Curate and Share Complex Biomolecular Simulation Workflows

Jas Kalayan[1], Kin Chao[1,2], Sarah Rouse[2], James Gebbie-Rayet [1]
[1]Scientific Computing Department, Science and Technology Facilities Council, Daresbury Laboratory, Warrington, United Kingdom
[2]Department of Life Sciences, Imperial College London, London, United Kingdom
Correspondence: jas.kalayan@stfc.ac.uk, james.gebbie@stfc.ac.uk

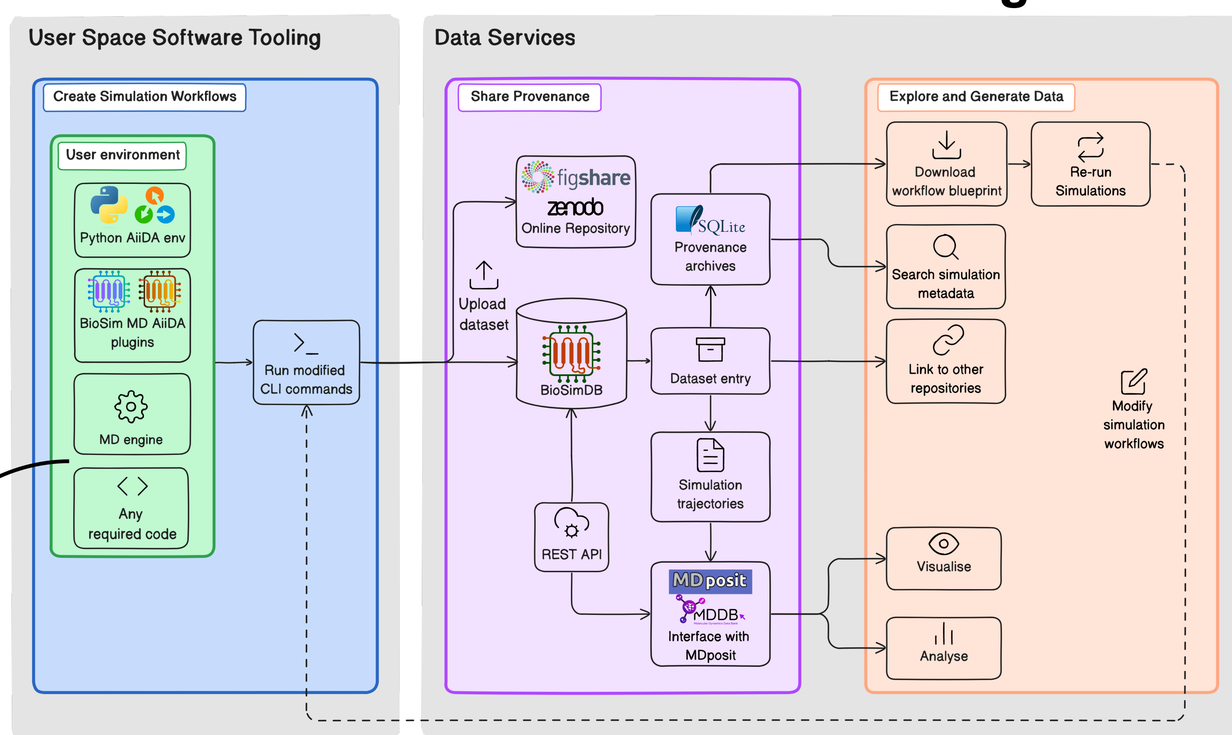
Project Aim
Use provenance tools to save all the steps taken to setup a simulation and share protocols via an open online database (BioSimDB).

Motivations
- Difficult to reproduce published research.
- No centralised databases for biomolecular simulation trajectories.
- Uneven computational resources for research groups across the world.
- No consensus on how simulation data should be stored.

Goals
- Make barriers to provenance capture as low as possible.
- Incorporate feedback from the biomolecular simulation community
- Align to FAIR principles [1].
- Reduce duplication of computational and scientific effort.
- Accelerate knowledge exchange with other domains (e.g. ML).

User tools being developed to collect MD provenance:
- aiida-gromacs [2]
- aiida-amber [3]
- aiida-namd
- aiida-genericMD

Vision for Biomolecular Simulation Data Provenance Collection and Storage



The aiida-gromacs plugin for AiiDA

`aiida-gromacs` is available at http://github.com/PSDI-UK/aiida-gromacs

- User guide
  - Installation
  - Start/Stop AiiDA
  - AiiDA CLI Cheatsheet
  - Plugin on the Commandline
  - Scripting with the Plugin
- Tutorials
  - Lysozyme tutorial
  - genericMD CLI
- Developer guide
  - Clone the repository
  - Running the tests
  - Automatic coding style checks
  - Continuous integration
  - Building the documentation
  - Putting it all together
  - Sending code contributions
  - Version Numbering
- API documentation
  - Subpackages
  - Submodules
  - aiida_gromacs.helpers module
  - Module contents

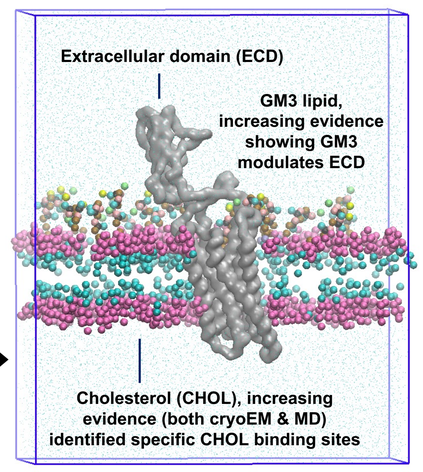
Case Study → Towards Building a GPCR Lipid Fingerprint Dataset

Research Aim
Investigate lipid binding sites and how they influence protein dynamics using coarse-grained (CG) MD simulations of B1 GPCRs in active and inactive states.

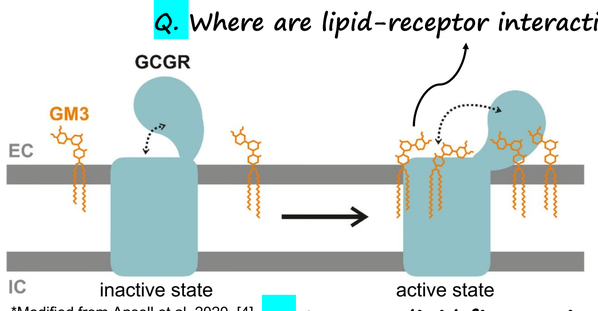
Motivations
- GLP-1R together with GIPR are largely **responsible for the secretion of insulin after eating**.
- **Emerging T2D drugs** Mounjaro and Wegovy are agonist for GLP-1R.
- Allosteric modulators allow for more refined spatiotemporal and/or pathway-specific. **signal modulation**, we are interested in **identifying allosteric site/ lipid modulations.**

Goals
- CG-MD to **identify lipid binding sites** in B1 GPCRs, study lipid modulation of extracellular domain (ECD), Post-Translational Modifications (PTM) and oligomerisation behaviours.
- **Collaborate with experimentalists** (Alejandra Tomas Lab at ICL Medicine) that specialise in various GPCR functional assays using a GM3 inhibitor (Eliglustat) to characterise *in vitro* the effect of the GM3 lipid on conformational dynamics of GLP-1R, GGCR and GIPR.
- **Create a B1 GPCRs dataset – stress-test aiida-gromacs and show importance of MD provenance,** allowing others to refine protocols and extend our dataset to other GPCR classes.

CG-MD setup using the latest Martini 3 GM3 [5] and CHOL lipids [6]. Diagram shows GM3 and CHOL lipids only in the membrane.

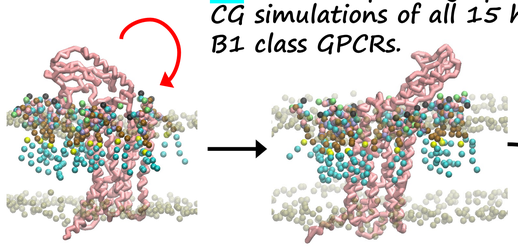


Q. Where are lipid-receptor interactions occurring?

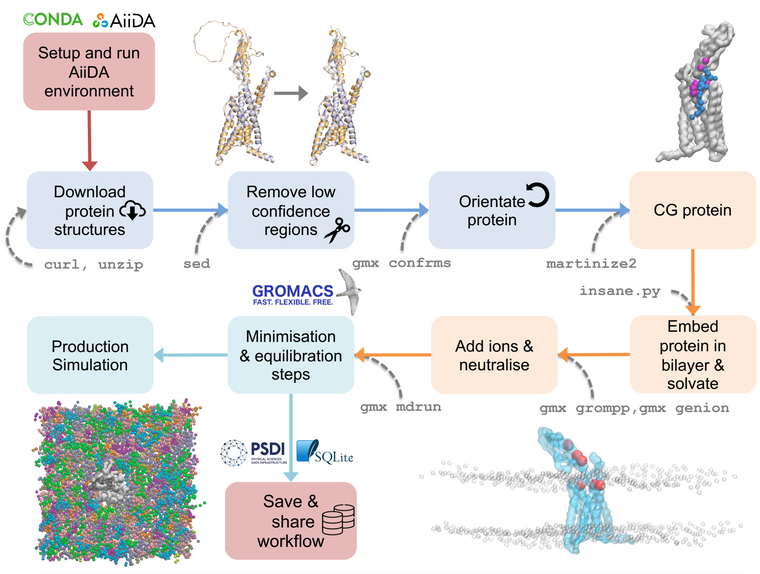


*Modified from Ansell et al, 2020. [4]

A. Create a lipid fingerprint from CG simulations of all 15 human B1 class GPCRs.



Simulation Workflow



Schematic example of the B1 GPCR simulation workflow steps captured when using aiida-gromacs. Orange boxes represent steps involved in enabling provenance capture and sharing, blue boxes represent steps for processing the protein structure, green boxes represent parametrisation and building of the molecular system, and purple boxes represent MD simulation steps. The CLI tools used in each workflow step are shown in grey font.

Notebook Tutorial

Advanced Data Provenance with aiida-gromacs

Who is this tutorial for?

This tutorial is aimed at molecular dynamics simulators who want to keep track of each step used to build and setup simulations of biomolecular systems.

Coarse-grained MD Simulation Data Provenance with aiida-gromacs

In this tutorial, we will go through how to set-up a coarse-grained molecular system of a PTH2R (Parathyroid hormone receptor type 2) protein embedded in a lipid bilayer membrane along with water and counter-ions. We will use the command-line tools provided in aiida-gromacs to track each step performed on the terminal.

Aquiring and tidying up the receptor protein structure

Our starting point is the PTH2R structure from the GPCRdb.

1. First, we download the PTH2R protein from GPCRdb using `curl`. We will track our commands as we perform them using `genericMD` command in aiida-gromacs:

run modified commands

```
! genericMD --code bash@localhost \
--command "curl https://gpcrdb.org/structure/homology_models/pth2r_human_active_full/download_pdb -o {PTH2R}.zip
--outputs {PTH2R}.zip
```

Saving and viewing all the steps used to build the system

We have built our starting configuration of an embedded protein in a lipid bilayer, hurray!

We can check again that all our commands ran successfully using `verdi`:

check commands

```
! verdi process list -a
```

```
  PK  Created    Process label        Process State    Process status
----  ---------  -------------------  ---------------  ----------------
   4  15m ago    GenericCalculation   ■ Finished [0]
```

display provenance

```
! verdi data provenance show
```

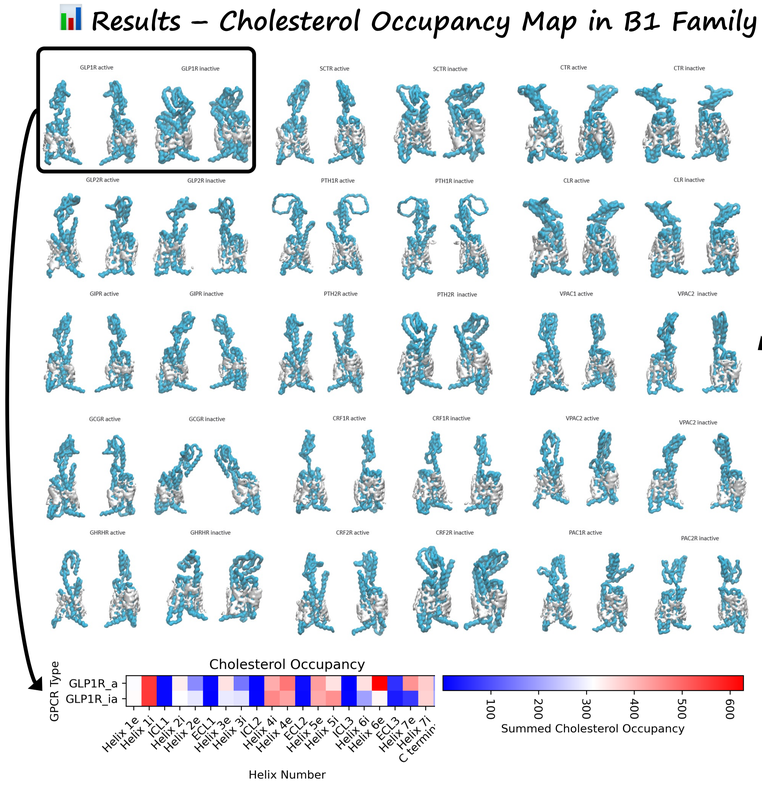
```
Step 1.
        command: curl https://gpcrdb.org/structure/homology_models/pth2r_human_active_full/download_pdb -o ClassB1
_pth2r_human_Active_AF_2022-08-16_GPCRdb.zip
        executable: bash
        input files:

        output files:
                ClassB1_pth2r_human_Active_AF_2022-08-16_GPCRdb.zip
```

provenance file creation

```
! verdi archive create --all archive.aiida
```

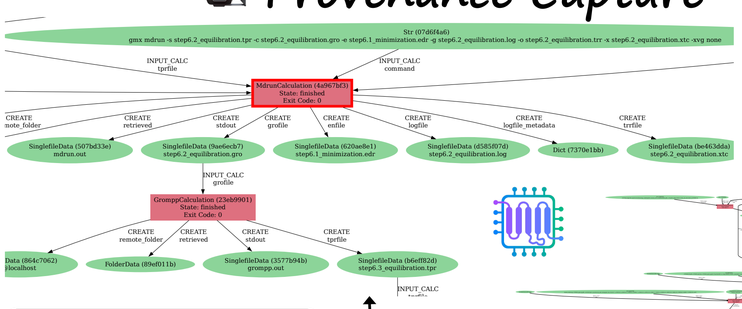
```
Report:
Archive Parameters
------------------  -------------
Path                archive.aiida
Version             main_0001
Compression         6

Inclusion rules
----------------------------  -----
Computers/Nodes/Groups/Users  All
Computer Authinfos            False
Node Comments                 True
Node Logs                     True

Report: Validating Nodes
Report: Creating archive with:
---------  ---
users        1
computers    1
nodes       98
links      108
---------  ---
Report: Finalizing archive creation...
Report: Archive created successfully
Success: wrote the export archive file to archive.aiida
```

Results – Cholesterol Occupancy Map in B1 Family



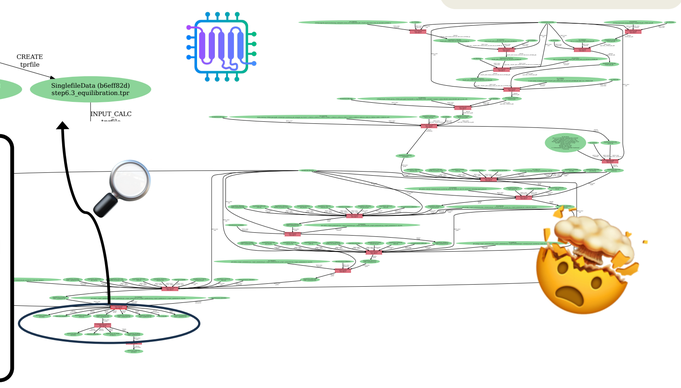
Background
- GPCRs (G protein-coupled receptors): **~800 in humans** and divided into 5 classes.
- Involved in cell signalling and are **important drug targets.**
- B1 Class GPCRs: family of **15 receptors**; peptide receptors; hormonal homeostasis.
- **Drug targets for Type 2 diabetes (T2D),** obesity, cardiovascular disease, and psychiatric disorders.
- GLP-1R (glucagon-like peptide-1 receptor), GGCR (Glucagon receptor) and GIPR (gastric inhibitory polypeptide receptor).

Provenance Capture



Modified command run via aiida-gromacs:

```
$ gmx_grompp
-f step6.3_equilibration.mdp
-o step6.3_equilibration.tpr
-c step6.2_equilibration.gro
-r solvated_ions.gro
-p system.top
-n index.ndx
-maxwarn 1

# NOTE: only difference in command is the underscore!
```



Squint to see the true graph representation of the complexity involved in B1 GPCR simulation workflow steps. Each process in the workflow is represented in red and each node of data is represented in green. Directed connections between nodes are represented with grey lines and arrows.

Feedback and Actions

- Examples needed for how to include commands run outside of GROMACS,
- Tutorials developed for complex workflows that include examples of commands from other programs.

- Ability to run commands using custom scripts,
- Incorporated use of custom scripts into aiida-gromacs, use of `insane.py` given as an example.

- Difficult to use on HPCs as pre-installation of AiiDA required,
- Created docker environments and will include future ability to run on HPC directly from local env using developed tools.

- Inclusion of analysis in workflows,
- Once simulations are performed, users can include analysis steps into a workflow and share how results were captured in publications.

References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
[2] https://aiida-gromacs.readthedocs.io
[3] https://aiida-amber.readthedocs.io
[4] T. Bertie Ansell, Wanling Song, Mark S.P. Sansom,The Glycosphingolipid GM3 Modulates Conformational Dynamics of the Glucagon Receptor, Biophysical Journal, 119, 2, (2020).
[5] Grünewald, Fabian, et al. Journal of Chemical Theory and Computation. 2022
[6] Grünewald, Fabian, et al. Journal of Chemical Theory and Computation. 2023