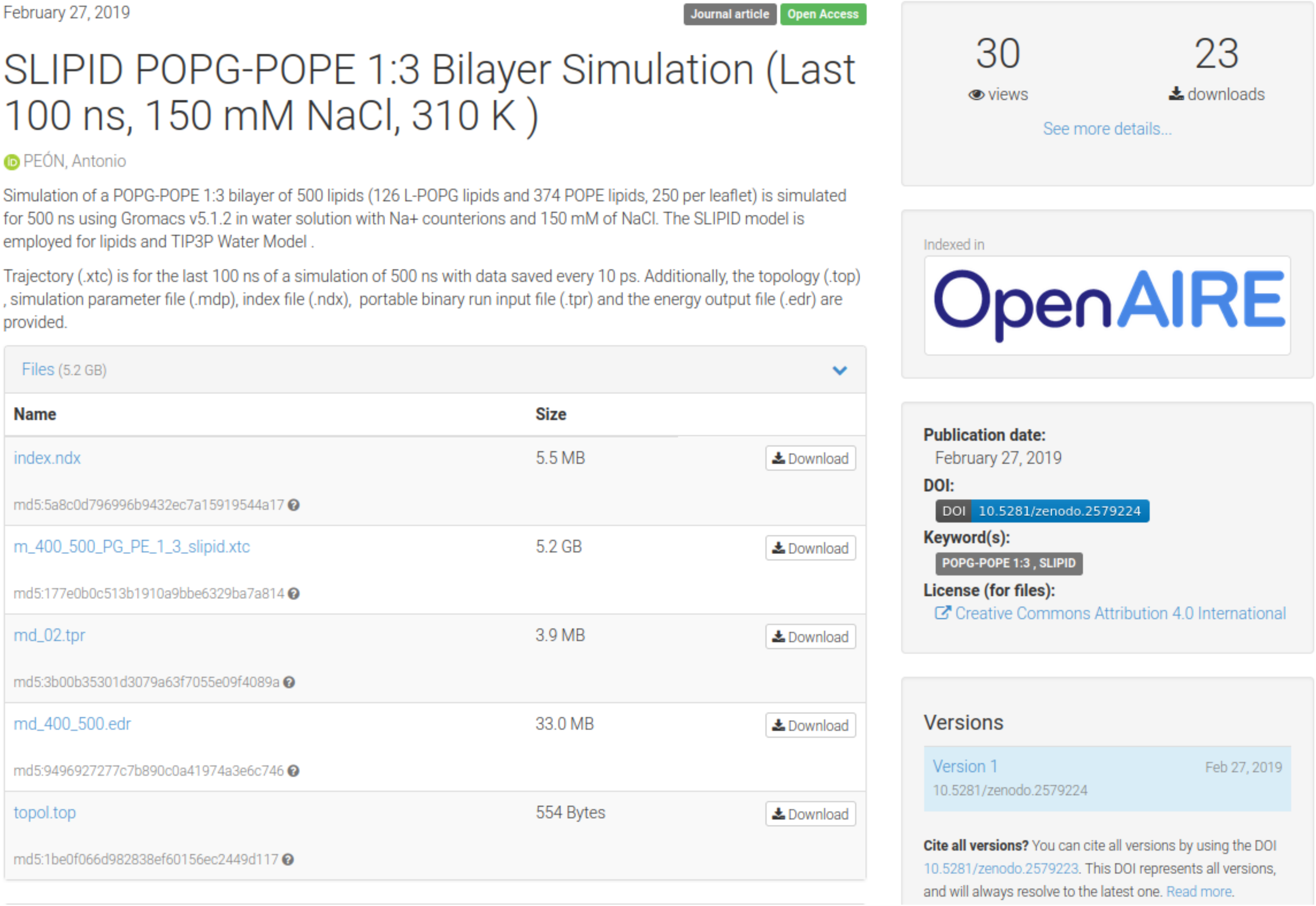
Ingredients: Tomato puree (water, tomato paste), water, wheat flour, sugar, Contains less than 2% of: salt, potassium salt, natural flavoring, citric acid, ascorbic acid (vitamin C), monopotassium phosphate, celery extract, garlic oil.
Contains: Wheat.

Controlled vocabulary
metadata (better) 🤖

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

Soup can idea from “Metadata for Effective Research Data Management”, Chiu et al., 2019.

Step 2: Sources of MD metadata



February 27, 2019

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download
md_02.tpr md5:3b00b35301d3079a63f7055e09f4089a	3.9 MB	Download
md_400_500.edr md5:9496927277c7b890c0a41974a3e6c746	33.0 MB	Download
topol.top md5:1be0f066d982838ef60156ec2449d117	554 Bytes	Download

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

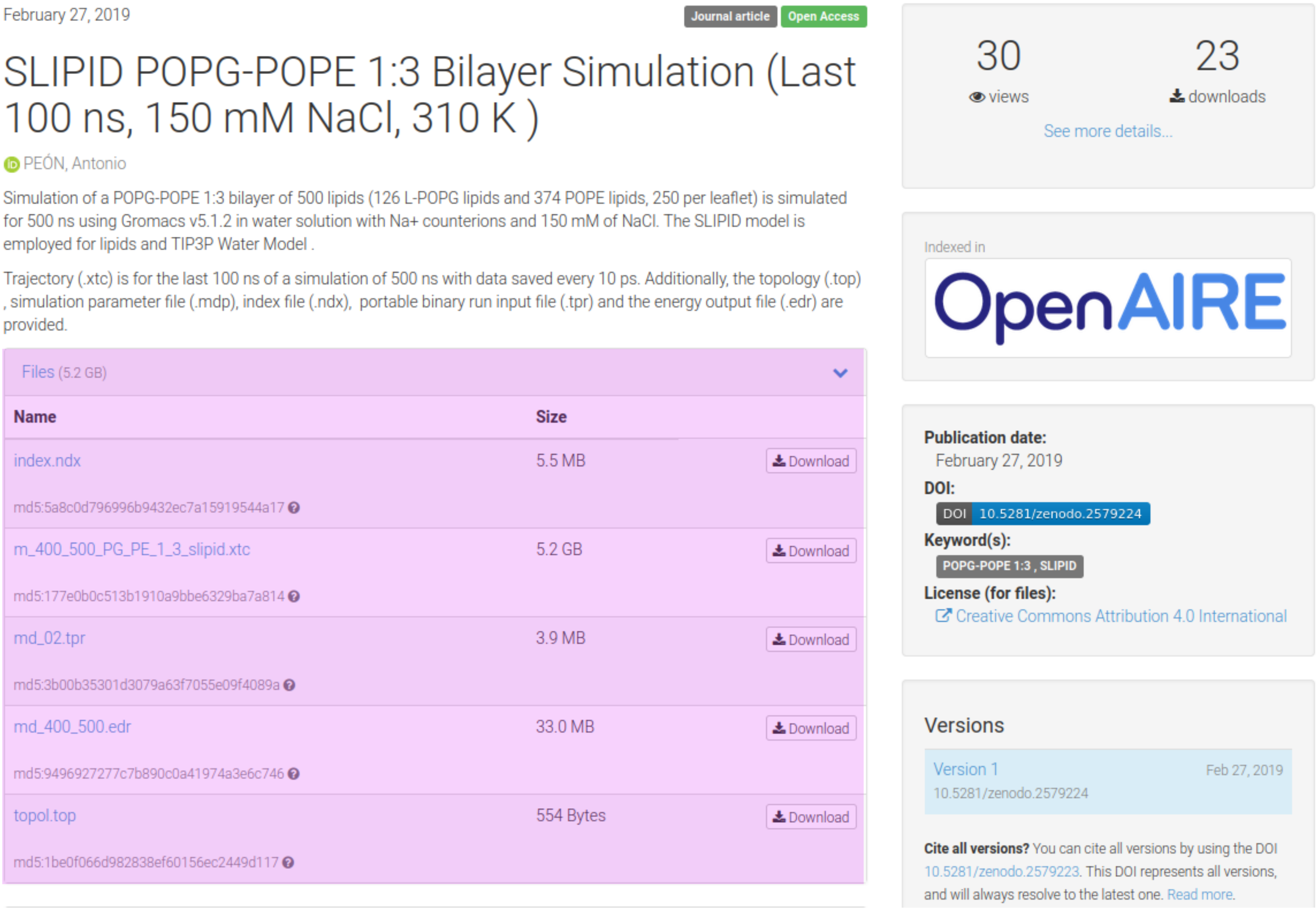
License (for files):
Creative Commons Attribution 4.0 International

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: Sources 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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download
md_02.tpr md5:3b00b35301d3079a63f7055e09f4089a	3.9 MB	Download
md_400_500.edr md5:9496927277c7b890c0a41974a3e6c746	33.0 MB	Download
topol.top md5:1be0f066d982838ef60156ec2449d117	554 Bytes	Download

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

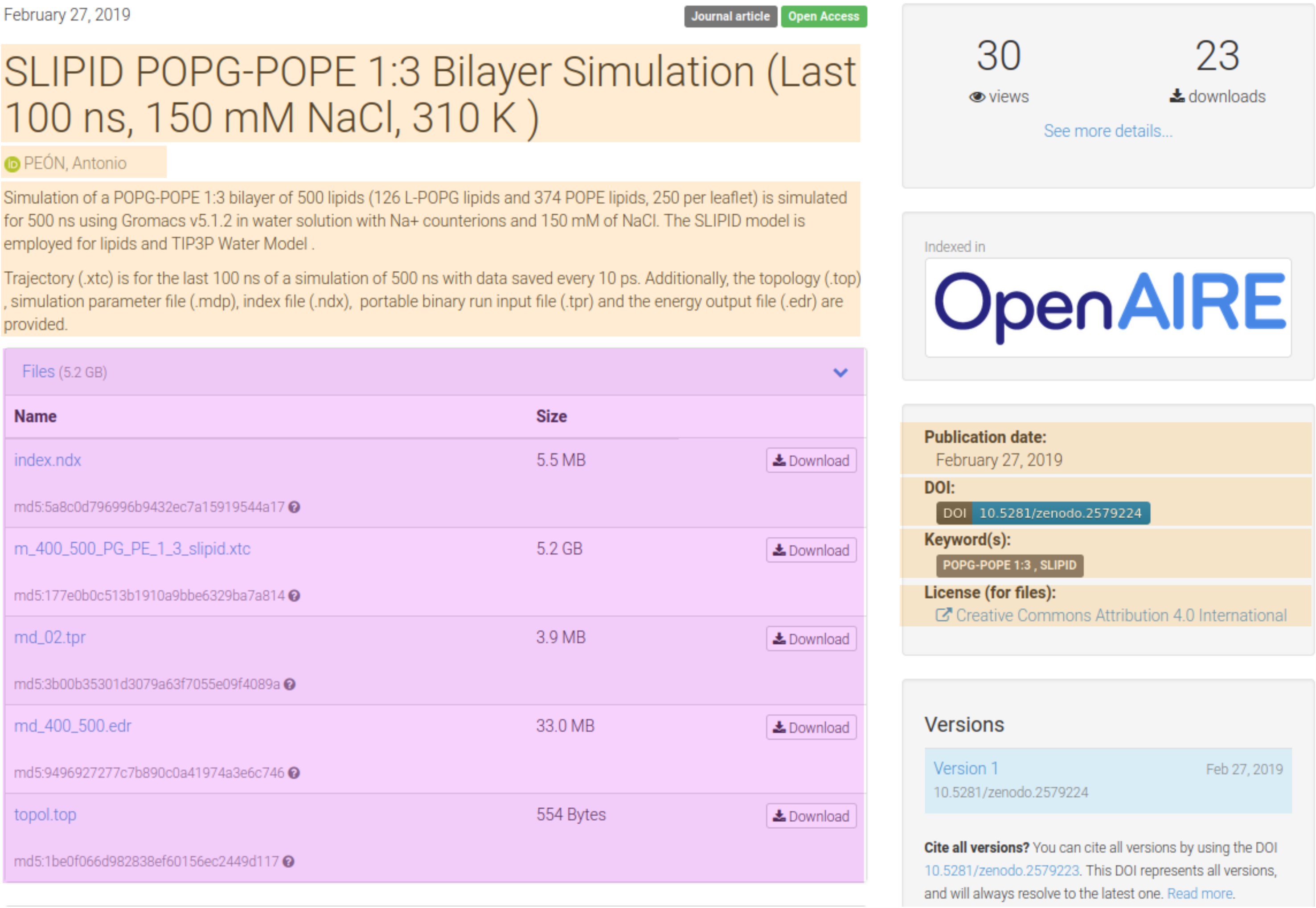
License (for files):
Creative Commons Attribution 4.0 International

Versions

Version 1 10.5281/zenodo.2579224	Feb 27, 2019

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: Sources of MD metadata



February 27, 2019

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download
md_02.tpr md5:3b00b35301d3079a63f7055e09f4089a	3.9 MB	Download
md_400_500.edr md5:9496927277c7b890c0a41974a3e6c746	33.0 MB	Download
topol.top md5:1be0f066d982838ef60156ec2449d117	554 Bytes	Download

30 views

23 downloads

See more details...

Indexed in

Publication date:
February 27, 2019

DOI:
DOI 10.5281/zenodo.2579224

Keyword(s):
POPG-POPE 1:3 , SLIPID

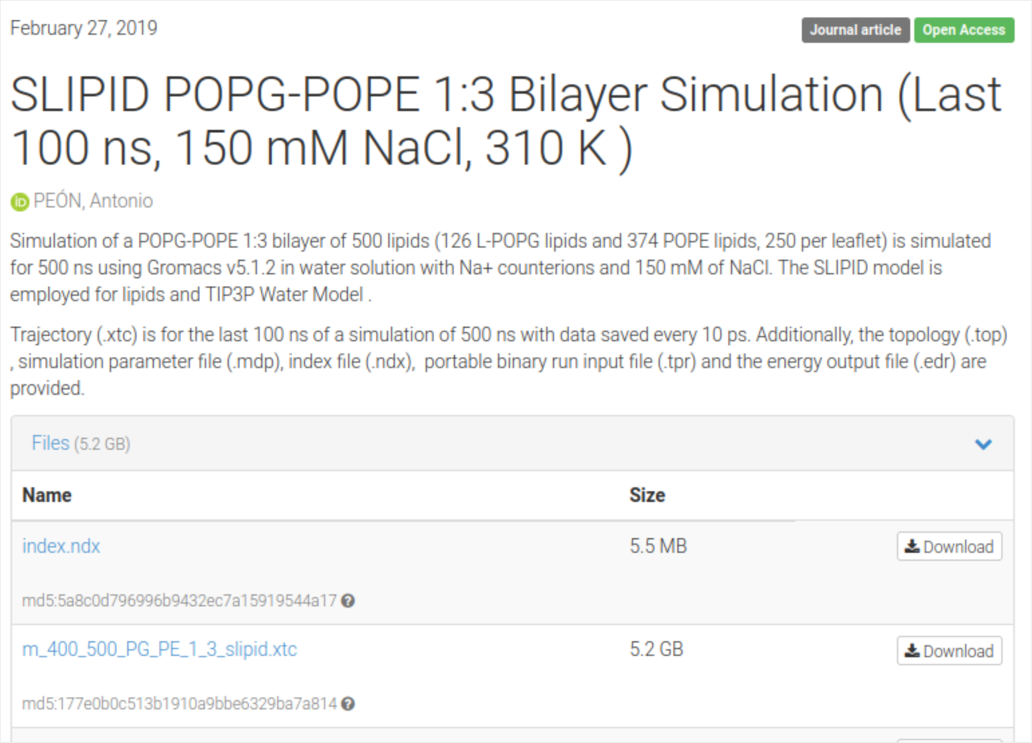
License (for files):
Creative Commons Attribution 4.0 International

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

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

PEÓN, Antonio

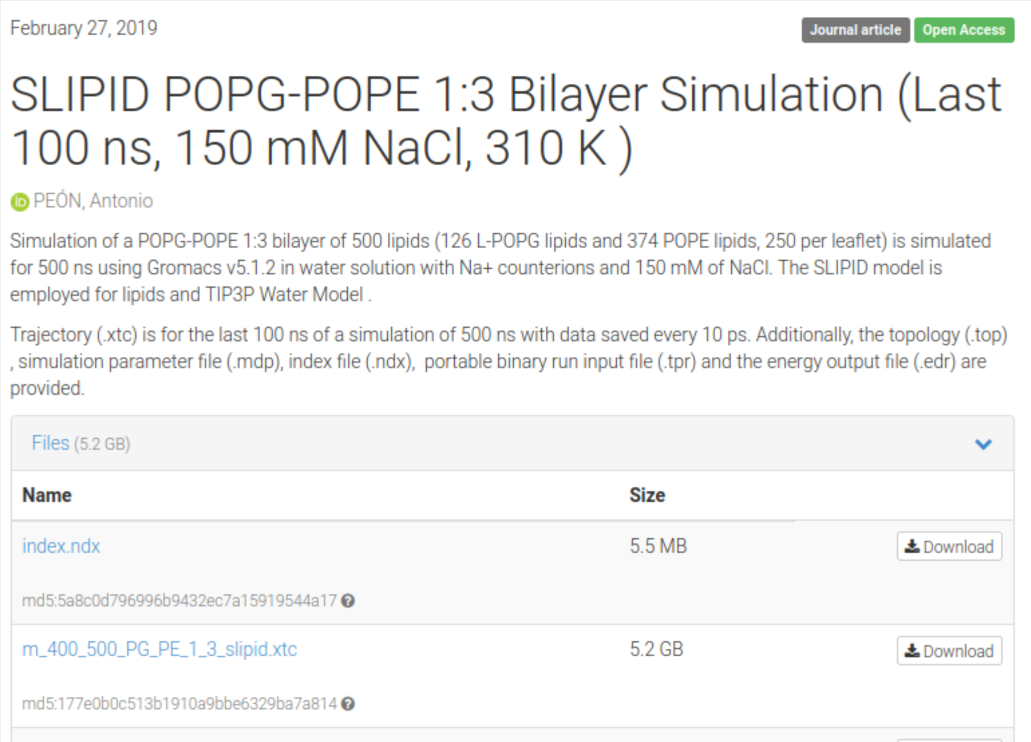
Simulation of a POPG-POPE 1:3 bilayer of 500 lipids (126 L-POPG lipids and 374 POPE lipids, 250 per leaflet) is simulated for 500 ns using Gromacs v5.1.2 in water solution with Na+ counterions and 150 mM of NaCl. The SLIPID model is employed for lipids and TIP3P Water Model .

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download

Step 2: Uneven amount of metadata



February 27, 2019

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

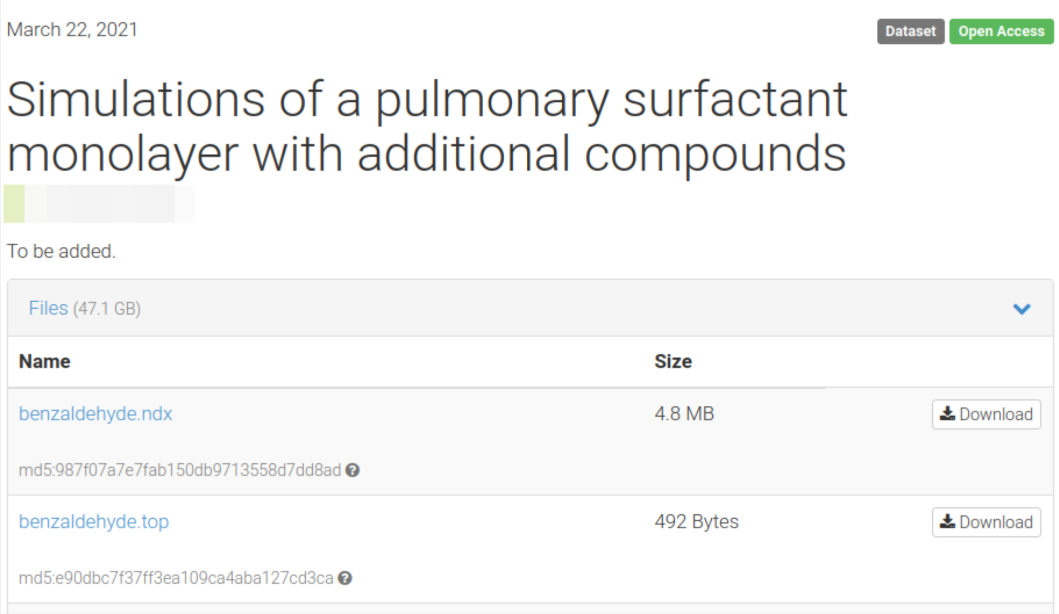
PEÓN, Antonio

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download



March 22, 2021

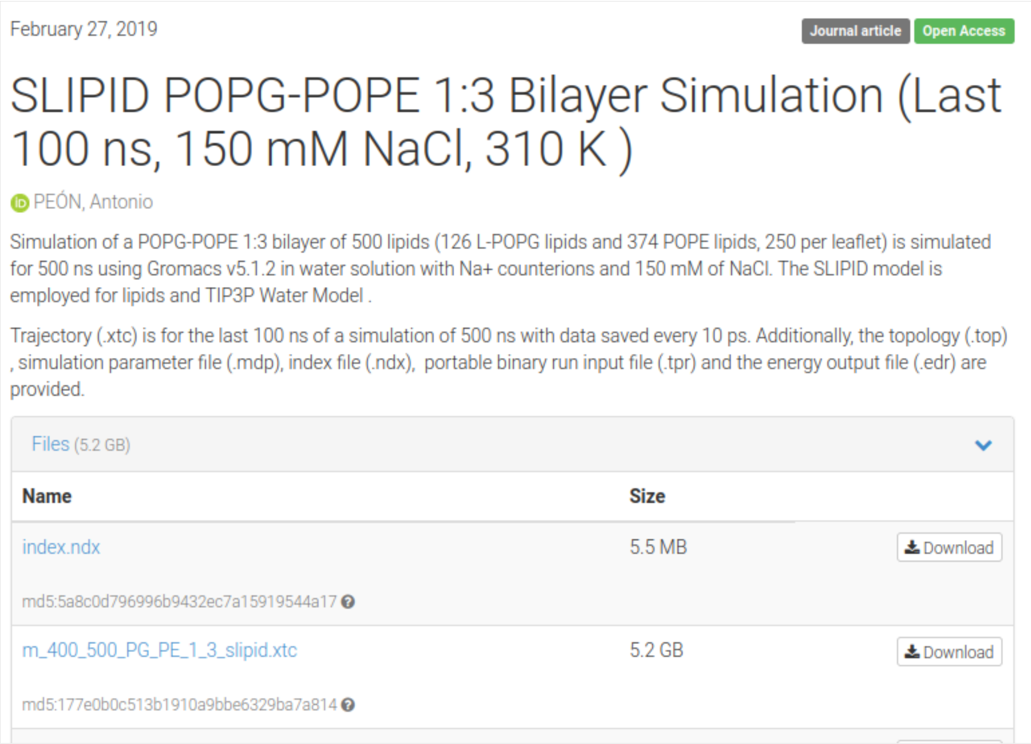
Simulations of a pulmonary surfactant monolayer with additional compounds

To be added.

Files (47.1 GB)

Name	Size	
benzaldehyde.ndx md5:987f07a7e7fab150db9713558d7dd8ad	4.8 MB	Download
benzaldehyde.top md5:e90dbc7f37ff3ea109ca4aba127cd3ca	492 Bytes	Download

Step 2: Uneven amount of metadata



February 27, 2019

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

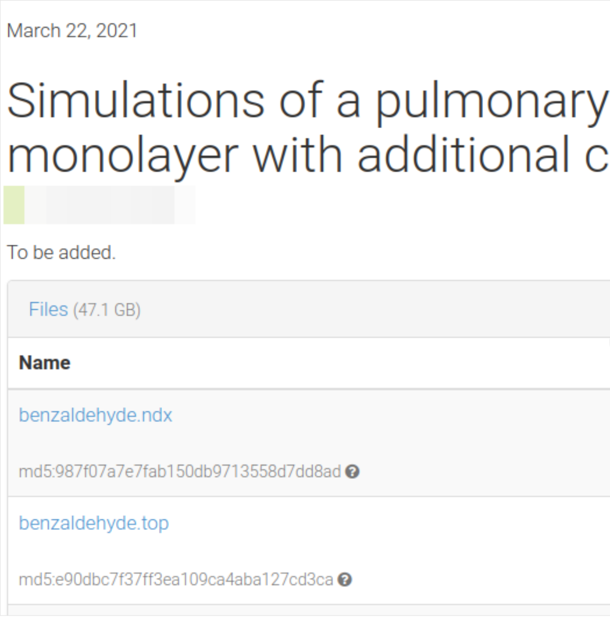
PEÓN, Antonio

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download



March 22, 2021

Simulations of a pulmonary surfactant monolayer with additional compounds

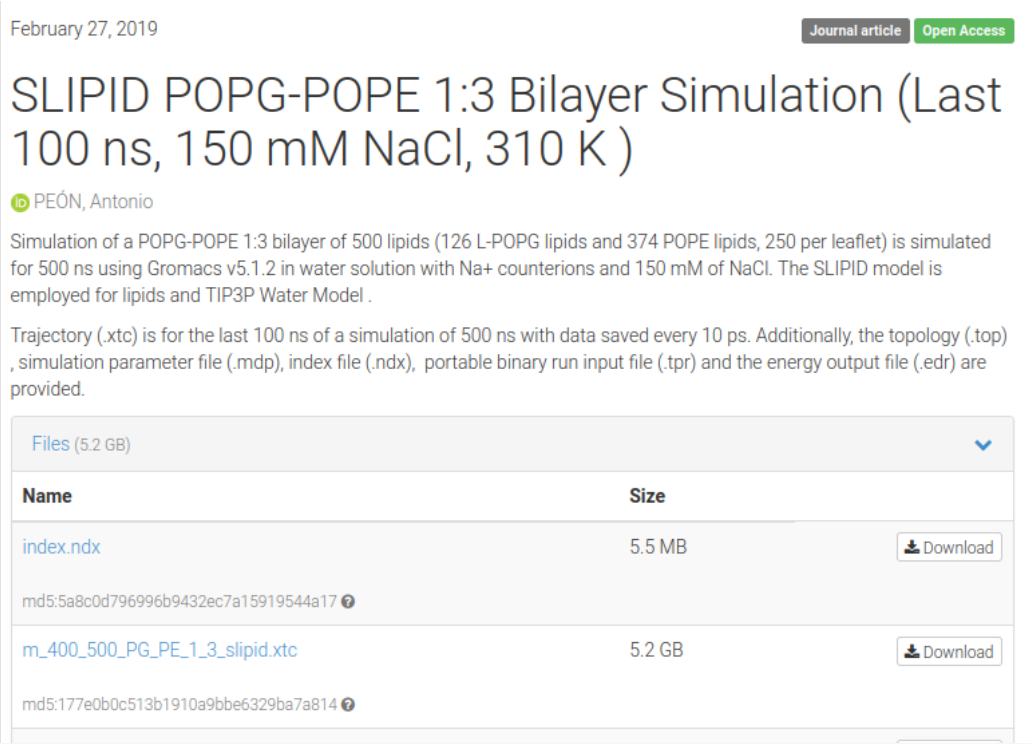
[redacted]

To be added.

Files (47.1 GB)

Name
benzaldehyde.ndx md5:987f07a7e7fab150db9713558d7dd8ad
benzaldehyde.top md5:e90dbc7f37ff3ea109ca4aba127cd3ca

Step 2: Uneven amount of metadata



February 27, 2019

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

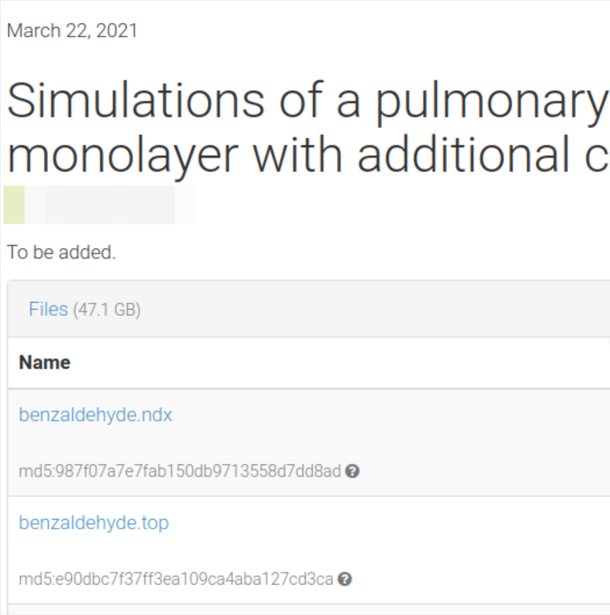
PEÓN, Antonio

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 md5:5a8c0d796996b9432ec7a15919544a17	5.5 MB	Download
m_400_500_PG_PE_1_3_slipid.xtc md5:177e0b0c513b1910a9bbe6329ba7a814	5.2 GB	Download



March 22, 2021

Simulations of a pulmonary surfactant monolayer with additional compounds

To be added.

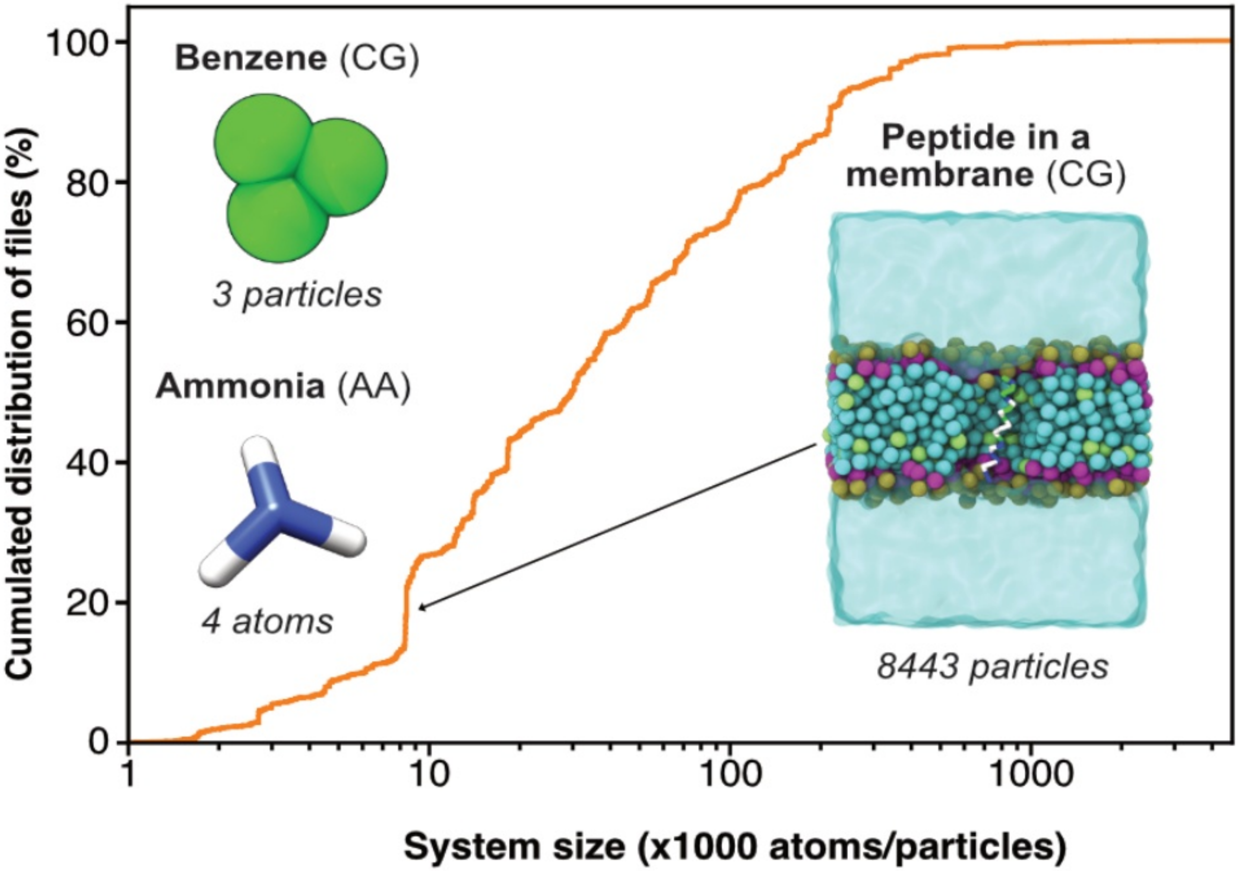
Files (47.1 GB)

Name
benzaldehyde.ndx md5:987f07a7e7fab150db9713558d7dd8ad
benzaldehyde.top md5:e90dbc7f37ff3ea109ca4aba127cd3ca

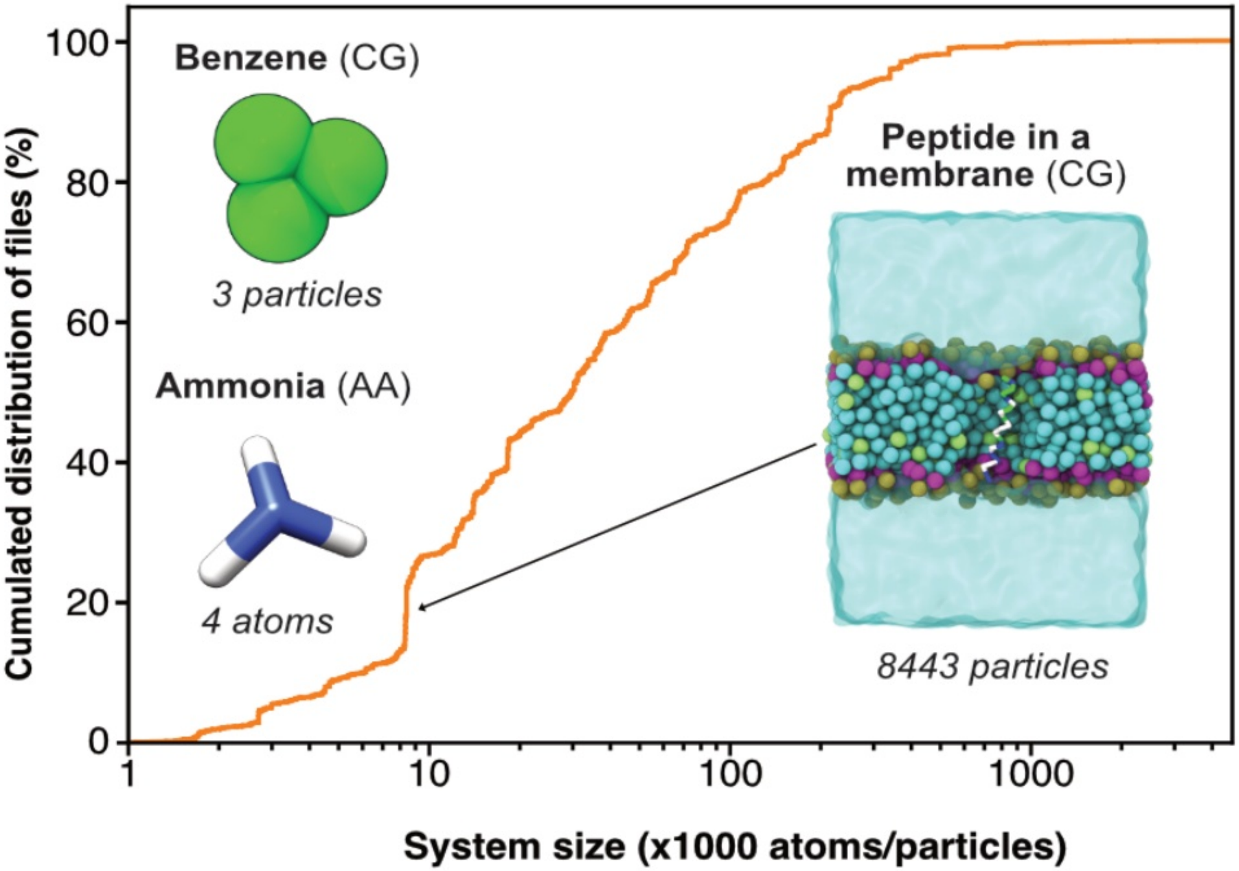
~~Open the can and taste the soup~~

Open the file and extract metadata

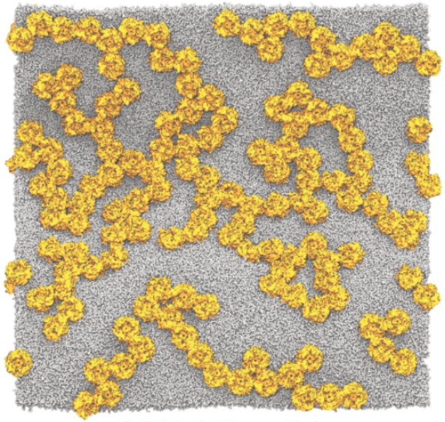
Step 2: Metadata from Gromacs (.gro) files



Step 2: Metadata from Gromacs (.gro) files

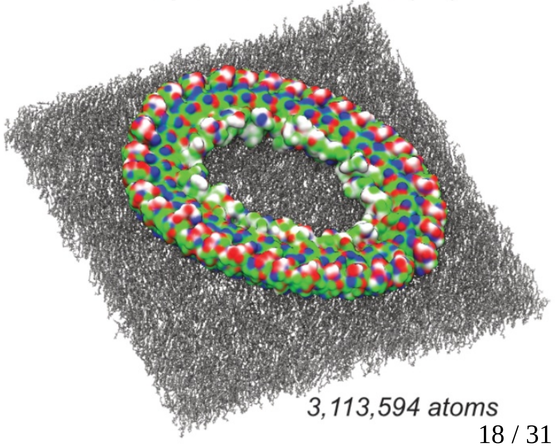


Kir Channels in a plasma membrane (CG)



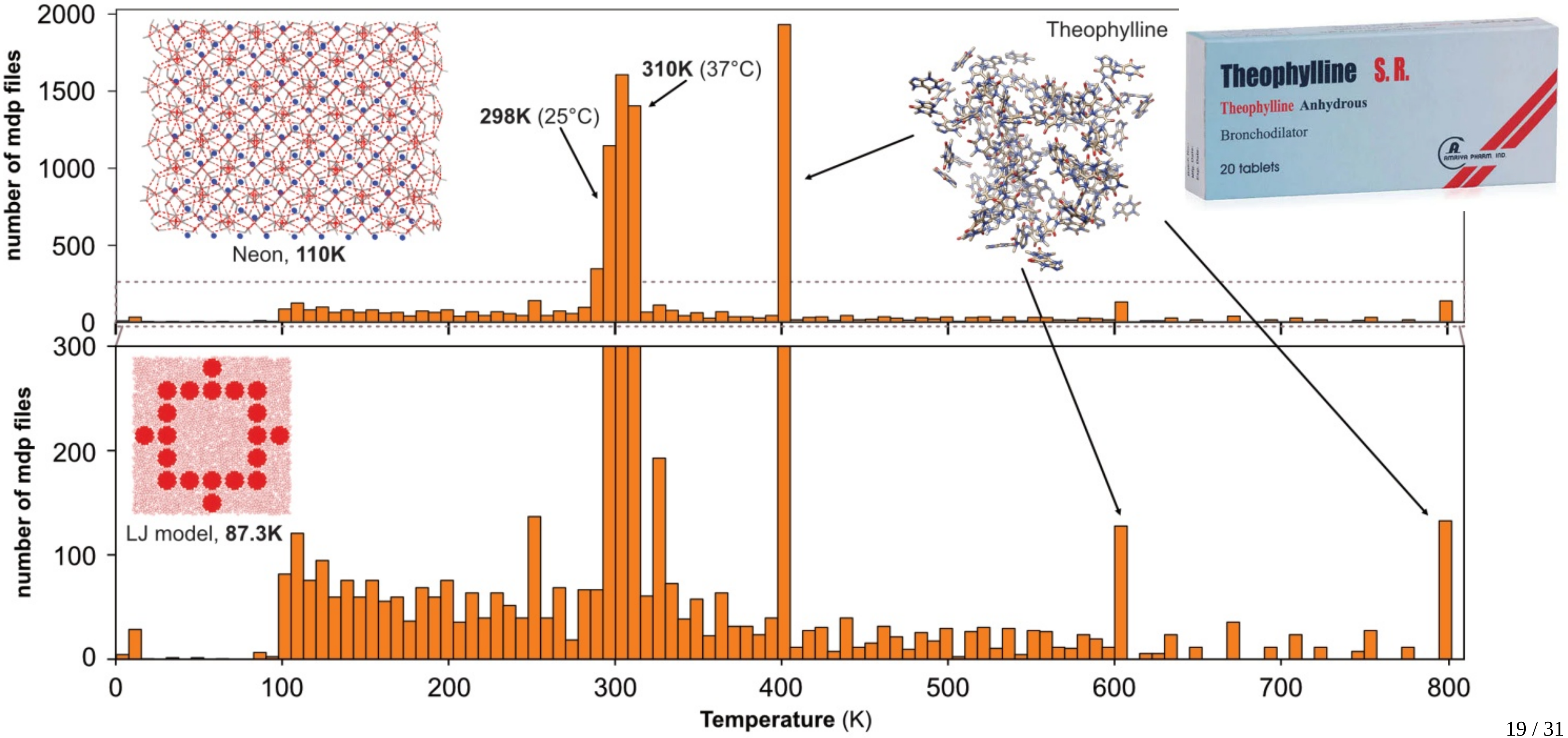
3,522,816 particles

Gasdermin prepore in a plasma membrane (AA)

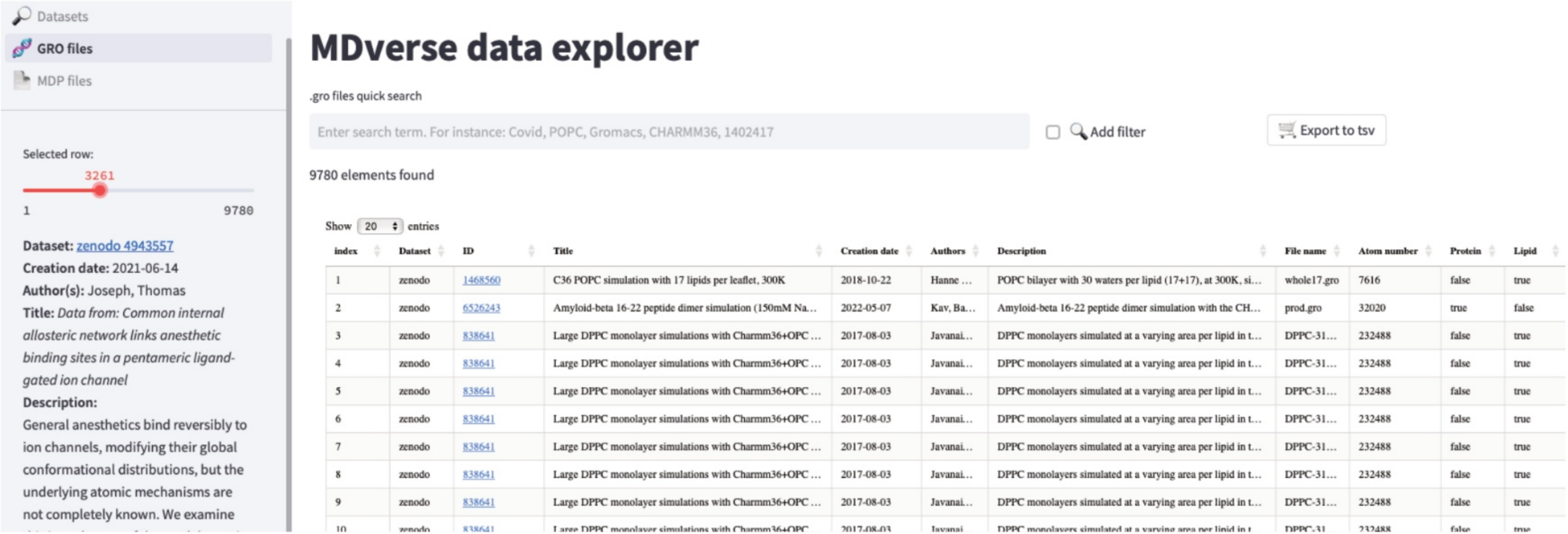


3,113,594 atoms

Step 2: Metadata from Gromacs (.mdp) files



Step 2: Explore metadata 🔍



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

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

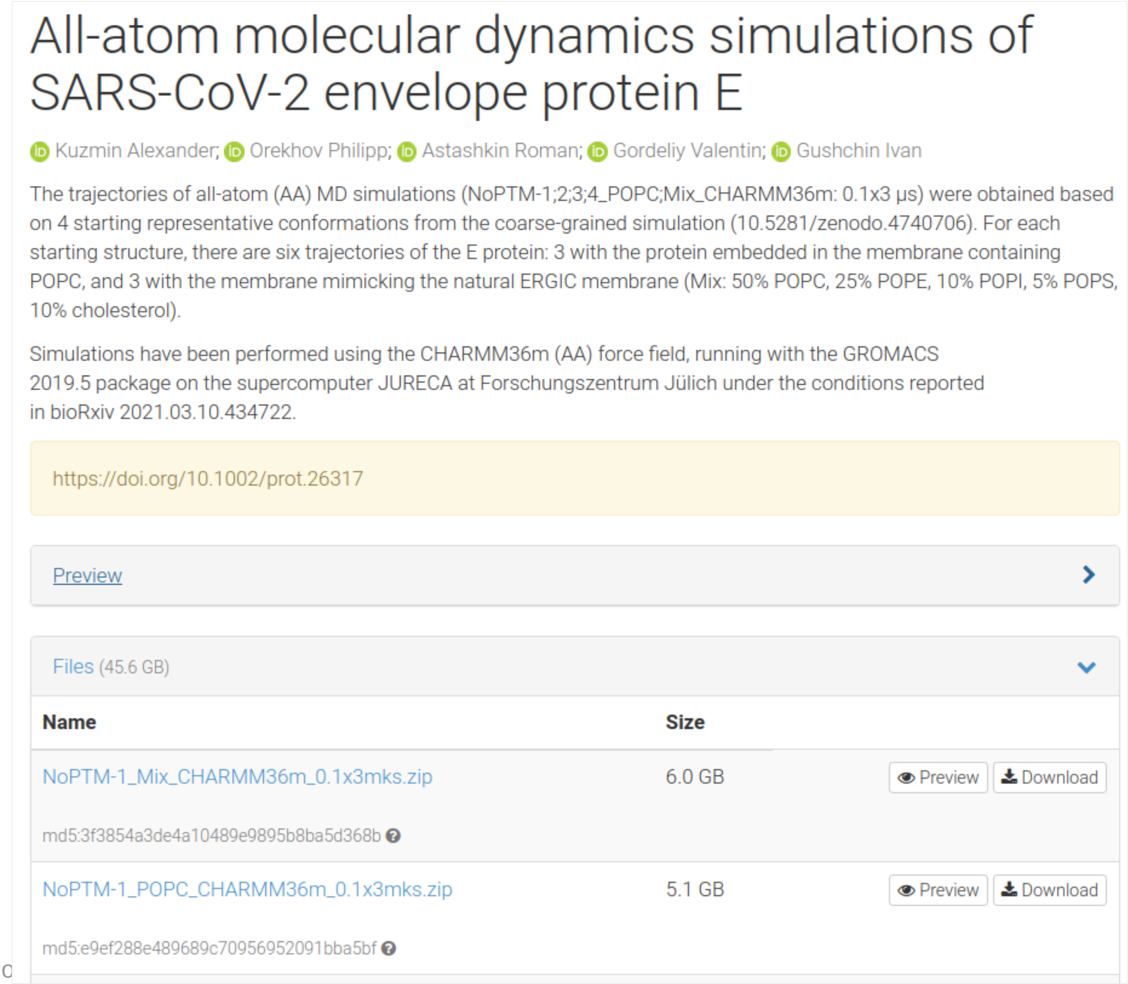
Step 3: Back to the FAIR principles

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 μs) 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.

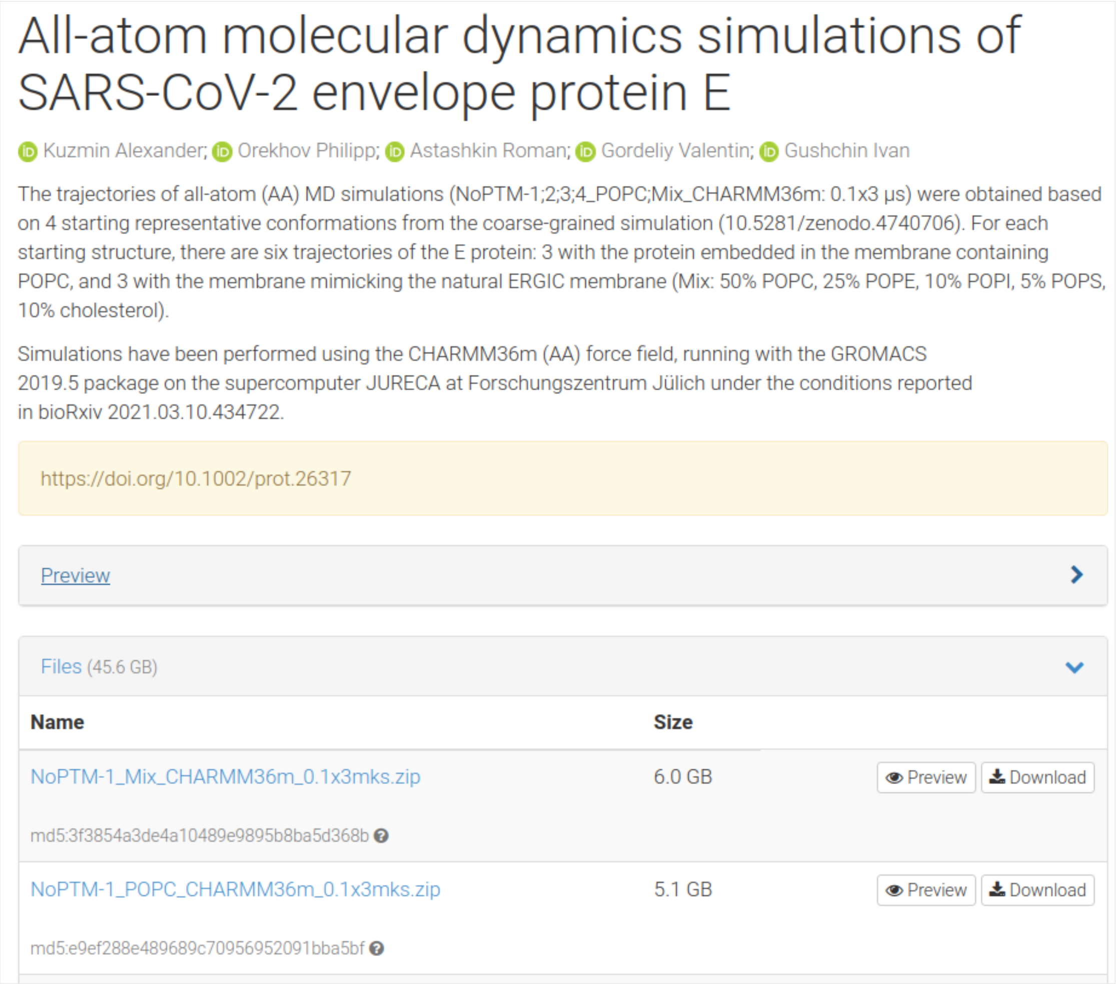
https://doi.org/10.1002/prot.26317

Preview

Files (45.6 GB)

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

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 μs) 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.

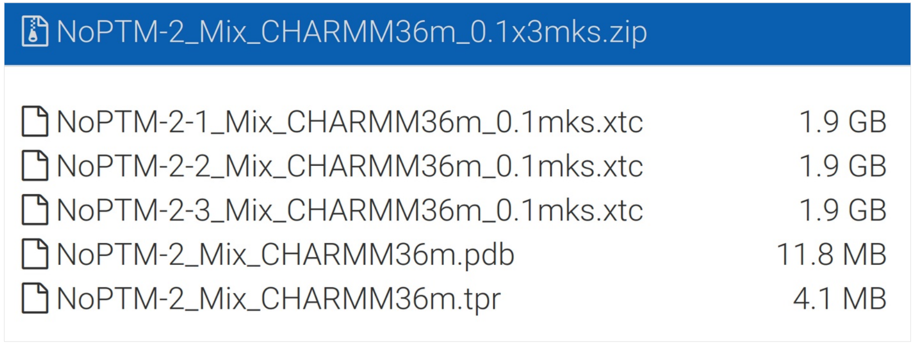
https://doi.org/10.1002/prot.26317

Preview

Files (45.6 GB)

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

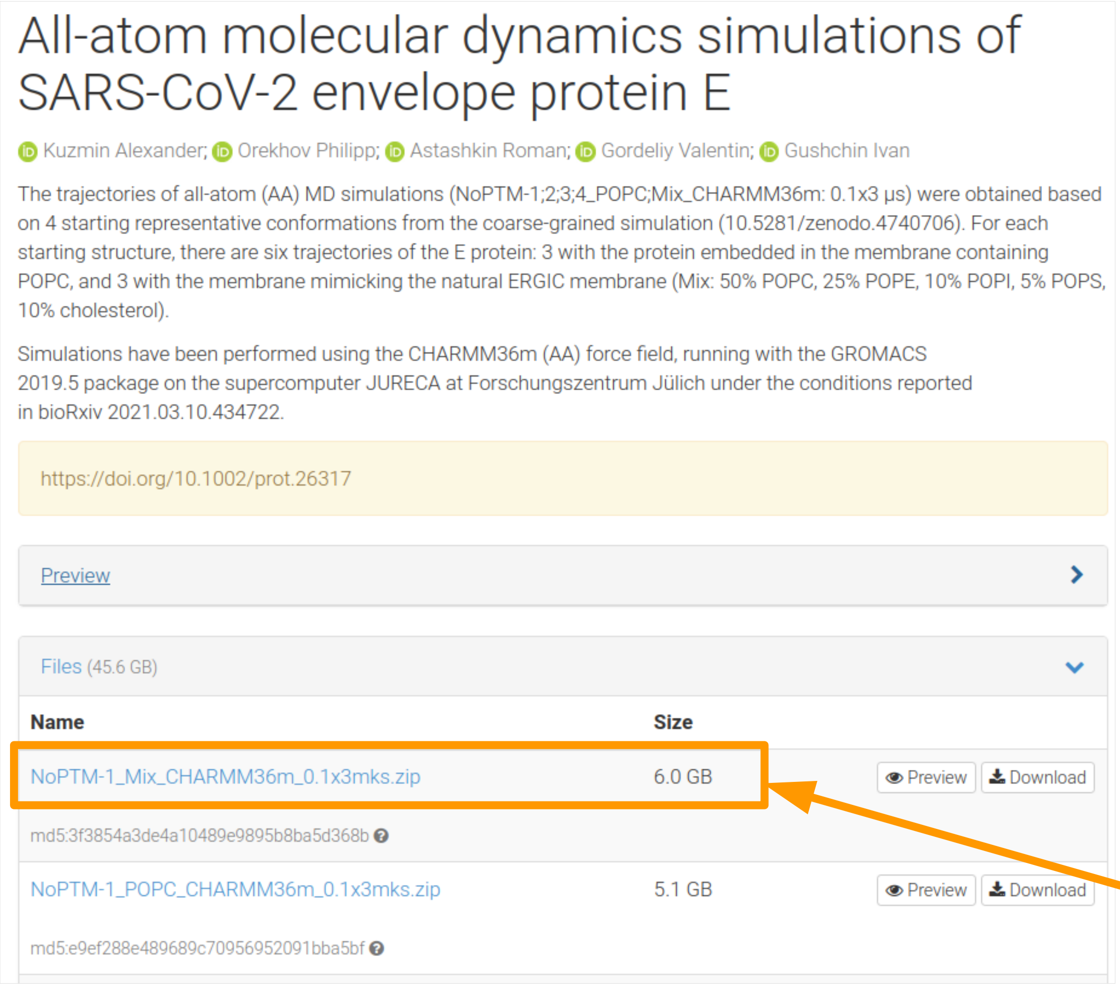
88 % of indexed files were in zip files



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 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 μs) 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.

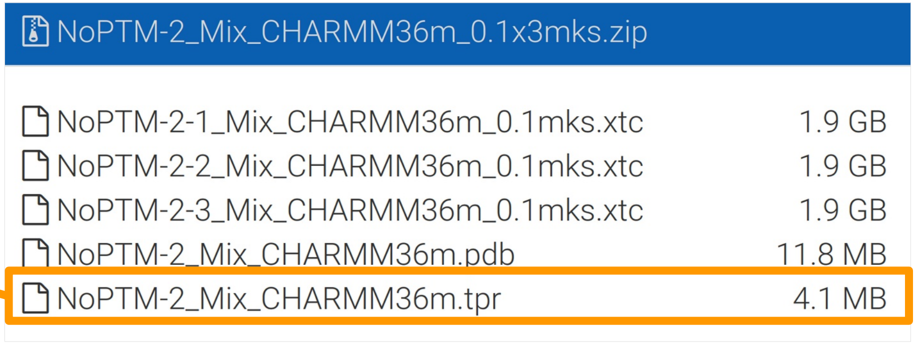
https://doi.org/10.1002/prot.26317

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:

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

Step 3: Assess reusability

A Gromacs trajectory file could be reused to:

- Analyze 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?

MD data sharing highlights

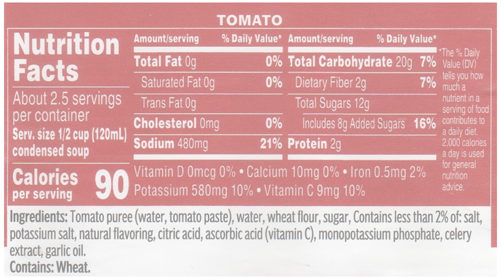
MD data sharing highlights

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

MD data sharing highlights

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

- Provide metadata (context)



TOMATO

Nutrition Facts

About 2.5 servings per container
Serv. size 1/2 cup (120mL) condensed soup

Calories 90
per serving

Amount/serving	% Daily Value*	Amount/serving	% Daily Value*
Total Fat 0g	0%	**Total Carbohydrate** 20g	7%
Saturated Fat 0g	0%	Dietary Fiber 2g	7%
Trans Fat 0g		Total Sugars 12g	
Cholesterol 0mg	0%	Includes 8g Added Sugars	16%
Sodium 480mg	21%	**Protein** 2g	

Vitamin D 0mcg 0% • Calcium 10mg 0% • Iron 0.5mg 2%
Potassium 580mg 10% • Vitamin C 9mg 10%

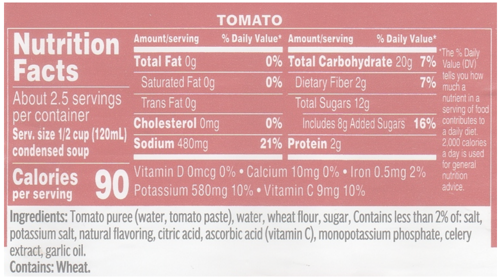
*The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Ingredients: Tomato puree (water, tomato paste), water, wheat flour, sugar, Contains less than 2% of: salt, potassium salt, natural flavoring, citric acid, ascorbic acid (vitamin C), monopotassium phosphate, celery extract, garlic oil.
Contains: Wheat.

MD data sharing highlights

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

- Provide metadata (context)
- Avoid zip files (no .tgz)



TOMATO

Nutrition Facts

About 2.5 servings per container
Serv. size 1/2 cup (120mL) condensed soup

Calories 90
per serving

Amount/serving	% Daily Value*	Amount/serving	% Daily Value*
Total Fat 0g	0%	**Total Carbohydrate** 20g	7%
Saturated Fat 0g	0%	Dietary Fiber 2g	7%
Trans Fat 0g		Total Sugars 12g	
Cholesterol 0mg	0%	Includes 8g Added Sugars	16%
Sodium 480mg	21%	**Protein** 2g	

Vitamin D 0mcg 0% • Calcium 10mg 0% • Iron 0.5mg 2%
Potassium 580mg 10% • Vitamin C 9mg 10%

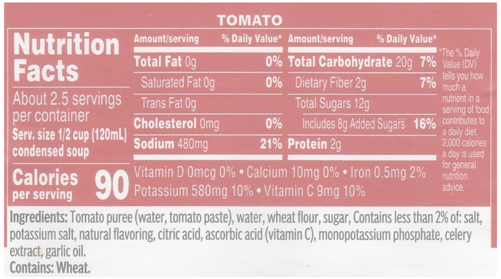
*The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Ingredients: Tomato puree (water, tomato paste), water, wheat flour, sugar, Contains less than 2% of: salt, potassium salt, natural flavoring, citric acid, ascorbic acid (vitamin C), monopotassium phosphate, celery extract, garlic oil.
Contains: Wheat.

MD data sharing highlights

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

- Provide metadata (context)
- Avoid zip files (no .tgz)



TOMATO

Nutrition Facts

About 2.5 servings per container
Serv. size 1/2 cup (120mL) condensed soup

Calories **per serving** **90**

Amount/serving	% Daily Value*	Amount/serving	% Daily Value*
Total Fat 0g	0%	**Total Carbohydrate** 20g	7%
Saturated Fat 0g	0%	Dietary Fiber 2g	7%
Trans Fat 0g		Total Sugars 12g	
Cholesterol 0mg	0%	Includes 8g Added Sugars	16%
Sodium 480mg	21%	**Protein** 2g	

Vitamin D 0mcg 0% • Calcium 10mg 0% • Iron 0.5mg 2%
Potassium 580mg 10% • Vitamin C 9mg 10%

*The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Ingredients: Tomato puree (water, tomato paste), water, wheat flour, sugar, Contains less than 2% of: salt, potassium salt, natural flavoring, citric acid, ascorbic acid (vitamin C), monopotassium phosphate, celery extract, garlic oil.
Contains: Wheat.

Why is it important?

- Sharing and storing data cost **money** and **energy**
- Dynamic generative deep-learning models? (Need for high quality, curated data)

What’s next?

Keep digging!

What’s next? Keep digging!

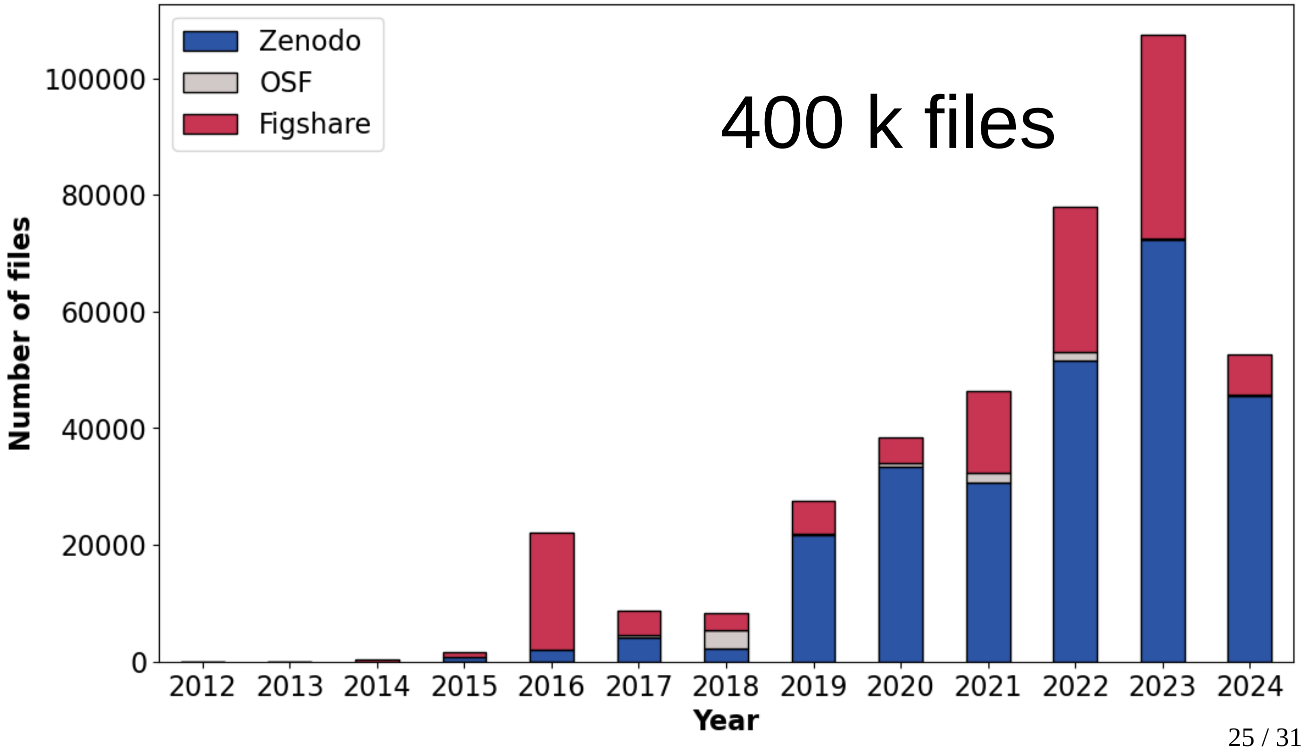
Explore other “data” repositories:

- Dryad, Science Data Bank
- ATLAS, MDDB

What’s next? Keep digging!

Explore other “data” repositories:

- Dryad, Science Data Bank
- ATLAS, MDDB



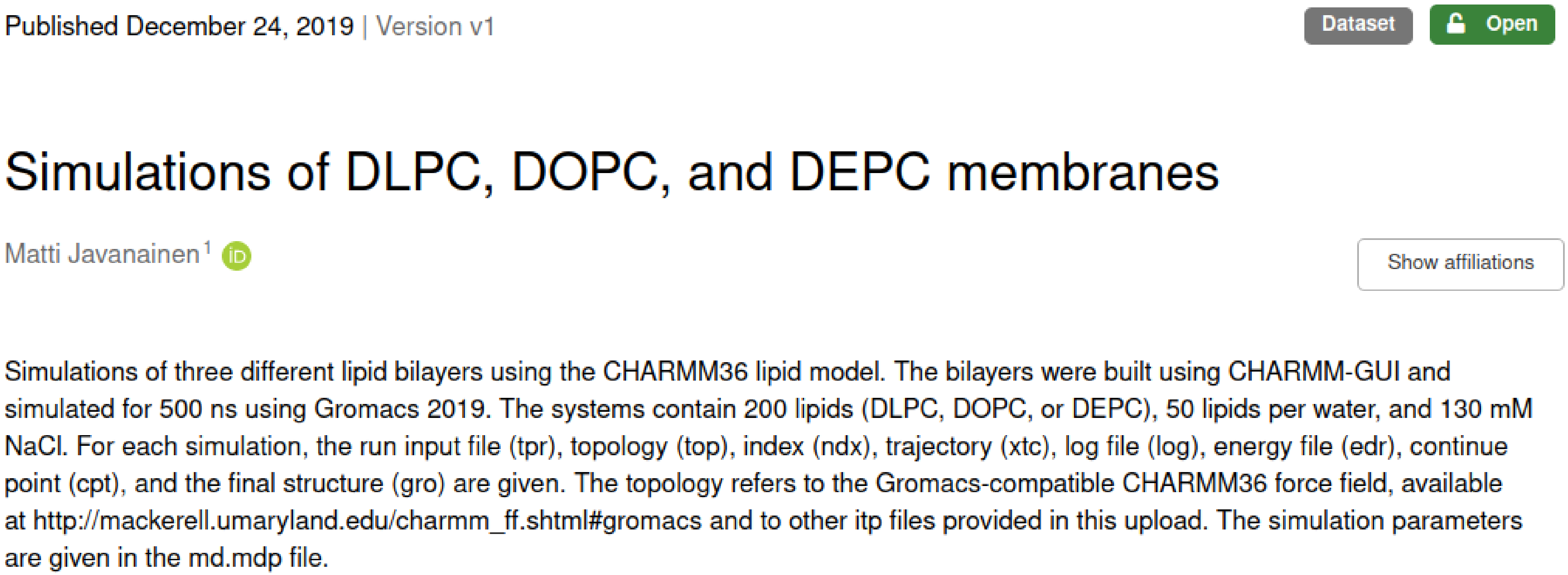
400 k files

What’s next?

Extract structured metadata from **raw text**

What’s next?

Extract structured metadata from **raw text**



Published December 24, 2019 | Version v1

Simulations of DLPC, DOPC, and DEPC membranes

Matti Javanainen[1]

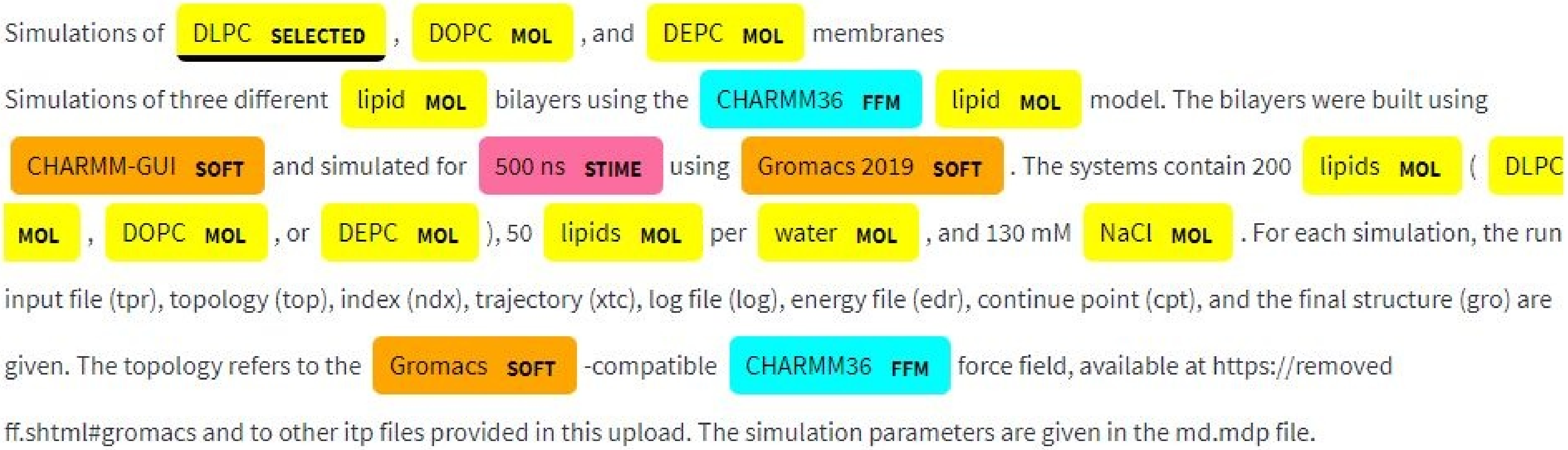
Simulations of three different lipid bilayers using the CHARMM36 lipid model. The bilayers were built using CHARMM-GUI and simulated for 500 ns using Gromacs 2019. The systems contain 200 lipids (DLPC, DOPC, or DEPC), 50 lipids per water, and 130 mM NaCl. For each simulation, the run input file (tpr), topology (top), index (ndx), trajectory (xtc), log file (log), energy file (edr), continue point (cpt), and the final structure (gro) are given. The topology refers to the Gromacs-compatible CHARMM36 force field, available at http://mackerell.umaryland.edu/charmm_ff.shtml#gromacs and to other itp files provided in this upload. The simulation parameters are given in the md.mdp file.

https://zenodo.org/records/3592499

What’s next?

Extract structured metadata from **raw text**

Text mining / Named Entity Recognition



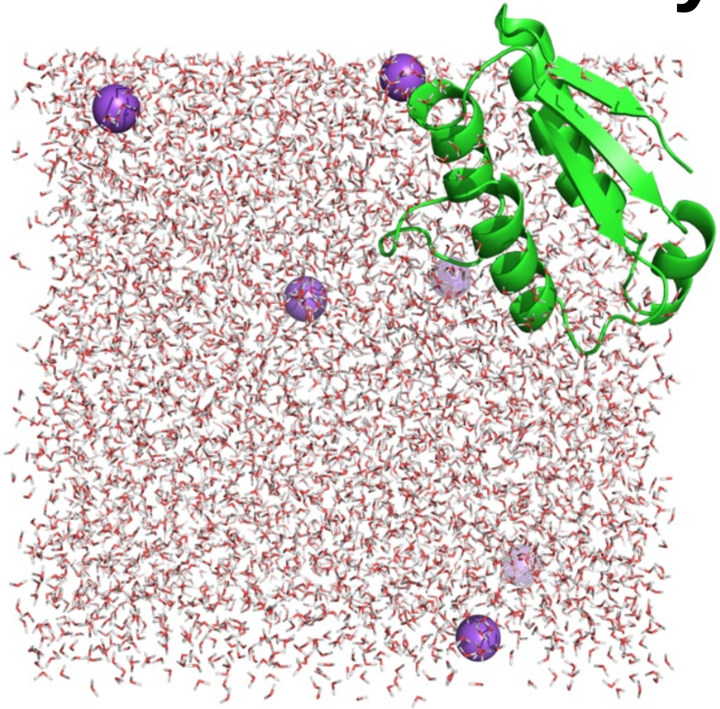
Simulations of DLPC SELECTED , DOPC MOL , and DEPC MOL membranes

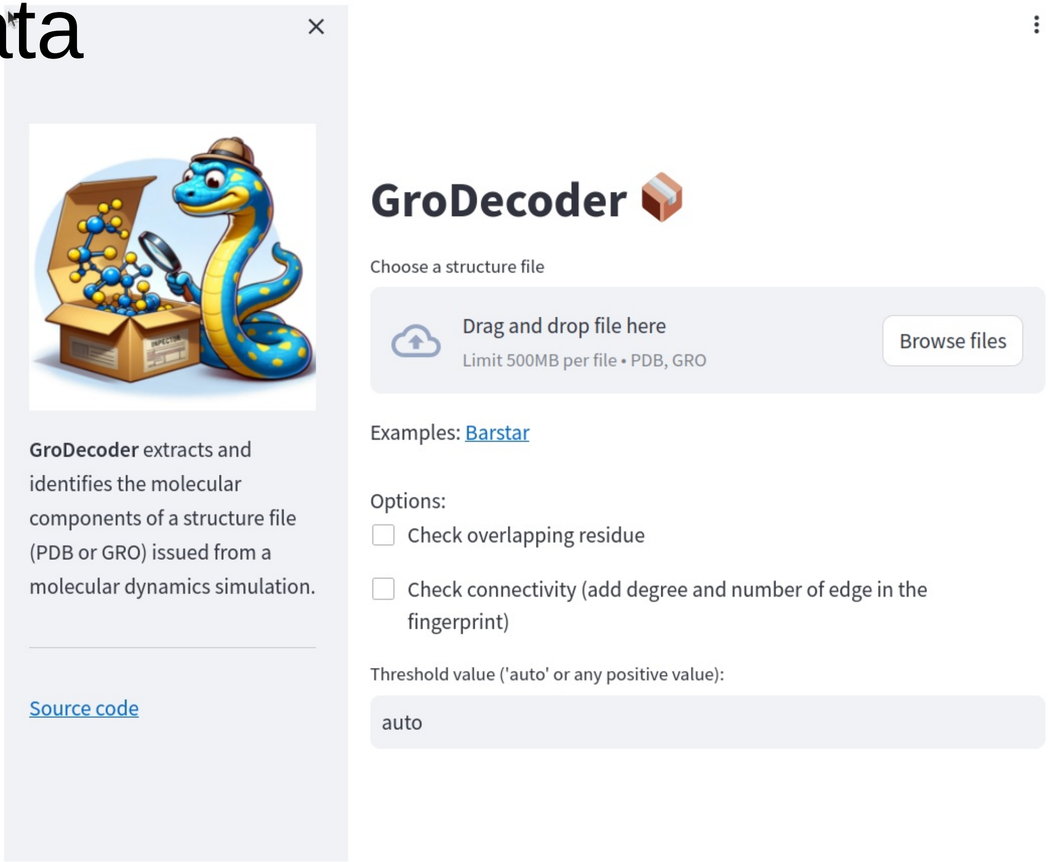
Simulations of three different lipid MOL bilayers using the CHARMM36 FFM lipid MOL model. The bilayers were built using CHARMM-GUI SOFT and simulated for 500 ns STIME using Gromacs 2019 SOFT . The systems contain 200 lipids MOL (DLPC MOL , DOPC MOL , or DEPC MOL), 50 lipids MOL per water MOL , and 130 mM NaCl MOL . For each simulation, the run input file (tpr), topology (top), index (ndx), trajectory (xtc), log file (log), energy file (edr), continue point (cpt), and the final structure (gro) are given. The topology refers to the Gromacs SOFT -compatible CHARMM36 FFM force field, available at https://removed ff.shtml#gromacs and to other itp files provided in this upload. The simulation parameters are given in the md.mdp file.

Mohamed Oussaren

What’s next?

Extract structured metadata from **molecular systems**





GroDecoder extracts and identifies the molecular components of a structure file (PDB or GRO) issued from a molecular dynamics simulation.

Source code

GroDecoder 📦

Choose a structure file

Drag and drop file here

Limit 500MB per file • PDB, GRO

Browse files

Examples: Barstar

Options:

□ Check overlapping residue

□ Check connectivity (add degree and number of edge in the fingerprint)

Threshold value ('auto' or any positive value):

auto

https://grodecoder.streamlit.app/

Karine Duong

What’s next?

Improve discoverability and data(sets) exploration

eosc | Lumen

Linked User-driven Multidisciplinary Exploration Network
2025 → 2027

Involved communities: SSH, Maths, Earth System, Molecular Dynamics

Thanks 🙏

IJM, Paris, France
Lisa Bouarroudj
Mohamed Oussaren

CBI, Toulouse, France
Magdalena Szczuka
Matthieu Chavent

LBT, Paris, France
Marc Baaden
Karine Duong

Amsterdam, Netherlands
Steven Garcia

Univ. Copenhagen, Denmark
Johanna K. S. Tiemann
Kresten Lindorff-Larsen

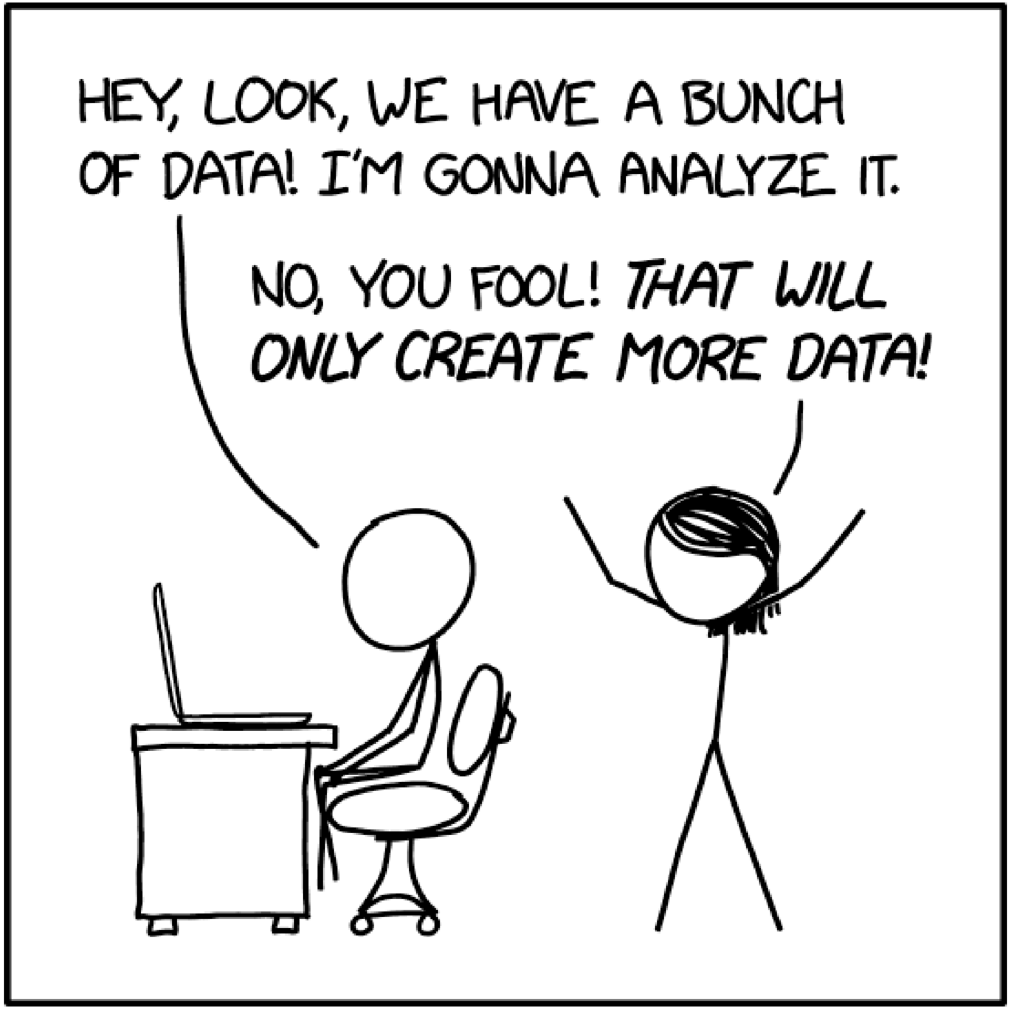
KTH Royal Inst. Tech. Stockholm, Sweden
Lucie Delemotte
Erik Lindahl

Stockholm Univ. , Sweden
Rebecca J. Howard

Questions? ✨

▲ This presentation ▲

DOI 10.5281/zenodo.14264571


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

Source: Data trap, XKCD, CC BY-NC