Status of the NMRlipids databank

NMRlipids winterschool 2021

December 17th 2021
Online

NMRlipids databank

https://github.com/NMRLipids/Databank

(www.databank.nmrlipids.fi)

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

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

Raw simulation data

Publicly available, e.g., in Zenodo

Experimental data
(git repository with yaml and data files)

Indexed experimental data (***Data/experiments***)

Databank builder
*(Python code: **AddData.py**)*

Indexes publicly available **simulation data**
based on information given by contributor

Quality evaluator
(Python code: **searchDATABANK.py**, **QualityEvaluation.py**)

Connects experimental and simulation
Datasets and calculates quality measures

NMRlipids Databank
(git repository with yaml files)

Folder of each simulation locating in **Data/Simulations** contains:

- **README.yaml file containing all relevant information of a simulation**
- **Quality evaluation of simulation based on C-H bond order parameters and form factors**
- **Area per lipid as a function of time**
- **Average thickness of the system**

Adding data

Detailed instructions: https://github.com/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**

A web app coming

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!**
- **Statistics:** ***https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/stats.ipynb***

Experimental data

- 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**
- Currently: DOI of the publication, temperature, molar fractions of lipids, ion concentration, total lipid concentration (or full hydration), information of counterions

Quality evaluation

- 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

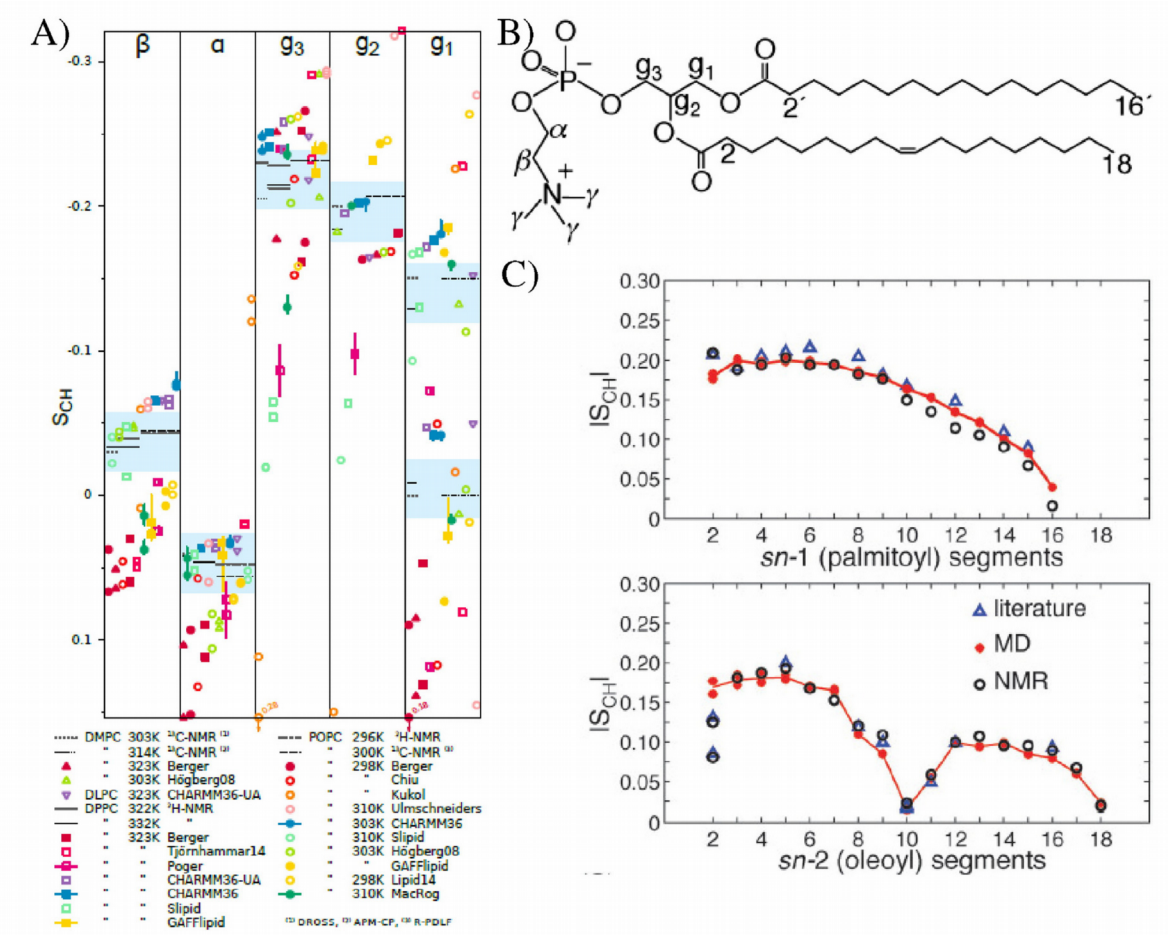
- Quality measure for simulation is determined by comparing the C-H bond order parameters and X-ray scattering form factors with experiments

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



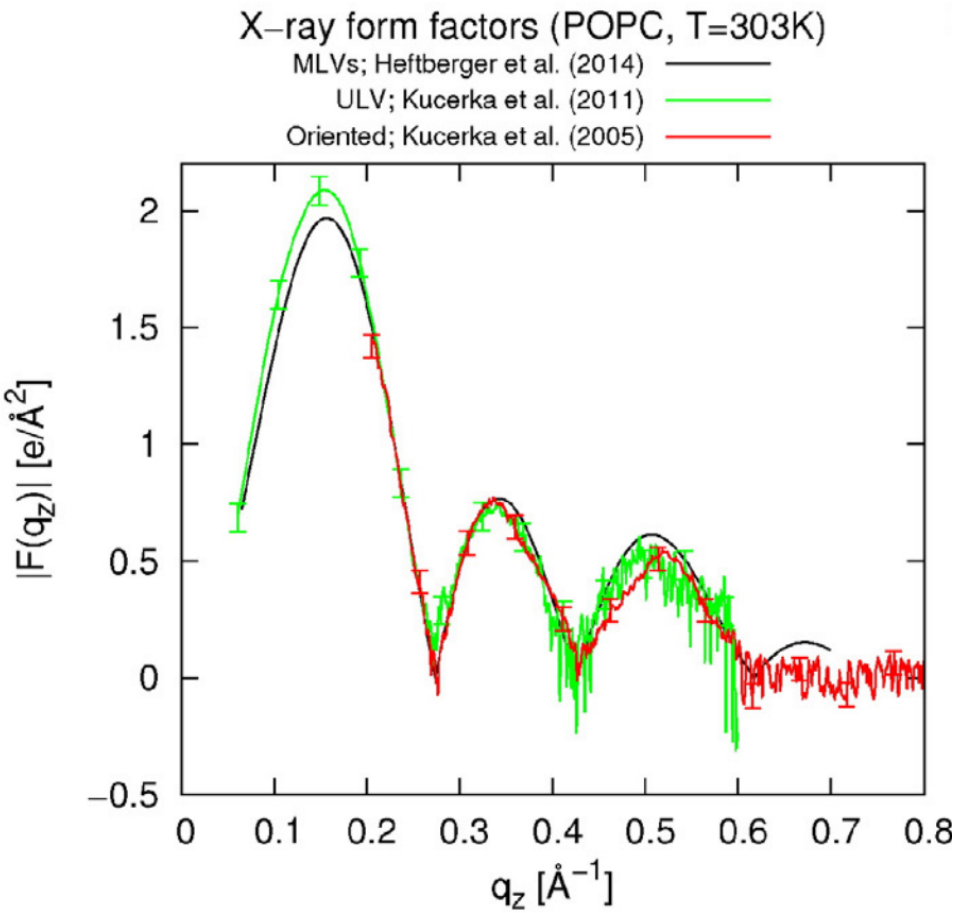
http://dx.doi.org/10.1016/j.bbamem.2016.01.019

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 quality measures

Individual order parameters:

$$S_q = -\log_{10}(P)$$

$$P = \int_{S_{exp}-\Delta S_{exp}}^{S_{exp}+\Delta S_{exp}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x$$

P = probability for simulation value to locate within experimental error bars
S_{exp} = experimental order parameter
Δs_{exp} = error bars of experimental order parameter
σ = order parameter from simulation
μ = standard deviation of order parameter from simulation

- S_q approaches zero when simulation order parameter approaches experimental value (P approaches 1)

NMRlipids quality measures

Individual lipid types in simulation:

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

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

Fragments within a simulation:

$$S_q^{\mathrm{frag}} = \sum_{\mathrm{lipid}} \chi_{\mathrm{lipid}} \langle S_q^{\mathrm{frag}}[\mathrm{lipid}]\rangle_{\mathrm{lipid}}$$

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

NMRlipids quality measures

Membrane dimension using form factor (in progress):

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

F_s = form factor from a simulation
F_e = form factor from experiment
ΔF_e = error of form factor from experiment
$(1,\ldots,N_q)$ = 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)

Ranking simulations based on quality measures

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

Sorted based on total quality

	headgroup	sn-1	sn-2	total	Forcefield	Molecules	Number of molecules	Temperature	DOI
0	7.546477	19.308345	17.561283	14.805369	CHARMM36	POPC:SOL	(256:9767)	300.0	10.5281/zenodo.1306800
1	62.126873	0.535656	1.682661	21.448397	Slipids	POPC:SOL	(512:23943)	298.0	10.5281/zenodo.166034
2	5.690801	39.470786	34.284259	26.481948	CHARMM36	POPS:SOL:SOD	(128:4480:128)	298.0	10.5281/zenodo.1129415
3	105.170638	0.356573	1.741698	35.756303	Berger	POPC:SOL	(128:7290)	298.0	10.5281/zenodo.4643875
4	170.675169	4.764065	7.588560	61.009265	Slipids	POPS:SOL:SOD	(128:4480:128)	298.0	10.5281/zenodo.1129441
5	171.442649	8.981030	8.384775	62.936151	Slipids	POPS:SOL:SOD	(128:4480:128)	298.0	10.5281/zenodo.1129441
6	4.283150	119.422468	85.873231	69.859616	CHARMM36	POPS:SOL:SOD	(128:4480:128)	298.0	10.5281/zenodo.1129415
7	25.051457	8.321688	8.162352	117.475004	slipids	CHOL:POPC:SOL	(256:256:20334)	298.0	10.5281/zenodo.159434

NMRlipids databank structure

https://github.com/NMRLipids/Databank

Raw simulation data

Publicly available, e.g., in Zenodo

Experimental data
(git repository with yaml and data files)

Indexed experimental data (***Data/experiments***)

Databank builder
*(Python code: **AddData.py**)*

Indexes publicly available **simulation data**
based on information given by contributor

Quality evaluator
(Python code: **searchDATABANK.py**, **QualityEvaluation.py**)

Connects experimental and simulation
Datasets and calculates quality measures

NMRlipids Databank
(git repository with yaml files)

Folder of each simulation locating in **Data/Simulations** contains:

- **README.yaml file containing all relevant information of a simulation**
- **Quality evaluation of simulation based on C-H bond order parameters and form factors**
- **Area per lipid as a function of time**
- **Average thickness of the system**

Analyzing simulations

- **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
- **Graphical access to the data: *www.databank.nmrlipids.fi***
- **Further analyses can be done with Python**

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

Analyzing simulations with Python

- **Practical steps:**

1. Get (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()
```

3. Loop over simulations and perform the analysis within the loop

- **Examples:**

- Area per lipid as function of membrane composition, correlations between area per lipid, thickness and form factor:
https://github.com/NMRLipids/DatabankExercises/blob/master/APL/AreaPerLipidAndThicknessExamples.ipynb

- Water diffusion in xy plane calculated form all simulations:
https://github.com/NMRLipids/DataBankManuscript/blob/main/scripts/calcWATERdiffusion.py
https://github.com/NMRLipids/DataBankManuscript/blob/main/scripts/plotWATERdiffusion.ipynb

Open issues

- **United atom simulations:**
  https://github.com/NMRLipids/Databank/issues/9
- **Stereospecific information on isomers:**
  https://github.com/NMRLipids/Databank/issues/1
- **Extending to other than Gromacs simulations:**
  https://github.com/NMRLipids/Databank/issues/5
- **“Sanity checks” for the data (Equlibration etc.):**
  https://github.com/NMRLipids/Databank/issues/3
- **NMRlipids III (systems with cholesterol):**
  https://github.com/NMRLipids/NmrLipidsCholXray
- **Web app for adding the data**
- **Cleaning, organizing and documenting the code**
- **Other than pure lipid bilayer simulations**

Work in progress

- Including form factors to quality evaluation 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 as showcases for the first publication

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.**