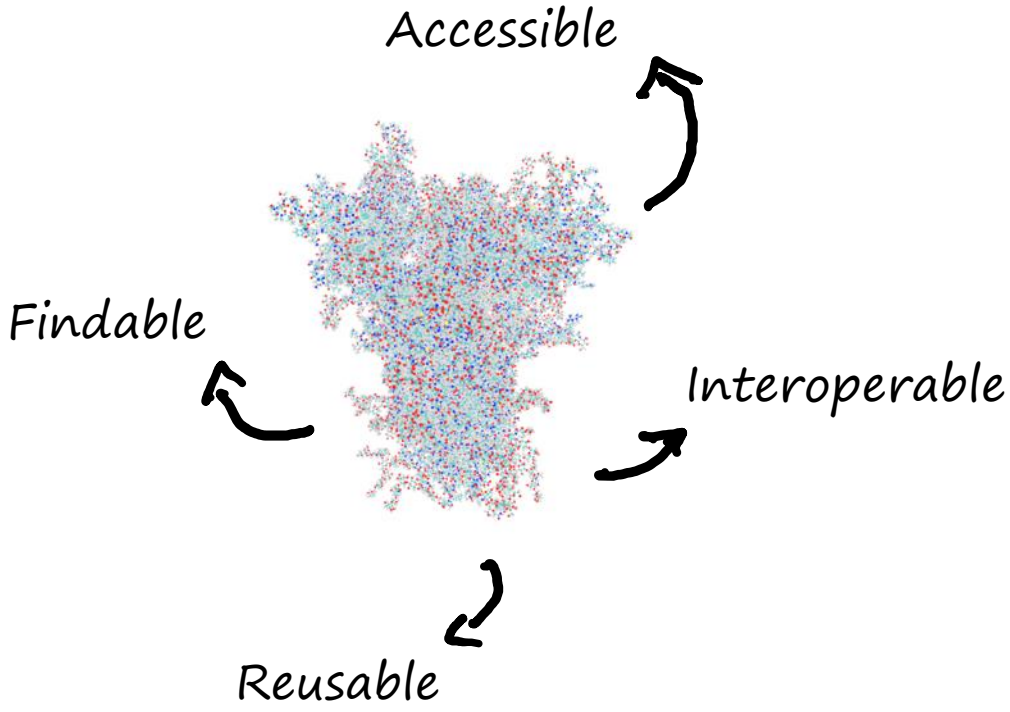
FAIR Data for the Biomolecular Simulation Community

Jas Kalayan

Senior Computational Scientist, STFC UKRI

jas.kalayan@stfc.ac.uk



*DESRES COVID19 spike protein

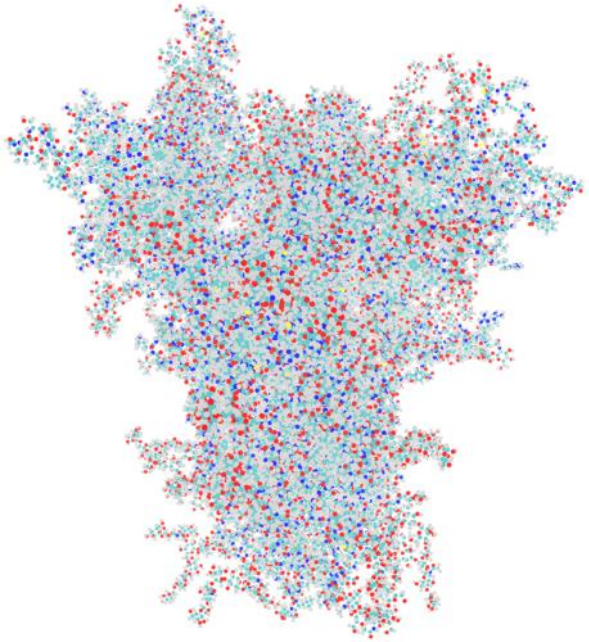
PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

UKRI Science and Technology Facilities Council

UKRI Engineering and Physical Sciences Research Council

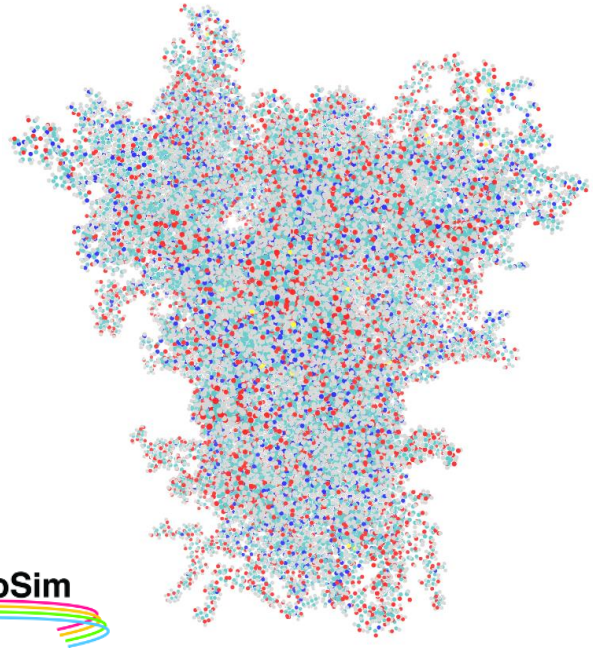
CCPBioSim

The Biomolecular Simulation Community



💾 Data we produce?

xyz coordinates for each atom
≈10^5 atoms,
≈10^3-10^4 snapshots,
≈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

WORLDWIDE PDB PROTEIN DATA BANK

AlphaFold

python

CHARMM-GUI
Effective Simulation Input Generator and More

ChemShell
multiscale computational chemistry

VMD

PACKMOL
Initial configurations for Molecular Dynamics Simulations by packing optimization

GROMACS
FAST. FLEXIBLE. FREE.

OpenMM

APBS & PDB2PQR
Software for biomolecular electrostatics and solvation

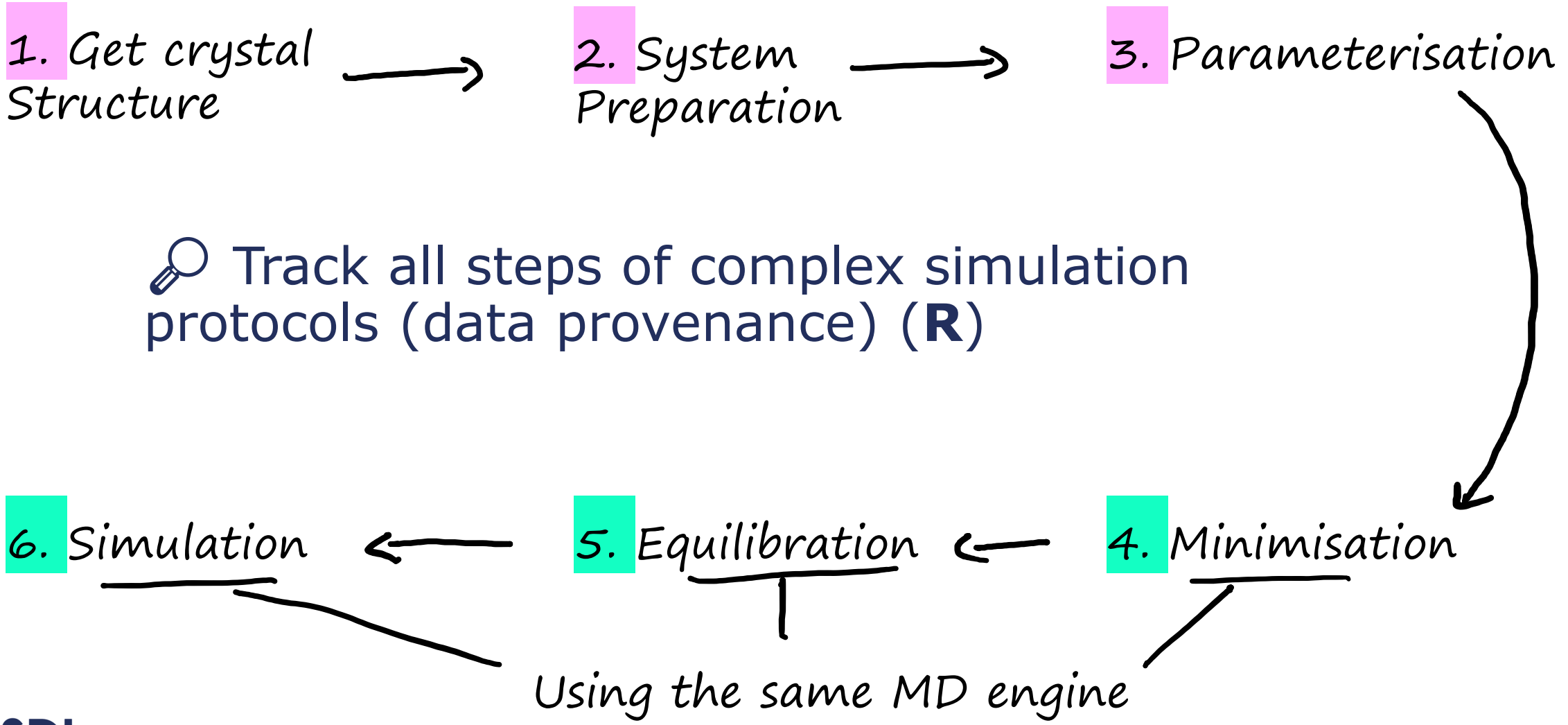
AMBER MD

4. Minimisation → 5. Equilibration → 6. Simulation

Using the same MD engine

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

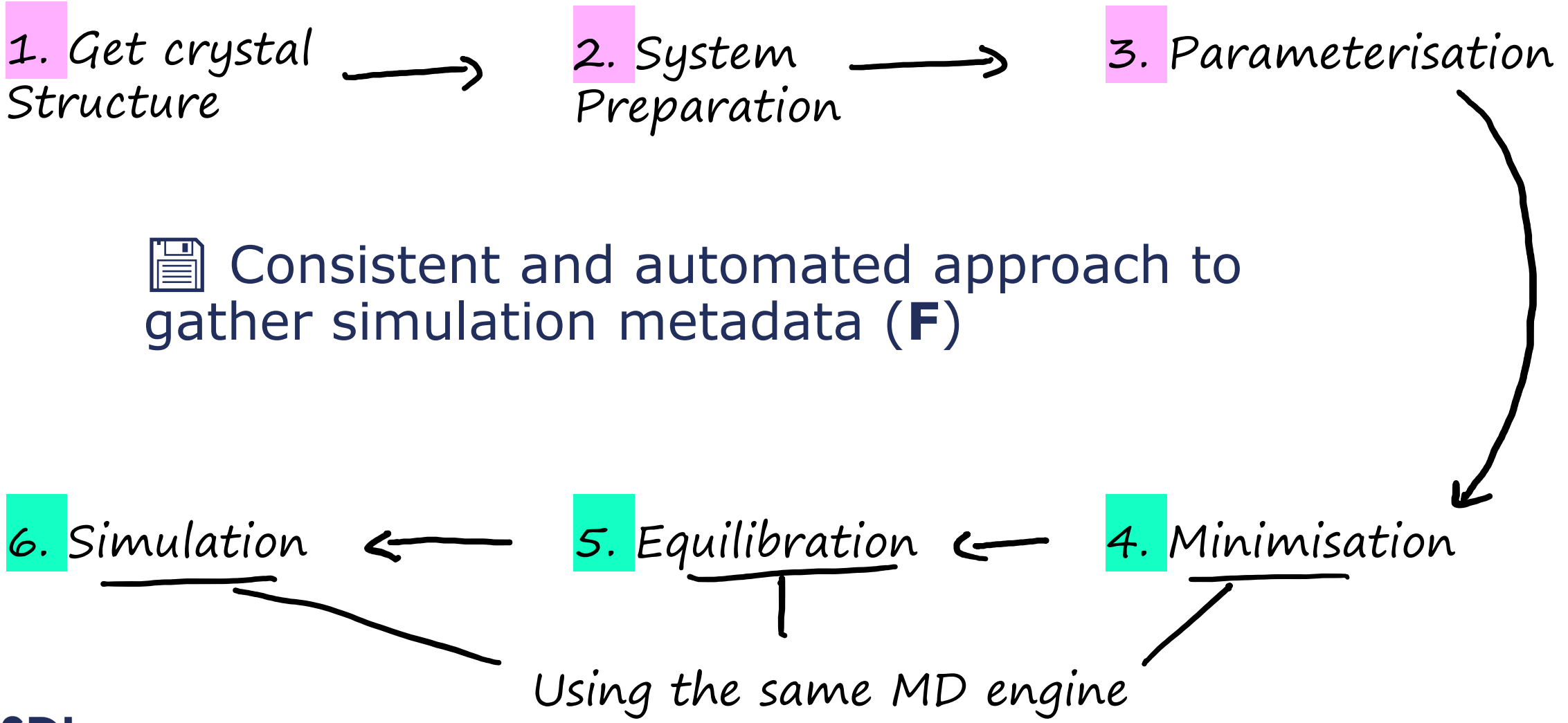
How to get **FAIR** Biomolecular MD Simulation Data?



🔎 Track all steps of complex simulation protocols (data provenance) (**R**)

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

How to get **FAIR** Biomolecular MD Simulation Data?



💾 Consistent and automated approach to gather simulation metadata (**F**)

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

PSDI: A Programme for National Infrastructure in the Physical Sciences

📎 Providing infrastructure that connects experimental and computational data services within Physical Sciences.

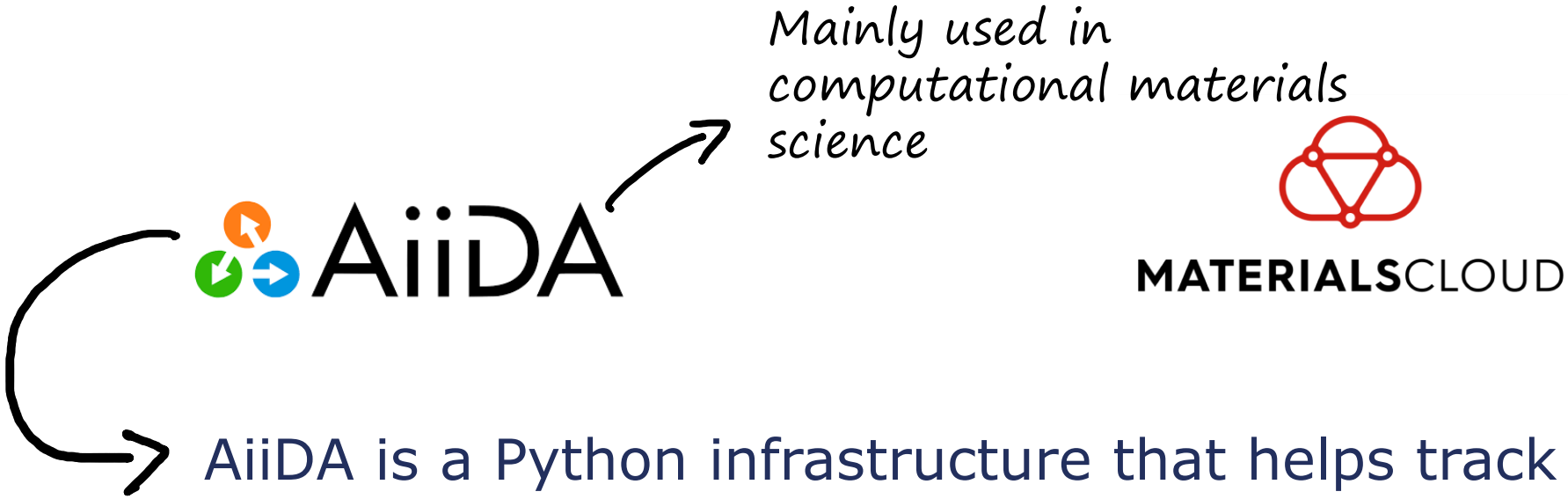
🔎 Various initial pathfinders to scope areas of development (catalysis/environment, **biomolecular science**, artificial intelligence).

Engineering and Physical Sciences Research Council

PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

https://www.psdi.ac.uk/

Biomolecular Simulation Data Provenance using AiiDA

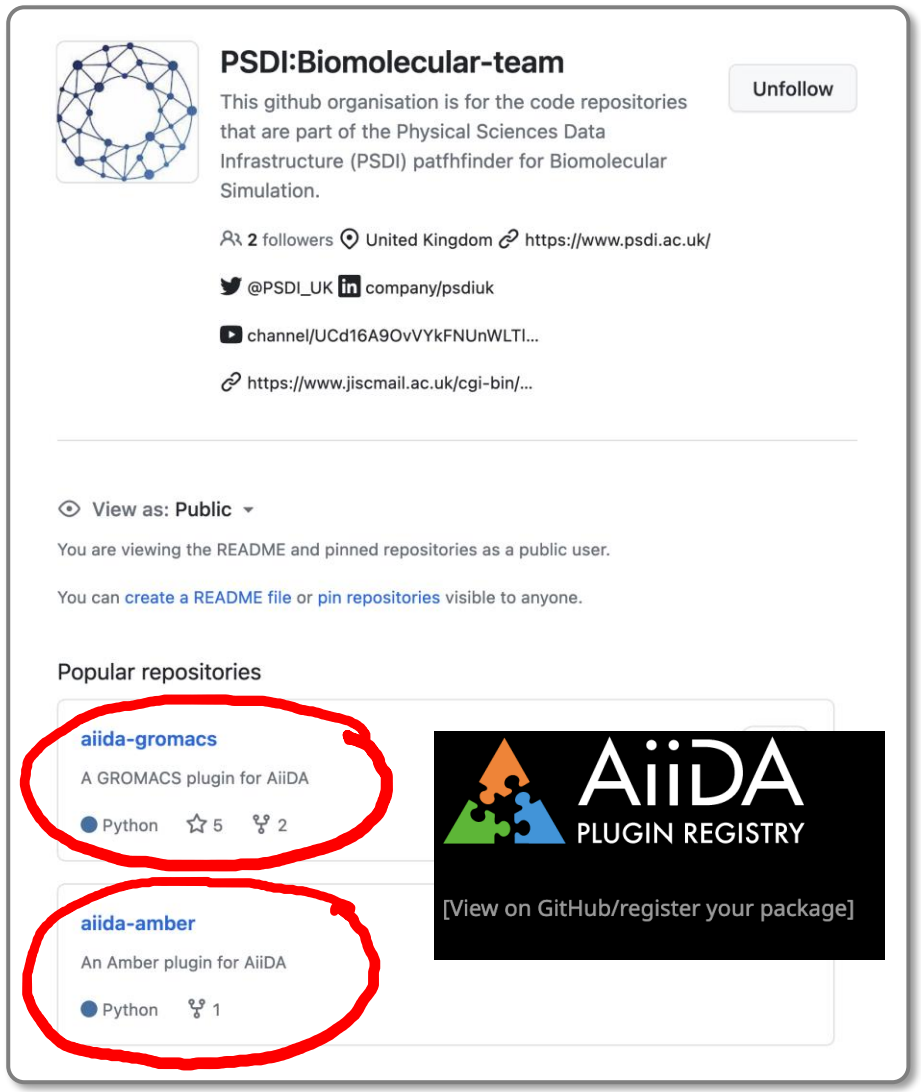


AiiDA is a Python infrastructure that helps track complex workflows used in computational science.
https://github.com/aiidateam/aiida-core

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

AiiDA Plugins for Data Provenance with Amber & GROMACS Simulations

Used by 90% of HECBioSim applicants



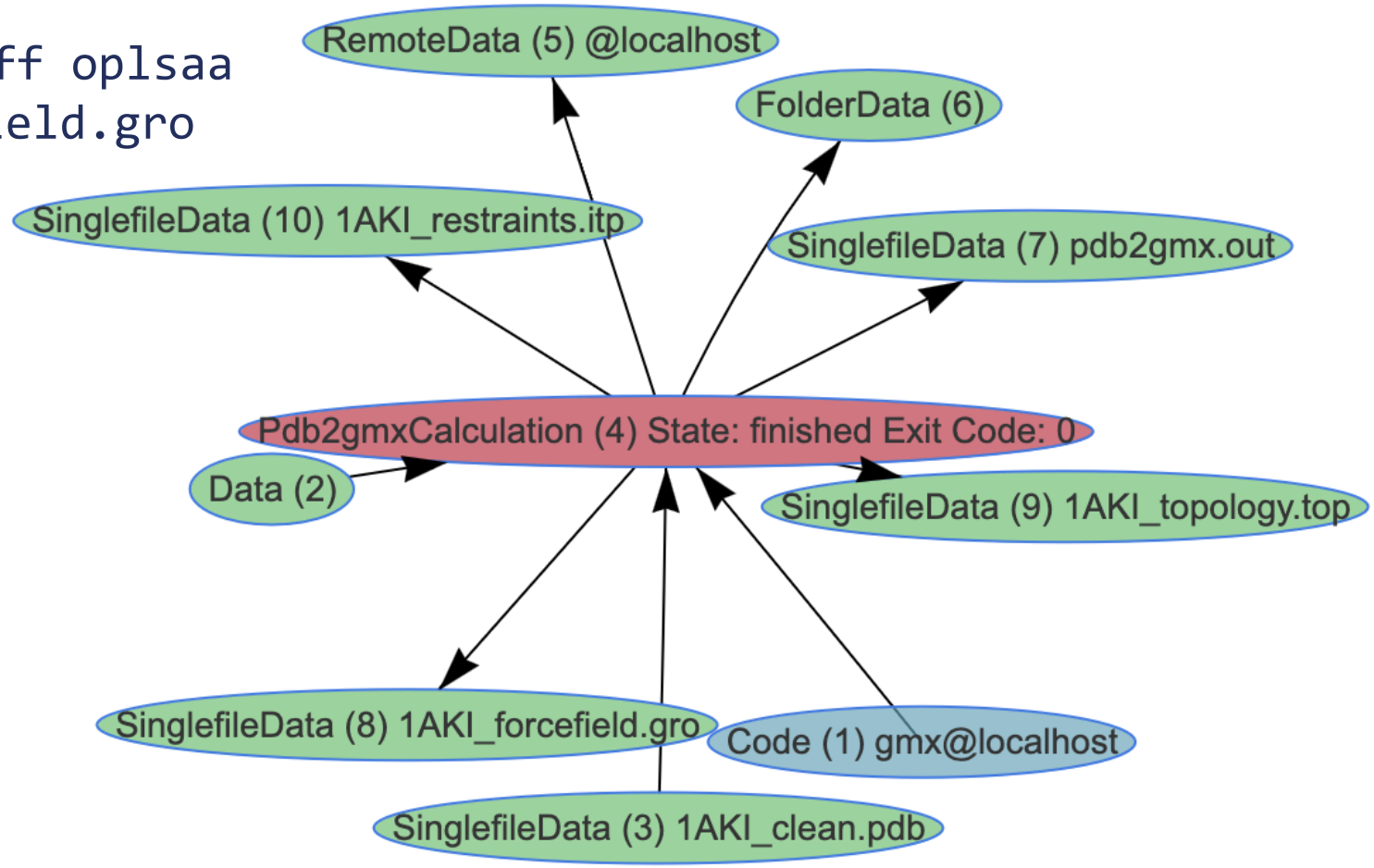
PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

https://github.com/PSDI-Biomolecular-team

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!

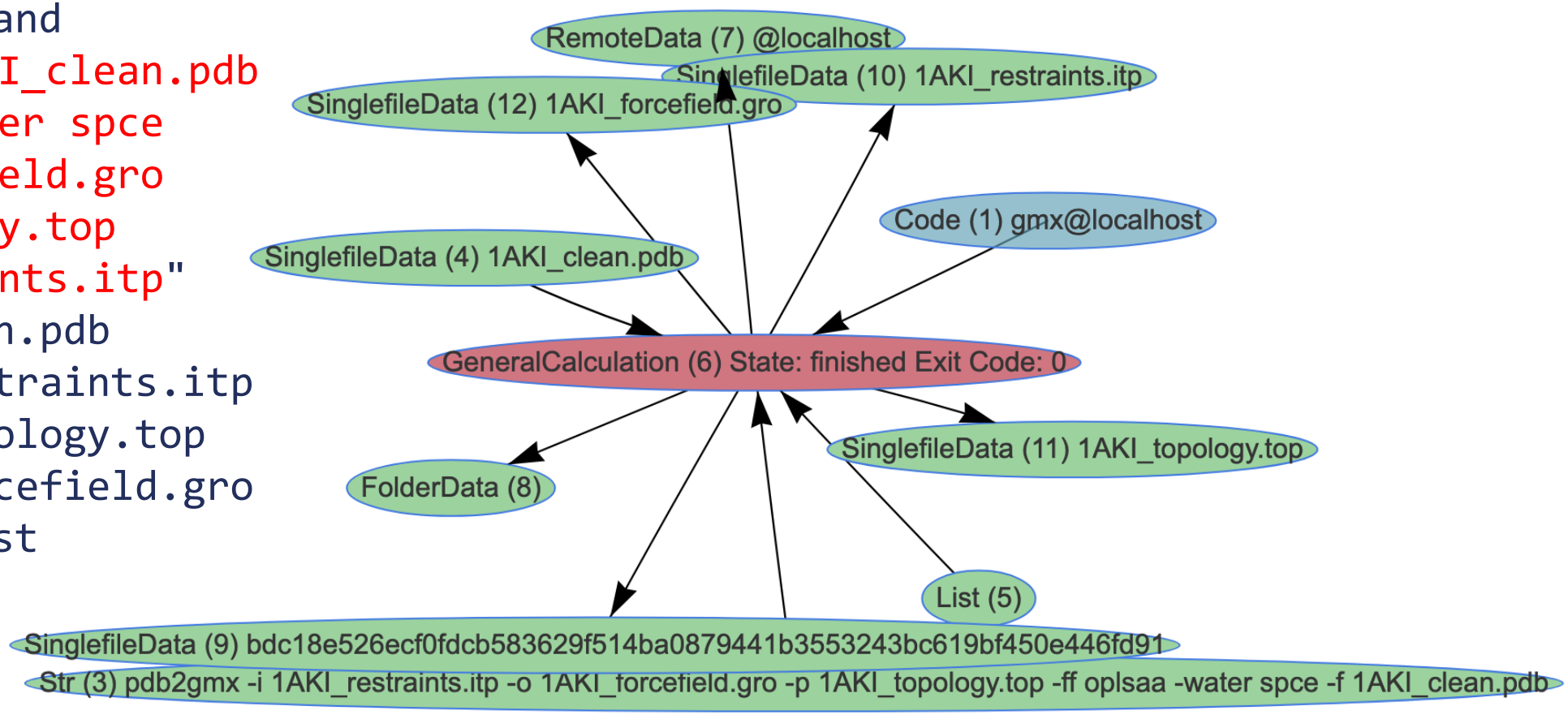


PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

https://aiida-gromacs.readthedocs.io

Example of Data Provenance with any CLI command

```
$ 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

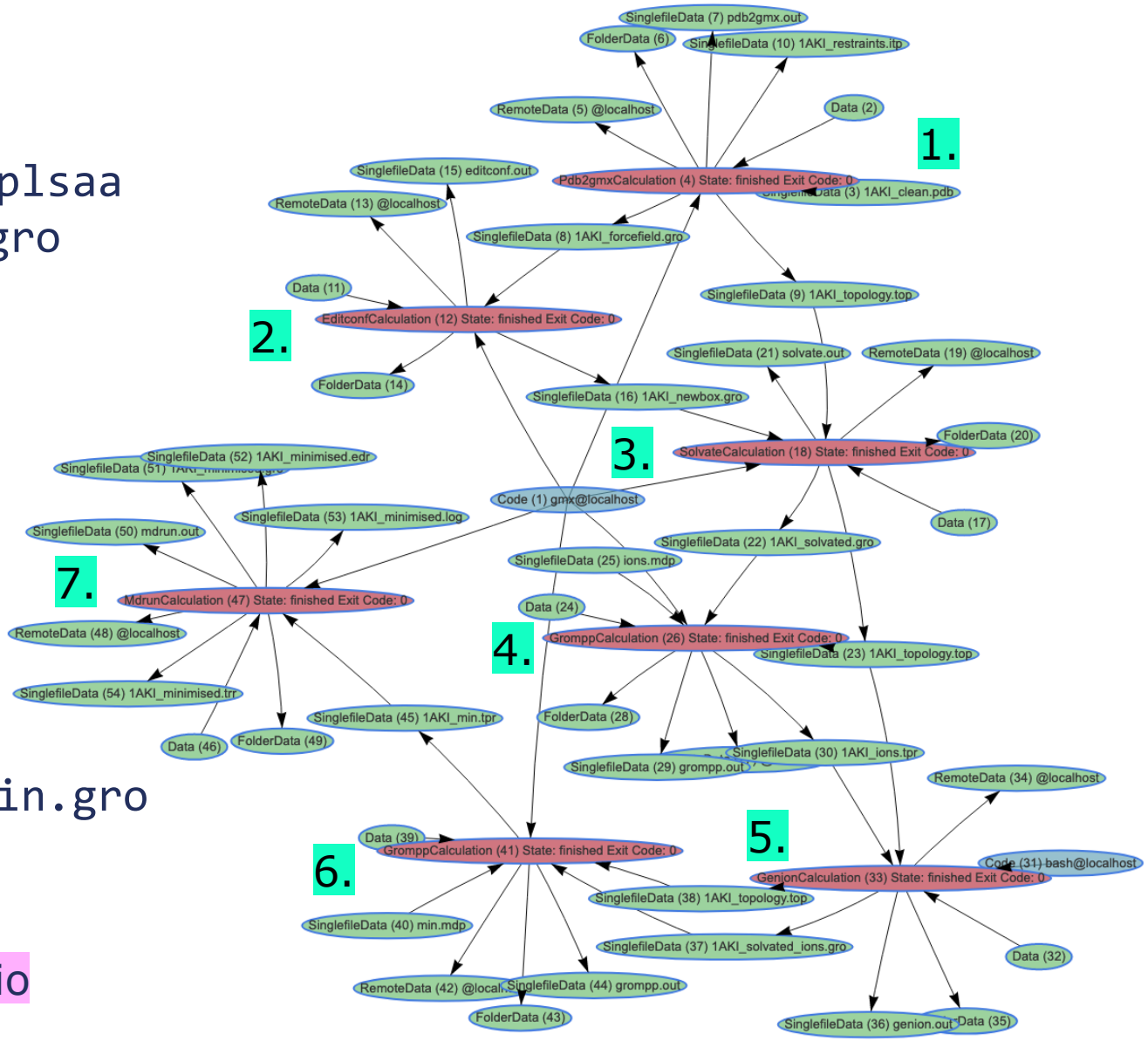
https://aiida-gromacs.readthedocs.io

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
```

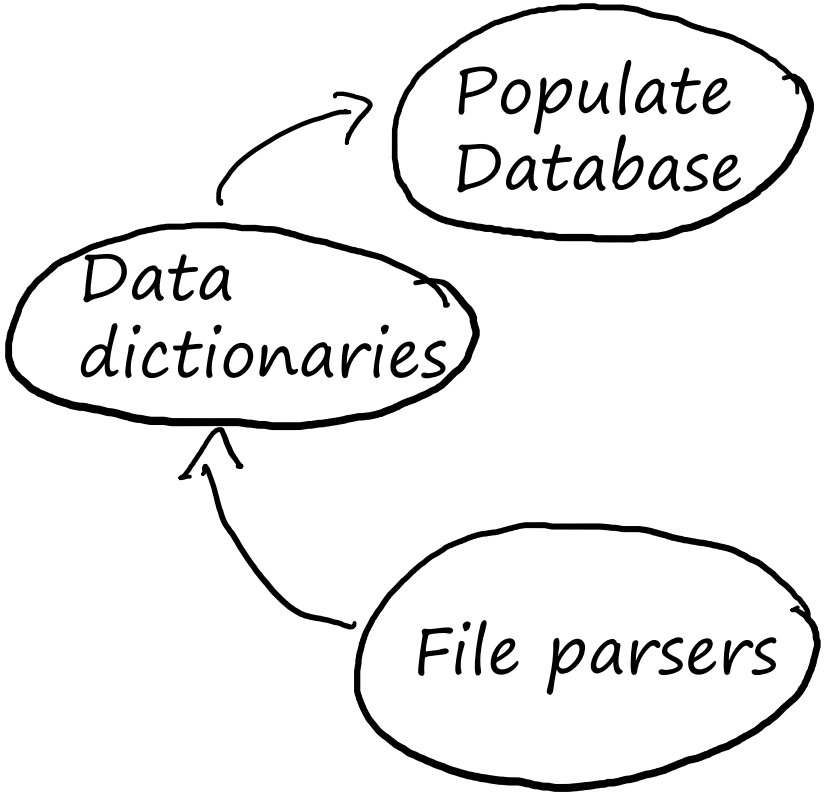
2. gmx_editconf,
3. gmx_solvate,
4. gmx_grompp,
5. gmx_genion,
6. gmx_grompp,

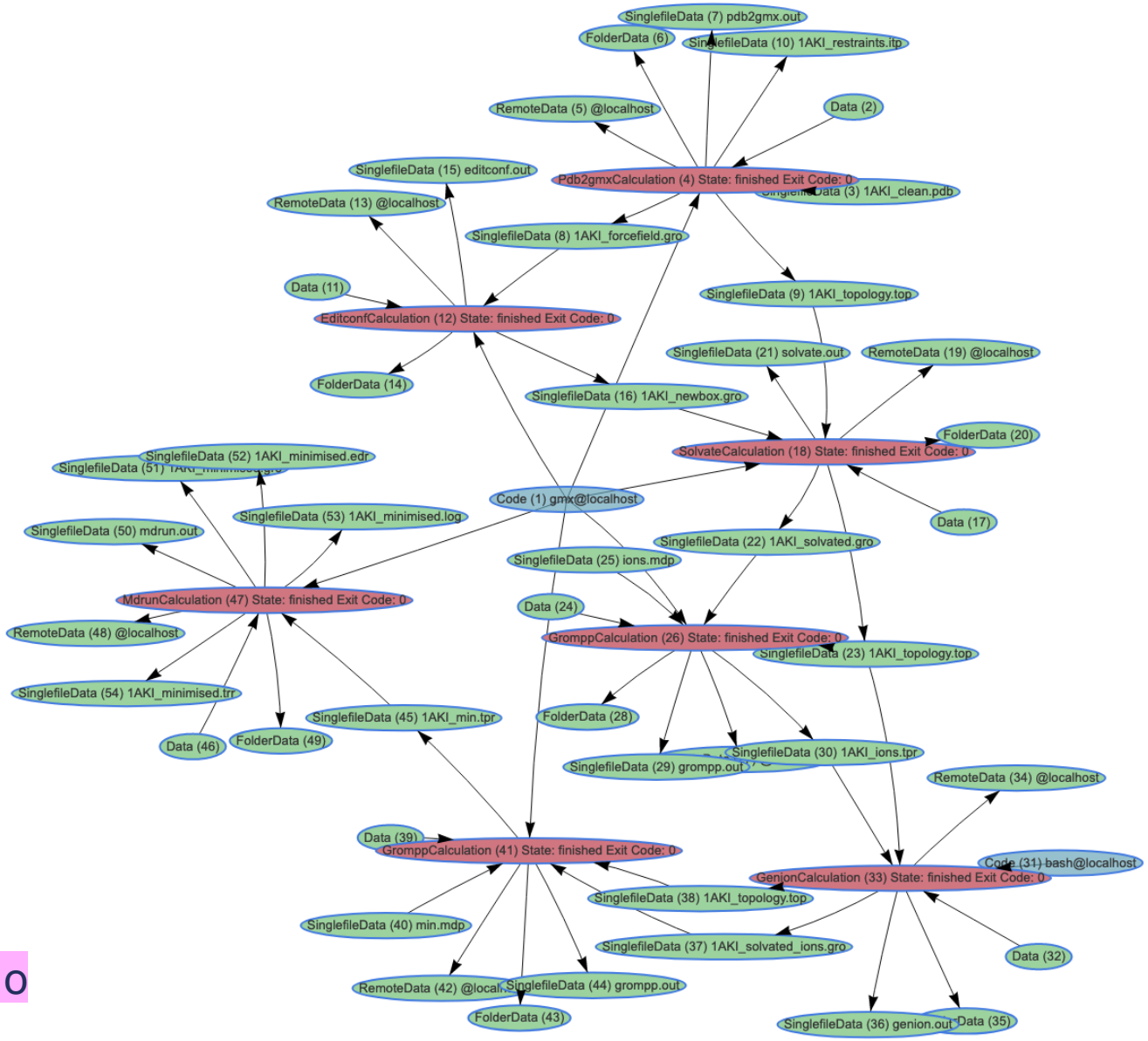
```
7. gmx_mdrun -s 1AKI_min.tpr -c 1AKI_min.gro
       -e 1AKI_min.edr -g 1AKI_min.log
       -o 1AKI_min.trr
```



https://aiida-gromacs.readthedocs.io

Automation of Metadata Extraction

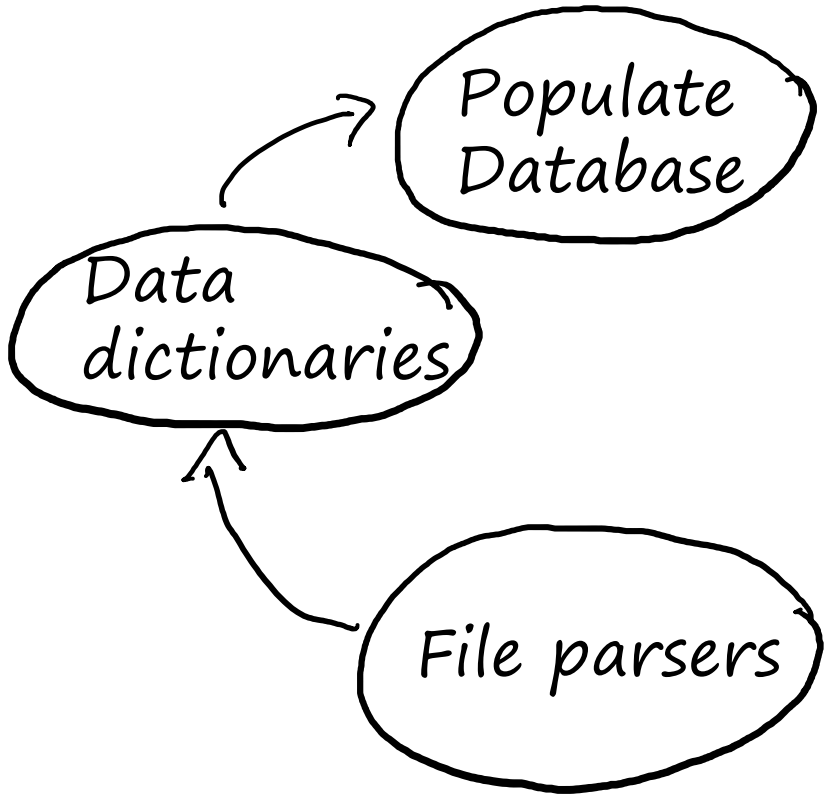




PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE

https://aiida-gromacs.readthedocs.io

Automation of Metadata Extraction



🗑️ Input and log files are usually excluded / discarded,

👀 But offer insight to quality of simulations, input parameters, compute,

🍭 Parsers needed for various file types with no defined formats,

🤝 Consensus needed on metadata standards.

PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE
AMBER MD
PLUMED
MD ANALYSIS
NAMD & VMD Molecular Dynamics Packages
CHARMM Chemistry at HARvard Macromolecular Mechanics
MDDB Molecular Dynamics Data Bank
EMBL-EBI
LAMMPS
GROMACS FAST. FLEXIBLE. FREE.

MD Simulation Input File Examples

Amber

```
Lysozyme NPT equilibration
 &cntrl
  imin=0,         ! No minimization
  ntx=5,          ! Read coords and velocities from coordinate file
  irest=1,        ! Restart simulation
  nstlim=10000,   ! Number of MD steps
  dt=0.002,       ! Time step (ps)
  ntf=2,          ! Bond interactions involving H omitted
  ntc=2,          ! Use SHAKE to constrain bond distances to H
  temp0=300.0,    ! Reference system temperature (K)
  ntpr=5000,      ! Print progress of min steps every ntpr steps
  ntwx=5000,      ! Print coordinates every ntpr steps
  cut=10.0,       ! Non-bonded cutoff value (Angstrom)
  ntb=2,          ! Constant pressure, PBCs for nb interactions
  ntp=1,          ! MD with isotropic position scaling
  ntt=3,          ! Langevin thermostat
  barostat=1,     ! Berendsen barostat
  gamma_ln=2.0,   ! Collision frequency (ps^-1)
  ig=-1,          ! random seed
 /
```

Group together shared terms:

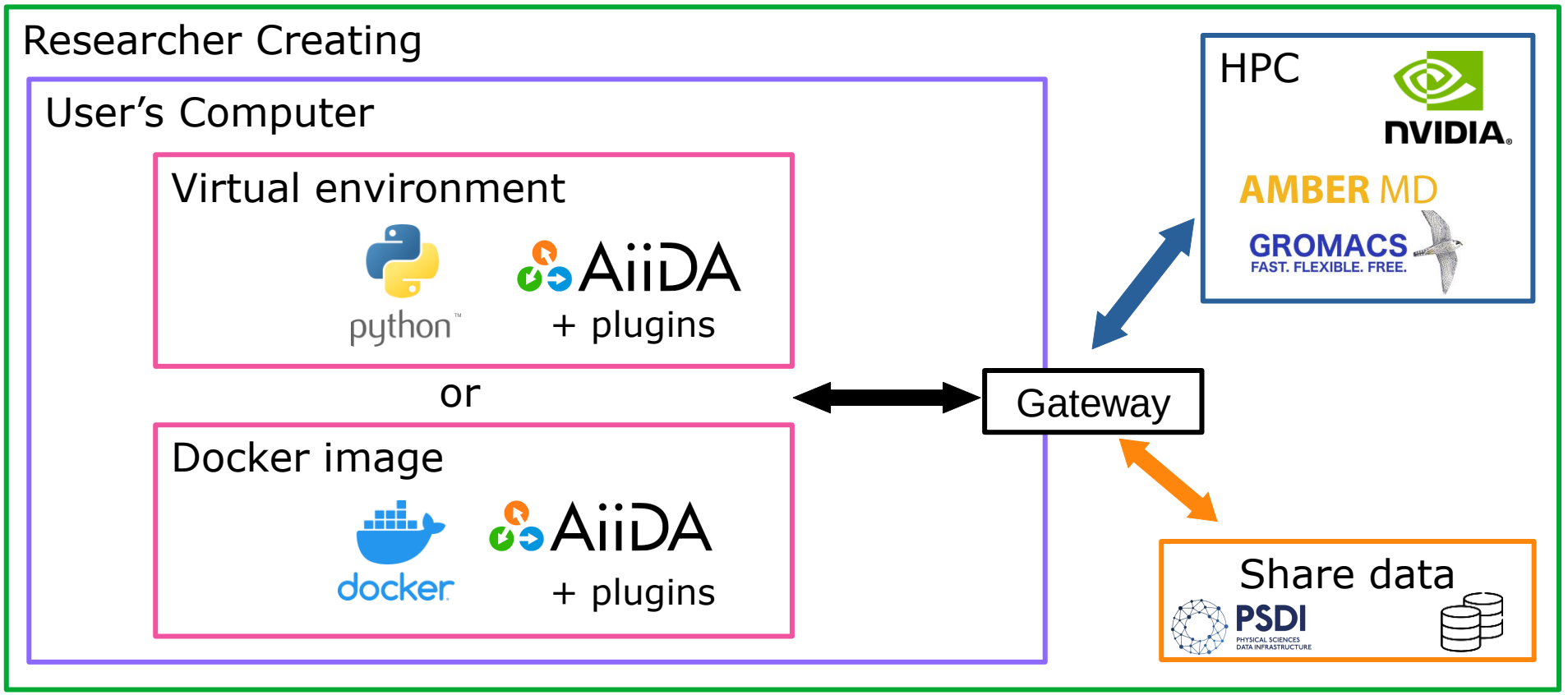
total steps: nstlim, nsteps = 10000
temperature: temp0, ref_t = 300 K
timestep: dt = 0.002 ps
coordinate steps: ntwx, nstxout = 5000
non-bonded cutoffs: cut, rcoulomb, rvdw = 1 nm
...
.
.

GROMACS

```
title                    = OPLS Lysozyme NPT equilibration
; Run parameters
integrator               = md        ; leap-frog integrator
nsteps                   = 10000     ; 2 * 500000 = 1000 ps (1 ns)
dt                       = 0.002     ; 2 fs
; Output control
nstxout                  = 5000      ; suppress bulky .trr file by specifying
nstvout                  = 5000      ; 0 for output frequency of nstxout,
nstfout                  = 5000      ; nstvout, and nstfout
nstenergy                = 5000      ; save energies every 10.0 ps
nstlog                   = 5000      ; update log file every 10.0 ps
; Bond parameters
continuation             = yes       ; Restarting after NPT
constraint_algorithm     = lincs     ; holonomic constraints
constraints              = h-bonds   ; bonds involving H are constrained
lincs_iter               = 1         ; accuracy of LINCS
lincs_order              = 4         ; also related to accuracy
; Neighborsearching
cutoff-scheme            = Verlet    ; Buffered neighbor searching
ns_type                  = grid      ; search neighboring grid cells
nstlist                  = 10        ; 20 fs, largely irrelevant with Verlet scheme
rcoulomb                 = 1.0       ; short-range electrostatic cutoff (in nm)
rvdw                     = 1.0       ; short-range van der Waals cutoff (in nm)
; Electrostatics
coulombtype              = PME       ; Particle Mesh Ewald for long-range electrostatics
pme_order                = 4         ; cubic interpolation
fourierspacing           = 0.16      ; grid spacing for FFT
; Temperature coupling is on
tcoupl                   = V-rescale             ; modified Berendsen thermostat
tc-grps                  = Protein Non-Protein   ; two coupling groups - more accurate
tau_t                    = 0.1     0.1           ; time constant, in ps
ref_t                    = 300     300           ; reference temperature, in K
; Pressure coupling is on
pcoupl                   = Parrinello-Rahman     ; Pressure coupling on in NPT
pcoupltype               = isotropic             ; uniform scaling of box vectors
tau_p                    = 2.0                   ; time constant, in ps
ref_p                    = 1.0                   ; reference pressure, in bar
```

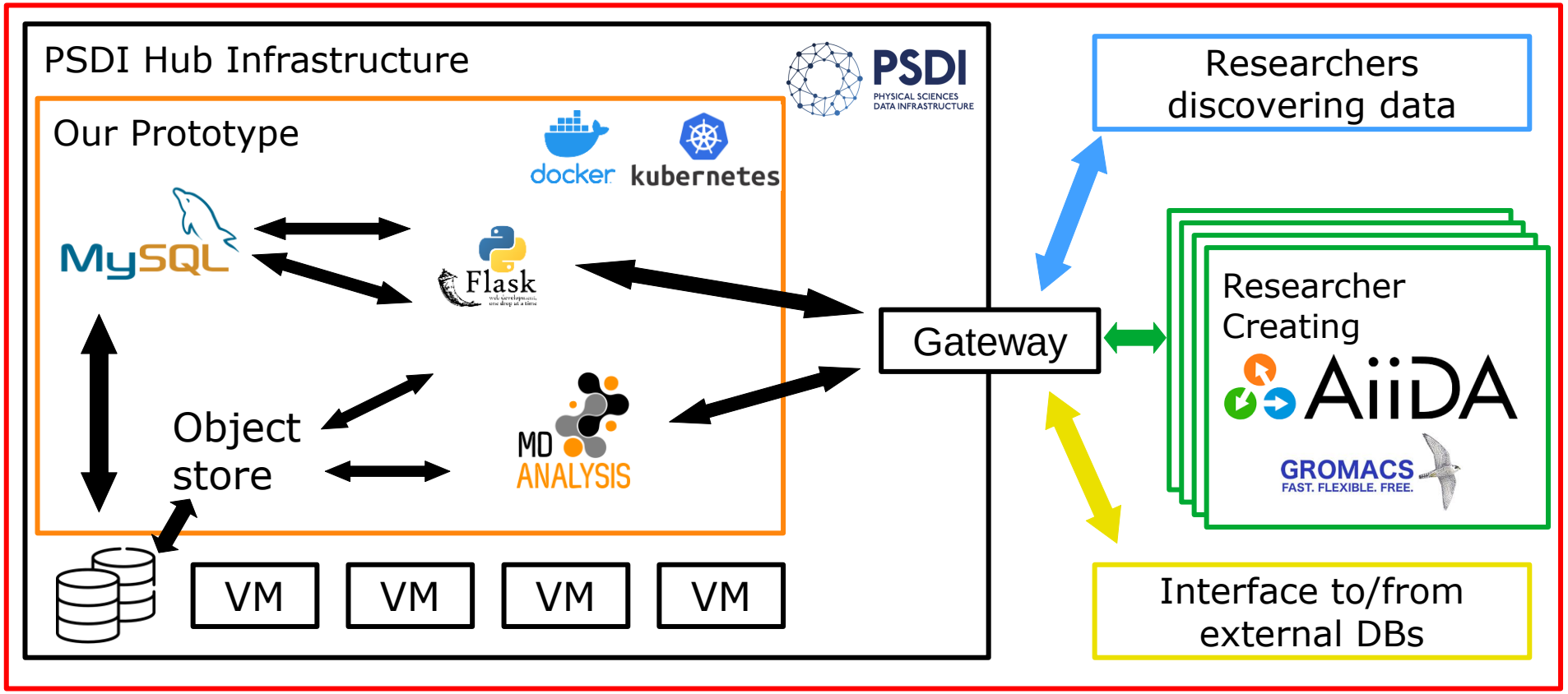
PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Our User Environment Prototype



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Our Infrastructure Prototype



Thank you for your attention!

Acknowledgments to:

Kin Chao, Sarah Rouse

Joel Greer, Tom Burnley, Martyn Winn

James Gebbie-Rayet (PI)

PSDI acknowledges the funding support by the EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1