NMRlipids Databank: Overlay Databank of Lipid Membrane Simulations Arising from Open Collaboration

Anne Kiirikki[1], **...**[2], **and O. H. Samuli Ollila**[1,*]

[1]University of Helsinki, Institute of Biotechonology, Helsinki, Finland
[2]Affiliation, department, city, postcode, country
[*]samuli.ollila@helsinki.fi

ABSTRACT

We present a databank of lipid bilayer simulations from the NMRlipids open collaboration project.

1 Introduction

The importance of sharing MD simulation data following the FAIR principles[1] has been widely recognized[2–9], and databanks are emerging[9–16]. The relevance of quality evaluation of simulation trajectories in databanks regarding technical details of simulations and accuracy of the underlying physical description of the system (force field) has become evident[3,7,10] and such quality evaluation has in some cases also been implemented[10,12]. However, straightforward quality comparisons between individual simulations or force fields within these databanks remain challenging. While importance of such databanks for MD simulations is widely recognized[2–9] and different kinds of approaches are emerging[9–16], generally accepted protocols and best practices are still under active development.

Here we present a solution for lipid bilayers based on overlay databank structure illustrated in Fig. 1. The concept of overaly databank is developed here to solve the practical challenges in generating databanks of MD simulation data enabling flexible analyses over large sets of simulation data, but it pontentially used for wide range of situations, particularly when storage of raw data requires significant resources and final outcomes or best practices are not yet clear, overlay databank approach lowers the barrier to start without compromizing the long term stability or scalability.

The practical relevance of the NMRlipids databank is exemplified by automatic quality evaluation and ranking of large amount of MD simulation data, data-driven analysis detecting correlations between properties of model cell membranes and analyses of rare phenomena that are beyond the scope of standard MD simulation studies. The NMRlipids databank provides new tools for researchers in wide range of fields in academia and industry from cell membrane biology to lipid nanoparticle formulations and data-driven computational chemistry and machine learning.

2 Results

2.1 Design of the NMRlipids databank

The key idea of the overlay databank is that the storage of raw data in layer 1 is distributed in publicly available repositories or other servers with long term stability and permanent links such as digital object identifiers. The core of the databank, layer 2, contains only information on the location and content of the raw data, thereby not requiring large resources to handle and maintain. This lowers the barrier for starting such databank as well as for long term storage. The NMRlipids databank is essentially a git repository containing information on the location of raw data and its indexing with universal naming convention. In addition to all computers where the databank is developed and used, the NMRlipids databank git is stored to Zenodo server, thereby enabling a very cost effective long term storage for the databank. The databank can be used in layer 3 by accessing the raw data and information stored in the databank by employing the universal naming conventions for atoms, molecules and simulation details. The applications can be linked to the core databank (layer 2) without actually including them, thereby enabling flexible extension of the data without compromising the simplicity and lightness of the core databank.

Currently the databank is composed of approximately 500 trajectories with the total length of approximately 231 microseconds of which most are contributed for the previous publications from the NMRlipids open collaboration[17–20]. The distribution of lipids, force fields, length of the trjajectories and available binary mixtures are shown in Fig. 1.

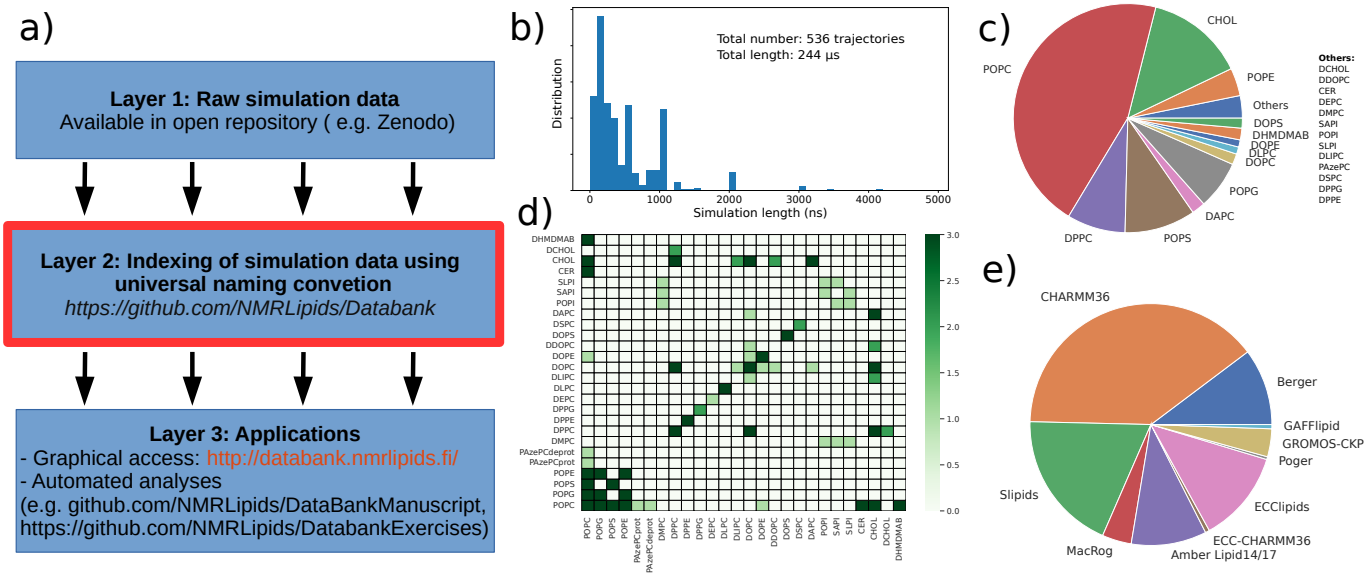


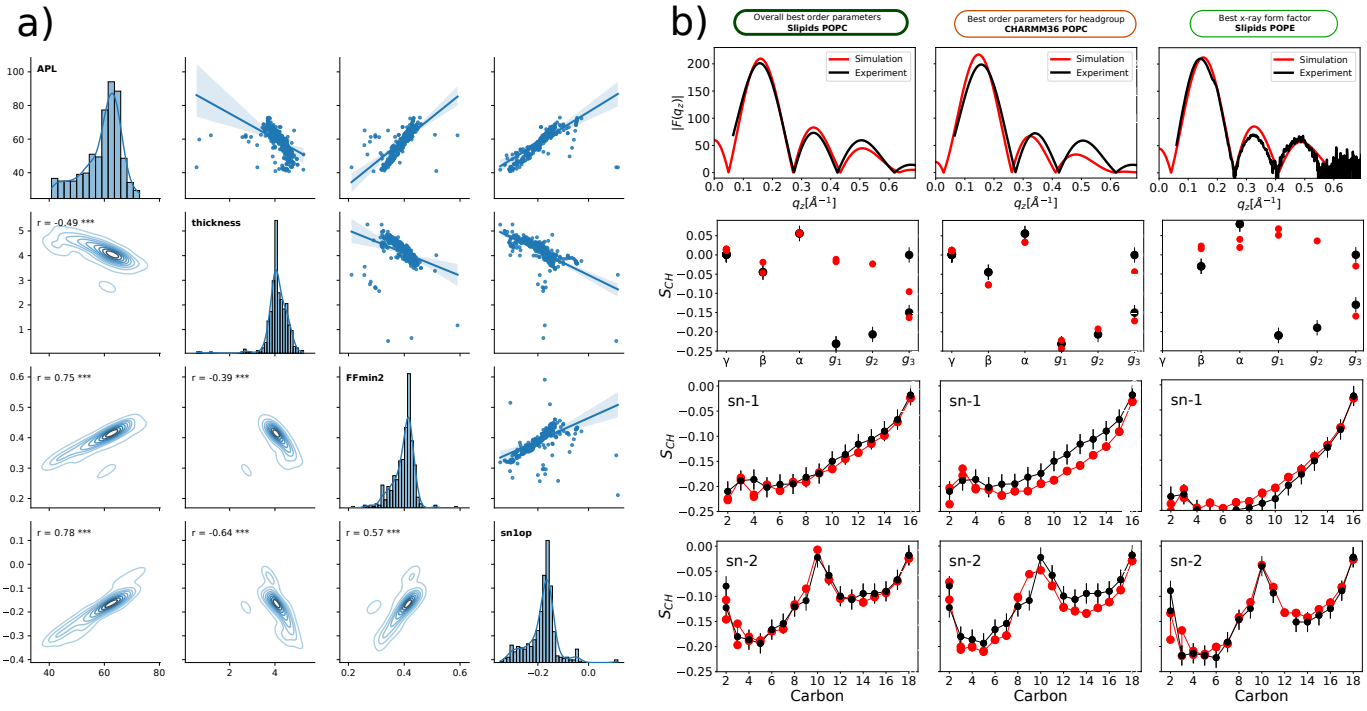
Figure 1. a) Structure of an overlay databank. More detailed structure of the layer 2 in the NMRlipids databank is illustrated in Fig. 3 in the SI. b) Distribution of the lengths of the trajectories, total number of trajectories and total lenght of the simulations in the NMRlipids databank. c) Distribution of lipids present in the trajectories in the NMRlipids databank. Lipids occuring in five or less simulations ('others') are listed in the right. d) Currently available binary mixtures in the NMRlipids databank. e) Distribution of force fields in the simulations in the NMRlipids databank. The figures and numbers are created on 9th of May 2022.

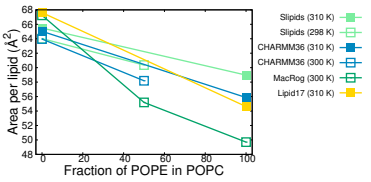
2.2 Quality evaluation of force fields

Quality of membrane simulations with different force fields have been evaluated against experimental data during parameterization and in separate comparison studies[17,18,21–24], but universal quality measure for membrane simulations is not defined and controversial results are often reported from simulations[?]. The unclarities in simulation quality complicate the selection of proper simulation models for specific applications and estimation of reliability of reported simulation results, thereby being a major obstacle in many applications of membrane MD simulations.

To enable rapid evaluation and comparison of membrane MD simulation qualities, we define the quality measure for membrane MD simulations using the C-H bond order parameters from NMR experiments and form factors from x-ray scattering, which are robust experimental measurables that can be directly connected to the simulation data[21]. C-H bond order parameters are related to the conformational ensemble of individual lipid molecules, while the form factor to arises from electron density thereby connecting to the overall structure of lipid bilayers. This is demonstrated in Fig. 2 a) showing correlations between membrane lateral density (area per lipid), thickness, minima of form factor and acyl chain order analyzed from all apporimately 500 trajectories in the NMRlipids databank. Membrane area per lipid has negative correlations with membrane order and thickness with Pearson coefficient of -0.78 and -0.49, respectively, because decreased area leads to more ordered, tightly packed and stretched acyl chains leading to thicker membranes. On the other hand, locations of form factor minima have positive correlation with the membrane area and negative correlation with the thickness. In conclusion, both x-ray scattering form factors and acyl chain order parameters from NMR are on average good proxies for membrane packing and thickness, while NMR order parameters give segmental resolution information on the quality of conformational ensembles of lipids.

NMRlipids databank contains also experimental order parameter and x-ray form factor data that is connected to corresponding simulations in order to define the simulation qualities and rank them to select the suitable simulation models for particular applications. The ranking of all simulations based on estimated average probability of order parameters to locate within experimental errors are shown in Fig. 2 b) and the comparison for the best models for overall quality, headgroup and glycerol backbone, and form factor are exemplified in Fig. 2 d). Full details of the NMRlipids quality measure for lipid bilayer simulations are given in the methods section and supplementary information. The top ranking simulations in Figs. 2 b) and d) demonstrate also the current complexity in lipid bilayer simulation quality. Simulation of POPC lipid bilayer with Slipids force field ranks first in overall score, glycerol backbone is not correcetly captured in this model. On the othere, simulation with the best quality for headroup and glycerol backbone, CHARMM36 POPC, predicts too ordered acyl chain, thereby not being within the top 10 simulations in the overall ranking. The best ranking for x-ray form factor is given by the Slipids POPE simulation, however, because the quantitative value for form factor quality depends on the quality of experimental data, form





	OPquality	Headgroup	sn-1	sn-2	FFquality	Force field	Molecules	Temperature	DOI
1	0.52	0.01	0.79	0.77	122.168365	Slipids	POPE:SOL (500:25000)	310.0	10.5281/zenodo.3231342
2	0.51	0.10	0.87	0.57	140.087165	Slipids	POPE:SOL (336:13460)	310.0	10.5281/zenodo.1293813
3	0.51	0.10	0.86	0.57	136.680061	Slipids	POPE:SOL (336:13460)	310.0	10.5281/zenodo.1293813
4	0.44	0.54	0.52	0.27	197.034217	CHARMM36	SOL:POPE (5760:144)	310.0	10.5281/zenodo.2641987
5	0.25	0.53	0.14	0.07	234.339081	CHARMM36	SOL:POPE (25000:500)	310.0	10.5281/zenodo.3237461
6	0.19	0.09	0.12	0.35	264.439235	Lipid17	POPE:SOL (500:25000)	310.0	10.5281/zenodo.4424292

d)

	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

Figure 2. a) Correlations between membrane area per lipid, thickness, second minima of form factor and average order parameter of the *sn*-1 acyl chain extracted from the NMRlipids databank. All Pearson correlation coefficients shown in bottom left corner have p-value below 0.001. b) Evaluation against experimental data exemplified for the simulation giving the best qualities in overall, for headgroup and x-ray scattering form factor. c) Change in area per lipid upon addition of POPE to POPC in different simulation extracted from the NMRlipids databank and quality evaluation table of POPE simulations. d) Quality evaluation table showing the best 13 simulations according to the overall order parameter quality.

factor qualities can be compared only between simulations evaluated against the same experimental dataset.

The power of NMRlipids quality metrics in selecting the best model for particular application is demontrated in Fig. 2 c) where the area per lipids of POPC:POPE mixtures from different force fields are shown and quality of POPE simulation in different force fields are evaluated. Among the simulations available in the NMRlipids databank, the Slipids force field gives the best quality in terms of acyl chain order parameters and form factor, and predicts the largest area per molecule for POPE and smallest difference with POPC. In conclusion, simulations with Slipids is the most reliable force field for membrane packing in POPC:POPE mixtures althought its glycerol backbone is not accurately modelled. Similar comparisons utilizing the preliminary data from the NMRlipids databank have concluded that Slipids is relatively good model also for mixtures of charged POPC:POPS membranes, although models with better counterion binding have higher quality, but CHARMM36 is the best to study differences in headgroup conformational ensembles between different lipid types. Understanding such complex picture of lipid bilayer MD simulation quality would not be possible without automatic quality evaluation of large sets of simulations enabled by the NMRlipids databank.

2.3 NMRlipids databank reveals correlations between membrane properties

The power of NMRlipids databank to find correlation between membrane properties is exemplified in Fig. **??** showing how water diffusion through and along the membrane correlate with different properties of systems. While NMRlipids databank enables similar analyses for any membrane properties, these are selected due to their relevance in the development of MRI imaging methods[?].

3 Discussion

Quality measure and automatic quality evaluation of lipid bilayer MD simulations introduced in the NMRlipids databank enables rapid ranking of available simulation models agains experimental NMR and x-ray scattering data. This provides a tool for researches to rapidly evaluate the credibility of MD simulations of their own and reported by other groups. Because such tool and quality measure has not been available, this will elaborate the quality of published MD simulations of lipid bilyers and reduce potentially misleading results[?]. The power of NMRlipids databank to select the best models for particular applications has been demontrated for PC/PE lipid mixtures (Fig. 2), PC/PS lipid mixtures[?], and lipid headgroups[20].

The increasing amount of MD simulation data with programmatic access in the NMRlipids databank opens up possibilities for wide range of applications utilizing the large set of accessible data. Extend of the data in the NMRlipids databank in terms of quantity (e.g., simulation length and number of conformations), content (e.g, lipid compositions and ion concentrations) and quality enables analyses that are not possible to conduct from MD simulation data produced by a single research group. Applications of the NMRlipids databank to understand how diffusion of water through and along cellular membrane depends on its physical properties are demonstrated in Fig. **??**. Permeation of water through membranes resembles the permeation of also other hydrophilic molecules, such as drugs, and detailed understanding of water dynamics through and along membranes is potentially useful for the development of MRI imaging methods[25]. These examples demonstrate the practical applications of NMRlipids databank on problems in the biological and biomedical sciences.

Building accessible databanks of molecular dynamics simulation data has been challenging due to the required long term support for hardware and software maintenance. In the overlay model used in the NMRlipids databank, the demand of hardware can be distributed and open collaboration model reduces the risk for ending software maintenance. Furthermore, the open collaboration model used in the NMRlipids project credits contributors by offering authorship in published articles, thereby creating an incentive for contributions. This model could be extended also to other fields where similar barriers to establish publicly accessible databanks exist. Emerging applications of machine learning are increasing the impact of such databanks. For example, the existing Protein Databank (PDB)[?] containing experimentally determined protein structures with programmatic access has enabled the development of machine learning based tools building on the data collected in the databank over the years[?]. NMRlipids and other databanks with open programmatic access have potential to lead similar unforeseen applications in the future.

4 Methods

4.1 Structure of the databank

Structure of the NMRlipids databank is illustrated in Fig. 3. The required input information to create an entry into the NMRlipids databank are listed in table 1. While the raw simulation data is not directly stored in the NMRlipids databank, permanent links from where the raw data can be accessed have to be given and are then stored in the README.yaml files at https://github.com/NMRLipids/Databank/tree/main/Data/Simulations. These files contain all the essential information listed in table 1 on each simulation entry that are needed for further applications. The raw MD simulation data can locate in any stable publicly available repository, although all the current data locates in Zenodo www.zenodo.org.

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 ??)	
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.	

Table 1. Keys stored in the README.yaml files of simulations.

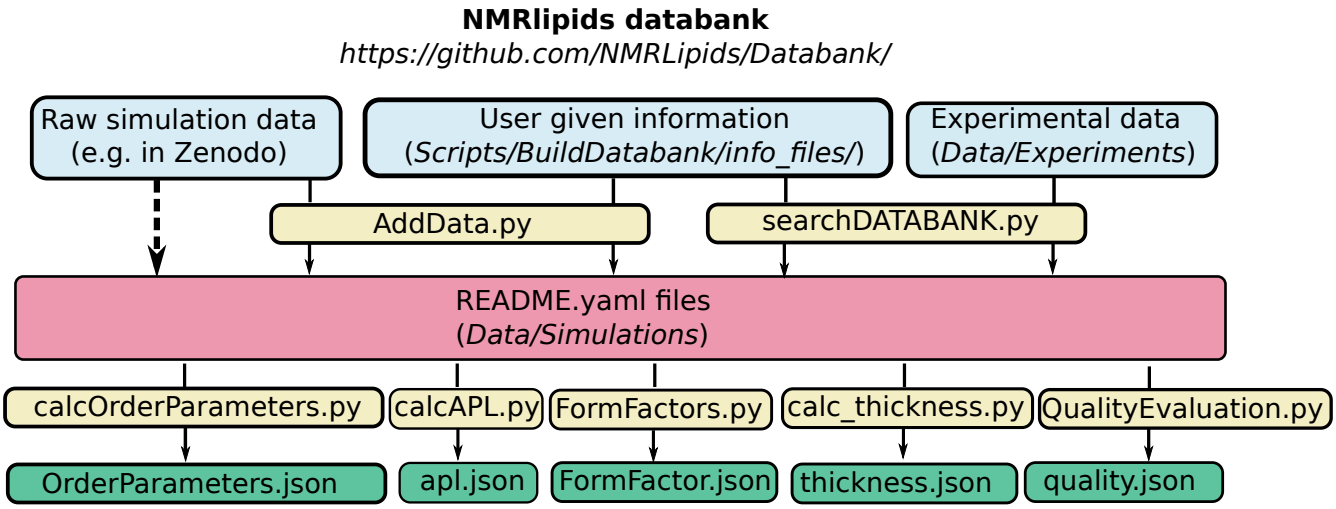


Figure 3. Structure of the NMRlipids databank. Manually added input data (blue boxes) includes basic information on the simulation, permanent links to the raw data, and experimental data if available. The databank entries (red box) and analysis results (green boxes), locating at https://github.com/NMRLipids/Databank/tree/main/Data/Simulations are automatically generated by the computer programs included in the NMRlipids databank (yellow boxes). Because raw data are not permanently stored but can be accessed based on the information in the databank, this connection is marked with the dashed line.

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 (defined as ??)
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.

Table 2. Keys stored in the README.yaml files of experiments.

In order to evaluate the quality of simulations, sets of C-H bond order parameters from NMR and from factors from x-ray scattering are included in the NMRlipids databank (https://github.com/NMRLipids/Databank/tree/main/Data/experiments). The required information for an experimental dataset are listed in table 2. A simulation is connected to a experimental data set if molar concentrations of all molecules are within ± 5 percentage units, charged lipids have the same counterions, and temperature is within ± 2 degrees. In such cases, a simulation and experimental data are paired by adding the path to the experimental data into the simulation README.yaml file.

Because README.yaml contains all the essential information on each simulation, arbitrary analyses can be automatically performed over all the simualtions in the NMRlipids databank. In addition to the order parameters for each C-bonds and x-ray scattering form factors used in the quality evaluation, the NMRlipids databank contains the area per lipid and thickness calculated from all simulations in the databank. These results are stored in the same folders as the README.yaml files in https://github.com/NMRLipids/Databank/tree/main/Data/Simulations.

4.2 Molecule and atom naming convention

Because universal convention for lipid molecules and atoms therein has not been defined, the naming conventions vary between authors and force fields. To enable automatic analyses over large sets of simulation in the NMRlipids databank, we have defined unique naming conventions for lipid molecules and atoms. The abbreviations of molecule names used in the NMRlipids

[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.

databank are listed in table 3. The unique atom names for each molecule and corresponding names in each simulation are defined using mapping files introduced in the NMRlipids project (https://nmrlipids.blogspot.com/2022/04/new-yaml-format-of-mapping-files.html).

Molecule and atom names in each simulation are connected to the unique naming convention with the COMPOSITION dictionary in the README.yaml file. Upon addition of a new entry in the databank, the molecule names in the simulation corresponding the unique names and mapping files (available at https://github.com/NMRLipids/Databank/tree/main/Scripts/BuildDatabank/mapping_files) are defined in the dictionary. The numbers of molecules in the system are then automatically calculated by the NMRlipids databank codes (AddData.py in Fig. 3) and stored in the README.yaml together with other content of the COMPOSITION dictionary. This information can be then used to find each molecule and atom from each simulation in the analysis codes.

Abbreviation	Molecule name
POPC	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
POPG	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol
POPS	1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine
POPE	1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine
CHOL	cholesterol
DHMDMAB	dihexadecyldimethylammonium
POT	potassium ion
SOD	sodium ion
CLA	chloride ion
CAL	calcium ion
SOL	water

Table 3. Abbreviations used in the databank

4.3 Quality evaluation

Quality of conformational ensembles of lipid molecules in simulations are evaluated by estimating the average probability of C-H bond order parameters to agree with the values from NMR experiments. Because each C-H bond order parameter is calculated as an average over individual lipids in a simulation and conformational ensembles of lipids are assumed to be independent, the random variable

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}, \tag{1}$$

where $\bar{X}$ is the sample mean, S is the sample variance and n is the sample size (number of lipids), has a Student's t-distribution with $n-1$ degrees of freedom and the mean of μ.

where the probability is calculated from the normal distribution

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

S_{exp} and ΔS_{exp} are experimental order parameter and its error, and μ and σ are the mean order parameter and its standard deviation from simulations. The quality measure S_q approaches to zero when probability for agreement between simulation and experimental results approach one, and increases when simulated and experimental values diverge. The accuracy of ± 0.02 is currently assumed for all experimental order parameters[21]. Because phospholipids sample their conformational ensemble within nanosecond timescale[26], all simulations in the databank would be sufficiently long to sample the realistic conformational phase of individual lipids. However, some force fields exhibit too slow dynamics which leads to large error bars in order parameter values[27]. Because large error bars widen the gaussian distribution in Eq. 2 thereby artificially increasing the probability to find the simulated value within experimental error bars, the order parameters with error bars larger than the experimental error 0.02 are not included in the quality evaluation.

The quality of separate fragments in each lipid type within a simulation are then evaluated by averaging individual order parameter qualities over C-H bond belonging to that fragment and dividing this with the percentage of order parameters for which the quality is available within the fragment, p

$$S_q^{\mathrm{frag}}[\mathrm{lipid}] = \frac{\langle S_q[\mathrm{lipid}]\rangle_{\mathrm{frag}}}{p_{\mathrm{frag}}[\mathrm{lipid}]}, \tag{3}$$

where frag can be *sn*-1, *sn*-2, headgroup or total (all order parameters within a molecule). The overall quality of different fragments in a simulation are then defined as a molar fraction weighted average over different lipid components

$$S_q^{\text{frag}} = \sum_{\text{lipid}} \chi_{\text{lipid}} \langle S_q^{\text{frag}}[\text{lipid}] \rangle_{\text{lipid}}, \tag{4}$$

where χ_{lipid} is the molar fraction of a lipid in the bilayer.

Qualities of form factors were evaluated with the same approach that was used in SIMtoEXP program[28]. Because experiments give form factors only in relative scale those were scaled to the simulation data with absolute scale using equation

$$k_e = \frac{\sum_{i=1}^{N_q} \frac{|F_s(q_i)||F_e(q_i)|}{(\Delta F_e(q_i))^2}}{\sum_{i=1}^{N_q} \frac{|F_e(q_i)|^2}{(\Delta F_e(q_i))^2}}. \tag{5}$$

The quality of each form factor was then calculated from the equation

$$\chi^2 = \frac{\sqrt{\sum_{i=1}^{N_q} (|F_s(q_i)| - k_e|F_e(q_i)|)^2/(\Delta F_e(q_i))^2}}{\sqrt{N_q - 1}}, \tag{6}$$

where F_s and F_e are form factors from a simulation and experiment, respectively, and summation goes over the experimentally available N_q points.

4.4 Analysing simulations in the NMRlipids databank

The README.yaml files contain all the essential information to perform arbitrary analyses of simulations in the NMRlipids databank, i.e., the permanent location of the original data and naming convention for all atoms and molecules in each system. In practise, the analyse codes contains a loop over all README.yaml files (i.e., simulations in the NMRlipids databank) which first downloads the raw simulation to a local computer and then uses the information about the atom and molecule naming conventions in README.yaml and mapping files to perform the desired analyses. For example, the code that calculates all C-H bond order parameters of all systems is available at https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/calcOrderParameters.py and minimal example of a analysis code is available at https://github.com/NMRLipids/Databank/blob/main/Scripts/AnalyzeDatabank/template.ipynb.

While the order parameters, form factors, area per lipid and thickness are stored within the NMRlipids databank (https://github.com/NMRLipids/Databank/tree/main/Data/Simulations), further analyses can be conventiently stored in separate repositories with the same folder structure based on hash identities of trajectory and topology files. For example, results from further analyses performed here are stored in folders at https://github.com/NMRLipids/DataBankManuscript/tree/main/Data. Such organization of the data enables further upcycling of the analyzed data as similarly to the original NMRlipids databank repository.

References

1. Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018, DOI: 10.1038/sdata.2016.18 (2016).
2. Feig, M., Abdullah, M., Johnsson, L. & Pettitt, B. Large scale distributed data repository: design of a molecular dynamics trajectory database. *Futur. Gener. Comput. Syst.* **16**, 101–110, DOI: https://doi.org/10.1016/S0167-739X(99)00039-4 (1999).
3. Tai, K. *et al.* Biosimgrid: towards a worldwide repository for biomolecular simulations. *Org. Biomol. Chem.* **2**, 3219–3221, DOI: 10.1039/B411352G (2004).
4. Silva, C. G. *et al.* P-found: The protein folding and unfolding simulation repository. In *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 1–8, DOI: 10.1109/CIBCB.2006.330978 (2006).
5. Abraham, M. *et al.* Sharing data from molecular simulations. *J. Chem. Inf. Model.* **59**, 4093–4099 (2019).
6. Hildebrand, P. W., Rose, A. S. & Tiemann, J. K. Bringing molecular dynamics simulation data into view. *Trends Biochem. Sci.* **44**, 902–913, DOI: 10.1016/j.tibs.2019.06.004 (2019).
7. Hospital, A., Battistini, F., Soliva, R., Gelpí, J. L. & Orozco, M. Surviving the deluge of biosimulation data. *WIREs Comput. Mol. Sci.* **10**, e1449, DOI: https://doi.org/10.1002/wcms.1449 (2020). https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1449.

8. Abriata, L. A., Lepore, R. & Dal Peraro, M. About the need to make computational models of biological macromolecules available and discoverable. *Bioinformatics* **36**, 2952–2954, DOI: 10.1093/bioinformatics/btaa086 (2020). https://academic.oup.com/bioinformatics/article-pdf/36/9/2952/33180880/btaa086.pdf.

9. Rodríguez-Espigares, I. *et al.* Gpcrmd uncovers the dynamics of the 3d-gpcrome. *Nat. Methods* **17**, 777–787, DOI: 10.1038/s41592-020-0884-y (2020).

10. Meyer, T. *et al.* Model (molecular dynamics extended library): A database of atomistic molecular dynamics trajectories. *Structure* **18**, 1399–1409, DOI: https://doi.org/10.1016/j.str.2010.07.013 (2010).

11. van der Kamp, M. W. *et al.* Dynameomics: A comprehensive database of protein dynamics. *Structure* **18**, 423–435, DOI: https://doi.org/10.1016/j.str.2010.01.012 (2010).

12. Hospital, A. *et al.* BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.* **44**, D272–D278, DOI: 10.1093/nar/gkv1301 (2016). https://academic.oup.com/nar/article-pdf/44/D1/D272/16661850/gkv1301.pdf.

13. Mixcoha, E., Rosende, R., Garcia-Fandino, R. & Piñeiro, A. Cyclo-lib: a database of computational molecular dynamics simulations of cyclodextrins. *Bioinformatics* **32**, 3371–3373, DOI: 10.1093/bioinformatics/btw289 (2016). https://academic.oup.com/bioinformatics/article-pdf/32/21/3371/7889578/btw289.pdf.

14. Newport, T. D., Sansom, M. S. & Stansfeld, P. J. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res.* **47**, D390–D397, DOI: 10.1093/nar/gky1047 (2018). https://academic.oup.com/nar/article-pdf/47/D1/D390/27437085/gky1047.pdf.

15. Bekker, G.-J., Kawabata, T. & Kurisu, G. The biological structure model archive (bsm-arc): an archive for in silico models and simulations. *Biophys. Rev.* **12**, 371–375, DOI: 10.1007/s12551-020-00632-5 (2020).

16. Suarez-Leston, F. *et al.* Supepmem: A database of innate immune system peptides and their cell membrane interactions. *Comput. Struct. Biotechnol. J.* **20**, 874–881, DOI: https://doi.org/10.1016/j.csbj.2022.01.025 (2022).

17. Botan, A. *et al.* Toward atomistic resolution structure of phosphatidylcholine headgroup and glycerol backbone at different ambient conditions. *J. Phys. Chem. B* **119**, 15075–15088 (2015).

18. Catte, A. *et al.* Molecular electrometer and binding of cations to phospholipid bilayers. *Phys. Chem. Chem. Phys.* **18**, 32560–32569 (2016).

19. Antila, H. *et al.* Headgroup structure and cation binding in phosphatidylserine lipid bilayers. *J. Phys. Chem. B* **123**, 9066–9079 (2019).

20. Bacle, A. *et al.* Inverse conformational selection in lipid–protein binding. *J. Am. Chem. Soc.* **143**, 13701–13709 (2021).

21. Ollila, O. S. & Pabst, G. Atomistic resolution structure and dynamics of lipid bilayers in simulations and experiments. *Biochim. Biophys. Acta* **1858**, 2512 – 2528 (2016).

22. Pluhackova, K. *et al.* A critical comparison of biomembrane force fields: Structure and dynamics of model dmpc, popc, and pope bilayers. *The J. Phys. Chem. B* **120**, 3888–3903 (2016).

23. Sandoval-Perez, A., Pluhackova, K. & Böckmann, R. A. Critical comparison of biomembrane force fields: Protein–lipid interactions at the membrane interface. *J. Chem. Theory Comput.* **13**, 2310–2321 (2017).

24. Leonard, A. N., Wang, E., Monje-Galvan, V. & Klauda, J. B. Developing and testing of lipid force fields with applications to modeling cellular membranes. *Chem. Rev.* **119**, 6227–6269 (2019).

25. Topgaard, D. Chapter 1 translational motion of water in biological tissues – a brief primer. In *Advanced Diffusion Encoding Methods in MRI*, 1–11 (The Royal Society of Chemistry, 2020).

26. Ferreira, T. M., Ollila, O. H. S., Pigliapochi, R., Dabkowska, A. P. & Topgaard, D. Model-free estimation of the effective correlation time for C-H bond reorientation in amphiphilic bilayers: ^{1}H-^{13}C solid-state NMR and MD simulations. *J. Chem. Phys.* **142**, 044905 (2015).

27. Antila, H. S., M. Ferreira, T., Ollila, O. H. S. & Miettinen, M. S. Using open data to rapidly benchmark biomolecular simulations: Phospholipid conformational dynamics. *J.Chem. Inf. Model.* **61**, 938–949 (2021).

28. Kučerka, N., Katsaras, J. & Nagle, J. Comparing membrane simulations to scattering experiments: Introducing the SIMtoEXP software. *J. Membr. Biol.* **235**, 43–50 (2010).

Acknowledgements

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information