Status of the NMRlipids databank

NMRlipids20 workshop

September 6th to 9th 2021
Prague, Czech Republic

Aims of the NMRlipids project

Open Collaboration to understand lipid systems in atomistic resolution

1) **Make atomistic details great again**

2) **Big Data, Big Success**

3) **Open collaboration is new black**

Aims of the NMRlipids project

Open Collaboration to understand lipid systems in atomistic resolution

1) **MD simulations that correctly capture not only general membrane features but also atomistic character of individual lipids** (correct area per lipid does not guarantee correct conformational ensemble)

2) **Foster novel applications by bringing quality evaluated MD simulations easily accessible for wide user base** (automatic analysis of hundreds MD simulation trajectories)

3) **Develop our open collaboration approach and explore its possibilities**

Quality evaluated atomistic resolution MD simulations of biologically relevant lipid mixtures in NMRlipids databank

	PC	PE	PG	PS	chol	CL	GM1	PIP	PA	DAG
PC	X	X	X	X	X					
PE		X								
PG			X							
PS				X						
chol										
CL										
GM1										
PIP										
PA										
DAG										

NMRlipids databank
general properties

- Overlay databank: *NMRlipids databank contains indexed links to the data. The actual MD simulation data is currently in Zenodo, but could be in any stable location.*

- Analysis of the data: *NMRlipids databank enables flexible analysis of the content.*

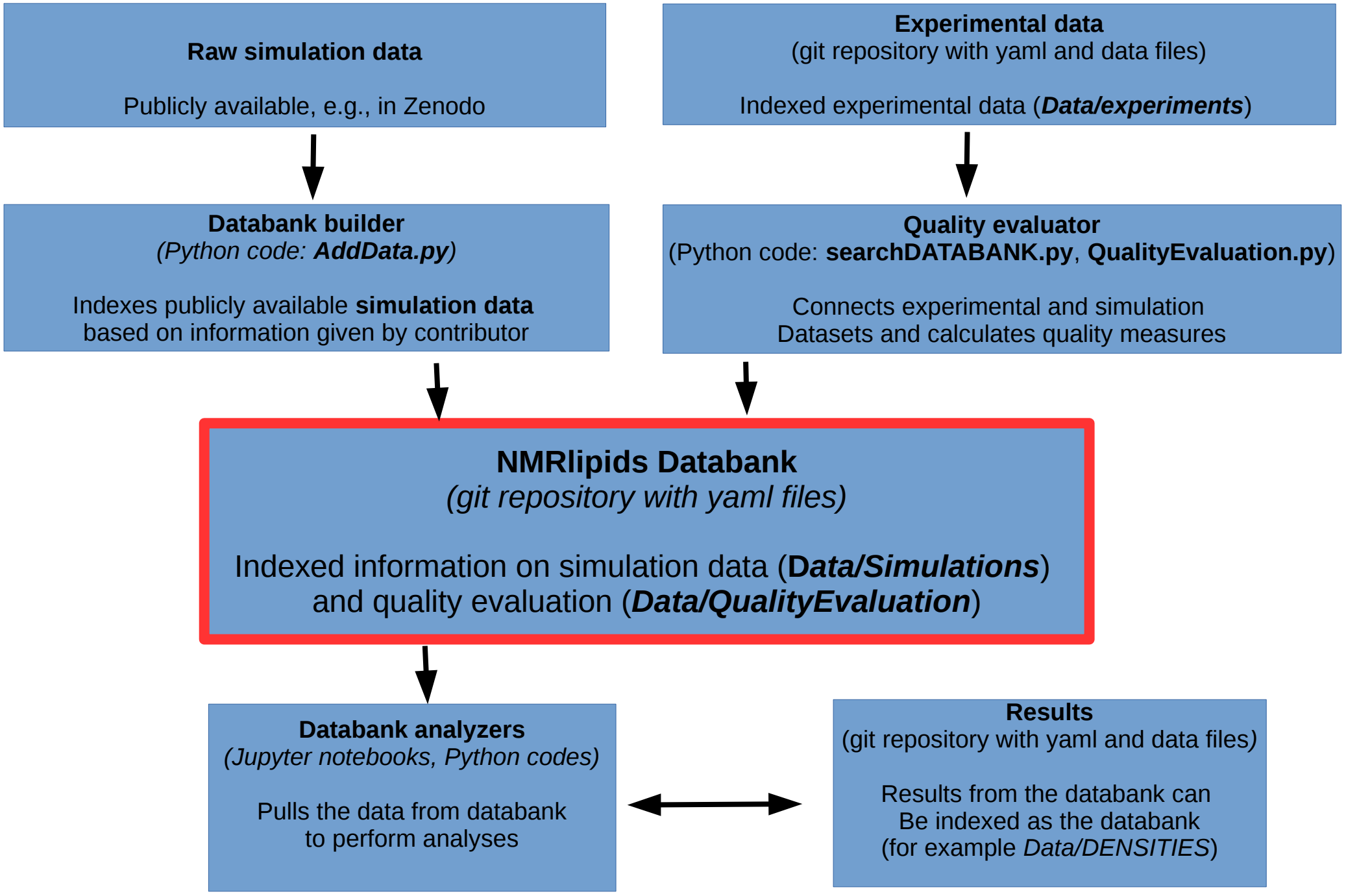
- Quality evaluation: *NMRlipids databank contains a quality evaluation protocol that is applied to all contributed datasets. Also the quality evaluation results are also stored in the databank.*

NMRlipids databank
expected applications

- Force field evaluation: *What is the best force field for my application?*
- Reference simulations: *For example, reference pure bilayer simulations for membrane-protein interaction studies.*
- Analysis of bilayer properties from large datasets: *For example, calculate P-N vector angle from all available PC and PG simulations.*
- Exercise and example for sharing simulation data: *“PDB” for simulations?*

NMRlipids databank structure

https://github.com/NMRLipids/Databank



Databank builder: AddData.py

https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/AddData.py

Instructions to add data: https://github.com/NMRLipids/Databank

Instructions to make info file:
https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/info_files/README.md

Usage of AddData.py

- **python3 AddData.py *InfoFile.yaml***

- **Info files currently available at:**
  https://github.com/NMRLipids/Databank/tree/main/Scripts/BuildDatabank/info_files

Output of AddData.py

- **Information from info files + automatically extracted information:** Number of molecules, temperature, length of trajectory, size of trajectory, number of atoms, date of running the AddData.py

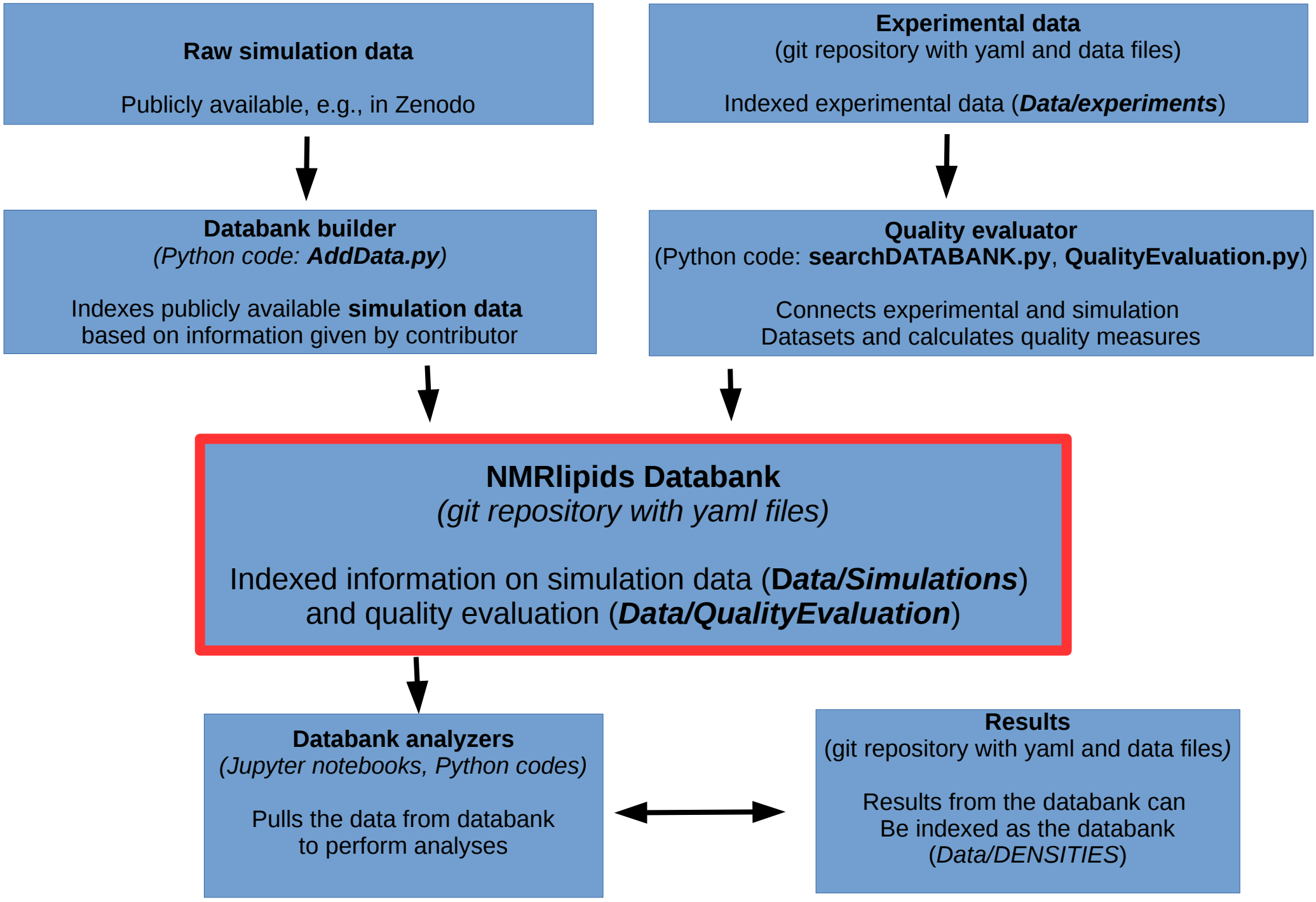
- **These are stored in README.yaml files located in folders based hash IDs:**

  https://github.com/NMRLipids/Databank/tree/main/Data/Simulations

Simulations in NMRlipids Databank

- https://github.com/NMRLipids/Databank/tree/main/Data/Simulations
- Each folder corresponds one simulation
- Folders are named according to the hash of trajectory and tpr file
- **Folders contain README.yaml which should contain all the relevant information on the simulation!**
- **Statistics:**
  https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/stats.ipynb

NMRlipids databank structure



Experimental data

- https://github.com/NMRLipids/Databank/tree/main/Data/experiments
- Each folder corresponds one experimental dataset
- Folders are named according to DOI of experimental data
- **Folders contain README.yaml which should contain all the relevant information to connect experimental and simulation datasets! Should we add something?**
- Currently: DOI of the publication, temperature, molar fractions of lipids, ion concentration, total lipid concentration (or full hydration), information of counterions

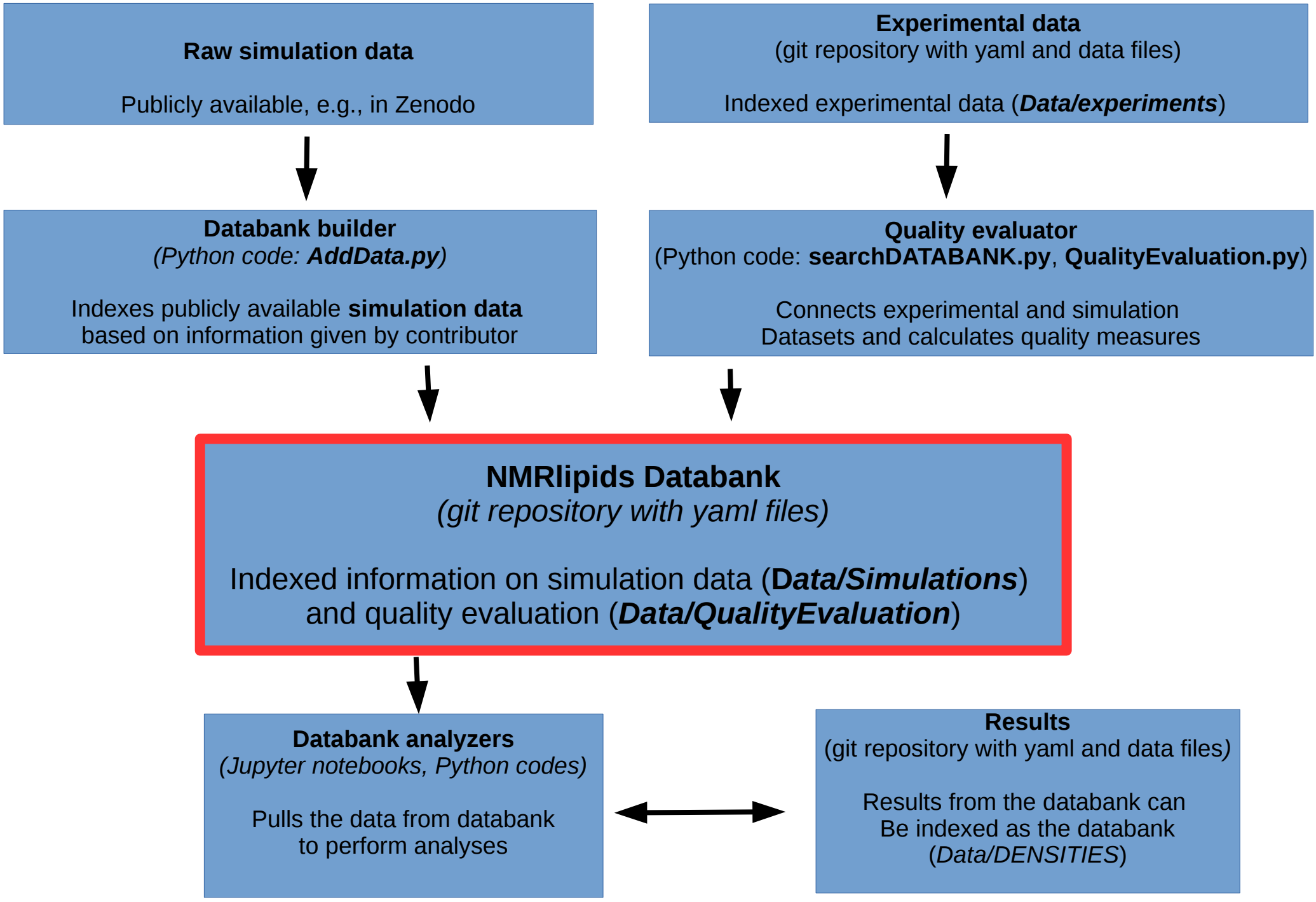
Connecting experimental and simulation data

- https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/searchDATABANK.py
- Searches simulation-experimental data pairs where
  - temperature is the same within ± 2 degrees
  - molar concentrations are within ± 5 percentage units
  - counterions are the same
- When pair is found, the path to experimental dataset is written in simulation README file, for example, see
  https://github.com/NMRLipids/Databank/tree/main/Data/Simulations/0c2/1a9/0c21a9be136ea0eb9df9e5c6cdc19f723a0af245/9ac73b6a98acb54a7a67a5d690794ad7f1e4a1d1

Quality evaluation

- https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/QualityEvaluation.py
- Calculates the order parameters for all simulations
- Evaluates the quality of simulations for which experimental data is available
- **Quality for each order parameter is calculated as a distance from experimental value and stored to *OrderParameters.json files in** https://github.com/NMRLipids/Databank/tree/main/Data/QualityEvaluation
- Also quality for headgroup+glycerol backbone, sn-1, and sn-2 acyl chains are evaluated and stored to *FragmentQuality.json files in same folders

NMRlipids databank structure



Databank analyzer

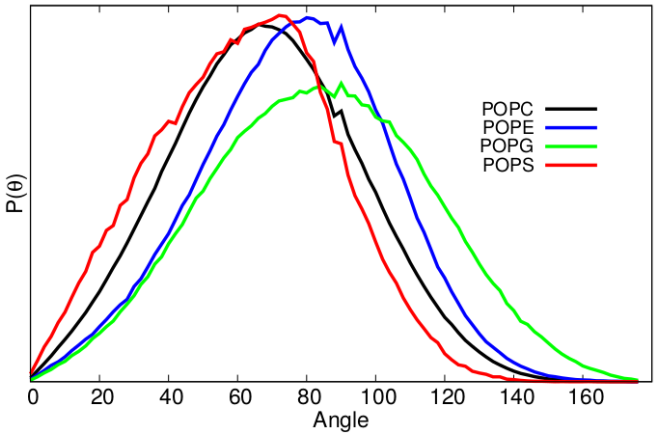
- Goes through the README.yaml files in the databank
- Performs wanted analysis for selected simulations
- Results can be saved in separate results databank with the same indexing
- Result databanks can be browsed in similar manner for plotting
- Template at https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/template.ipynb

Databank Analyzer Examples

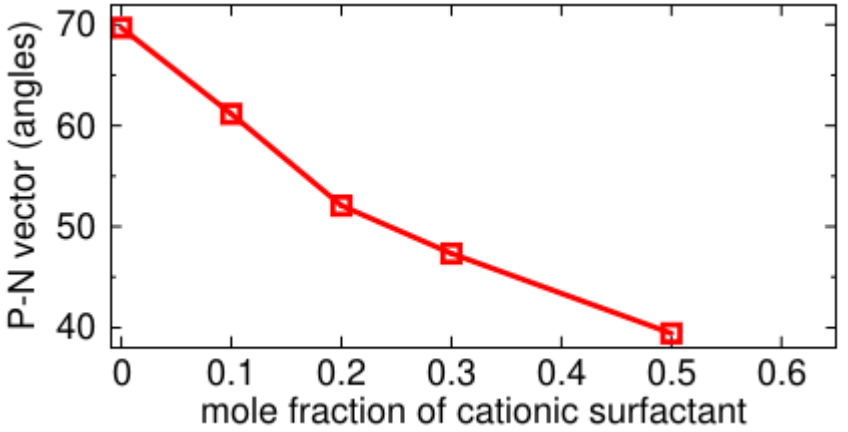
1) Calculate the P-N vector angles of POPS, POPE, POPG and POPC lipids from each simulation

Code: https://github.com/NMRLipids/NMRlipidsIVPEandPG/blob/master/scripts/calcPNvectors.py

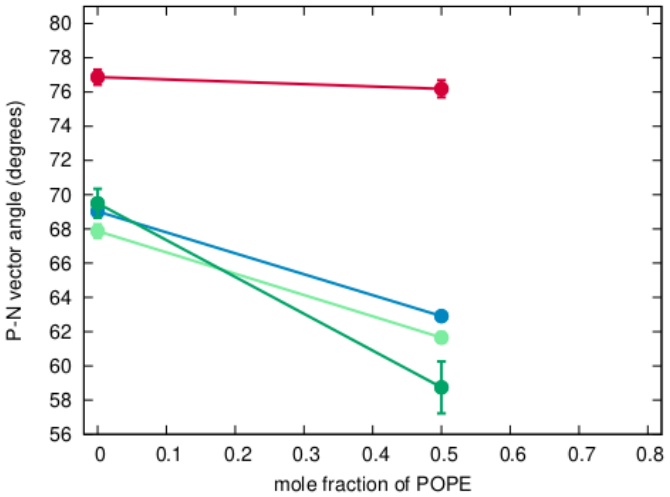
Results: https://github.com/NMRLipids/NMRlipidsIVPEandPG/tree/master/Data/HGorientation

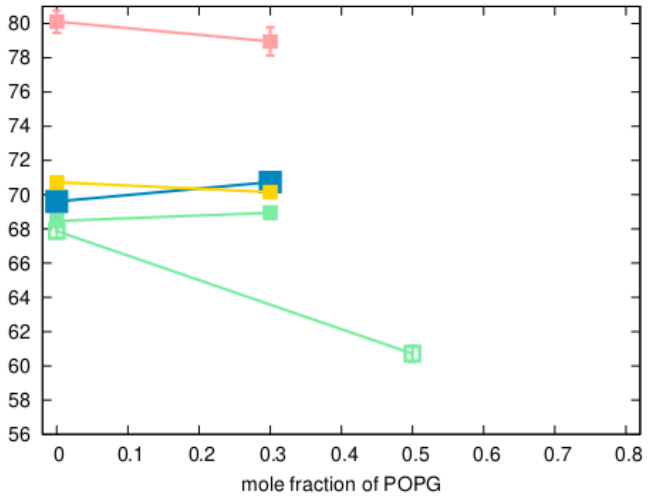


POPC

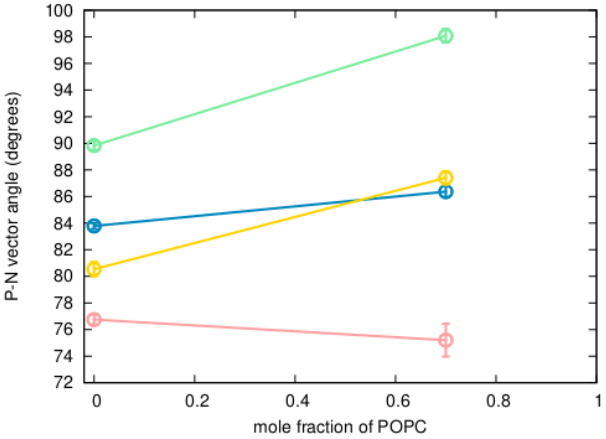


POPC





POPG



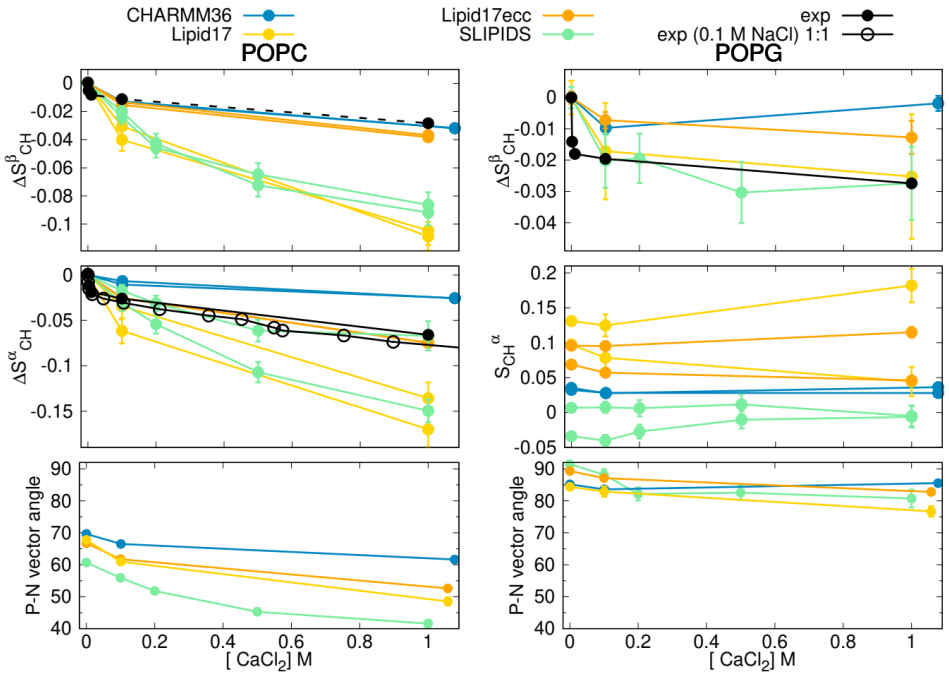
Databank Analyzer Examples

2) Find data for order parameter changes upon addition of CaCl_2 from all available POPC:POPG mixtures

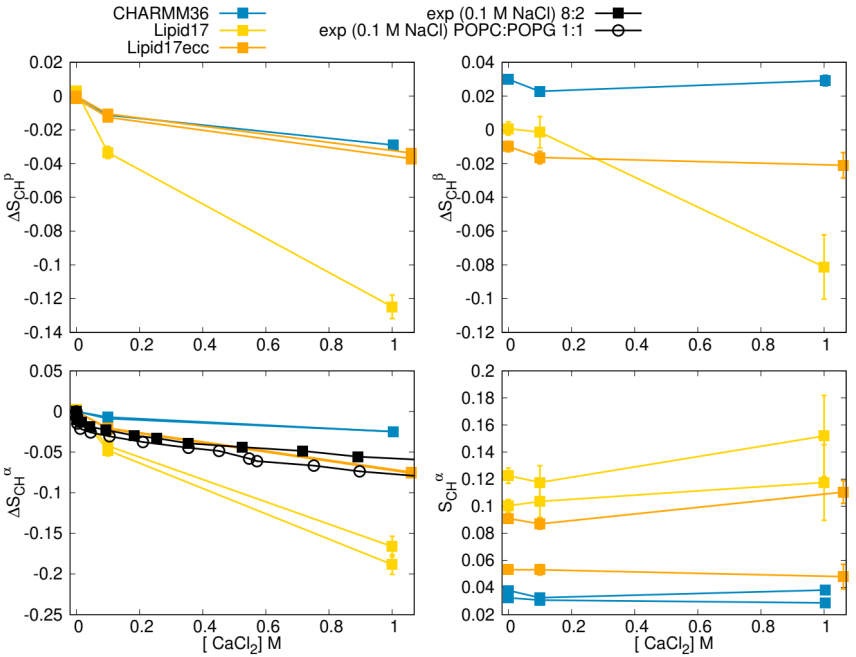
Code: https://github.com/NMRLipids/NMRlipidsIVPEandPG/blob/master/scripts/plotOPsWITHsalt.ipynb

Results:

POPC:POPG (1:1)



POPC:POPG (4:1)



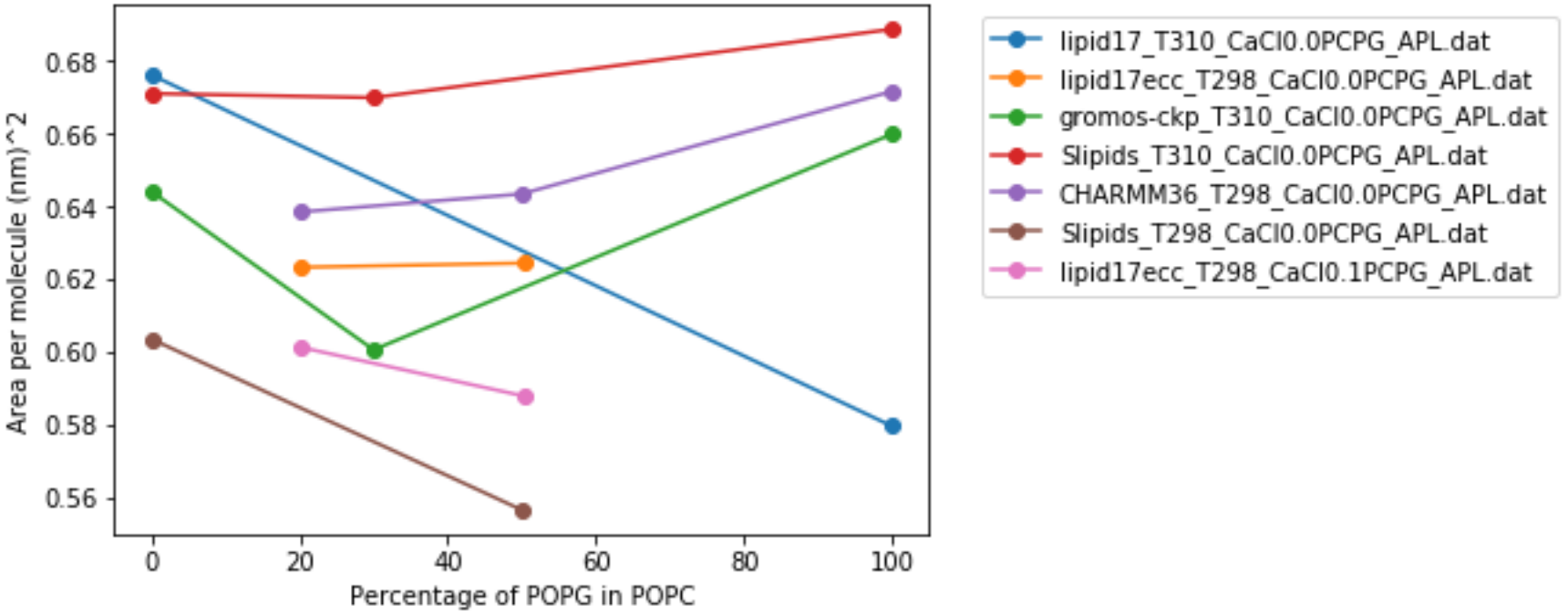
Databank Analyzer Examples

3) How area per lipid changes in PC:PG lipid mixtures as a function of PG concentration

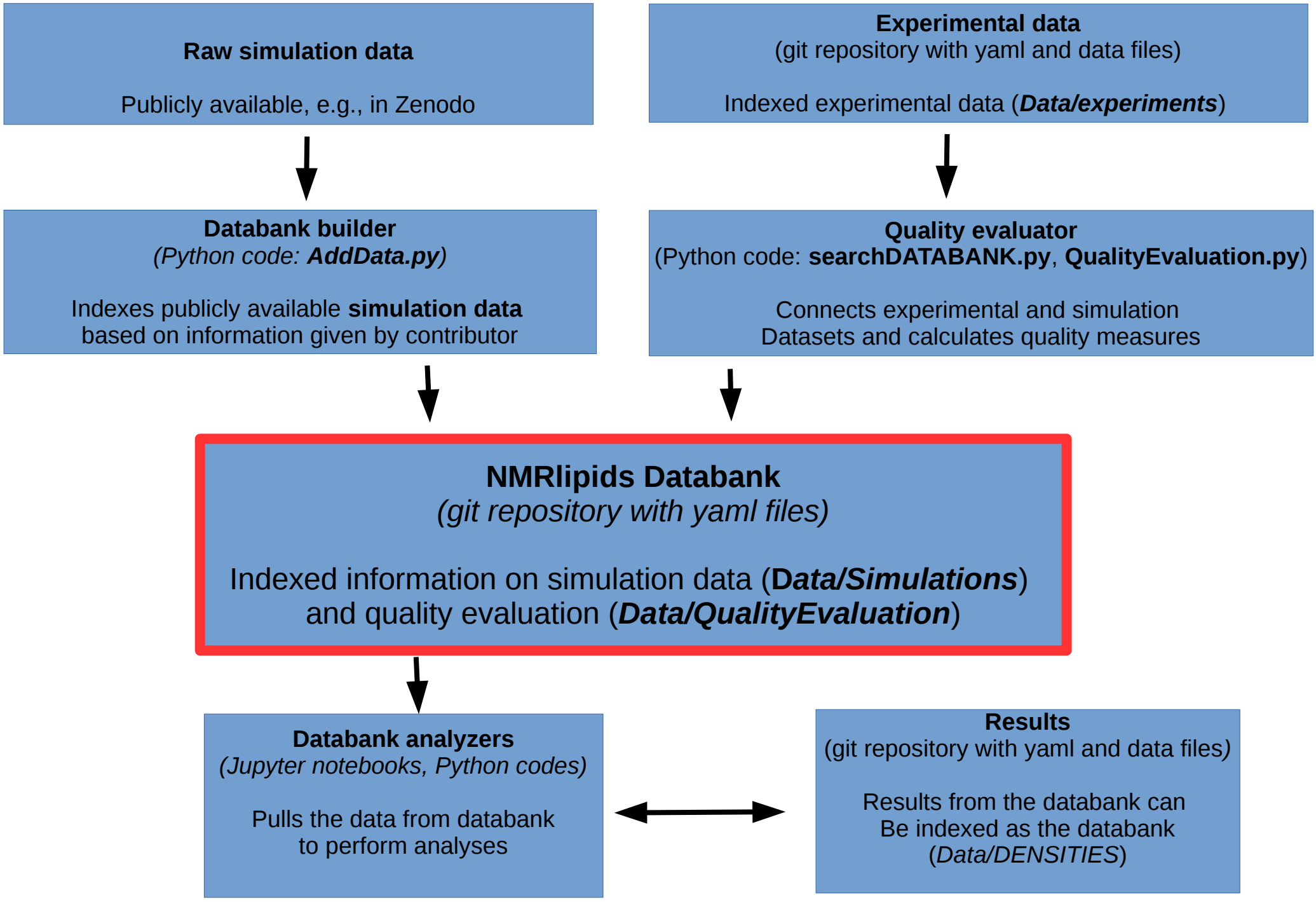
Code: https://github.com/NMRLipids/NMRlipidsIVPEandPG/blob/master/scripts/calcAPL.py

https://github.com/NMRLipids/NMRlipidsIVPEandPG/blob/master/scripts/plotAPLs.ipynb

Results:



NMRlipids databank structure



NMRlipids databank publication plan

- **Article describing the databank and highlight applications will be prepared.**
- **At least all trajectories contributed to the NMRlipids will be included (approximately 300-400 trajectories currently).**
- **Possible highlight applications:**

  - Quality ranking of all simulation

  - Analysis of rare phenomena using large datasets, such as water permeation through bilayers or lipid flip-flops

  - Example of analysis useful for community who are typically not using MD simulations, such as T_1 spin relaxation times of water near membranes that are used in MRI imaging

- **NMRlipids authorship rules will be applied in the first publication of the databank** (authorship will be offered to all contributors and order is alphabetical) **with two exceptions: Samuli Ollila will the last author and Anne Kiirikki will be the first.**

Work in progress

- Quality evaluation and ranking in progress by Anne Kiirikki and Samuli Ollila
- Addition of available data in Zenodo into the databank by Lara Bort
- Codes to analyze lipid flip-flops, water diffusion, and water spin relaxation times in progress by Anne Kiirikki and Samuli Ollila
  https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/calcWATERdiffusion.py
  https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/plotWATERdiffusion.ipynb

Open issues

- **United atom simulations:** Should be maybe translated to all atom before addition?
- **Should we include stereospecific information on isomers?**
- **Extending to other than Gromacs simulations:** OpenMM and possibly Amber. Others?
- **Calculation of form factors**
- **“Sanity checks” for the data:** Equlibration etc.
- **NMRlipids III (systems with cholesterol):** Should we publish the reported data within the databank publication?

Topics tacked in this workhop

1) **Quality evaluation:** Define the quality measures for order parameters, find robust code for form factor and include this into the quality measure
2) **Extension to other programs:** OpenMM and Amber
3) **Analysis of the data:** How to present the available simulations, interesting analyses, etc
4) **Addition of simulation and experimental data**
5) **NMRlipids VI:** Can we do something to advance the project?