FAIR Molecular Dynamics

Adrián Rošinec

CEITEC CF Biodata

adrian@muni.cz

Why?

- MD has evolved in last decades
- System size by a factor of 10^9
- Trajectory length by a factor of 10^9
- Ensemble size increased by a factor of 10^11
- From dozens of groups to thousands
  - Approx. 15% of all HPC use dedicated to MD

MD matured, but…

in terms of data management behind other fields

- Simulation efforts are lost
- No build-up on existing research
- No metanalysis possible
- Lack of reproducibility and quality checks
- Integration of AI / ML methods
- Poor interaction with other fields

ELIL5: FAIR

- Principles on using and sharing (scientific) data
- FAIR
  - **Findable**: easy and transparent to search data
  - **Accessible**: clear rules how to access the data
  - **Interoperable**: common file formats, software inputs/outputs
  - **Reusable**: data usable for more purposes

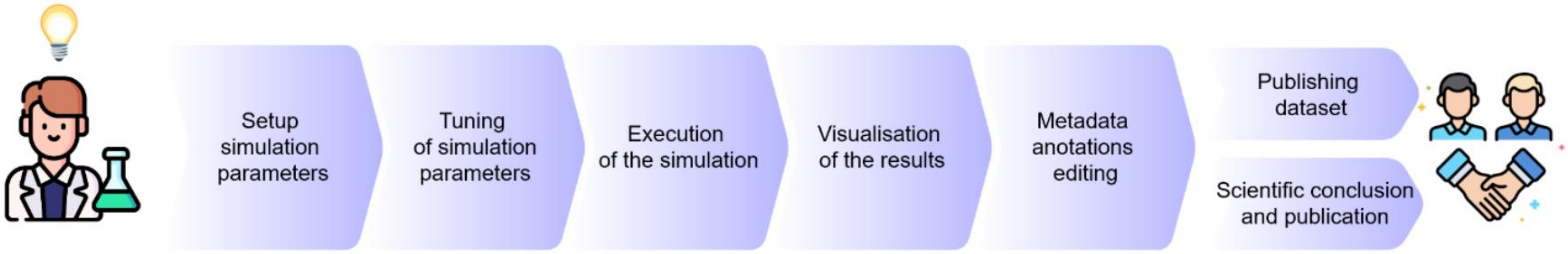
More at: https://www.go-fair.org/fair-principles/

Challenges

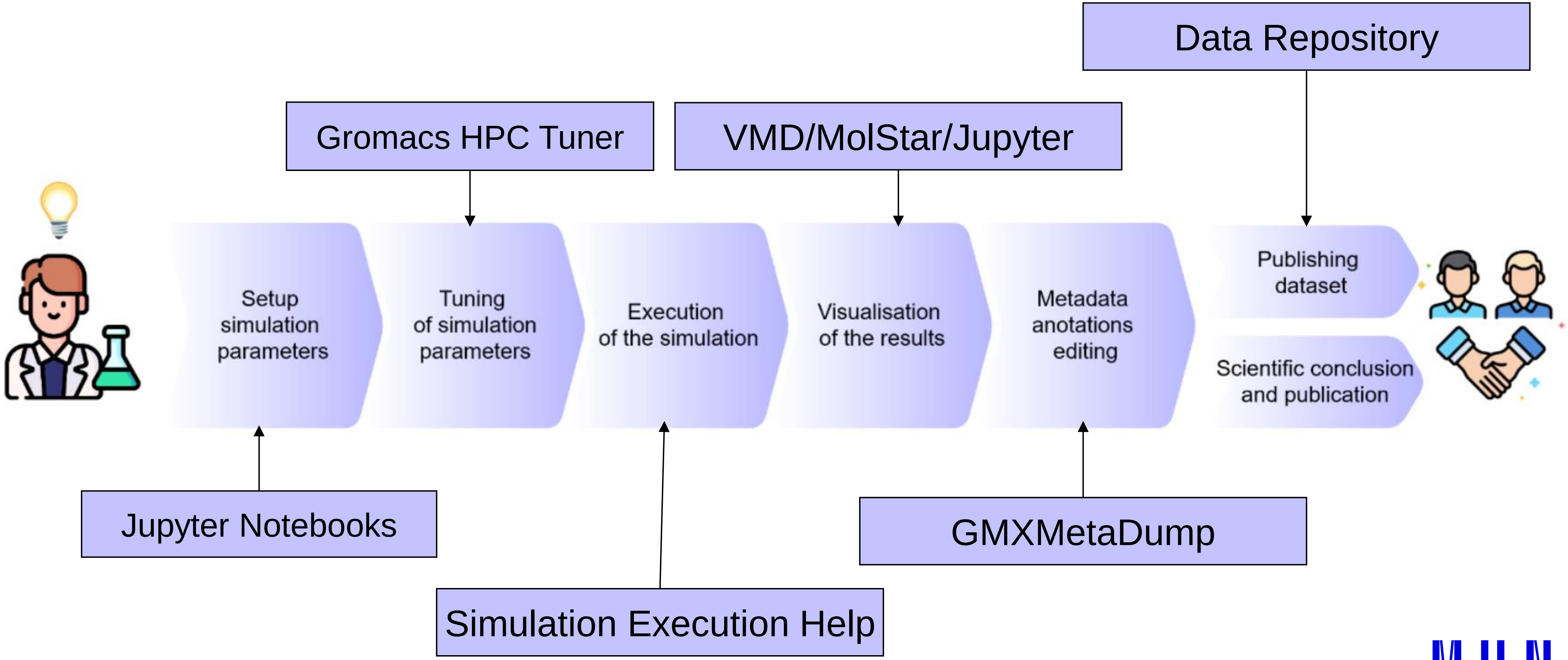
- Standard MD data exchange formats
  - Trajectory and traj. compression formats
  - Trajectory identification – atom/residue names
  - Full simulation settings-parameters of sim.
- Establish metadata ontologies and semantics
  - Search based on contents (biomolecules, …) / parameters (thermal, …) / purpose of simulation
- Provenance – how was trajectory generated, hashes of files, steps to create trajectory, attached additional files
  - Custom named / missing residues, non-standard force-fields or molecules,
- Quality control mechanisms and metrics
- Sharing – PIDs, community repositories
- Cost of storing large amounts of data

Broader context

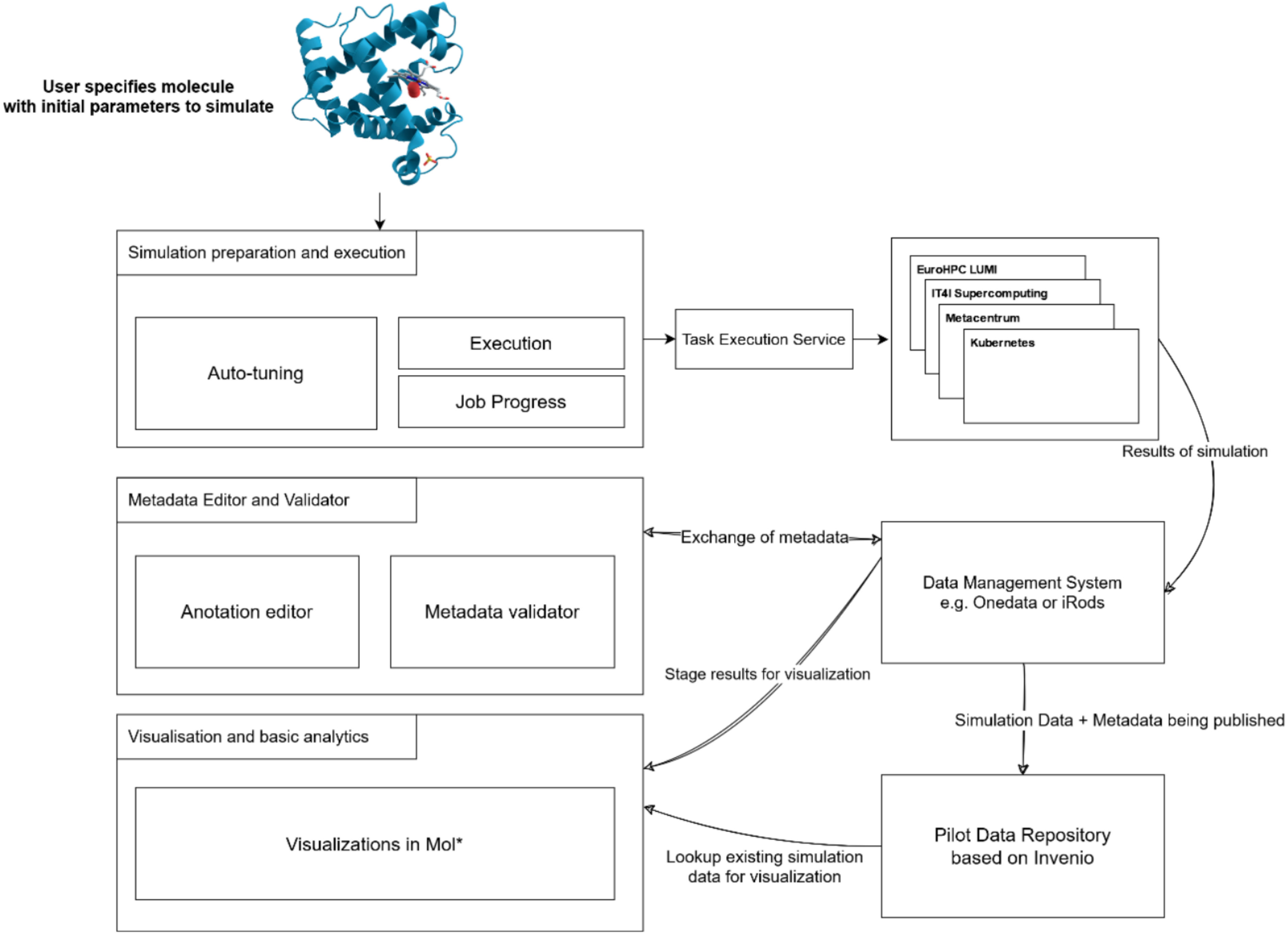
- The typical MD scientific process


Setup simulation parameters
Tuning of simulation parameters
Execution of the simulation
Visualisation of the results
Metadata anotations editing
Publishing dataset
Scientific conclusion and publication

Broader context



Data Repository
Gromacs HPC Tuner
VMD/MolStar/Jupyter
Setup simulation parameters
Tuning of simulation parameters
Execution of the simulation
Visualisation of the results
Metadata anotations editing
Publishing dataset
Scientific conclusion and publication
Jupyter Notebooks
GMXMetaDump
Simulation Execution Help



User specifies molecule with initial parameters to simulate

Simulation preparation and execution

Auto-tuning

Execution

Job Progress

Task Execution Service

EuroHPC LUMI

IT4I Supercomputing

Metacentrum

Kubernetes

Results of simulation

Metadata Editor and Validator

Anotation editor

Metadata validator

Exchange of metadata

Data Management System e.g. Onedata or iRods

Stage results for visualization

Simulation Data + Metadata being published

Visualisation and basic analytics

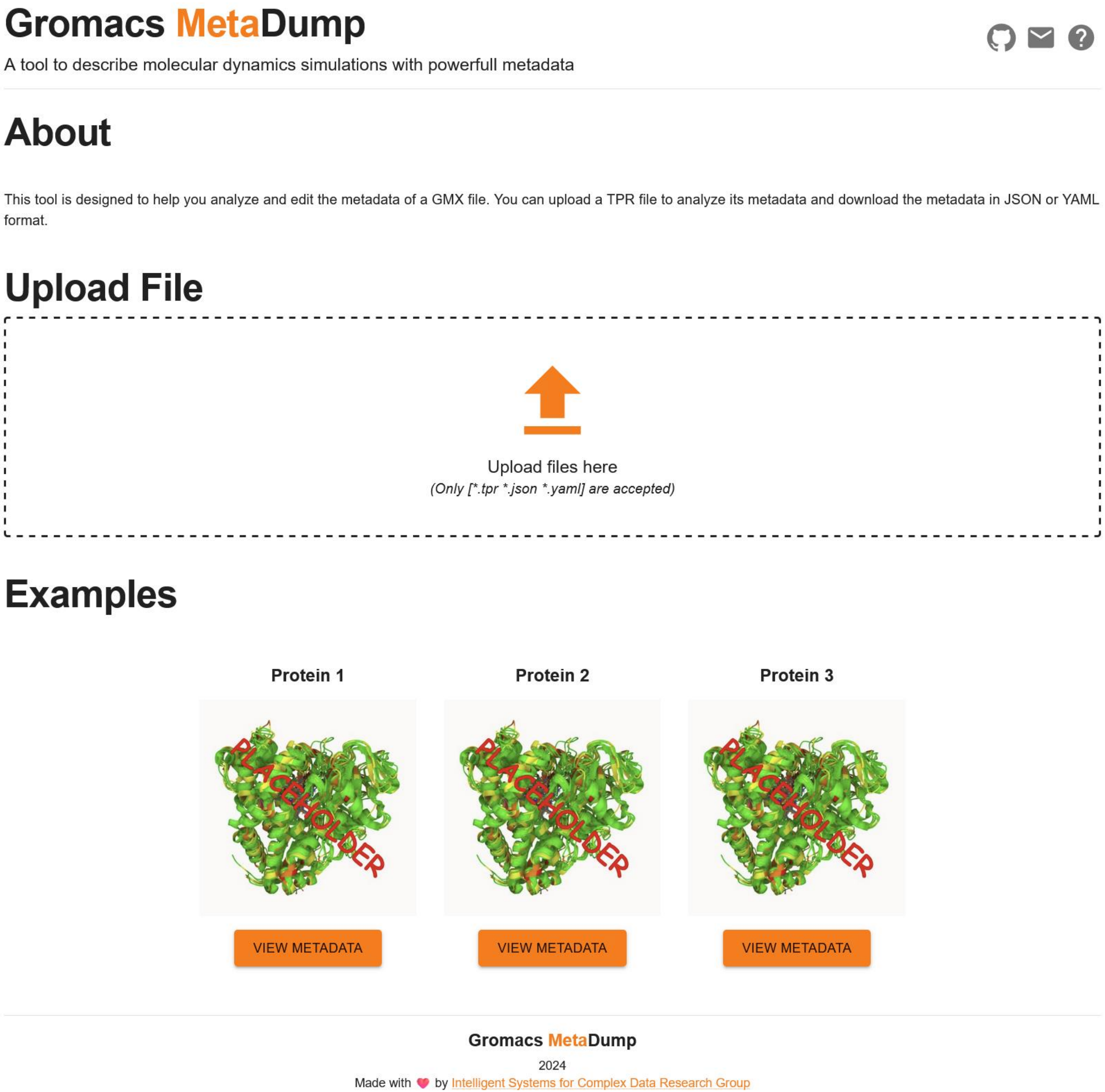
Visualizations in Mol*

Lookup existing simulation data for visualization

Pilot Data Repository based on Invenio

“FAIRification” tools

- Automatization
- Remove burden of manual annotations
- Ensuring completeness
  - Automatic harvesting from Gromacs/… file formats
  - Linking to the biomolecule databases (e.g. PDB)
  - Administrative metadata such as publishing institution, authors, funding information, …
- Validations – quality control
  - Meeting repositories and community requirements for accepting datasets
  - e.g. has subject of simulation – biomolecule, has environmental conditions (temp, press), has used force field + attached custom FF
  - Based on ontologies



Gromacs MetaDump

A tool to describe molecular dynamics simulations with powerfull metadata

About

This tool is designed to help you analyze and edit the metadata of a GMX file. You can upload a TPR file to analyze its metadata and download the metadata in JSON or YAML format.

Upload File

Upload files here

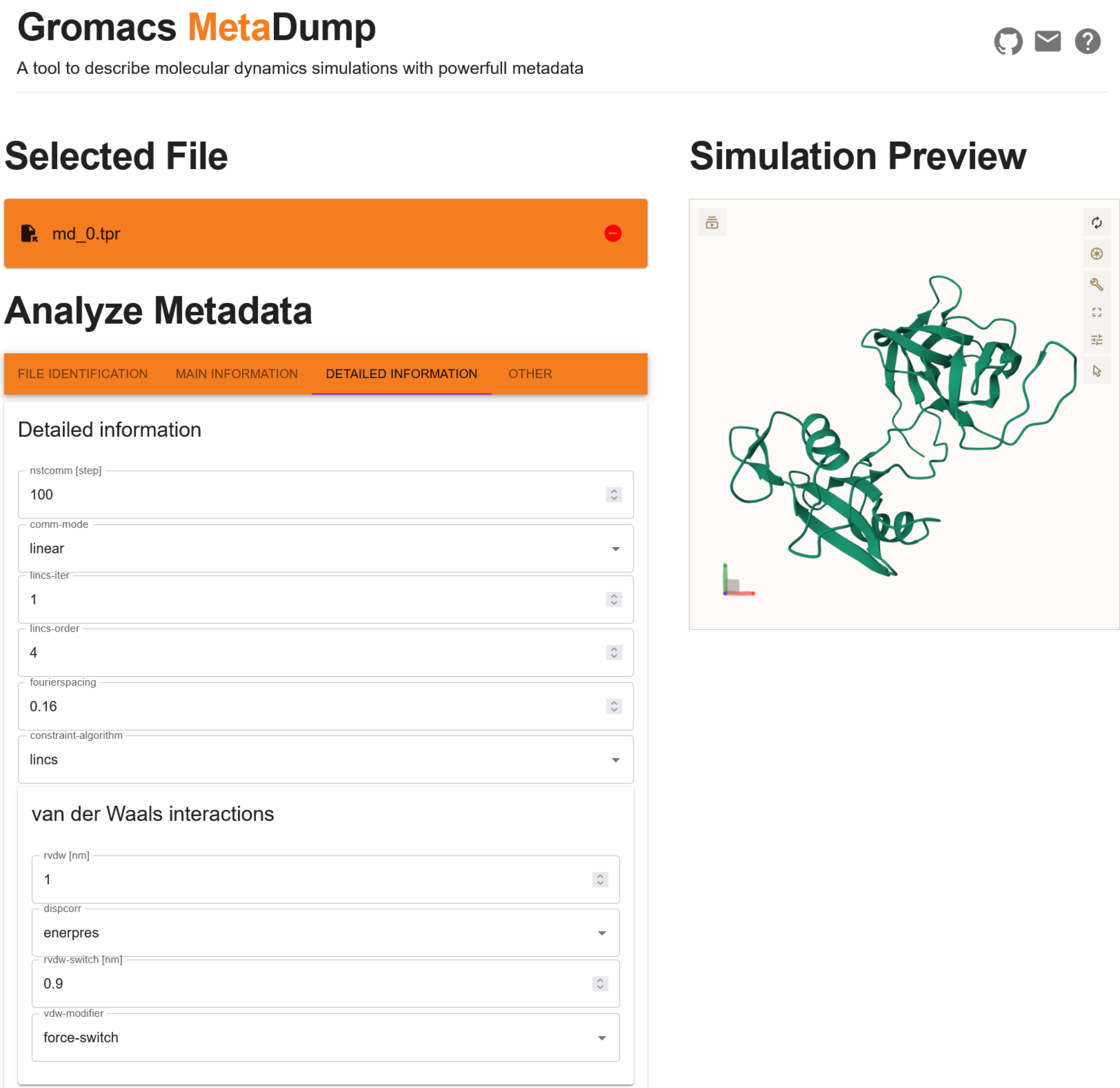
(Only [.tpr *.json *.yaml] are accepted)*

Examples

Gromacs MetaDump

2024

Made with ❤ by Intelligent Systems for Complex Data Research Group



Gromacs MetaDump

A tool to describe molecular dynamics simulations with powerfull metadata

Selected File

md_0.tpr

Simulation Preview

Analyze Metadata

FILE IDENTIFICATION | MAIN INFORMATION | DETAILED INFORMATION | OTHER

Detailed information

nstcomm [step]: 100

comm-mode: linear

lincs-iter: 1

lincs-order: 4

fourierspacing: 0.16

constraint-algorithm: lincs

van der Waals interactions

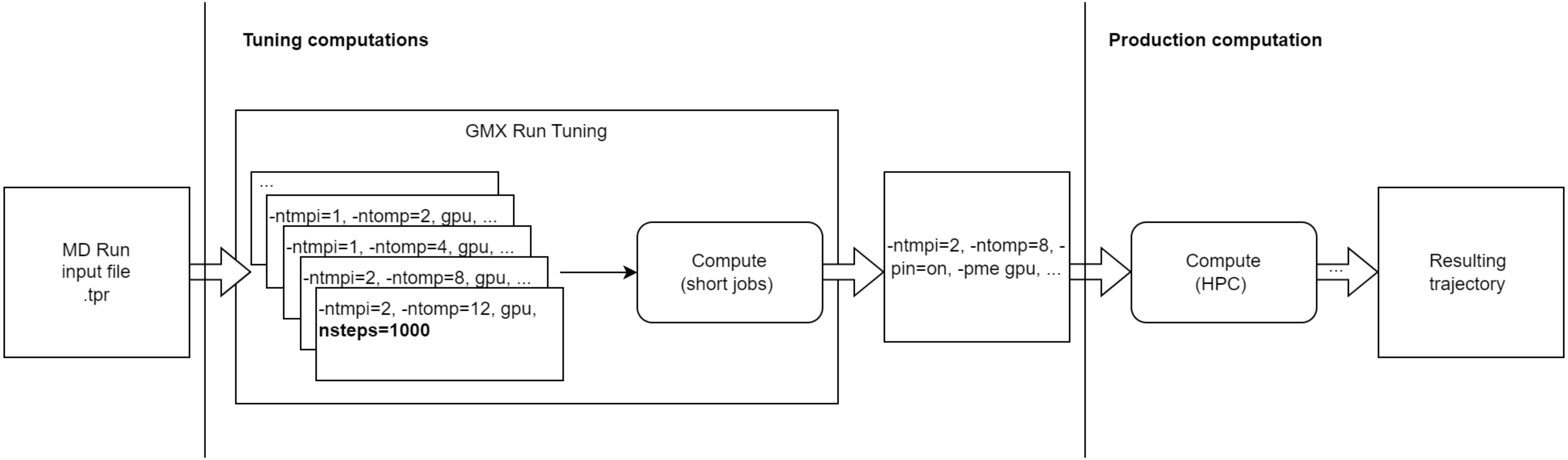
rvdw [nm]: 1

dispcorr: enerpres

rvdw-switch [nm]: 0.9

vdw-modifier: force-switch

Tuning and execution

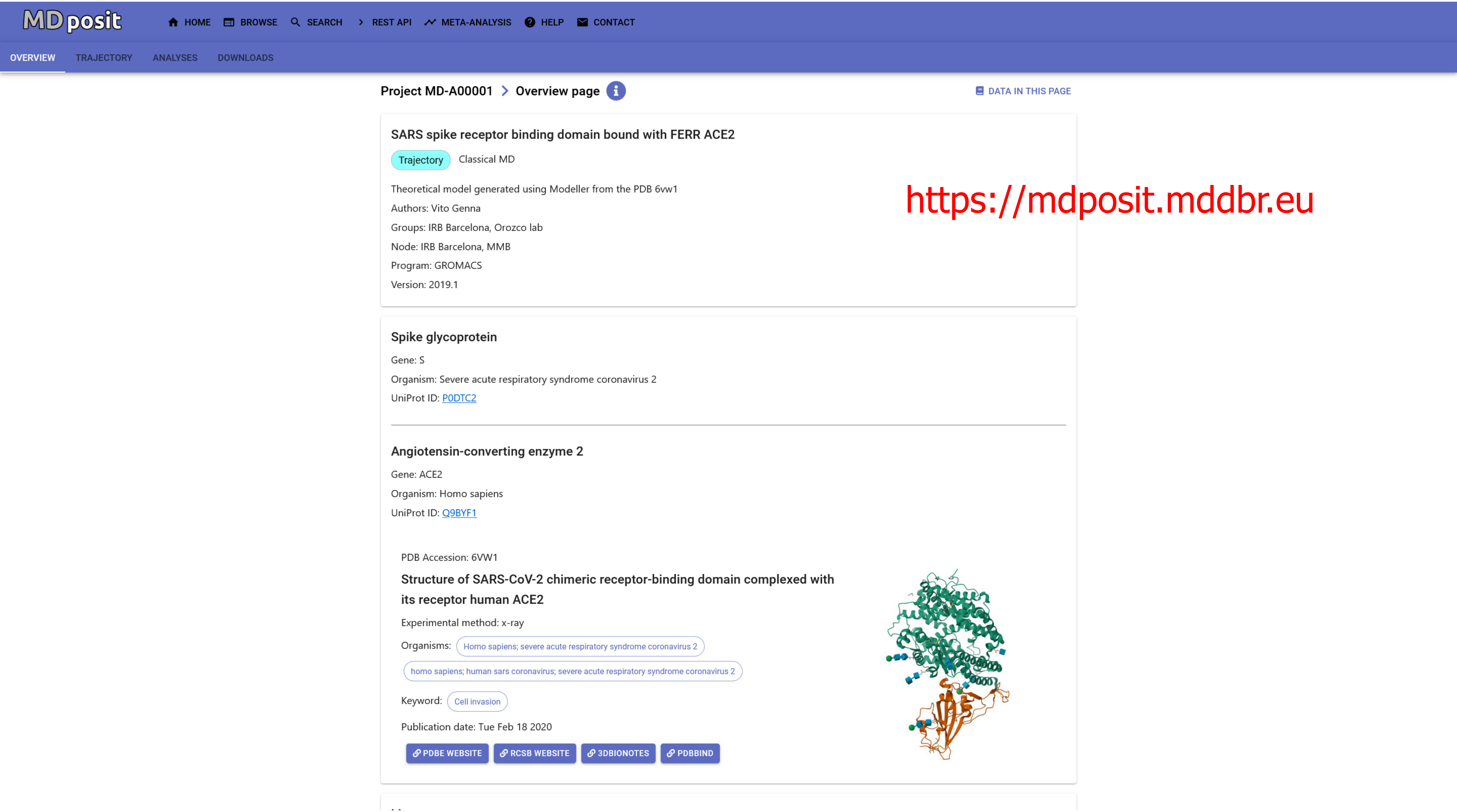

Tuning computations
Production computation
MD Run
input file
.tpr
GMX Run Tuning
...
-ntmpi=1, -ntomp=2, gpu, ...
-ntmpi=1, -ntomp=4, gpu, ...
-ntmpi=2, -ntomp=8, gpu, ...
-ntmpi=2, -ntomp=12, gpu,
nsteps=1000
Compute
(short jobs)
-ntmpi=2, -ntomp=8, -pin=on, -pme gpu, ...
Compute
(HPC)
...
Resulting
trajectory

Experiment Tuning and Exec

- Interactive experiment setup
  - Choosing biomolecule, Equilibration, sampling process, setting simulation parameters
- (Auto)Tuning production run parameters
  - Avoiding common mistakes:
  - e.g. when running in paralell slows the computation
  - too small time-steps, not exploiting GPU or overexploiting them, right force fields, not minimized/equilibrated system, and many more
  - Running several small jobs (several seconds per job) to find better parameters before production run (several weeks)
- Remove burden of interacting with various computing interfaces
  - k8s, batch systems, …
- Provenance/Protocols – capture operations used to get trajectories
- Controlled => Simple => Correct

EU Efforts on building MDREPO

- Federated architecture
  - Method of sustainable funding to hold increments of tens of PBs per year
- Central component is metadata catalogue
  - One (features rich) interface to browse all MD data
  - Enables findability of datasets and points to “downstream” storage a.k.a repository (several nodes within nation, one national, …)
- Implemented quality control mechanisms
- Built as a custom solution
  - Is the CZ variant of repo platform (within NDI) more sustainable?
  - Is it worth extend functionality of CZ repo platform?
  - …



https://mdposit.mddbr.eu

Project MD-A00001 > Overview page

SARS spike receptor binding domain bound with FERR ACE2

Trajectory Classical MD

Theoretical model generated using Modeller from the PDB 6vw1

Authors: Vito Genna

Groups: IRB Barcelona, Orozco lab

Node: IRB Barcelona, MMB

Program: GROMACS

Version: 2019.1

Spike glycoprotein

Gene: S

Organism: Severe acute respiratory syndrome coronavirus 2

UniProt ID: P0DTC2

Angiotensin-converting enzyme 2

Gene: ACE2

Organism: Homo sapiens

UniProt ID: Q9BYF1

PDB Accession: 6VW1

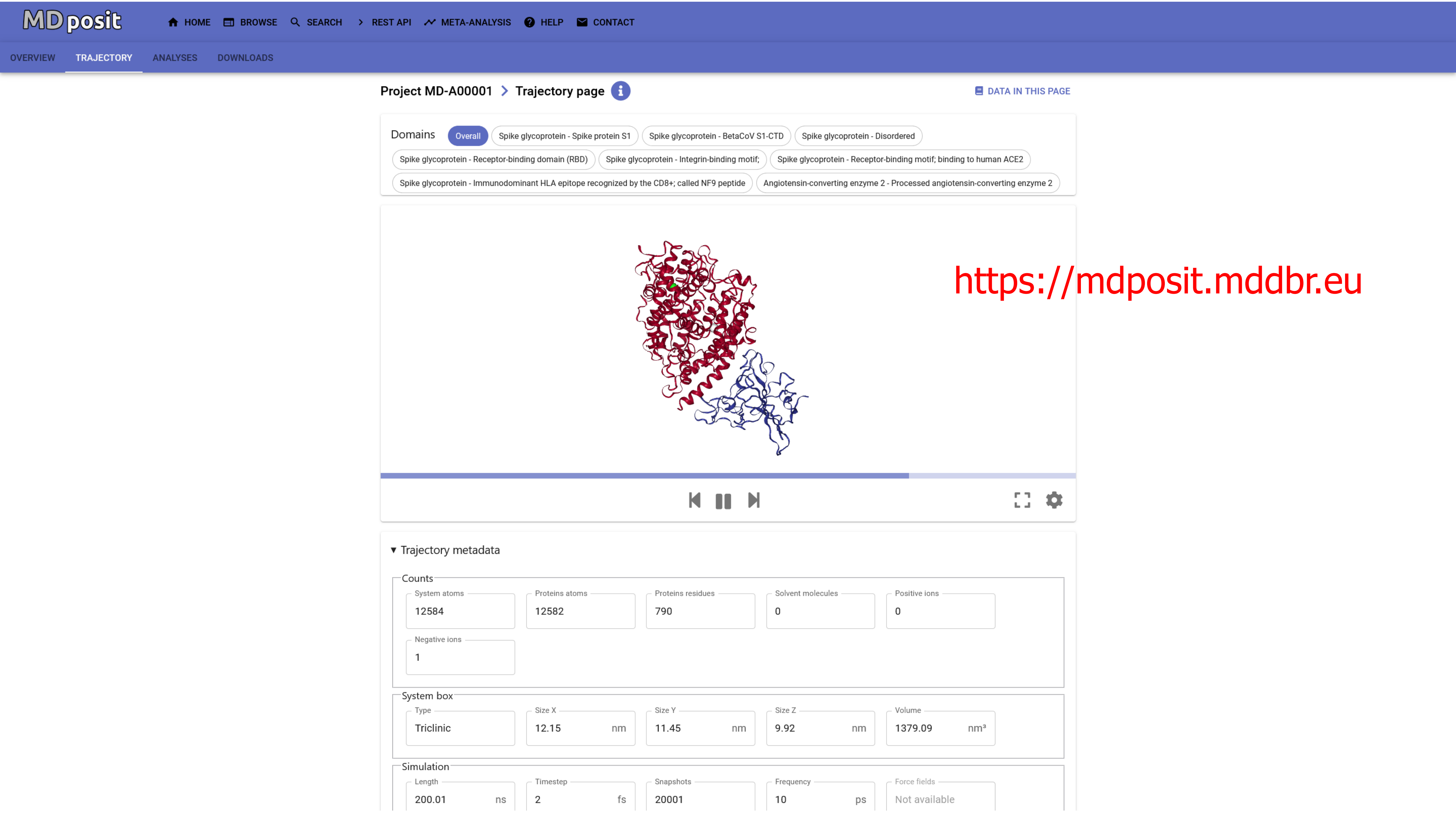
Structure of SARS-CoV-2 chimeric receptor-binding domain complexed with its receptor human ACE2

Experimental method: x-ray

Organisms: Homo sapiens; severe acute respiratory syndrome coronavirus 2 | homo sapiens; human sars coronavirus; severe acute respiratory syndrome coronavirus 2

Keyword: Cell invasion

Publication date: Tue Feb 18 2020



https://mdposit.mddbr.eu

Project MD-A00001 > Trajectory page

Domains: Overall | Spike glycoprotein - Spike protein S1 | Spike glycoprotein - BetaCoV S1-CTD | Spike glycoprotein - Disordered | Spike glycoprotein - Receptor-binding domain (RBD) | Spike glycoprotein - Integrin-binding motif; | Spike glycoprotein - Receptor-binding motif; binding to human ACE2 | Spike glycoprotein - Immunodominant HLA epitope recognized by the CD8+; called NF9 peptide | Angiotensin-converting enzyme 2 - Processed angiotensin-converting enzyme 2

▾ Trajectory metadata

Counts

System atoms: 12584

Proteins atoms: 12582

Proteins residues: 790

Solvent molecules: 0

Positive ions: 0

Negative ions: 1

System box

Type: Triclinic

Size X: 12.15 nm

Size Y: 11.45 nm

Size Z: 9.92 nm

Volume: 1379.09 nm³

Simulation

Length: 200.01 ns

Timestep: 2 fs

Snapshots: 20001

Frequency: 10 ps

Force fields: Not available

National Node of MDREPO

- Need to join the EU activity and build national node
- WIP mdrepo.eu
- Several groups publishing MD results
  - Robert Vácha at CEITEC, Vojtěch Spiwok at UOCHB, Michal Kolář at VŠCHT, Karel Berka at UPOL, …
- Need for (shared) data curator, quality control, support (steward) and tools
- Storage for CZ MD data (PBs/year)
- Publishing to the metadata catalogues – EU MDDB node

Sources

- https://mddbr.eu/first-mddb-webinar-recording-now-online/
- http://arXiv:2407.16584
- Own project proposal