NMRLipids Databank

Internal structure and functioning

Alexey M. Nesterenko
comcon1@pm.me
https://github.com/comcon1

NMRLipids2025, Bergen, Norway

Coding | Building | Using

Part II: Building

Deploying the Databank

```
$:~/repo/test$ git clone https://github.com/NMRLipids/Databank
Cloning into 'Databank'...
remote: Enumerating objects: 4870, done.
remote: Counting objects: 100% (450/450), done.
remote: Compressing objects: 100% (41/41), done.
remote: Total 4870 (delta 412), reused 417 (delta 408), pack-reused 4420 (from 1)
Receiving objects: 100% (4870/4870), 25.57 MiB | 8.91 MiB/s, done.
Resolving deltas: 100% (3351/3351), done.
$:~/repo/test$ cd Databank
$:~/repo/test/Databank$ ls Data
$:~/repo/test/Databank$ git submodule update --init
Submodule 'Data' (https://github.com/NMRLipids/BilayerData.git) registered for path
  'Data'
Cloning into '/home/comcon1/repo/test/Databank/Data'...
Submodule path 'Data': checked out 'd8e07b41c5b562f81e4154b8dcb45ed35ce657ef'
$: ~/repo/test/Databank$ ls Data
LICENSE  Molecules  Ranking  Simulations  experiments  info_files  lipid_json_buildH
```

Data is now under separated version control.

Switching to your Data-fork

```
$:~/repo/test/Databank$ cd Data
$:~/repo/test/Databank/Data$ git remote set-url origin https://github.com/comcon1/BilayerData.git
$:~/repo/test/Databank/Data$ git checkout main
Previous HEAD position was d8e07b41 Inherit license from Databank
Switched to branch 'main'
Your branch is up to date with 'origin/main'.
$:~/repo/test/Databank/Data$ git status
On branch main
Your branch is up to date with 'origin/main'.

nothing to commit, working tree clean
$:~/repo/test/Databank/Data$ git remote -v
origin  https://github.com/comcon1/BilayerData.git (fetch)
origin  https://github.com/comcon1/BilayerData.git (push)
```

Adding a simulation with existing molecules

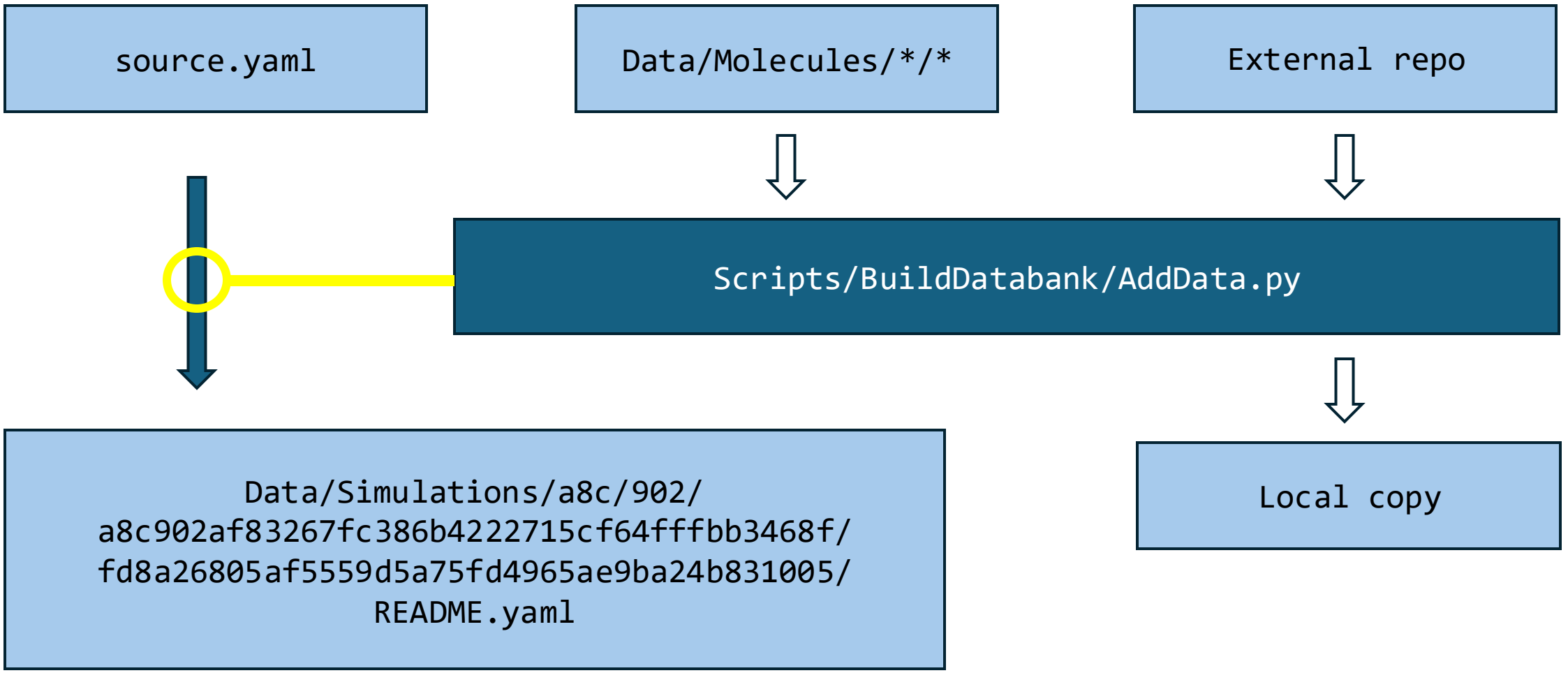
source.yaml

```yaml
DOI: 10.5281/zenodo.1487895
SOFTWARE: gromacs
AUTHORS_CONTACT: Melcr, Josef
SYSTEM: 12POPS_60POPC_3483SOL_51SOD_39CLA_298K
SOFTWARE_VERSION: 2018.0
FF: Lipid17 and Dang ions
TRJ: lipid17_5PC-1PS-mix_39NaCl_298K.xtc
TPR: lipid17_5PC-1PS-mix_39NaCl_298K.tpr
PREEQTIME: 0
TIMELEFTOUT: 0
COMPOSITION:
 POPS:
  NAME: POPS
  MAPPING: mappingPOPSlipid17_POPSres.yaml
 POPC:
  NAME: POPC
  MAPPING: mappingPOPCcharmm.yaml
 SOL:
  NAME: SOL
  MAPPING: mappingTIP3Pwater.yaml
 SOD:
  NAME: NA
  MAPPING: mappingNA.yaml
 CLA:
  NAME: CL
  MAPPING: mappingCL.yaml
```

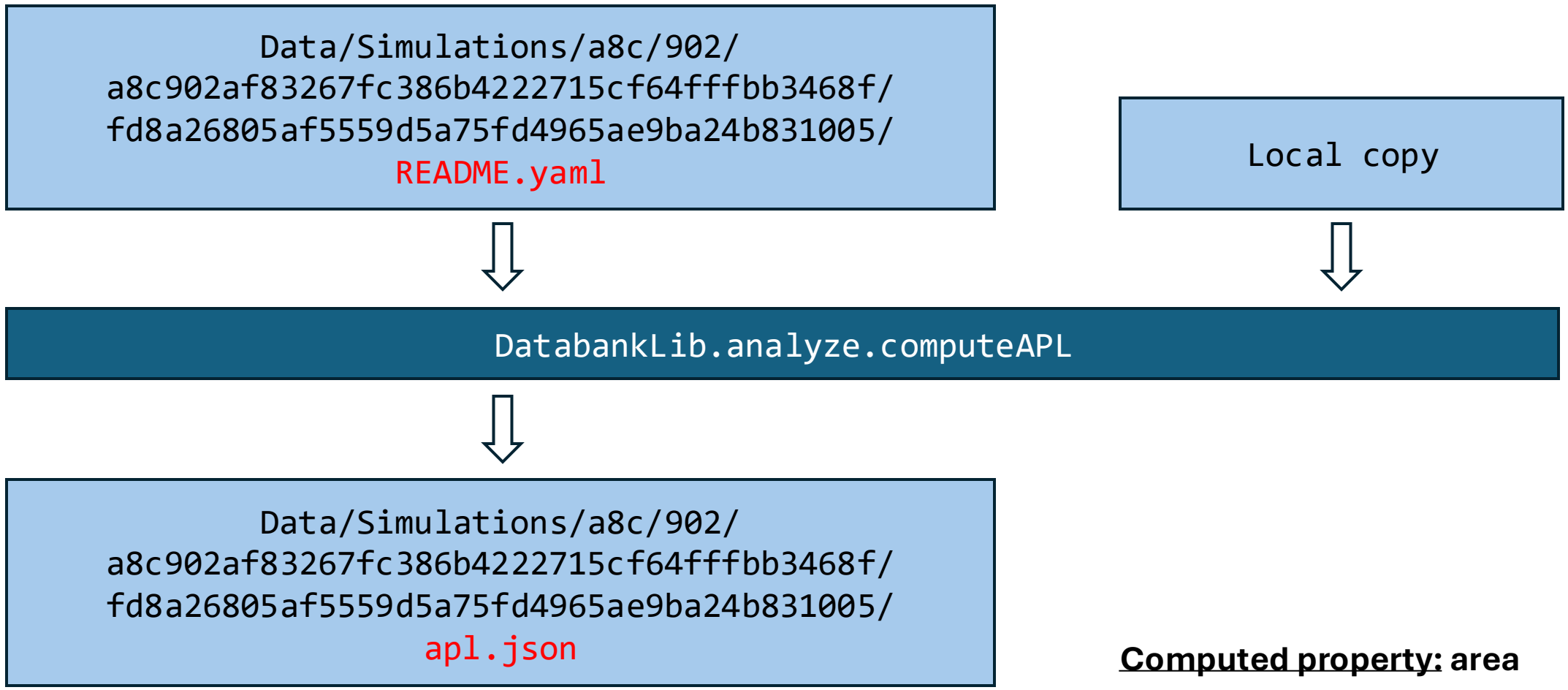
Mappings

```
Molecules/
|-- membrane
|   |-- CHOL
|   |   `-- mappingCHOLESTEROLlipid14.yaml
|   |-- DPPC
|   |   `-- mappingDPPCberger.yaml
|   |-- POPC
|   |   |-- mappingPOPCcharmm.yaml
|   |   `-- mappingPOPClipid14.yaml
|   |-- POPE
|   |   |-- mappingPOPEGROMOS43A1-S3.yaml
|   |   `-- mappingPOPEcharmm.yaml
|   `-- TOCL
|       `-- mappingTOCLcharmm.yaml
`-- solution
    |-- CLA
    |   `-- mappingCL.yaml
    |-- SOD
    |   |-- mappingNA.yaml
    |   `-- mappingSOD.yaml
    `-- SOL
        |-- mappingSPCwater.yaml
        |-- mappingTIP3PCHARMMgui.yaml
        `-- mappingwaterlipid14.yaml
```

Building script 1: *AddData.py*



Analysis through `DatabankLib.analyze` library



Computed property: area

Building script 2: *calcProperties.sh*

```python
from DatabankLib.core import *
from DatabankLib.analyze import computeAPL

systems = initialize_databank()
result_dict = {
    RCODE_COMPUTED: 0,
    RCODE_SKIPPED: 0,
    RCODE_ERROR: 0}

for system in systems:
    logger.info("System title: " + system['SYSTEM'])
    logger.info("System path: " + system['path'])
    res = computeAPL(system, logger)
    result_dict[res] += 1

print(f"""
    COMPUTED: {result_dict[RCODE_COMPUTED]}
    SKIPPED: {result_dict[RCODE_SKIPPED]}
    ERROR: {result_dict[RCODE_ERROR]}
    """)
```

+Computed properties

```
DRPCOrderParameters.json

FormFactor.json
README.yaml
TotalDensity.json
apl.json
eq_times.json
thickness.json
```

Building script 3: *searchDATABANK.py*

README.yaml

```yaml
COMPOSITION:
  POPC:
    NAME: POPC
    MAPPING: mappingPOPC-OPLS4.yaml
    COUNT:
    - 100
    - 100
  SOL:
    NAME: SPC
    MAPPING: mappingwaterlipid14.yaml
    COUNT: 1500

...

TEMPERATURE: 300.0
```

```
$: ls experiments/OrderParameters/10.1021/acs.jpcb.4c04719/2/ -1
POPC_Order_Parameters.dat
POPC_Order_Parameters.json
README.yaml

$: cat experiments/OrderParameters/10.1021/acs.jpcb.4c04719/2/README.yaml
DOI: 10.1021/acs.jpcb.4c04719
TEMPERATURE: 298.0
MOLAR_FRACTIONS:
  POPC: 1
ION_CONCENTRATIONS:
  POT: 0
  SOD: 0
  CLA: 0
  CAL: 0
TOTAL_LIPID_CONCENTRATION: 7.4
COUNTER_IONS:
```

Experiment's README.yaml

`Scripts/BuildDatabank/searchDATABANK.py`

README.yaml

```yaml
EXPERIMENT:
  ORDERPARAMETER:
    POPC:
      10.1021/acs.jpcb.4c04719:
10.1021/acs.jpcb.4c04719/2
  FORMFACTOR: {}
```

Building script 3: *QualityEvaluation.py*

```
EXPERIMENT:
  ORDERPARAMETER:
    POPC:
      10.1021/acs.jpcb.4c04719:
10.1021/acs.jpcb.4c04719/2
  FORMFACTOR: {}
```

Data/experiments/OrderParameters/
10.1021/acs.jpcb.4c04719/2/
POPC_OrderParameter.json

Data/Simulations/a8c/902/
a8c902af83267fc386b4222715cf64fffbb3468f/
fd8a26805af5559d5a75fd4965ae9ba24b831005/
POPC_OrderParameter.json

```
apl.json
eq_times.json
FormFactor.json
LipidDensity.json
POPC_FragmentQuality.json
POPCOrderParameters.json
POPC_OrderParameters_quality.json
README.yaml
SYSTEM_quality.json
thickness.json
TotalDensity.json
WaterDensity.json
```

Scripts/BuildDatabank/QualityEvaluation.py

+Quality

Simulation record is done!

Data/Simulations/a8c/902/
a8c902af83267fc386b4222715cf64fffbb3468f/
fd8a26805af5559d5a75fd4965ae9ba24b831005/

```
apl.json
eq_times.json
FormFactor.json
LipidDensity.json
POPC_FragmentQuality.json
POPCOrderParameters.json
POPC_OrderParameters_quality.json
README.yaml
SYSTEM_quality.json
thickness.json
TotalDensity.json
WaterDensity.json
```

README + Computed prop-s + Quality

Push to a branch in your fork

⇩

Open pull-request

⇩ Pre-merge actions

Get reviewed

⇩

Rebase&merge

⇩

Post-merge actions

Part III: Usage

Getting started

```python
import os
os.environ["NMLDB_ROOT_PATH"] = '/home/comcon1/repo/Databank'

from DatabankLib.core import *
from DatabankLib.databankLibrary import *

systems = initialize_databank()
print("Loaded ", len(systems))
print(systems.loc(803))
```

Path to the data storage
is set *before* module initialization

Construction simulation array

```
Databank initialized from the folder:
/home/comcon1/repo/Databank/Data/Simulations
813
System(803):
d8c/596/d8c5962a1961606bf11d5d8d159df45b5da9a31d/4514cc6549784a6dc5129c
8082c492175e7a35b7/
```

Hash-path allows to locate the system in `Data/Simulations` folder

System entity: access to README.yaml

```
{'AUTHORS_CONTACT': 'Javanainen, Matti',
 'COMPOSITION': {'CAL': {'COUNT': 73,
                         'MAPPING': 'mappingCA.ya
                         'NAME': 'CA'},
                 'CLA': {'COUNT': 146,
                         'MAPPING': 'mappingCL.ya
                         'NAME': 'CL'},
                 'POPC': {'COUNT': [100, 100],
                          'MAPPING': 'mappingPOP
                          'NAME': 'POPC'},
                          'COUNT': 9000,
                          'MAPPING': 'mappingTIP3PCHARMMgui.yaml',
                          'NAME': 'TIP3'}},
                          2021',
                          ort/Databank',
 'DOI': '10.5281/zenodo.45008',
 'EXPERIMENT': {'FORMFACTOR': {}, 'ORDERPARAMETER': {'POPC': {}}},
 'FF': 'CHARMM36 with ECC-scaled ions',
 'FF_DATE': None,
 'FF_SOURCE': 'http://www.charmm-gui.org for lipids and '
              'https://bitbucket.org/hseara/ions/ for ions',
 'ID': 353,
 'LOG': None,
 'NUMBER_OF_ATOMS': 54019,
...
 'TRJ': [['md.xtc']],
 'TRJLENGTH': 100100.0,
 'TYPEOFSYSTEM': 'lipid bilayer',
 'path':
'02e/922/02e9220d27c2d6f4aa76992ae66014ac8d5557e0/4237bd2c8e11c1b483253
86fd672fd6fd3e75894/'}
```

Composition information

Link to trajectory storage

Description of `REAMDE.yaml` standard:
https://nmrlipids.github.io/READMEcontent.html

https://www.databank.nmrlipids.fi/trajectories/**353**

Path to the folder

System is more than a dictionary

```python
from DatabankLib.core import System
from pprint import pprint

s: System = systems[12]

print(s['path'])
pprint(s.content)
```

`content` property gives access to information about molecules

```
02e/922/02e9220d27c2d6f4aa76992ae66014ac8d5557e0/4237bd2c8e11c1b48325386fd672fd6fd3e75894/
{'CAL': NonLipid(CAL),
 'CLA': NonLipid(CLA),
 'POPC': Lipid(POPC),
 'SOL': NonLipid(SOL)}
```

Mirroring interfaces:

```python
print(s.readme['FF'])
print(s['FF'])
```

Molecule handling in **DatabankLib**

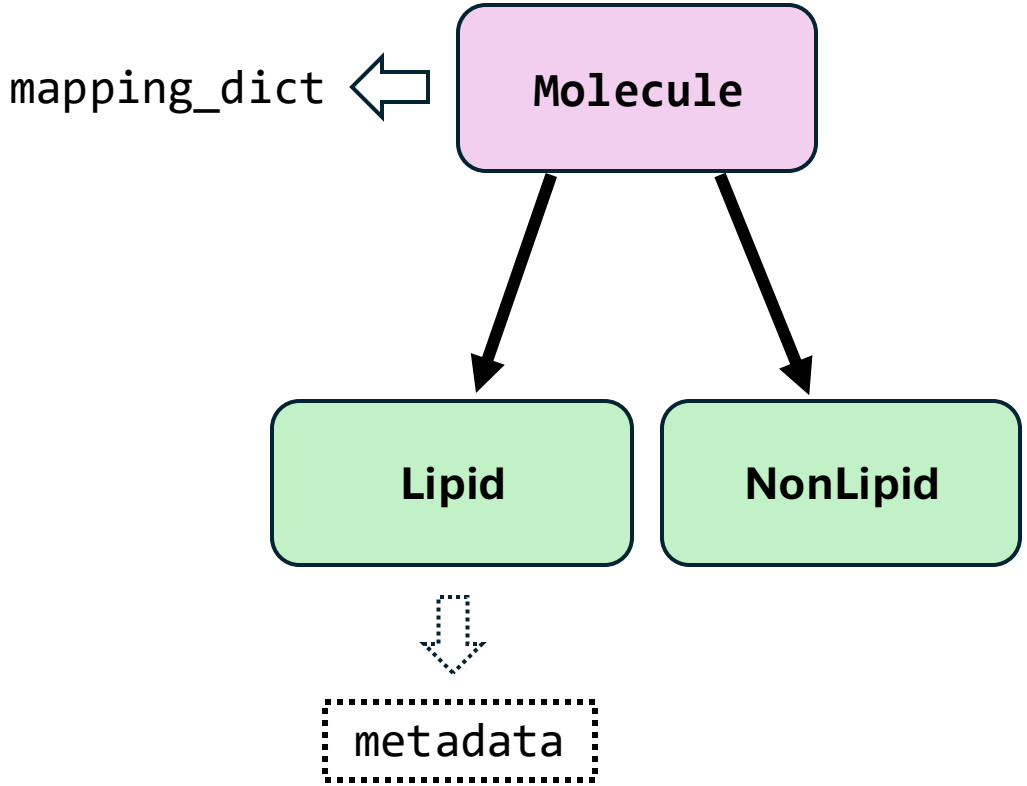
```python
from DatabankLib.settings.molecules import Molecule

m: Molecule = s.content['POPC']

print("m = ", m)
print("m.name = ", m.name)
print("mol's g3: ", m.mapping_dict['M_G3_M'])
```

```
m =  Lipid(POPC)
m.name =  POPC
mol's g3:  {'ATOMNAME': 'C1', 'FRAGMENT': 'glycerol backbone'}
```

Access to mapping information through **System.content** property

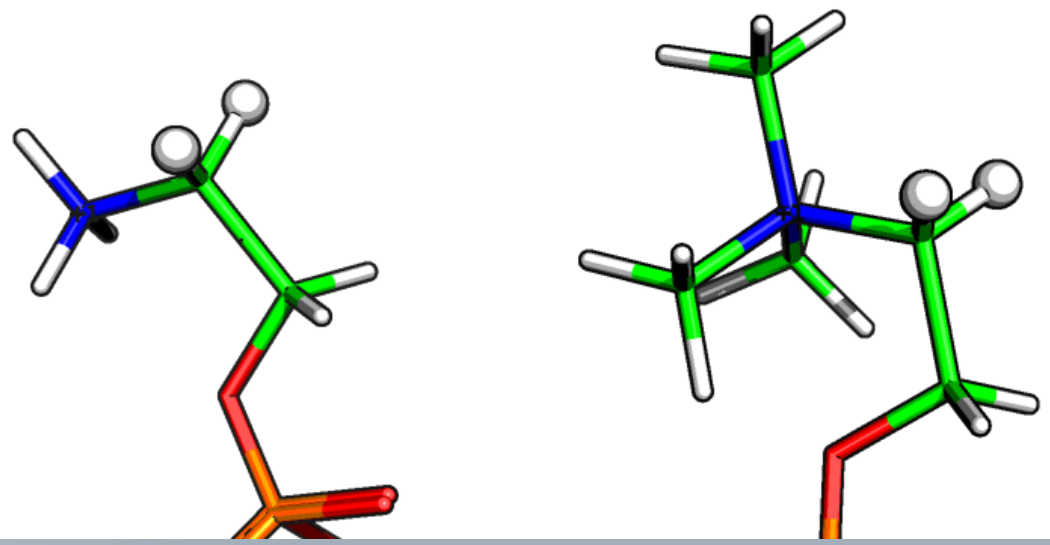


Example using **`DatabankLib`** library to access the databank

```python
ss = initialize_databank()
n_lips = []
for s in ss:
    lips = [m for _,m in s.content.items() if (type(m) is Lipid)]
    n_lips.append(len(lips))

i_sort = np.argsort(n_lips)[::-1]
c = 1
for i in i_sort:
    s = ss[i]
    print(f"System #{s['ID']} ({s['AUTHORS_CONTACT']})")
    print("Content: " + ",".join(s.content.keys()))
    c += 1
    if c == 10:
        print("...")
        break
```

```
Databank initialized from the folder:
/home/comcon1/repo/Databank/Data/Simulations
System #776 (Reza Talandashti and Nathalie Reuter)
Content: POPC,POPE,POPS,POPI,CER180,CHOL,SOL,SOD
System #774 (Reza Talandashti and Nathalie Reuter)
Content: POPC,POPE,POPS,SAPI24,SM16,CHOL,SOL,SOD
System #768 (Anusha Lalitha)
Content: POPE,POPC,POPS,CHOL,SM16,SOL,CAL,CLA,GM1
System #769 (Suman Samantray)
Content: POPE,POPC,POPS,CHOL,SM16,SOL,GM1
System #767 (Suman Samantray)
Content: POPE,POPC,POPS,CHOL,SM16,GM1,SOL,SOD,CLA
System #3 (Matti Javanainen (), Hector Martinez-Seara ())
Content: DMPC,CLA,SAPI,SOD,POPI,SOL,SLPI
System #491 (Javanainen, Matti)
Content: DPPC,CHOL,DOPC,SOL,SOD,CLA
System #487 (Javanainen, Matti; Martinez-Seara, Hector;
Vattulainen, Ilpo)
Content: DPPC,CHOL,DOPC,SOL,SOD,CLA
System #762 (Fernando Favela)
Content: POPC,CHOL,SM16,SOL
...
```



```python
def getBetaOP(s, lname, hn=1):
    opfn = os.path.join(
      NMLDB_SIMU_PATH, s['path'],
      f"{lname}OrderParameters.json")
    with open(opfn) as fd:
        opdict = json.load(fd)
    v = float(opdict[f"M_G3C5_M M_G3C5H{hn}_M"][0][0])
    return v

peList1 = []; pcList1 = []
peList2 = []; pcList2 = []
for s in tqdm(systems):
    for mname, mol in s.content.items():
        if type(mol) is Lipid:
```

```
100%|██████████| 813/813 [00:00<00:00, 1311.41it/s]
0.021 +- 0.004
0.013 +- 0.004
-0.047 +- 0.002
-0.047 +- 0.002
```

PE

PC

Thank you for your attention!