NMRlipids databank

NMRlipids summer school 2022

June 2nd 2022
Espoo, Finland

NMRlipids databank

https://github.com/NMRLipids/Databank

(www.databank.nmrlipids.fi)

- **Databank containing quality evaluated molecular dynamics (MD) simulations of lipid bilayers with atomic resolution**
- **Overlay databank with programmatic access**
- **Initiated from the NMRlipids project (nmrlipids.blogspot.fi)**
- **Open for submissions**

NMRlipids databank: Expected applications

- **Force field evaluation:** What is the best force field for my application?
- **Analysis of bilayer properties from large datasets:** For example, correlations between lipid bilayer properties.
- **Reference simulations:** For example, reference pure bilayer simulations for membrane-protein interaction studies.
- **Exercise and example for sharing simulation data:** “PDB” for simulations?

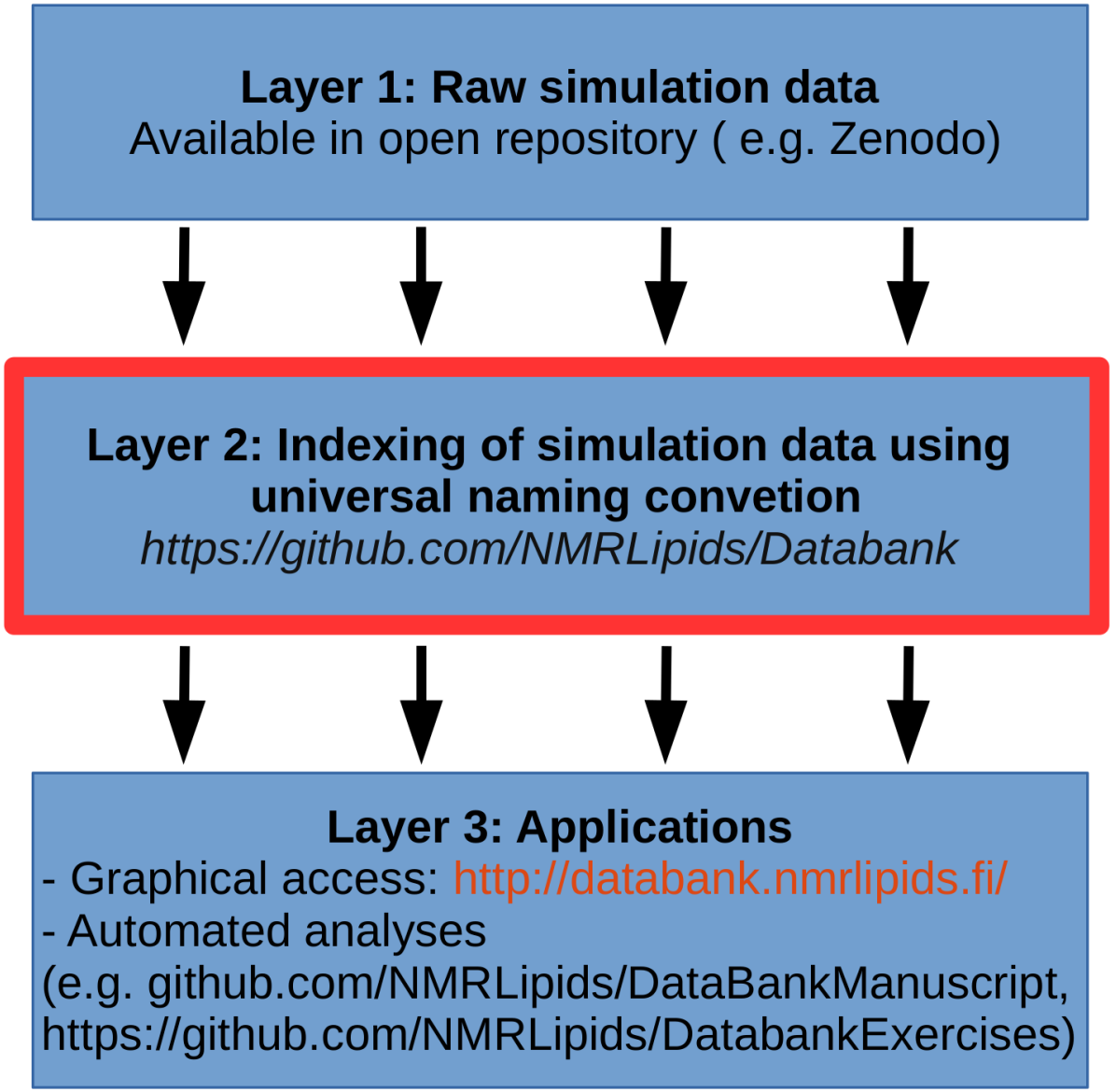
NMRlipids databank: Key features

- **Overlay databank:** NMRlipids databank contains links to the raw data, but data is stored in public repositories (currently in Zenodo).

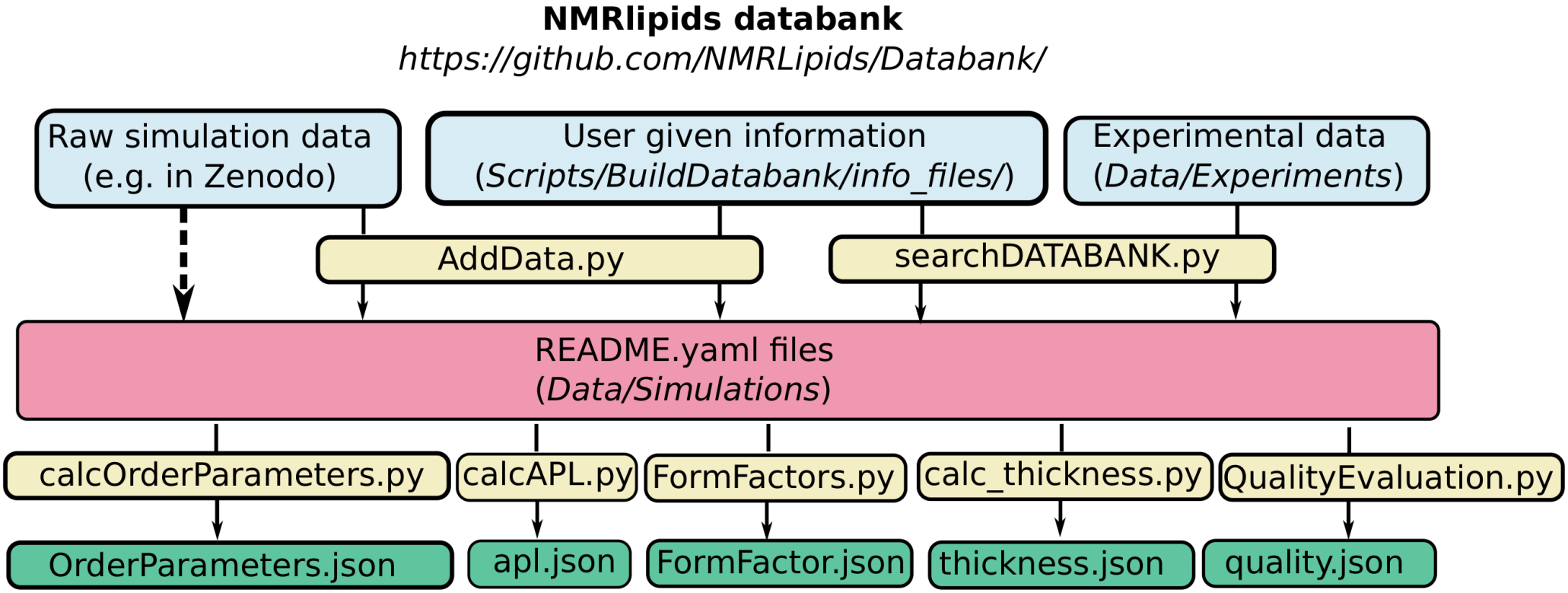
- **Programmatic access:** Programmatic access enables flexible analysis of the content.

- **Quality evaluation:** Automatic quality evaluation is made against available NMR order parameters and x-ray scattering form factors, and the results are stored in the databank.

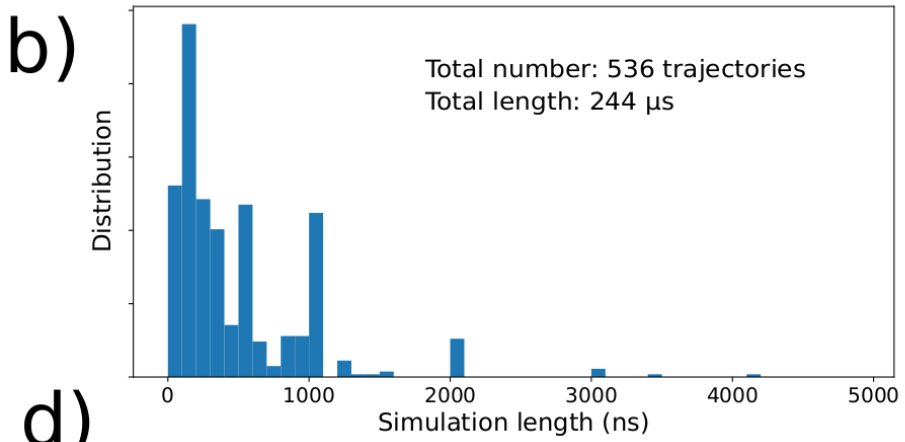
Overlay structure of the NMRlipids databank

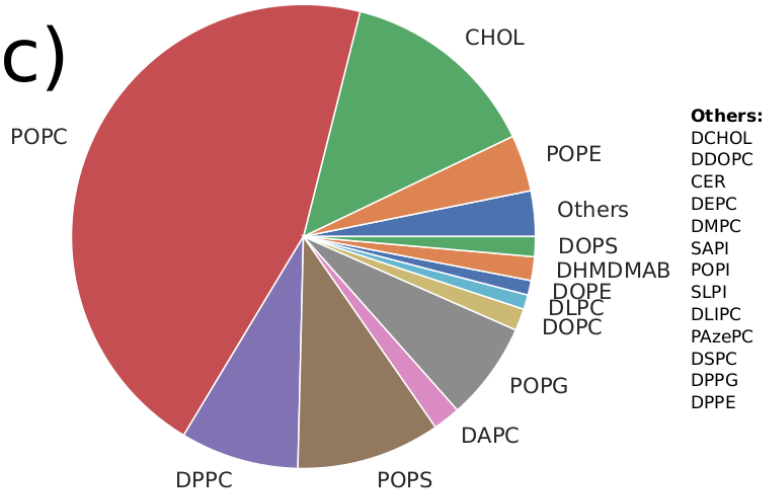


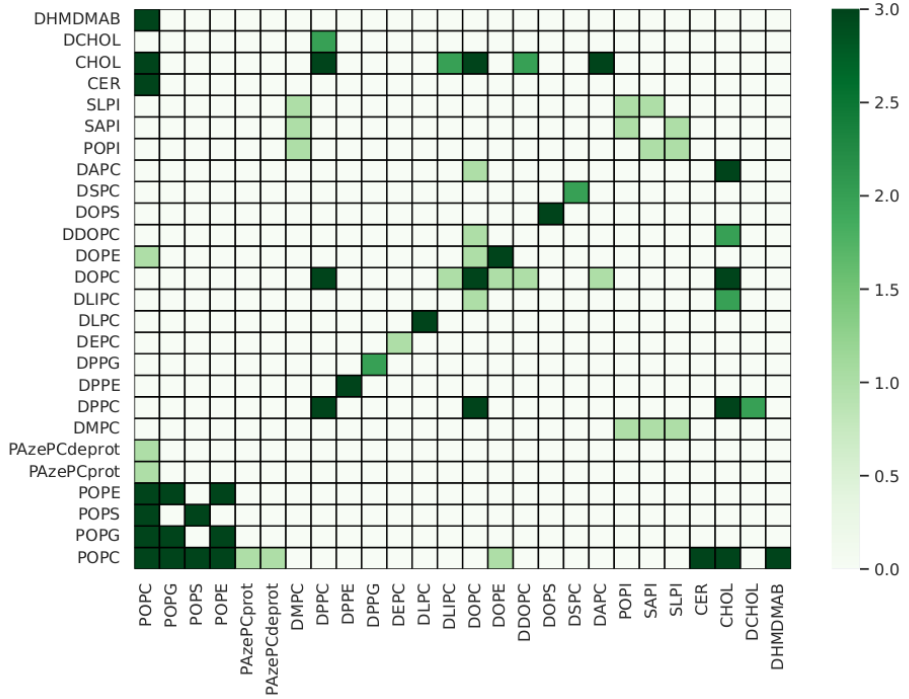
Overlay structure of the NMRlipids databank

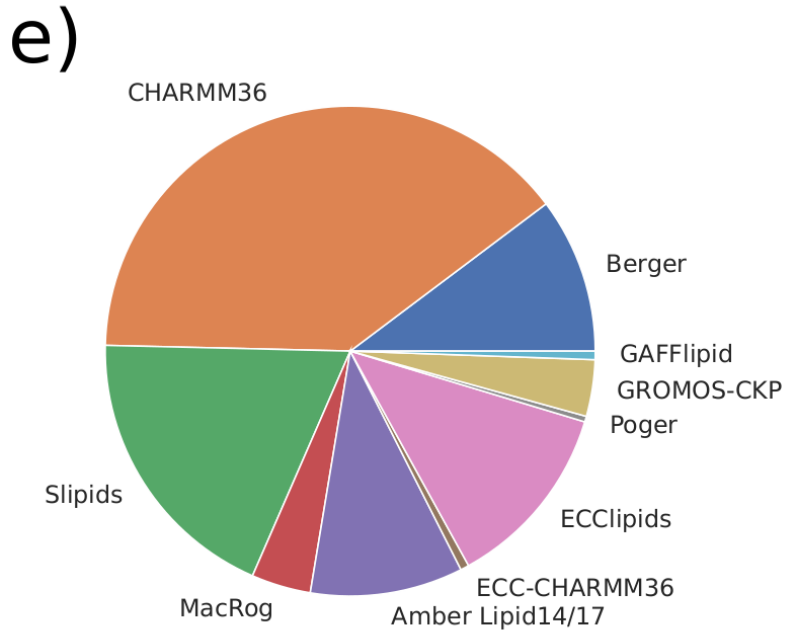


NMRlipids databank: Current content









Simulations in the NMRlipids Databank

- Each folder in *https://github.com/NMRLipids/Databank/tree/main/Data/Simulations* corresponds one simulation
- Folders are named according to the hash of trajectory and tpr file
- **README.yaml in each folder contains all the relevant information on the simulation!**

Simulations in the NMRlipids Databank

key	description	type
DOI	DOI from where the raw data is found	user given (compulsory)
SOFTWARE	Software used to run the simulation (e.g. Gromacs, Amber, NAMD, etc.)	
TRJ	Name of the trajectory file found from DOI	
TPR	Name of the topology file found from DOI (trp file in the case of Gromacs)	
PREEQTIME	Pre-equilibrate time simulated before the uploaded trajectory in nanoseconds. [1]	
TIMELEFTOUT	Equilibration period in the uploaded trajectory that should be discarded in analyses. [2]	
COMPOSITION	Molecules names used in the simulation and corresponding mapping files (see section 4.2)	
DIR_WRK	Temporary working directory in your local computer.	
UNITEDATOM_DICT	Information for constucting hydrogens for united atom simulations, empty for all atom simulations	
TYPEOFSYSTEM	Lipid bilayer or something else	
PUBLICATION	Give reference to a publication(s) related to the data.	User given (optional)
AUTHORS_CONTACT	Name and email of the main author(s) of the data.	
SYSTEM	System description on free text format	
SOFTWARE_VERSION	Version of the used software	
FF	Name of the used force field	
FF_SOURCE	Source of the force field parameters, e.g, CHARMM-GUI, webpage, citation to a publication, etc.	
FF_DATE	Date when force field parameters were accessed on the gives source (day/month/year).	
FFmolename	Molecule specific force field information, e.g., water model with FFSOL and sodium parameters with FFSOD.	
CPT	Name of the Gromacs checkpoint file.	
LOG	Name of the Gromacs log file.	
TOP	Name of the Gromacs top file.	
GRO	Name of the Gromacs gro file.	
TRAJECTORY_SIZE	Size of the trajectory file in bytes	automatically extracted data.
TRJLENGTH	Lenght of the trajectory (ps).	
TEMPERATURE	Temperature of the simulation.	
NUMBER_OF_ATOMS	Number of atoms in the simulation.	
DATEOFRUNNIG	Date when added into the databank	
EXPERIMENT	Potentially connected experimental data	
COMPOSITION	Numbers of lipid molecules (NPOPC, NPOPG, etc.) per membrane leaflet are calculated by determining on which side of the center of mass of the membrane the center of mass of the head group of each lipid molecule is located. Numbers of other molecules such as solvent and ions (NSOL, NPOT, NSOD, etc.) are read from the topology file.	

[1] For example, if you upload 100-200 ns part of total 200 ns simulation, this should value should be 100.

[2] For example, if you upload 0-200 ns part of total 200 ns simulation where the first 100 ns should be considered as an equilibration, this value should be 100.

Unique atom and molecule naming conventions in COMPOSITION

Universal molecule name

Name of molecule in force field

Definition of universal atom names (mapping file)

```
COMPOSITION:
 CHOL:
  NAME: CHL1
  MAPPING: mappingCHOLESTEROLcharmm.txt
 POPC:
  NAME: POPC
  MAPPING: mappingPOPCcharmm.txt
 SOL:
  NAME: TIP3
  MAPPING: mappingTIP3PCHARMMgui.txt
 SOD:
  NAME: SOD
  MAPPING: mappingSOD.txt
 CLA:
  NAME: CLA
  MAPPING: mappingCLA.txt
```

Table of unique molecule names:
https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/info_files/README.md
Mapping files:
https://github.com/NMRLipids/Databank/tree/main/Scripts/BuildDatabank/mapping_files
https://nmrlipids.blogspot.com/2022/04/new-yaml-format-of-mapping-files.html
http://nmrlipids.blogspot.com/2015/03/mapping-scheme-for-lipid-atom-names-for.html

Adding simulation data to the NMRlipids databank

Short instructions:

1) **Clone** *https://github.com/NMRLipids/Databank* **repository**
2) **Create info.yaml file based on instructions at:**
   https://github.com/NMRLipids/Databank/blob/main/Scripts/BuildDatabank/info_files/README.md
3) **Run: python3 AddData.py *InfoFile.yaml***
4) **Add and commit the resulting README.yaml file into the repository and make a pull request to the master branch**

Detailed instructions: https://github.com/NMRLipids/Databank

A web app coming

Experimental data in the NMRlipids databank

- Each folder in *https://github.com/NMRLipids/Databank/tree/main/Data/experiments* corresponds one experimental dataset
- Folders are named according to the DOI of experimental data
- **All the relevant information to connect experimental and simulation datasets are found from README.yaml files within these folders**

key	description
DOI	DOI of the publication related to the experimental data.
TEMPERATURE	Temperature of the experiment.
MOLAR_FRACTIONS	Dictionary of molar fractions of bilayer components
ION_CONCENTRATIONS	Dictionary of ion concentrations of the system
TOTAL_LIPID_CONCENTRATION	Total concentration of lipid components. If exact concentration is not known, but experiments are performed in excess water, 'full hydration' can be given.
COUNTER_IONS	Type of counter ions if present.

Quality evaluation in the NMRlipids databank

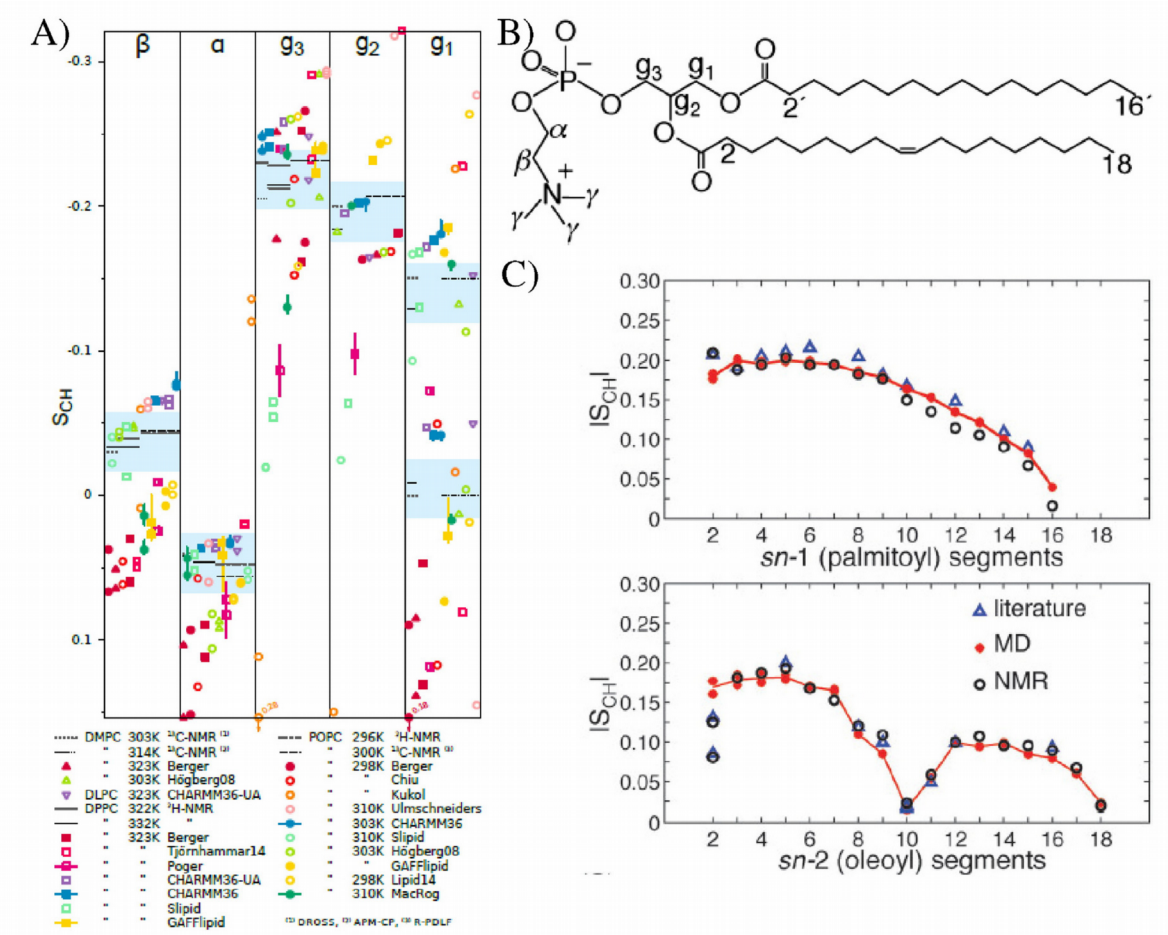
- Simulations (*Data/Simulations/*) are paired with experimental data (*Data/experiments*) when:
  - temperature is the same within ± 2 degrees
  - molar concentrations are within ± 5 percentage units
  - counterions are the same
- If a pair is found, the path to the experimental data is added to the "EXPERIMENT" key in the simulation README.yaml (*searchDATABANK.py*)
- Quality measures are calculated for simulations paired with experimental data (*QualityEvaluation.py*)
- Quality is defined against C-H bond order parameters from NMR and x-ray scattering form factors

Quality evaluation: Order parameters

- Order parameters are sensitive to the conformational ensembles of individual lipids
- Acyl chain order correlates with lipid packing (area per lipid)

$$S_{CH} = \frac{1}{2}\langle 3\cos^2\theta - 1\rangle$$

θ = angle between C-H bond and membrane normal
<..> = average over conformational ensemble



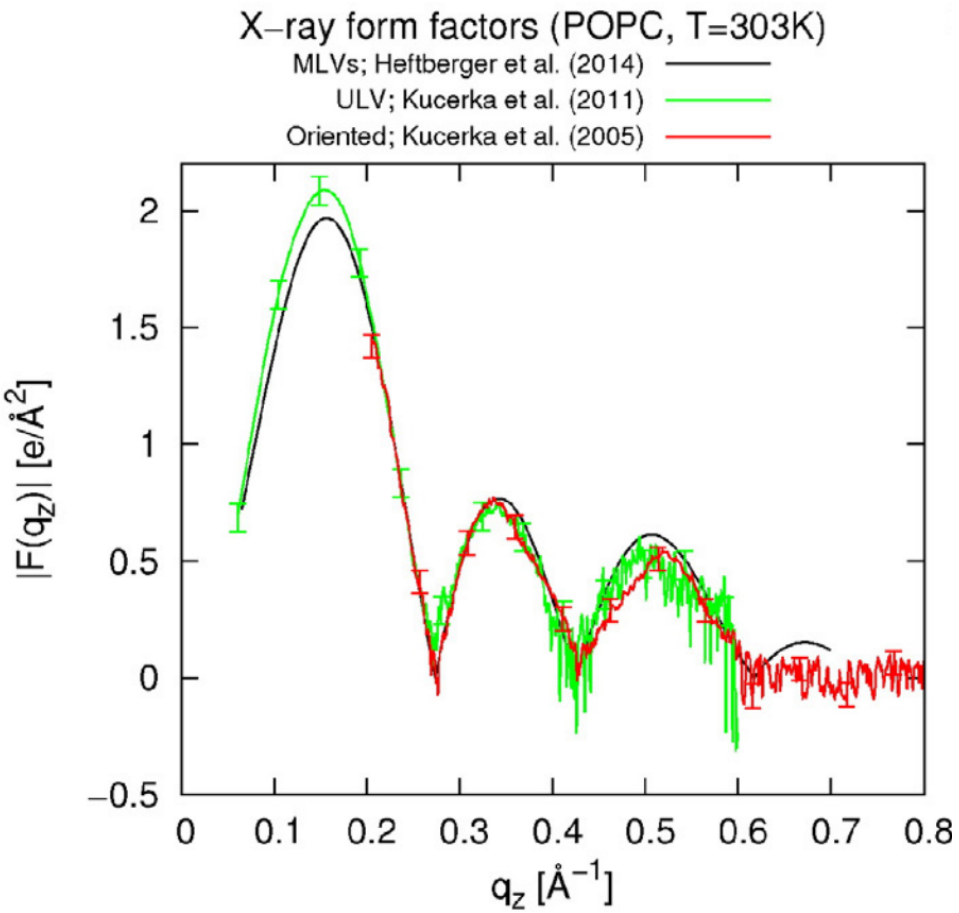
http://dx.doi.org/10.1016/j.bbamem.2016.01.019

Quality evaluation: Form factor

- Scattering form factors are sensitive to membrane dimensions (electron density profile, thickness and area per molecule)

$$F(q) = \int_{-D/2}^{D/2} \Delta\rho_e(z) \cos(zq_z) \mathrm{d}z$$

$\rho_e(z)$ = electron density difference with respect to bulk water
D/2 = beginning of bulk water region



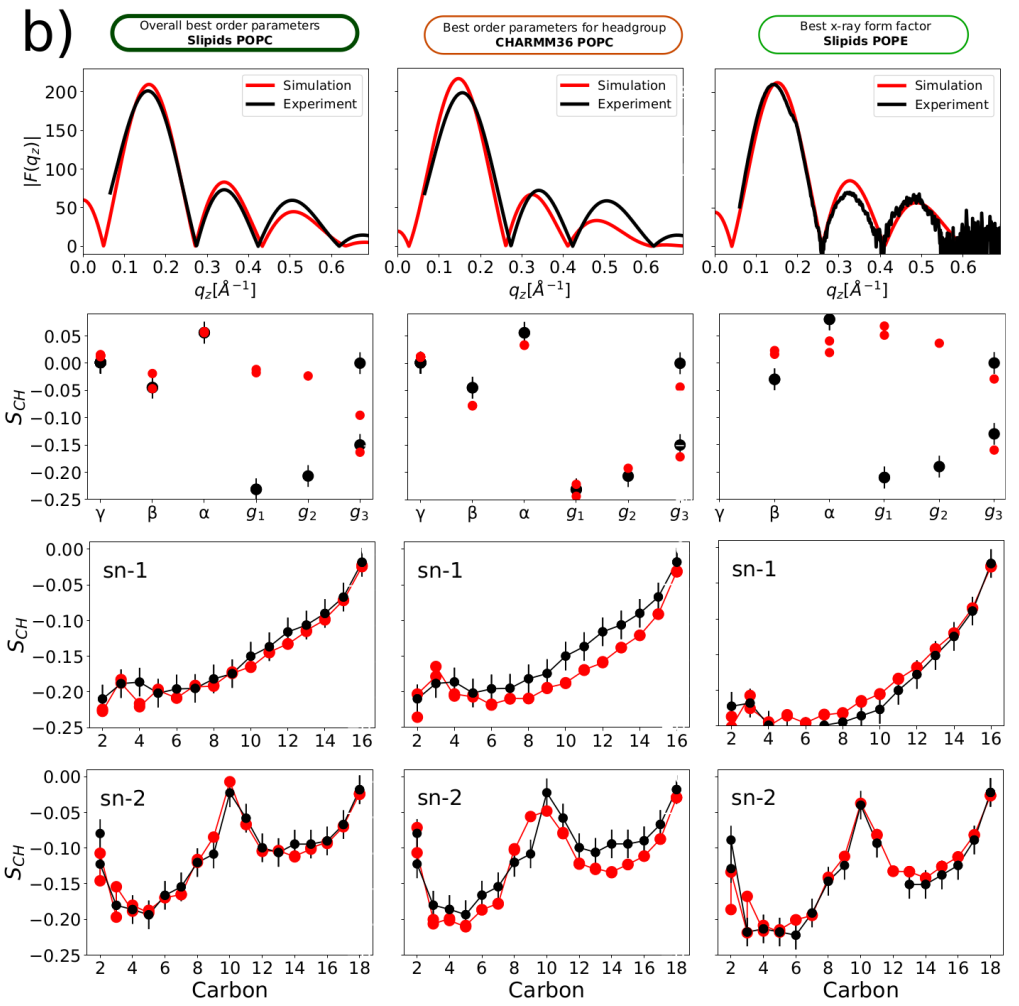
http://dx.doi.org/10.1016/j.bbamem.2016.01.019

NMRlipids databank quality evaluation

Top 13 ranking based on total order parameter quality

	OPquality	Headgroup	sn-1	sn-2	FFquality	Force field	Molecules	Temperature	DOI
1	0.76	0.61	0.84	0.82	515	Slipids	POPC:SOL (512:23943)	298.00	10.5281/zenodo.166034
2	0.69	0.63	0.70	0.76	713	MacRog	POPC:SOL (128:5120)	300.00	10.5281/zenodo.3741793
3	0.61	0.64	0.48	0.72	674	MacRog	POPC:SOL (288:14400)	298.00	10.5281/zenodo.13498
4	0.60	0.65	0.74	0.42	576	ECC-lipids	POPC:SOL (128:6400)	300.00	10.5281/zenodo.3335503
5	0.58	0.01	0.88	0.84		Berger	POPC:SOL (256:10342)	300.00	10.5281/zenodo.1402417
6	0.55	0.02	0.83	0.78		Berger	POPC:SOL (128:7290)	298.00	10.5281/zenodo.4643875
7	0.54	0.65	0.54	0.43	613	ECC-lipids	POPC:SOL (128:6400)	300.00	10.5281/zenodo.1118980
8	0.52	0.01	0.79	0.77	122	Slipids	POPE:SOL (500:25000)	310.00	10.5281/zenodo.3231342
9	0.51	0.10	0.87	0.57	140	Slipids	POPE:SOL (336:13460)	310.00	10.5281/zenodo.1293813
10	0.51	0.10	0.86	0.57	137	Slipids	POPE:SOL (336:13460)	310.00	10.5281/zenodo.1293813
11	0.50	0.16	0.64	0.69	297	ECClipids	POPS:SOL:SOD (72:3600:72)	298.00	10.5281/zenodo.1488094
12	0.48	0.67	0.36	0.40	690	CHARMM36	POPC:SOL (256:9767)	300.00	10.5281/zenodo.1306800
13	0.47	0.67	0.37	0.38	903	CHARMM36	POPC:SOL (1024:51200)	298.15	10.5281/zenodo.5767451

Selected comparisons



Full quality evaluations of over 100 simulations available from here:

https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/plotQuality.ipynb

NMRlipids quality measure of lipid conformational ensemble

- Probability for each order parameter to locate within experimental error bars calculated from the Student's t-distribution

$$P = f\left(\frac{S_{\mathrm{CH}} - (S_{\mathrm{exp}} + \Delta S_{\mathrm{exp}})}{s/\sqrt{n}}\right) - f\left(\frac{S_{\mathrm{CH}} - (S_{\mathrm{exp}} - \Delta S_{\mathrm{exp}})}{s/\sqrt{n}}\right)$$

P = probability for simulation value to locate within experimental error bars
S_{exp} = experimental order parameter
Δs_{exp} = error bars of experimental order parameter (0.02)
S_{CH} = order parameter from simulation
s = standard deviation of order parameter from simulation
n = number of lipids in the simulation
f = Student's t-distribution

NMRlipids quality measure of lipid conformational ensemble

- Quality of fragments of each lipid estimated by averaging over order parameters
- Divided by the fraction of available order parameters to penalize from missing data

$$P^{\mathrm{frag}}[\mathrm{lipid}] = \frac{\langle P[\mathrm{lipid}]\rangle_{\mathrm{frag}}}{p_{\mathrm{frag}}[\mathrm{lipid}]}$$

frag = headgroup, *sn*-1 chain, *sn*-2 chain, or total (all order parameter in a molecule)
$\mathbf{p}_{\mathbf{frag}}$ = fraction of experimentally available order parameters for the fragment

- The overall quality of fragments in each simulation is estimated by averaging over lipids weighted by molar fraction

$$P^{\mathrm{frag}} = \sum_{\mathrm{lipid}} \chi_{\mathrm{lipid}} P^{\mathrm{frag}}[\mathrm{lipid}]$$

$\boldsymbol{\chi}_{\mathbf{lipid}}$ = molar fraction of a lipid in the simulation

NMRlipids quality measures for lipid bilayer dimensions

- Form factor quality evaluated as in SIMtoEXP program

$$\chi^2 = \frac{\sqrt{\sum_{i=1}^{N_q} \left(|F_s(q_i)| - k_e|F_e(q_i)|\right)^2 / \left(\Delta F_e(q_i)\right)^2}}{\sqrt{N_q - 1}}$$

F_s = form factor from a simulation
F_e = form factor from experiment
ΔF_e = error of form factor from experiment
(1,…,N_q) = Experimental datapoints

$$k_e = \frac{\sum_{i=1}^{N_q} \frac{|F_s(q_i)||F_e(q_i)|}{\left(\Delta F_e(q_i)\right)^2}}{\sum_{i=1}^{N_q} \frac{|F_e(q_i)|^2}{\left(\Delta F_e(q_i)\right)^2}}$$

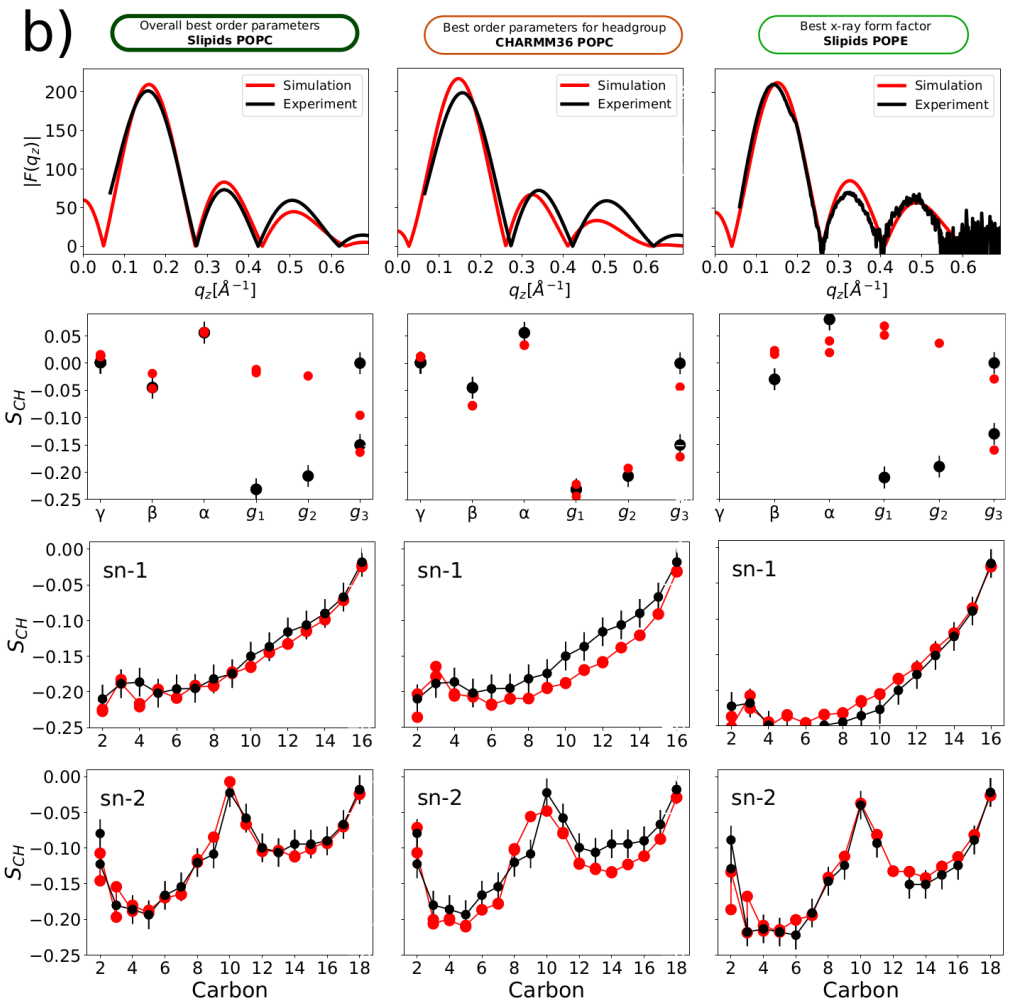
https://doi.org/10.1007/s00232-010-9254-5 (SIMtoEXP publication)

NMRlipids databank quality evaluation

Top 13 ranking based on total order parameter quality

	OPquality	Headgroup	sn-1	sn-2	FFquality	Force field	Molecules	Temperature	DOI
1	0.76	0.61	0.84	0.82	515	Slipids	POPC:SOL (512:23943)	298.00	10.5281/zenodo.166034
2	0.69	0.63	0.70	0.76	713	MacRog	POPC:SOL (128:5120)	300.00	10.5281/zenodo.3741793
3	0.61	0.64	0.48	0.72	674	MacRog	POPC:SOL (288:14400)	298.00	10.5281/zenodo.13498
4	0.60	0.65	0.74	0.42	576	ECC-lipids	POPC:SOL (128:6400)	300.00	10.5281/zenodo.3335503
5	0.58	0.01	0.88	0.84		Berger	POPC:SOL (256:10342)	300.00	10.5281/zenodo.1402417
6	0.55	0.02	0.83	0.78		Berger	POPC:SOL (128:7290)	298.00	10.5281/zenodo.4643875
7	0.54	0.65	0.54	0.43	613	ECC-lipids	POPC:SOL (128:6400)	300.00	10.5281/zenodo.1118980
8	0.52	0.01	0.79	0.77	122	Slipids	POPE:SOL (500:25000)	310.00	10.5281/zenodo.3231342
9	0.51	0.10	0.87	0.57	140	Slipids	POPE:SOL (336:13460)	310.00	10.5281/zenodo.1293813
10	0.51	0.10	0.86	0.57	137	Slipids	POPE:SOL (336:13460)	310.00	10.5281/zenodo.1293813
11	0.50	0.16	0.64	0.69	297	ECClipids	POPS:SOL:SOD (72:3600:72)	298.00	10.5281/zenodo.1488094
12	0.48	0.67	0.36	0.40	690	CHARMM36	POPC:SOL (256:9767)	300.00	10.5281/zenodo.1306800
13	0.47	0.67	0.37	0.38	903	CHARMM36	POPC:SOL (1024:51200)	298.15	10.5281/zenodo.5767451

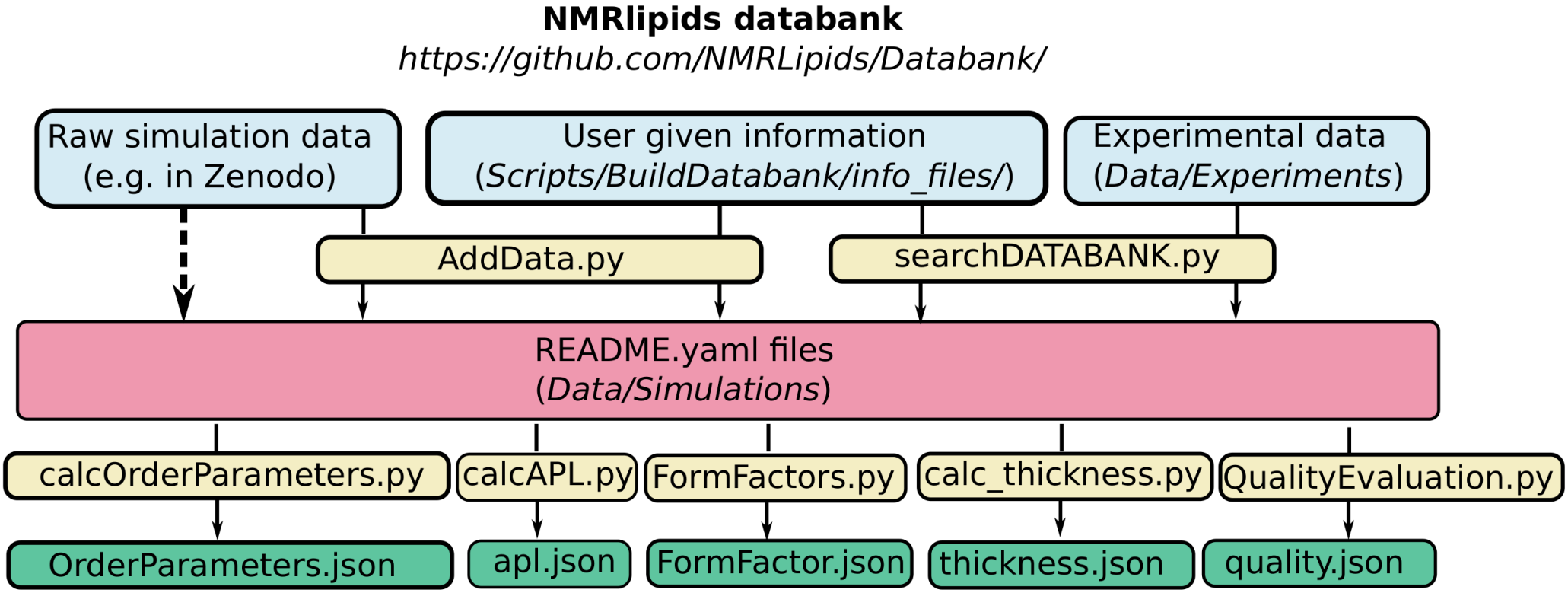
Selected comparisons



Full quality evaluations of over 100 simulations available from here:

https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/plotQuality.ipynb

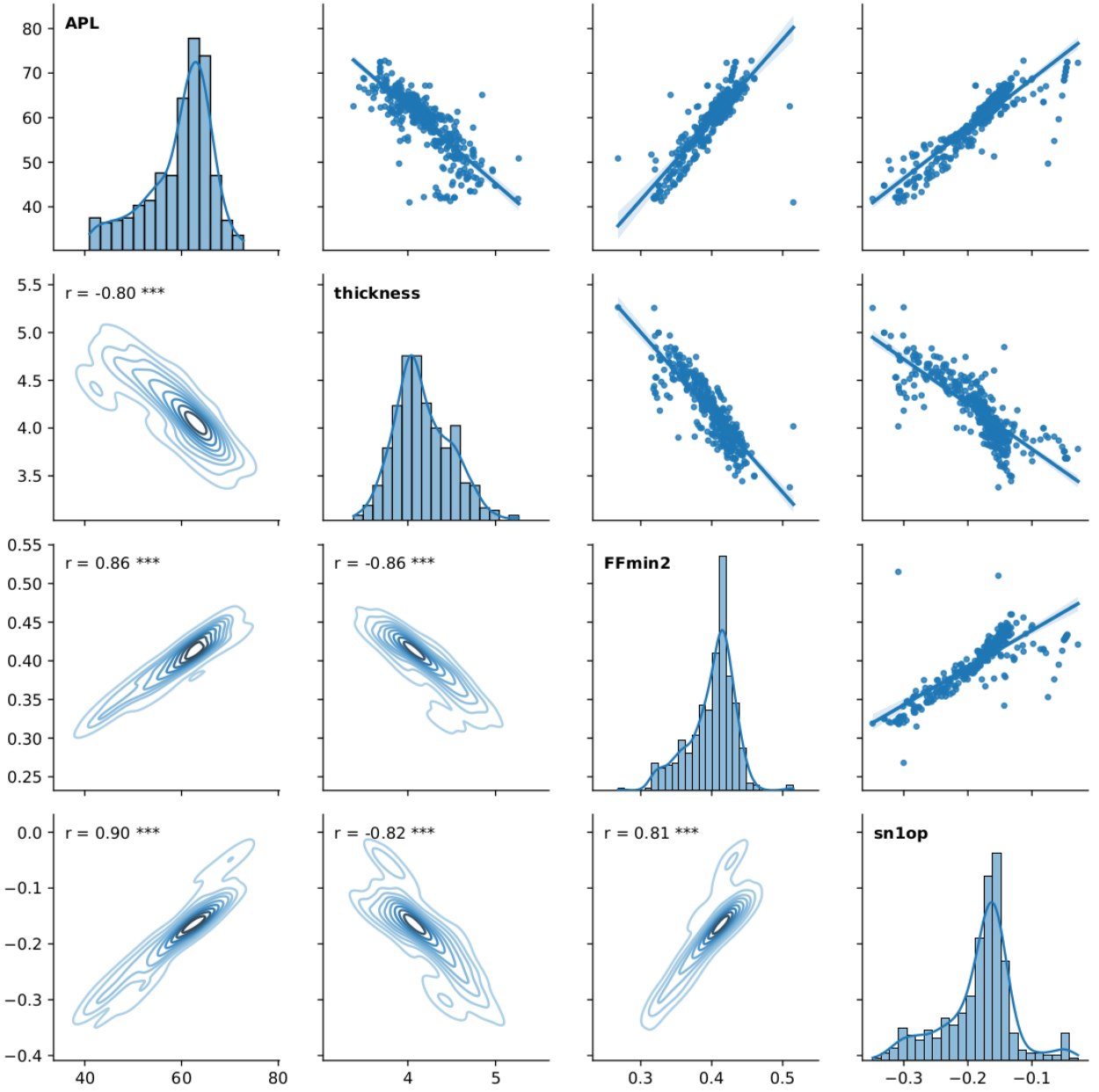
Overlay structure of the NMRlipids databank



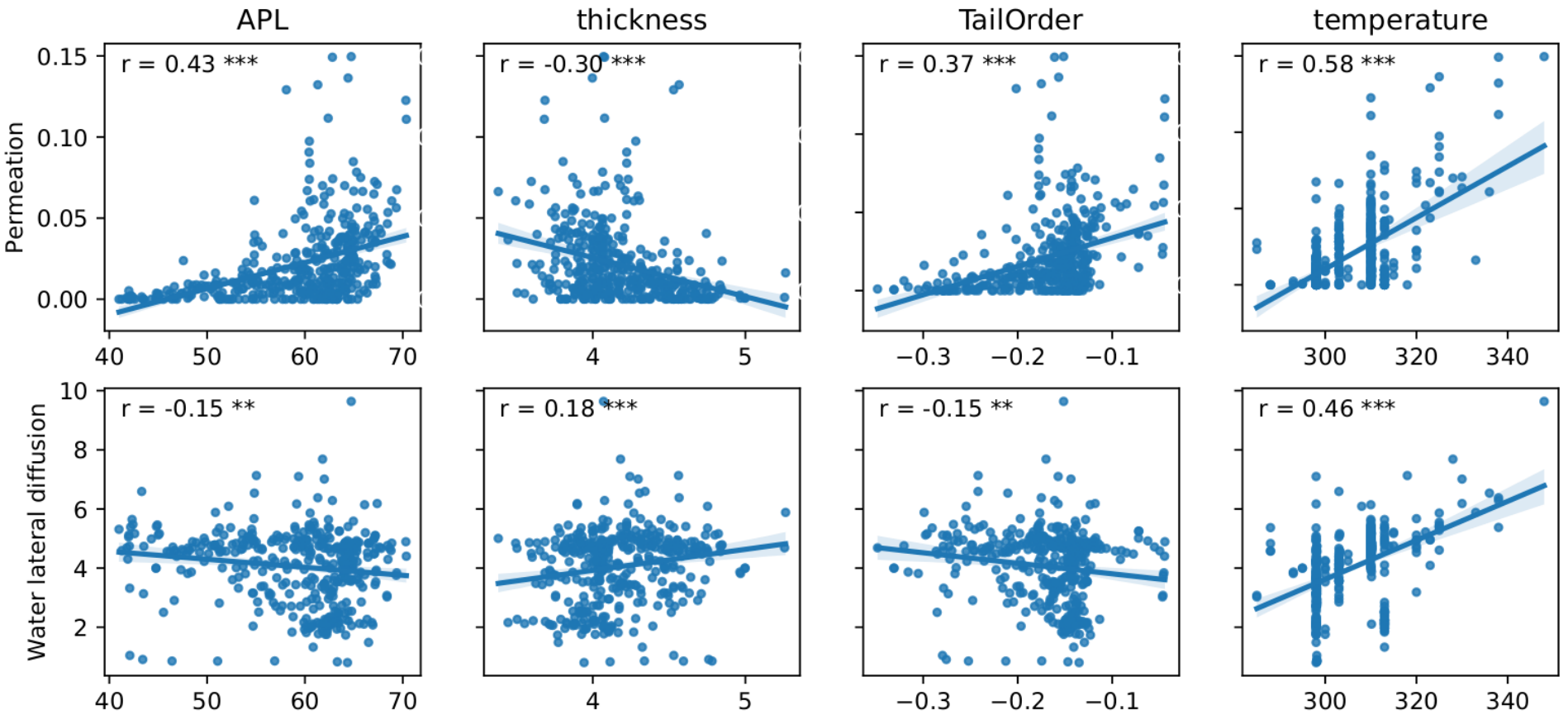
Analyzing simulations in the NMRlipids databank

- **Properties analyzed from all simulations and stored to the databank (Data/Simulations):**
  - C-H bond order parameters
  - X-ray scattering form factors
  - Area per lipid as a function of time
  - Membrane thickness from intersection of water and lipid densities
- **Further analyses can be done with Python**

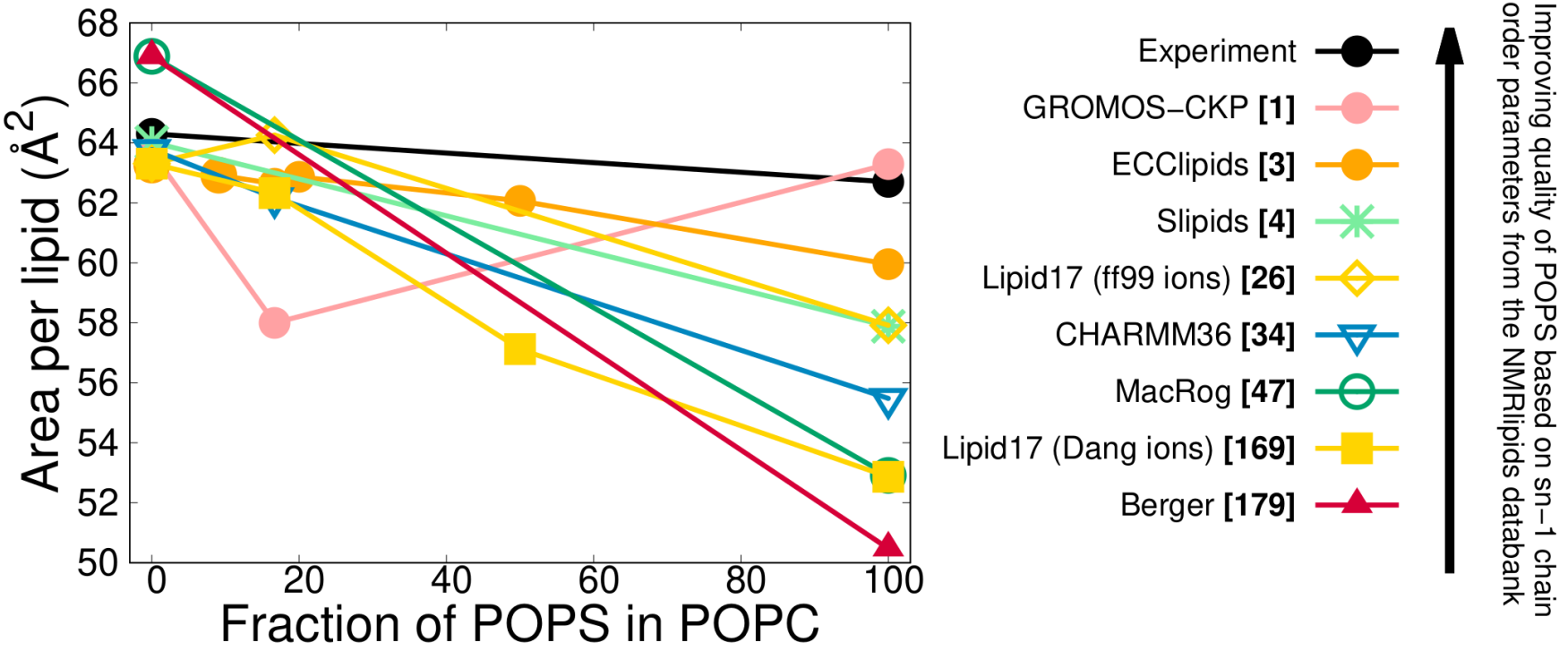
Correlations between area per lipid, thickness, form factor minima and acyl chain order parameters from the NMRlipids databank



Correlations of water diffusion through and along the membrane with membrane properties from the NMRlipids databank



Changes in the area per lipid upon addition of PS lipids to PC matrix from the NMRlipids databank



Analyzing simulations with Python

- Initializing databank:

In [2]:
```
import sys

sys.path.insert(1, '../../Databank/Scripts/BuildDatabank/')
from databankLibrary import download_link, lipids_dict, databank

path = '../../Databank/Data/Simulations/'
db_data = databank(path)
systems = db_data.get_systems()
```

- Loop over simulations:

In [7]:
```
for system in systems:
    print(system)
```

```
{'AUTHORS_CONTACT': 'Javanainen, Matti', 'FF_DATE': None, 'SYSTEM': '22CHOL_200POPC_9000SOL_310K', 'TYPEOFSYSTEM':
'lipid bilayer', 'TEMPERATURE': 310.15, 'PUBLICATION': None, 'NUMBER_OF_ATOMS': 55428, 'EXPERIMENT': {'CHOL': {},
'POPC': {}}, 'FF_SOURCE': 'CHARMM-GUI', 'COMPOSITION': {'CHOL': {'NAME': 'CHL1', 'COUNT': [10, 12], 'MAPPING': 'ma
ppingCHOLESTEROLcharmm.txt'}, 'POPC': {'NAME': 'POPC', 'COUNT': [100, 100], 'MAPPING': 'mappingPOPCcharmm.txt'}, '
SOL': {'NAME': 'TIP3', 'COUNT': 9000, 'MAPPING': 'mappingTIP3PCHARMMgui.txt'}}, 'TIMELEFTOUT': 0, 'CPT': [['chol10
_500ns.cpt']], 'TRJLENGTH': 500100.0, 'TRAJECTORY_SIZE': 1025713704, 'SOFTWARE_VERSION': 5.0, 'FF': 'CHARMM36', 'T
OP': [['chol10.top']], 'PREEQTIME': 0, 'DOI': '10.5281/zenodo.3237420', 'DATEOFRUNNING': '05/10/2021', 'TPR': [['c
hol10.tpr']], 'TRJ': [['chol10_500ns.xtc']], 'LOG': None, 'SOFTWARE': 'gromacs', 'DIR_WRK': '/media/osollila/Data/
tmp/DATABANK/', 'path': '../../Databank/Data/Simulations/006/559/006559139e730fc43b244726992145c2f37a1461/3c99810c
45a83b4ba0e54a69fdea8817498a8930/'}
{'AUTHORS_CONTACT': 'Javanainen, Matti', 'FF_DATE': 'pre-2020', 'SYSTEM': '200POPC_9000SOL_81SOD_81CLA_310K', 'TYP
EOFSYSTEM': 'lipid bilayer', 'TEMPERATURE': 310.0, 'PUBLICATION': None, 'NUMBER_OF_ATOMS': 53962, 'EXPERIMENT': {'
POPC': {}}, 'FF_SOURCE': 'http://mmkluster.fos.su.se/slipids/ and https://bitbucket.org/hseara/ions/', 'COMPOSITIO
N': {'CLA': {'NAME': 'CL', 'COUNT': 81, 'MAPPING': 'mappingCL.txt'}, 'POPC': {'NAME': 'POPC', 'COUNT': [100, 100],
'MAPPING': 'mappingPOPCslipids.txt'}, 'SOL': {'NAME': 'SOL', 'COUNT': 9000, 'MAPPING': 'mappingTIP3PwaterSlipids.t
xt'}, 'SOD': {'NAME': 'NA', 'COUNT': 81, 'MAPPING': 'mappingNA.txt'}}, 'TIMELEFTOUT': 0, 'CPT': None, 'TRJLENGTH':
100100.0, 'TRAJECTORY_SIZE': 199059704, 'SOFTWARE_VERSION': 4.6, 'FF': 'Slipids for lipids, Kohagen for NaCl', 'TO
P': [['500.top']], 'PREEQTIME': 0, 'DOI': '10.5281/zenodo.35193', 'DATEOFRUNNING': '12/10/2021', 'TPR': [['500.tpr
']], 'TRJ': [['500.xtc']], 'LOG': None, 'SOFTWARE': 'gromacs', 'DIR_WRK': '/usr/home/bort/Databank', 'path':
```

https://github.com/NMRLipids/DatabankExercises/blob/master/APL/Minimum_example.ipynb

Practical steps to analyze simulations from the NMRlipids databank

1. Clone (or download) the NMRlipids databank repository:
https://github.com/NMRLipids/Databank

2. Set the paths of *Databank/Scripts/BuildDatabank/* and *Databank/Data/Simulations/* folders when initializing the databank

In [2]:
```
import sys

sys.path.insert(1, '../../Databank/Scripts/BuildDatabank/')
from databankLibrary import download_link, lipids_dict, databank

path = '../../Databank/Data/Simulations/'
db_data = databank(path)
systems = db_data.get_systems()
```

Practical steps to analyze simulations from the NMRlipids databank

3. Loop over simulations, get the required files and perform the analysis

```
for system in systems:
    doi = system['DOI']
    trj = system.get('TRJ')
    tpr = system.get('TPR')
    trj_url = download_link(doi, trj[0][0])
    tpr_url = download_link(doi, tpr[0][0])
    if (not os.path.isfile(tpr_name)):
        response = urllib.request.urlretrieve(tpr_url, tpr_name)

    if (not os.path.isfile(trj_name)):
        response = urllib.request.urlretrieve(trj_url, trj_name)

     analysis command
```

4. Results can be saved with the same file organization as the original data

```
outputFOLDERS = subdir.replace("../../Databank/Data/Simulations/","../Data/WATERdiffusion/")
os.system('cp ' + READMEfilepath + ' ' + outputFOLDERS)
```

For example: https://github.com/NMRLipids/DataBankManuscript/blob/main/scripts/calcWATERdiffusion.py

Examples of codes analysing the NMRlipids databank

- Correlations between area per lipid, thickness and form factor, and area per lipid as function of membrane composition:

https://github.com/NMRLipids/DatabankExercises/blob/master/APL/AreaPerLipidAndThicknessExamples.ipynb

- Area per lipid:

https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/calcAPL.py

- Order parameters:

https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/calcOrderParameters.py

- Water diffusion in xy plane calculated from all simulations:

https://github.com/NMRLipids/DataBankManuscript/blob/main/scripts/calcWATERdiffusion.py
https://github.com/NMRLipids/DataBankManuscript/blob/main/scripts/plotWATERdiffusion.ipynb

Acknowledgements

DOCTORAL PROGRAMME IN MATERIALS RESEARCH AND NANOSCIENCE
(MATRENA)

NMRlipids contributors

Hanne Antila
Amelie Bacle
Lara Bort
Alexandru Botan
Pavel Buslaev
Rebeca García Fandiño
Andrea Catte
Sneha Dixit
Olle Edholm
Fernando Favela

Tiago Mendes Ferreira (NMR experiments)
Patrick Fuchs
Lukasz Cwiklik
Michael Girych
Ivan Gushchin
Peter Heftberger
Matti Javanainen
Pavel Jungwirth
Matej Kanduc
Jon Kapla

Batuhan Kav
Anne Kiirikki
Waldemar Kulig
Antti Lamberg
Claire Loison
Alexander Lyubartsev
Jesper Madsen
Hector Martinez-Seara
Josef Melcr
Blake Mertz

Markus S. Miettinen
Luca Monticelli
Jukka Määttä
Ricky Nencini
Vasily Oganesyan
O. H. Samuli Ollila
Georg Pabst
Chris Papadopoulos
Antonio Peón
Thomas Piggot

Ángel Piñeiro
Pierre Poulain
Pengyu Ren
Marius Retegan
Paula Milan Rodriguez
Tomasz Rog
Suman Samantray
Hubert Santuz
Otto Schullian
Peter Tieleman

Joona Tynkkynen
Sergey Vilov
Salla Virtanen
Alexander Vogel
Alex de Vries
Mark Wilson