PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

PSDI – Pathfinder 4

FAIR Data for the Biomolecular Simulation Community

James Gebbie-Rayet & Jas Kalayan

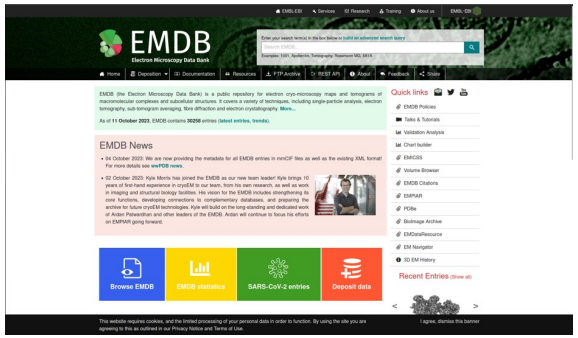
Webinar Structure

- Overview of domain (James)
- Aims of the project (James)
- Technical overview of platform so far (Jas)
- Technical demonstration (Jas)
- Questions (You)

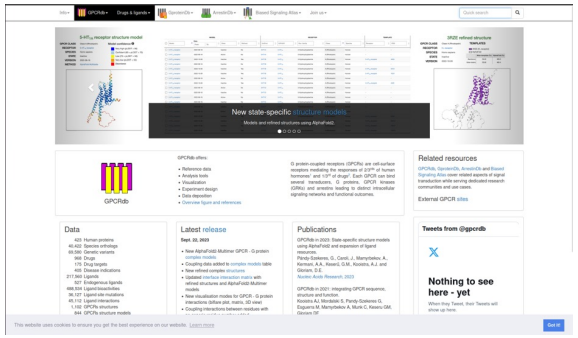
Biologists are brilliant with data!

Other areas of biology are very organised about data!

https://www.rcsb.org/

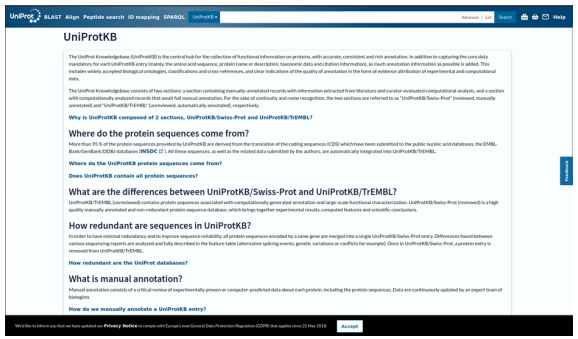


https://www.ebi.ac.uk/emdb/

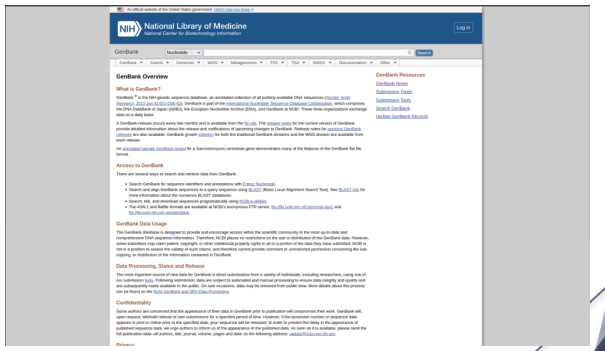


https://gpcrdb.org/

https://www.ebi.ac.uk/empiar/



https://www.uniprot.org/help/uniprotkb



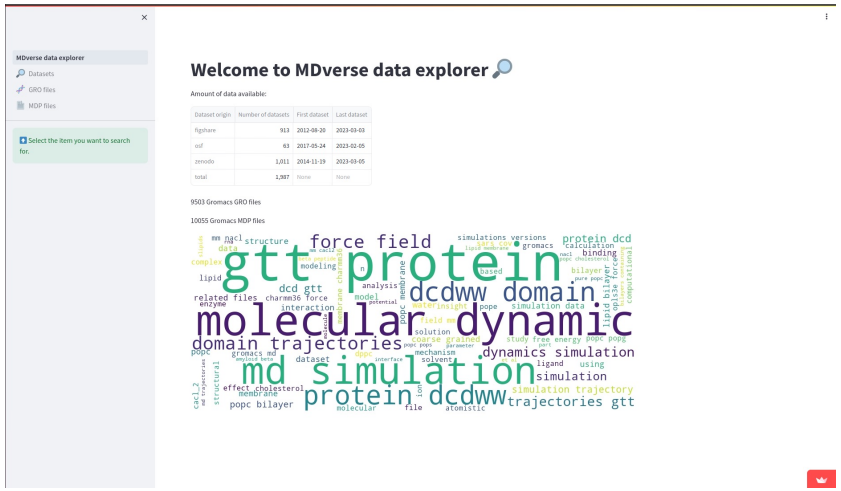
https://www.ncbi.nlm.nih.gov/genbank/

What about Biomolecular Simulation?

Not much in the way of established production databases or services in UK!!

What's there already/coming?

EU Initiatives focussed purely on biomolecular simulation:



https://mdverse.streamlit.app/

Focuses on mining common file shares

Difficult to FAIR-ise data

https://mddbr.eu/

Large EU project bringing together several small Dbs

Challenging issue around integrating many sources, unclear how FAIR data will be

Existing Generic Sharing Platforms

Some existing ways we as researchers share our data:

https://zenodo.org/

https://figshare.com/

https://osf.io/

- Great for sharing files
- Great for getting persistent identifiers
- No requirement to share data that is FAIR
- Researchers left to own devices on what is shared!

https://www.mendeley.com/

What's the Main Issues in Domain?

The main issues that our pathfinder aims to tackle are:

- No consistent approach to storing or sharing simulation data across the community.
- No real infrastructure or route to make simulation data accessible.
- Usage of existing or emerging platforms not FAIR
- The funding, research, publish cycle in the field currently discourages researchers to think or do anything about sharing data.
- Research papers in biomolecular simulation do not often contain information to allow fully reproducible studies.
- Even publications where simulations are well described, are difficult to reproduce due to full provenance of model creation not being present.
- Difficult to know exactly which experimental data sets/sources involved in research studies.

What is PSDI Pathfinder Doing?

Objective is to establish data infrastructure prototype and tools to improve data practices in field without major cultural shifts:

- Main aim is to improve data practices in domain – align with FAIR principles
- Prototype tools to capture full data provenance for model creation, simulation and analytics (FAIR)
- Prototype infrastructure tools to store, access, find and share data (FAIR)
- Establish long term collaboration with other data initiatives (EBI, EU and US funded)
- Establish hard data links to experimental data sources
- "I" (FAIR) not yet in scope of this pathfinder (excellent projects in wider community)

We want data to be easy!

Our philosophy is that data practice should be baked into the way we work, if it adds extra burden then we won't do it!

- We want to eliminate having to fill in web forms with huge numbers of fields
- Our tooling is designed to harvest data and metadata in an automated way
- We do this in a way that avoids you having to change much about how you work
- You should be in control of your data at all times and choose if and when to share
- It should be easy to generate overviews of your data to make publishing simpler
- Much more is under development!

What could it enable?

We think that better data practices unlock enormous potential!

- As a research field we can avoid duplication of effort
- Having our full data out in the open will enable exploitation by other domains (AI/ML)
- Collecting and sharing full model provenance will enable full reproducibility
- Sharing simulation data can improve statistical/physical understanding of systems when studied with different codes/methods (we saw this during SARS-COV2)
- Sharing simulation data is climate friendly – simulations on HPC are energy intensive!
- We can use high quality datasets to inform novel method development and evaluate the quality of existing and emerging forcefield, physics etc
- Knowledge will bleed between research groups since model creation can be shared
- Councils and other funders will be able to see the true cost of research, by linking with experimental databases and collecting contributors in the chain

Over to Jas!

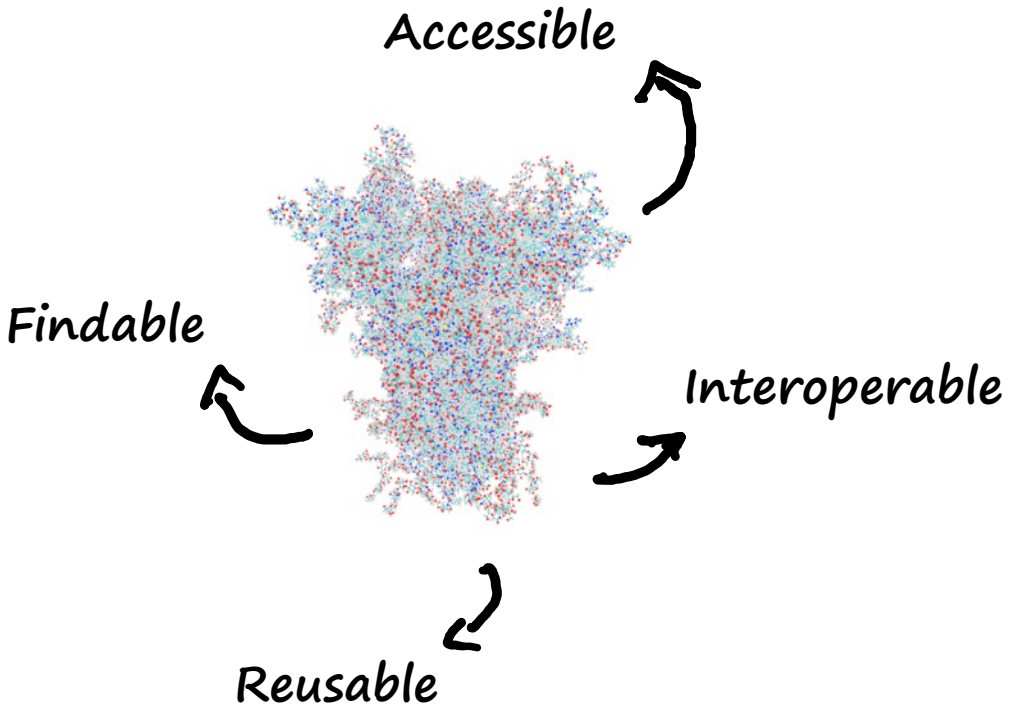
PSDI Webinar: FAIR Data for the Biomolecular Simulation Community

Acknowledgments to:

Kin Chao,
Joel Greer, Tom Burnley,
Martyn Winn

Imperial College London

CCP-EM



*DESRES COVID19 spike protein

18.10.2023

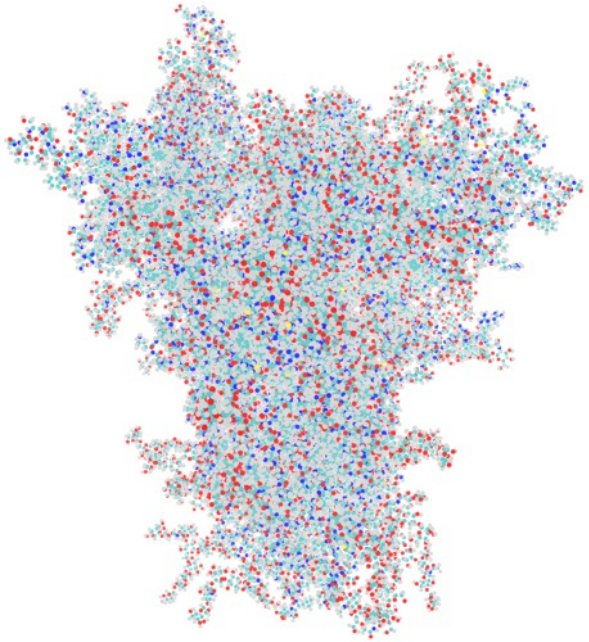
PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

UKRI Science and Technology Facilities Council

UKRI Engineering and Physical Sciences Research Council

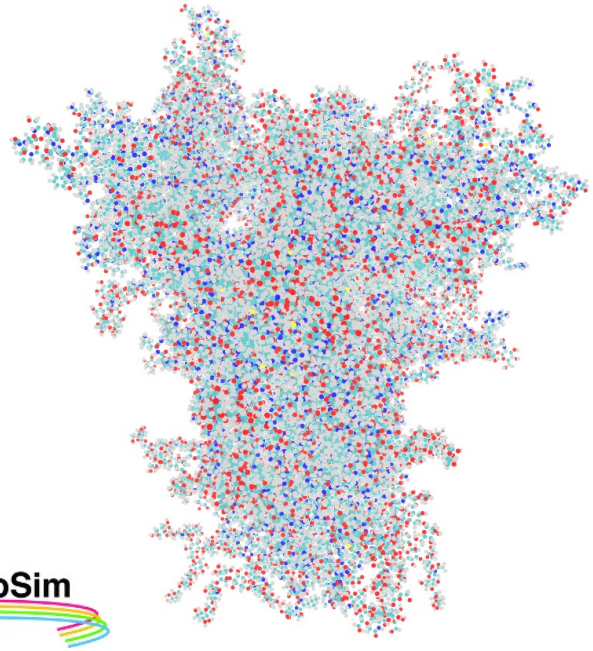
CCPBioSim

The Biomolecular Simulation Community



Data we produce?

xyz coordinates for each atom
≈10^5 atoms,
≈10^3-10^4 frames,
≈weeks to perform
≈TBs of data per simulation



JADE Tier 2 HPC · archer2 · N8 Bede COMPUTATIONALLY INTENSIVE RESEARCH · HECBioSim

Start → ...various protocols... → Simulation

PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

*DESRES COVID19 spike protein simulation

Various Protocols in MD Simulation

1. Get crystal Structure → 2. System Preparation → 3. Parameterisation

PACKMOL

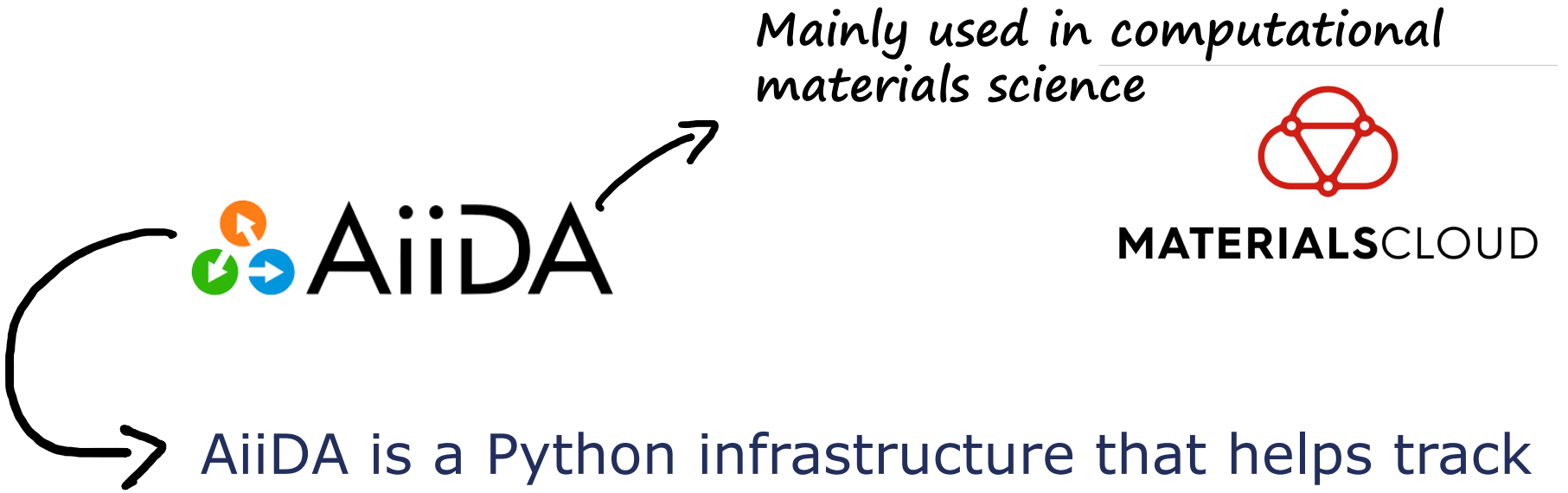
Initial configurations for Molecular Dynamics Simulations by packing optimization

4. Minimisation → 5. Equilibration → 6. Simulation

Using the same MD engine

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Biomolecular Simulation Data Provenance using AiiDA



AiiDA is a Python infrastructure that helps track complex workflows used in computational science.
https://github.com/aiidateam/aiida-core

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

AiiDA-GROMACS: A plugin for Data Provenance with GROMACS Simulations

GROMACS is used by 70% of HECBioSim users, But more plugins to come!

AiiDA

AiiDA PLUGIN REGISTRY

[View on GitHub/register your package]

AiiDA plugin package "aiida-gromacs"

< back to the registry index

General information

Current state: status alpha
Short description: A plugin for using GROMACS with AiiDA for molecular dymanics simulations.
How to install: `pip install git+https://github.com/jimboid/aiida-gromacs`
Source code: Go to the source code repository
Documentation: Go to plugin documentation

Detailed information

Author(s): James Gebbie-Rayet
Contact: james.gebbie@stfc.ac.uk
How to use from python: `import aiida_gromacs`
Most recent version:
Compatibility: AiiDA >=2.0,<3

Plugins provided by the package

Calculations 7 | Parsers 7 | Data 6 | Workflows 1

PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

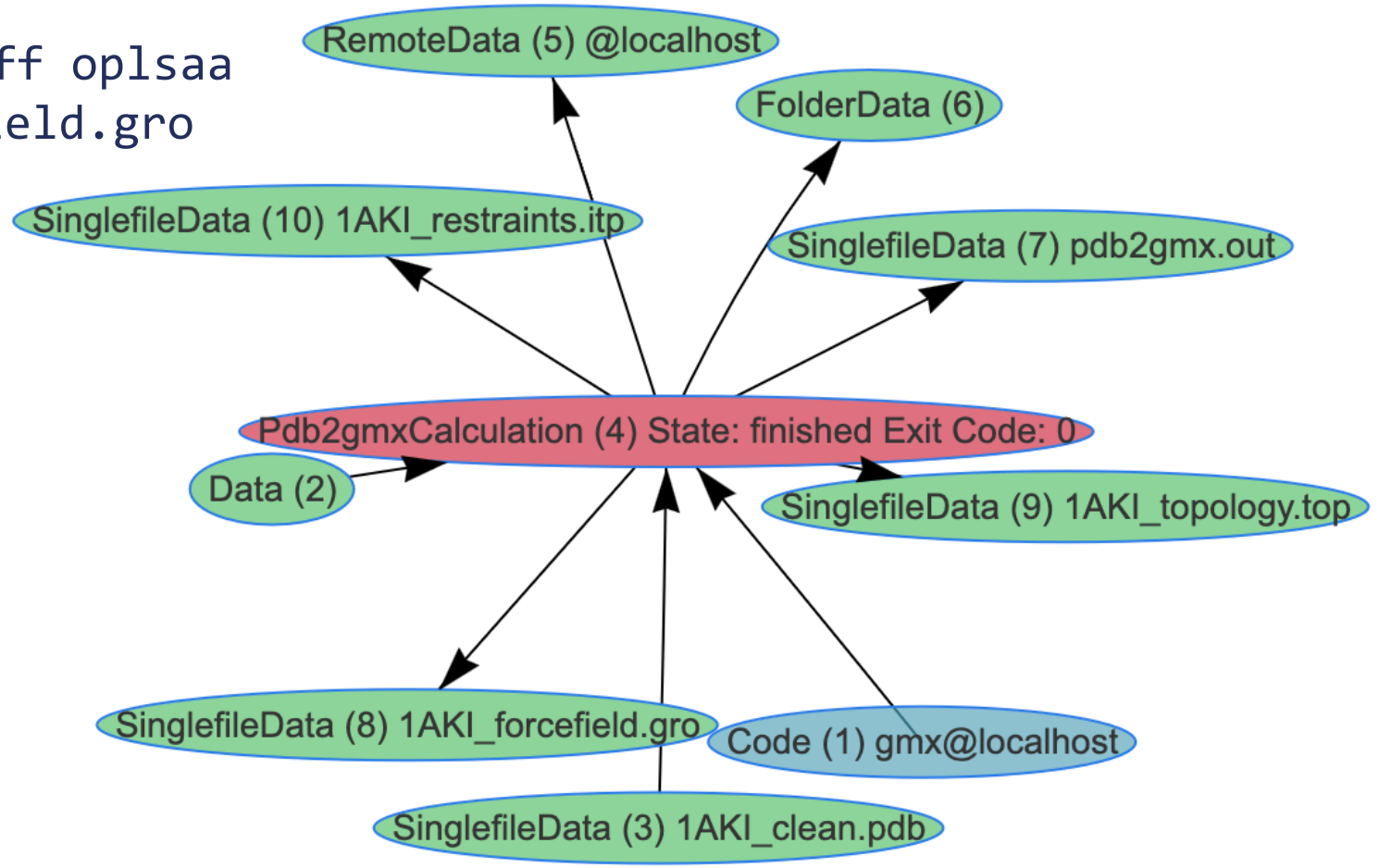
Basics of using AiiDA for Data Provenance

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Example of Data Provenance with aiida-gromacs

```
$ gmx_pdb2gmx -f 1AKI_clean.pdb -ff oplsaa
    -water spce -o 1AKI_forcefield.gro
    -p 1AKI_topology.top
    -i 1AKI_restraints.itp
```

Note the underscore!

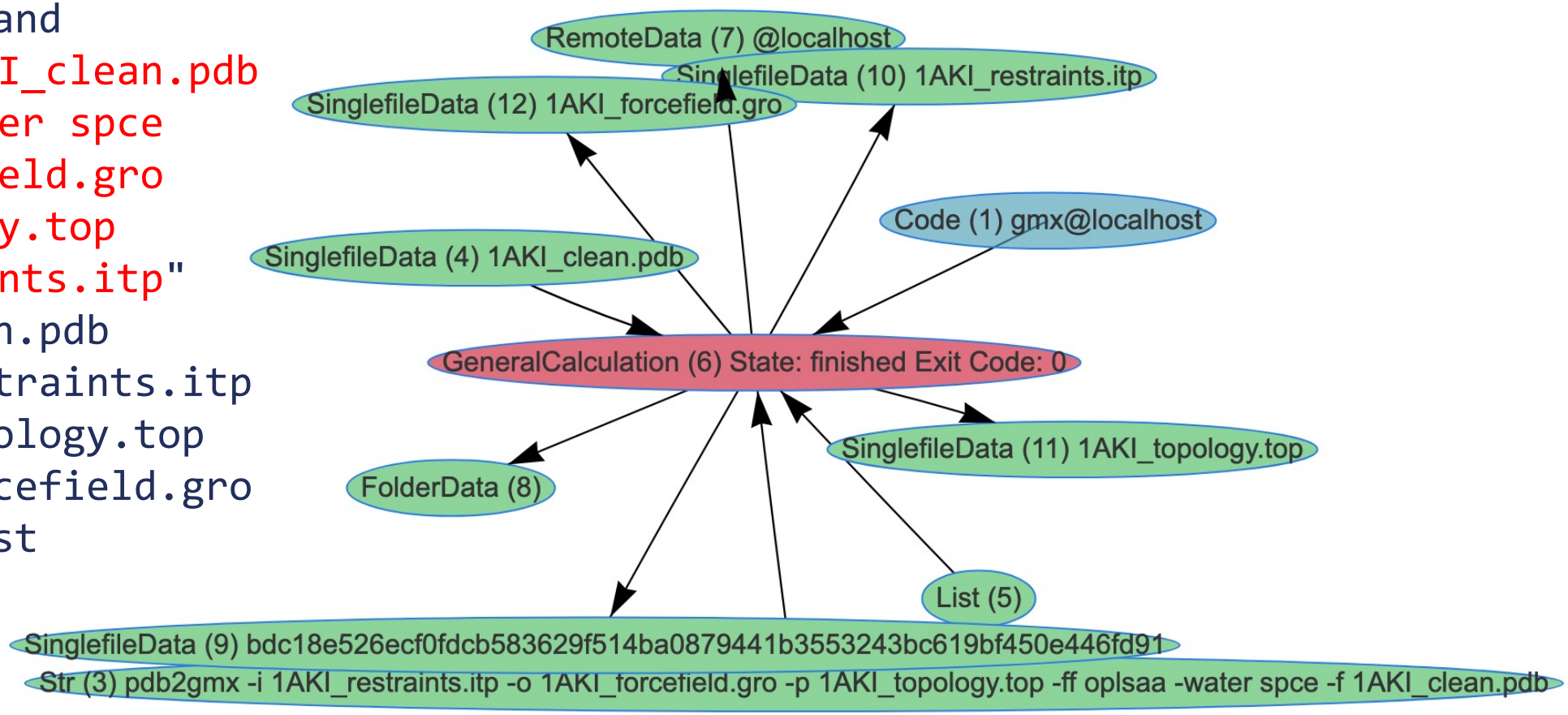


PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

https://aiida-gromacs.readthedocs.io

Example of Data Provenance with aiida-gromacs

```
$ genericMD --command
    "pdb2gmx -f 1AKI_clean.pdb
    -ff oplsaa -water spce
    -o 1AKI_forcefield.gro
    -p 1AKI_topology.top
    -i 1AKI_restraints.itp"
--inputs 1AKI_clean.pdb
--outputs 1AKI_restraints.itp
--outputs 1AKI_topology.top
--outputs 1AKI_forcefield.gro
--code gmx@localhost
```



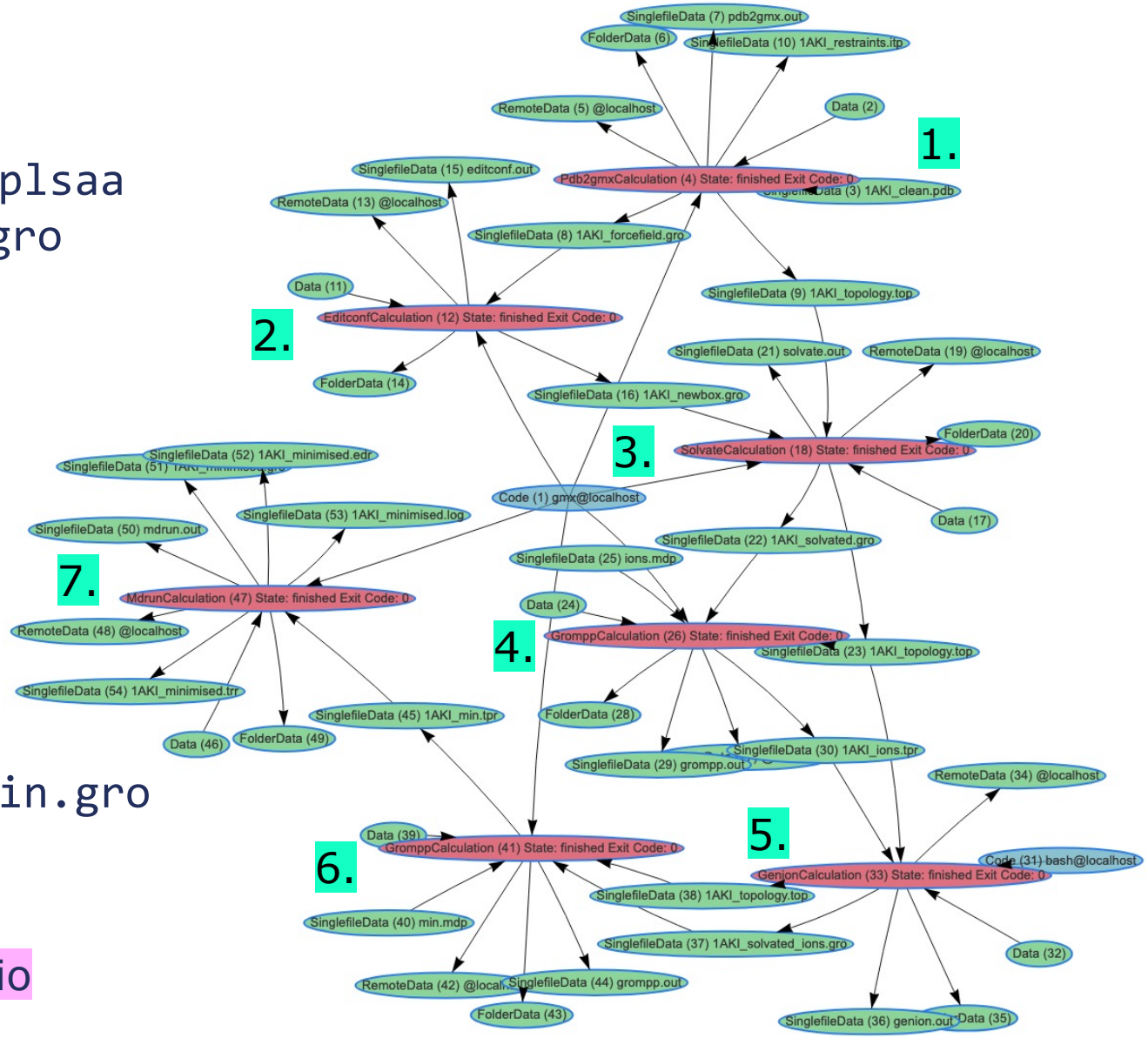
PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Example of Data Provenance with aiida-gromacs

1. gmx_pdb2gmx -f 1AKI_clean.pdb -ff oplsaa
   -water spce -o 1AKI_forcefield.gro
   -p 1AKI_topology.top
   -i 1AKI_restraints.itp

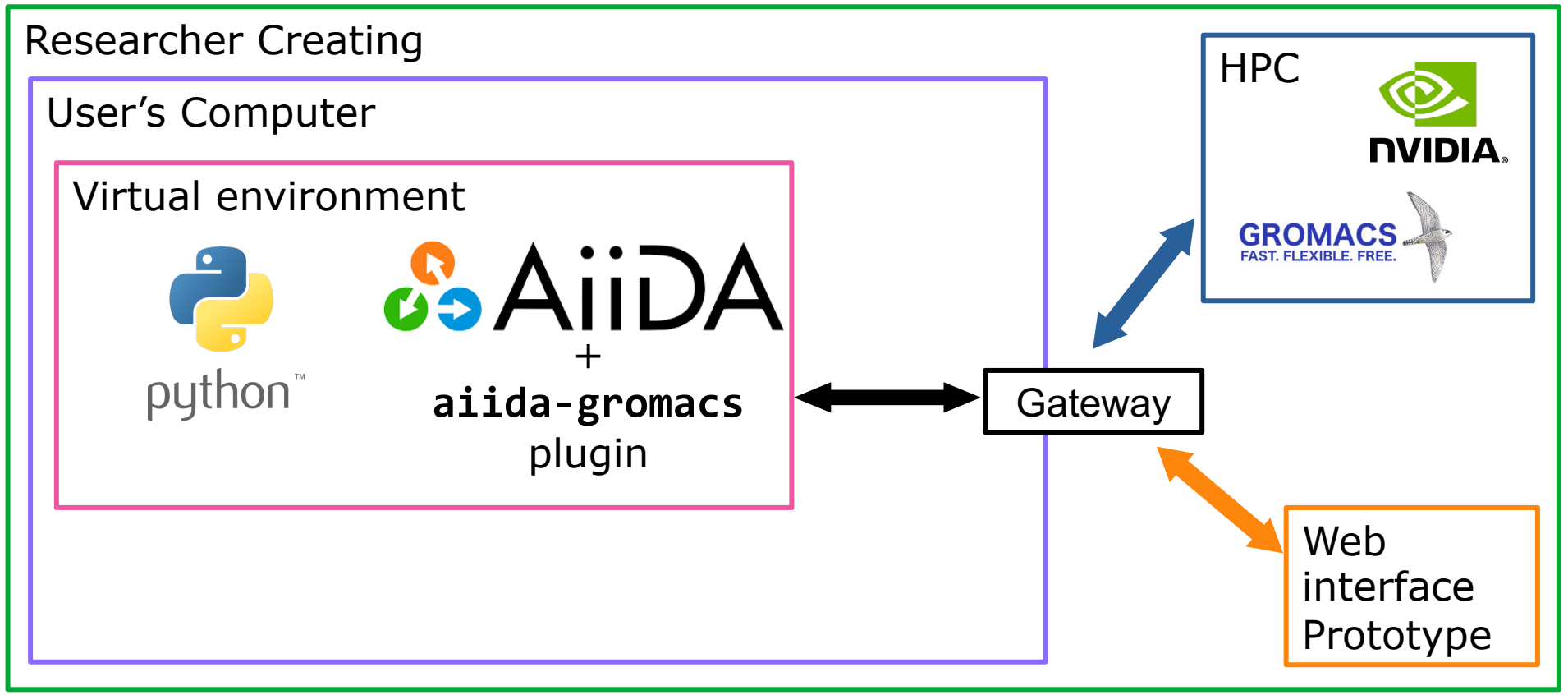
2. gmx_editconf,
3. gmx_solvate,
4. gmx_grompp,
5. gmx_genion,
6. gmx_grompp,

7. gmx_mdrun -s 1AKI_min.tpr -c 1AKI_min.gro
   -e 1AKI_min.edr -g 1AKI_min.log
   -o 1AKI_min.trr



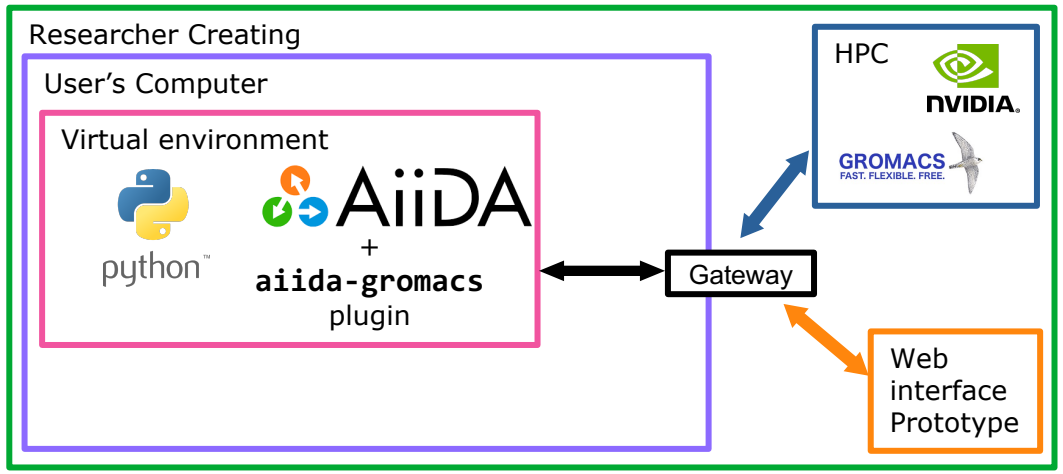
https://aiida-gromacs.readthedocs.io

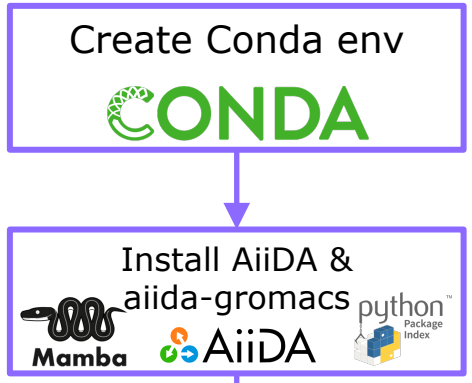
Our User Environment Prototype

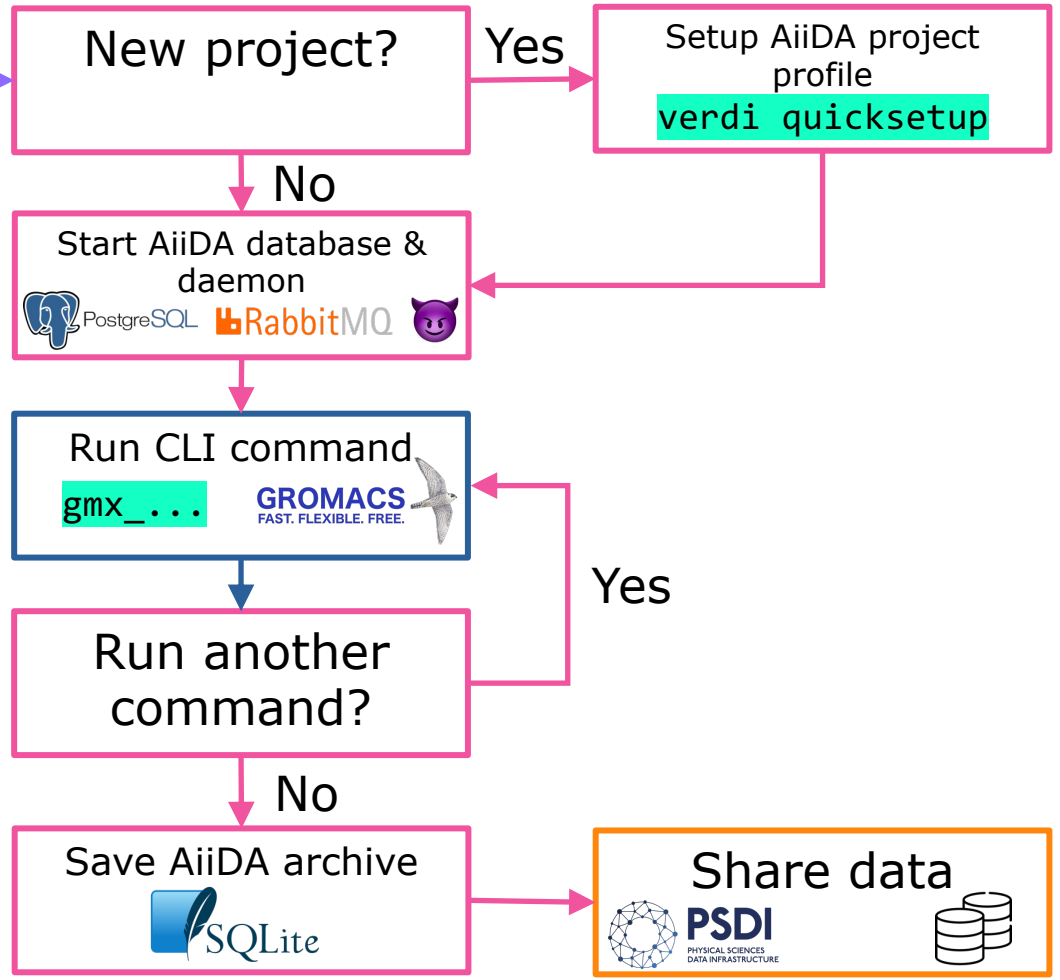


PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Our User Environment Prototype

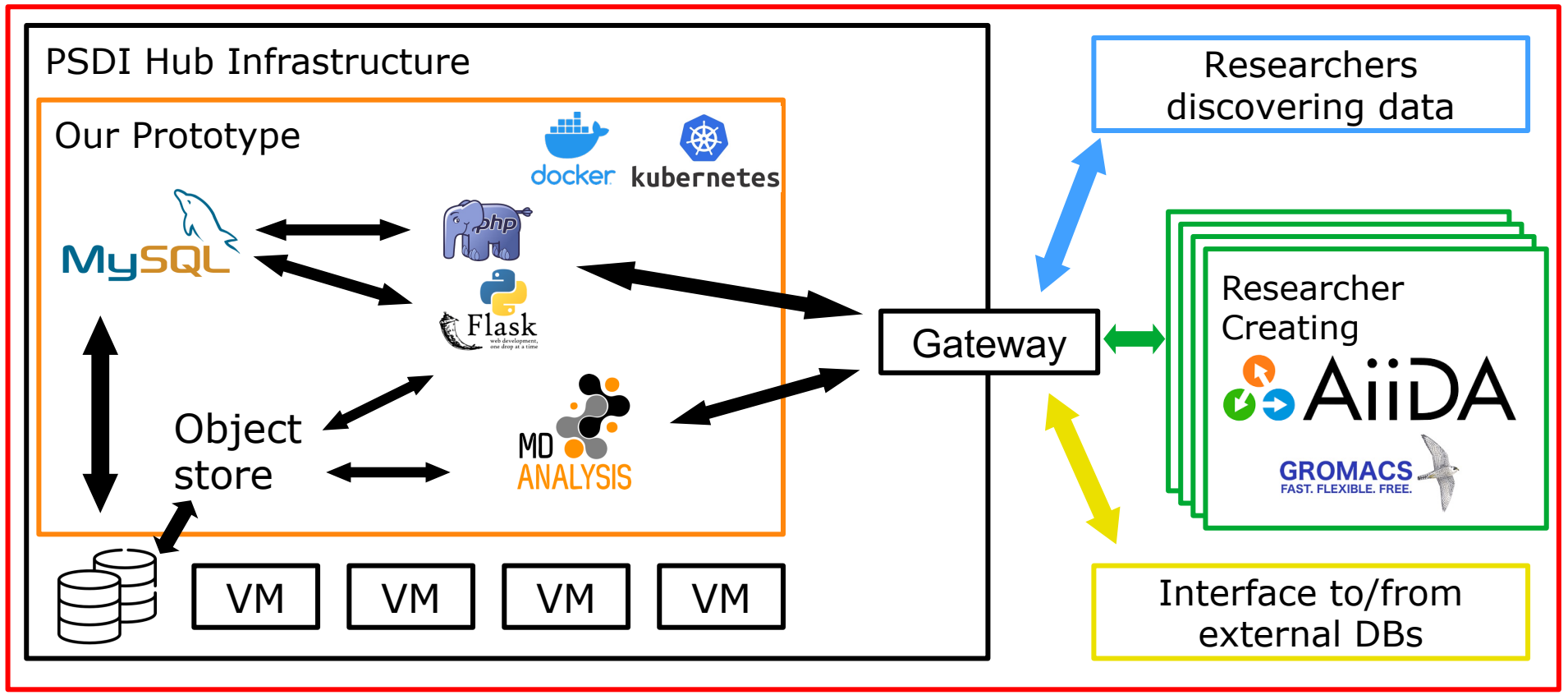






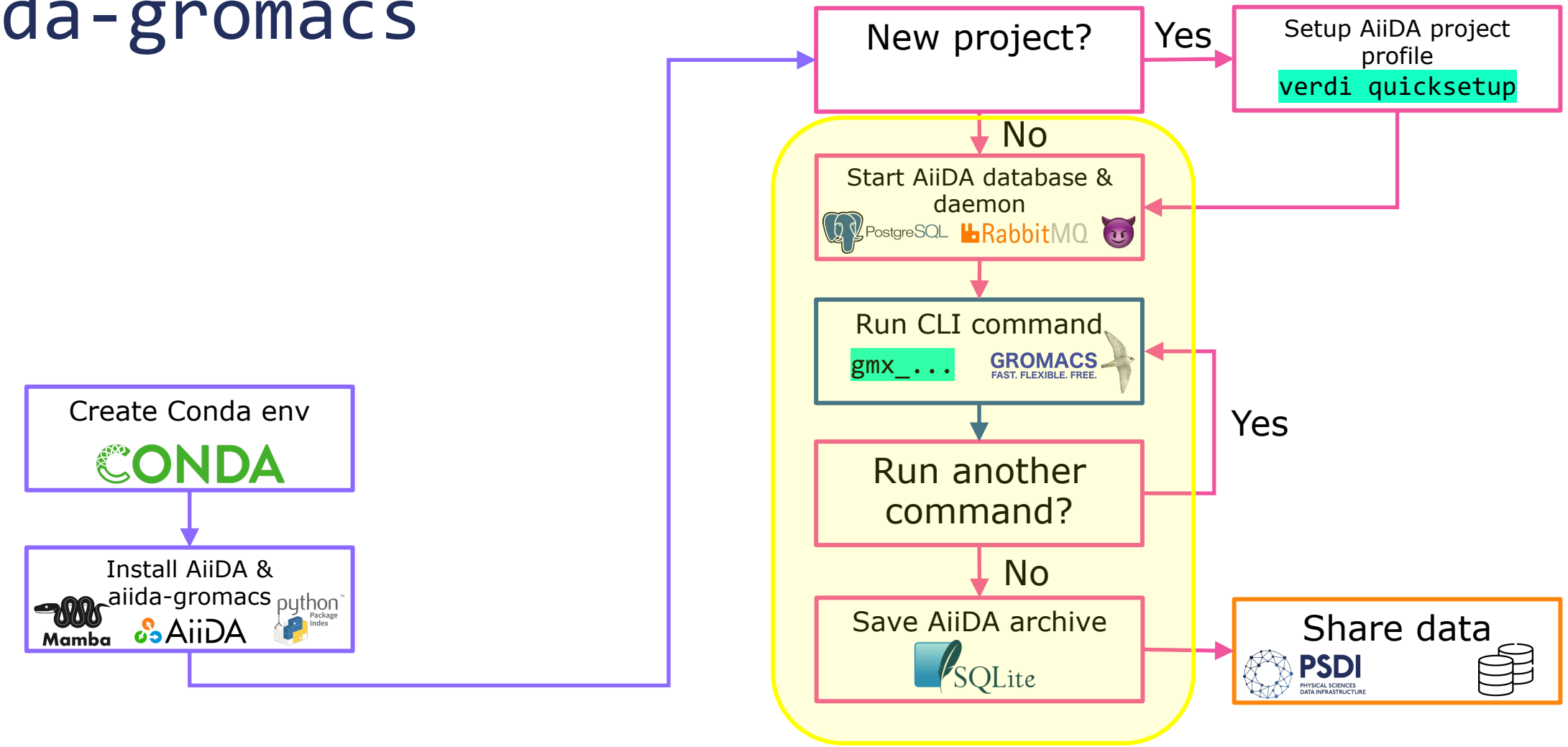
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Our Infrastructure Prototype



PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Demo: Lysozyme Minimisation with aiida-gromacs



https://www.ccpbiosim.ac.uk/notebooks/hub/spawn?profile=aiida-demo