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 overlay 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 can be potentially 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 Overlay structure of the NMRlipids databank

The three layer structure of the NMRlipids databank, illustrated in Fig. 1 a, enables an efficient way to share and upcycle large MD simulation trajectories. Layer 1 contains raw simulation data that can locate in any publicly available repository or server offering long term stability and permanent links to the data, such as digital object identifiers. Layer 2 contains all the relevant information on the simulations, including links to the raw data in layer 1 and relevant metadata describing the systems. Layer 3 is the application layer containing outputs from the databank, such as graphical user interface (`http://www.databank.nmrlipids.fi/`) and results from analyses desribed in sections below.

Layer 2 is the core of the databank. In addition to the metadata and links to raw simulation data, it contains universal naming conventions for the molecules and atoms in the databank, computer programs to generate and analyse the databank, basic properties calculated from all simulations (area per lipid, C-H bond order parameters, x-ray scattering form factors and membrane thickness), and quality evaluation of simulations against experimental data. In practise, this information is stored is the git repository which is currently available at `https://github.com/NMRLipids/Databank`. Detailed description on its content is in the supplementary information. Applications in layer 3 uses the information in layer 2 to access raw data in layer 1 to perform automatic analyses over large sets of simulations in the databank.

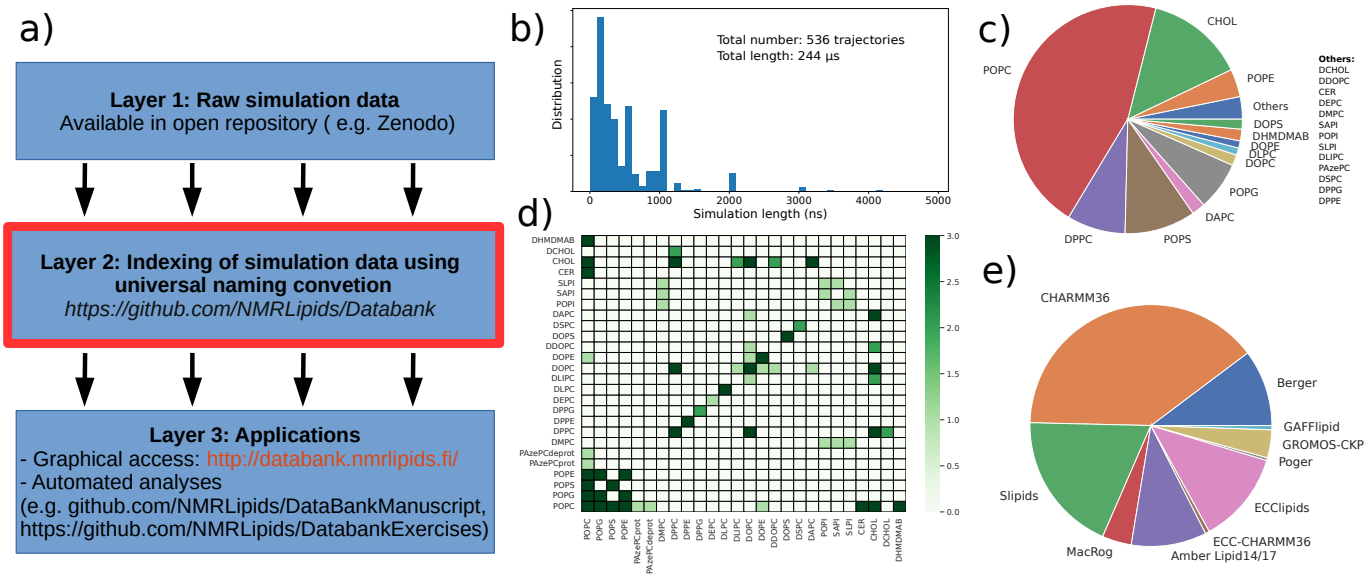

Figure 1. a) Structure of an overlay databank. More detailed structure of the layer 2 in the NMRlipids databank is illustrated in Fig. 5 in the SI. b) Distribution of the lengths of the trajectories, total number of trajectories and total lenght of the simulations in the NMRlipids databank. c) Distribution 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 Content of the 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.

2.3 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 lack of universal quality measure for membrane simulations complicates the selection of proper simulation parameters and the estimation of reliability in simulations, thereby being a major obstacle in many applications of membrane MD simulations. To enable the rapid quality evaluation of membrane MD simulations, a quantitative quality measure based on C-H bond order parameters from NMR experiments is defined in the NMRlipids databank. This is complemented by comparing the locations of x-ray scattering form factor minima against experiments. Both of these are robust experimental measurables that can be directly connected to the simulation data[21]. The used quality measures are defined in detail in the supplementary information.

Simulations with the highest scores based on order parameter evaluation for all simulations are shown in Fig. 2 A and for POPE lipids in Fig. 2 B. The direct comparison with experiments for the best simulation, POPC with OPLS3e, is shown in Fig. 2 C, where discrepancies with experiments are observed only for the third carbon in the *sn*-1 chain, and carbons 8 and 9 close to the double bond in *sn*-2 chain. The best overall quality for POPE is found in GROMOS-CKP simulation, for which the direct comparison in Fig. 2 D show minor differences in acyl chains, but major discrepancies in glycerol backbone and in the first carbon of *sn*-2 chain. These parts are better described for POPE by CHARMM36 parameters (direct comparison shown in Fig. 2 E), but its overall quality suffers from the overestimated order in acyl chains.

To exemplify the finding of optimal model for a certain lipid, the top ranking simulation for the order parameter quality of POPE lipid are shown in Fig. 2 B.

Differences in the predictions of area per lipids for POPC and POPE bilayers between different force fields are demonstrated in Fig. 3. Because area per lipid is strongly correlated with the average order of *sn*-1 chain ((Fig. 2 F)), its order parameters can used as a proxy for the area per lipid to evaluate its quality against experiments. Simulations with the largest area per lipids for POPE in Fig. 3, GROMOS-CKP and Slipids, have significantly higher qualities in Fig. 2 B than lipid17 and CHARMM36 with smaller are per lipids. Among the simulations available in the NMRlipids databank, the Slipids force field gives the

A)

	OPquality	Headgroup	tails	FFquality	Force field	Molecules	Temperature	ID
1	0.83	0.76	0.86	0.15	OPLS3e	POPC:SOL (200:8859)	300.00	1
2	0.76	0.61	0.83	0.45	Slipids	POPC:SOL (512:23943)	298.00	617
3	0.69	0.63	0.73	0.55	MacRog	POPC:SOL (128:5120)	300.00	457
4	0.61	0.64	0.60	0.55	MacRog	POPC:SOL (288:14400)	298.00	63
5	0.60	0.65	0.58	0.55	ECC-lipids	POPC:SOL (128:6400)	300.00	573
6	0.59	0.59	0.59	1.03	CHARMM36	POPC:CHOL:SOL (1024:128:57600)	298.15	426
7	0.58	0.01	0.86	0.95	Berger	POPC:SOL (256:10342)	300.00	115
8	0.56	0.52	0.58	1.13	CHARMM36	POPC:CHOL:SOL (64:8:3600)	298.15	109
9	0.55	0.02	0.81	1.05	Berger	POPC:SOL (128:7290)	298.00	497

B)

	OPquality	Headgroup	sn-1	sn-2	FFquality	Force field	Molecules	Temperature	ID
1	0.53	0.29	0.83	0.48	0.4	GROMOS-CKP	POPE:SOL (500:25000)	310.0	400
2	0.52	0.01	0.79	0.77	0.1	Slipids	POPE:SOL (500:25000)	310.0	414
3	0.51	0.10	0.87	0.57		Slipids	POPE:SOL (336:13460)	310.0	29
4	0.51	0.10	0.86	0.57		Slipids	POPE:SOL (336:13460)	310.0	74
5	0.44	0.54	0.52	0.27	1.3	CHARMM36	SOL:POPE (5760:144)	310.0	430
6	0.38	0.32	0.56	0.26	0.3	CHARMM36-UA	POPE:SOL (336:15254)	310.0	352
7	0.37	0.27	0.57	0.28	0.3	CHARMM36-UA	POPE:SOL (336:15254)	310.0	233
8	0.25	0.53	0.14	0.07	1.7	CHARMM36	SOL:POPE (25000:500)	310.0	67
9	0.19	0.09	0.12	0.35	1.6	Lipid17	POPE:SOL (500:25000)	310.0	195

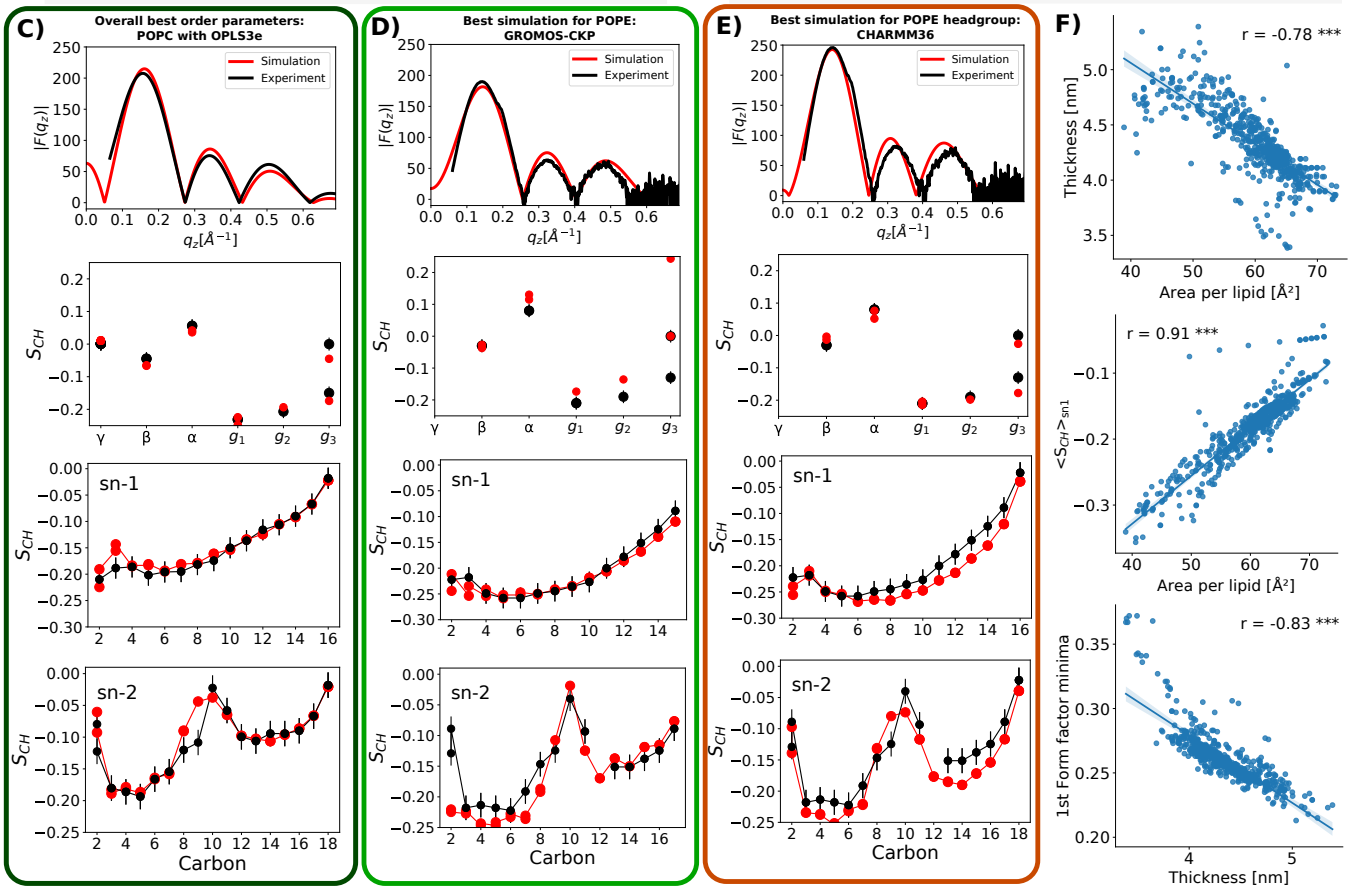

Figure 2. A) Best simulations ranked based on overall order parameter quality. B) Best simulations ranked based on the overall order parameter quality of POPE lipid. C) Scatter plots and Pearson correlation coefficients for the area per lipid, thickness, and experimental observables. All correlation coefficients have p-value below 0.001. For more correlations see Fig. S1. D)-F) Evaluation against experimental data exemplified for a simulation with the best overall order parameter quality (D), the best quality for POPE lipid (E), and the headgroup quality for POPE (F).

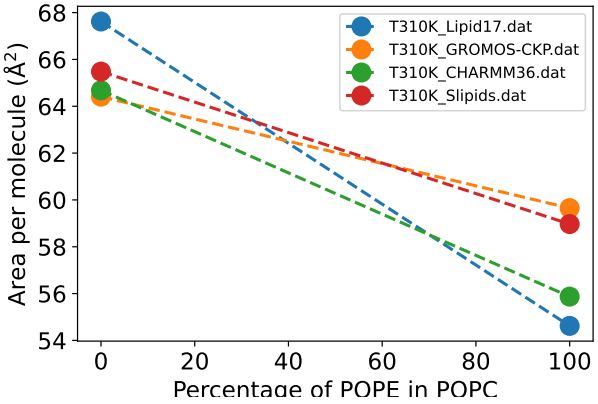

Figure 3. Differences in area per lipids of POPC and POPE bilayers between different force fields.

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.

On the other hand, the first form factor minima is correlated stronger with membrane thickness (Fig. 2 C) than with the area per lipid (Fig. S1). Scatter plots and Pearson correlation coefficients in Fig. 2 C reveal strong correlation between membrane area per lipid and thickness.

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.

2.4 Water diffusion anisotropy in membrane systems

The anisotropic water diffusion through and along of membranes is utilized in MRI imaging[?] and plays a role in the drug translocation through biological material, particularly in skin[?,25–28]. Dependencies between membrane properties, water permeablity, and anisotropic water diffusion have been recognized[?] but consensus on how water permeability depends on membrane properties, such as area per lipid and thickness, or molecular composition, has not yet been reached despite of several available models[?,?,?,29,30]. MD simulations can be used to elucidate such dependencies, but systematic collection of sufficient amount of data has been challenging because few water permeation events are typically observed in a single MD simulation trajectory[?]. Nevertheless, averaging over large amounts of data available in the NMRlipids databank enables to study the trends of water permeation upon changing memebrane properties.

Water permeabilities calculated from all the simulations in the NMRlipids databank as a function of area per lipids, thickness, and acyl chain order are shown in Fig. 4 (a-c). To illustrate how permeability depends these properties, average over systems with fixed values of area per lipids, thickness, and acyl chain order are also shown in the figure. Permeability increases with increasing area per lipid and decreases with increasing membrane thickness and acyl chain order. Previous simulation and experimental studies have reported linear dependence of permeability on both area per lipid and thickness for limited set of lipid mixtures but not always for all types of lipid mixtures[?,?]. Our results suggest stronger dependence for thicknesses below 3.9 nm and areas above 69 Å^2, but the dependence may appear linear above and below these values as shown in the insets of Figs. a) and b). As expected, the permeability increases with increasing temperature in Figs. d). The Arrhenius plot gives $17 \pm 4\,k_bT$ for the average energy barrier for the water permeation. The values for water permeabilities from simulations in the databank vary between 0.3 and 322 μm/s with the mean and median of 14 μm/s and 8 μm/s, respectively. These values have the same order of magnitude as the experimentally determined diffusive permeability coefficients, but are on average below the values reported for PC lipids in liquid crystalline phase, 190-330 μm/s[31]. Clear dependencies of permeability on hydration level, charged lipid fraction, cholesterol fraction or POPE fraction were not observed, although weak decrease for the latter two may be visible.

To analyze the anisotropic water diffusion, we also calculated the water diffusion along the membrane. The results at different hydration levels are shown in Fig. 4 h) together with the experimental data[32]. The experimental water diffusion increases with increasing hydration level toward the value for bulk water ($3.1 \cdot 10^{-9}$ m^2/s at 313 K)[33]. Simulations overestimate the experimental bulk value approximately with the factor of 2 at high hydration levels which is not surprising as the most

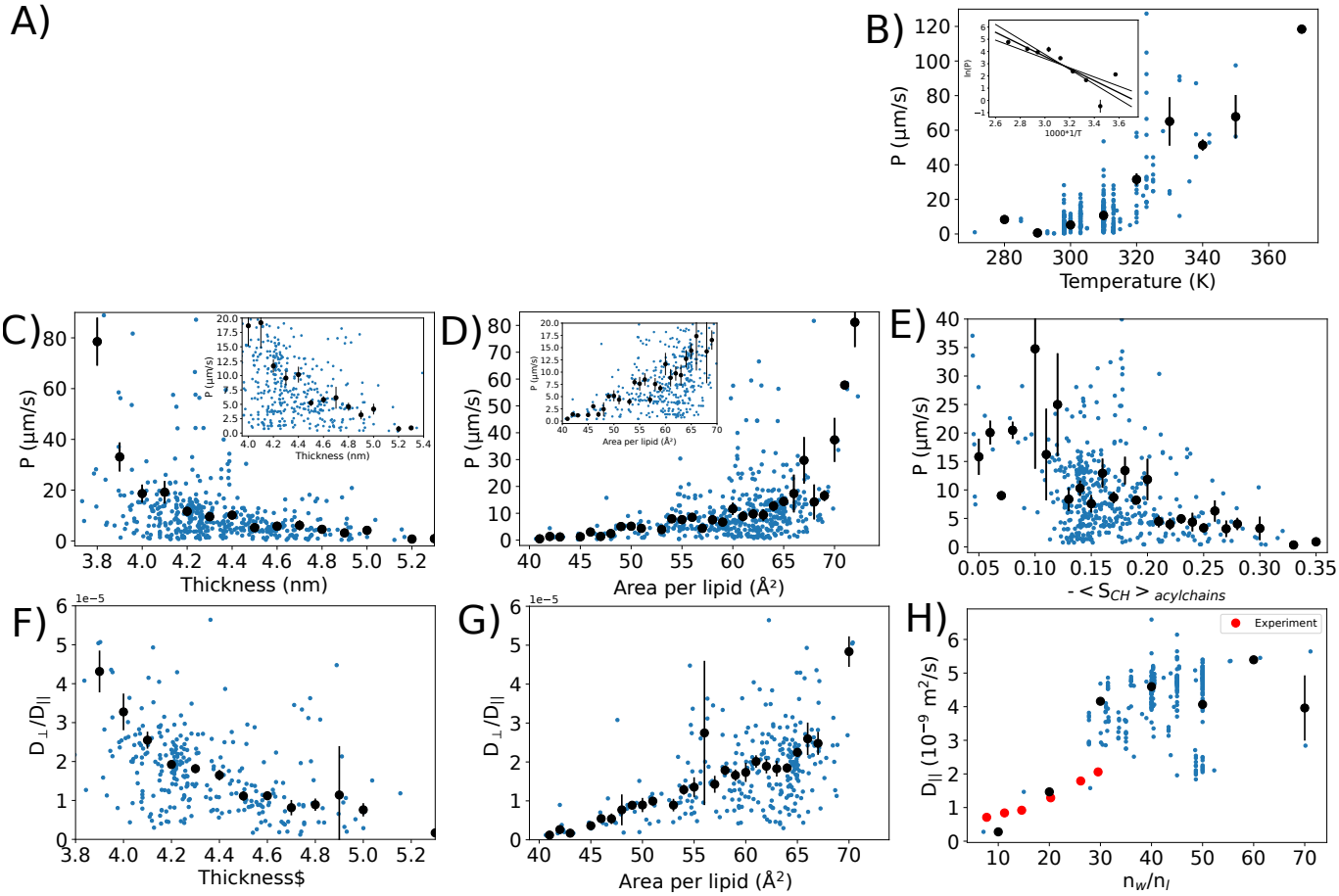

Figure 4. (b-c) Water permeation through lipid membranes analyzed from the databank as a function of temperature, thickness, area per lipid, and acyl chain order. Inserts in c) and d) show the region where the dependence could be considered approximately linear. Insert in a) shows the Arrhenius plot of permeation (ln (P) vs. 1/T) that gives $17 \pm 4\,k_bT$ for the average activation energy for water permeation through lipid bilayer. (f-g) Diffusion anistoropy of water as a function of thickness and area per lipid. (h) Lateral diffusion of water as a function of hydration level. Experimental points for DMPC bilayers at 313 K at different hydration levels are shown[32].

common water model used in membrane simulations, TIP3P, overestimates the bulk water diffusion. However, simulation results are closer to experiments with low hydration levels. To estimate the diffusion anistoropy, we translated the water permeation to diffusion coefficient through multilamellar stack using the Tanner equation[?]. The resulting diffusion coefficients trough the membranes are approximately five orders of magnitude slower than along the membrane which is at the upper limits of the anisotropic estimated from the experimental data[27]. Relatively high anisotropy is understandable as simulations give slightly slower permeation rates and higher lateral diffusion rates than experiments. Significant dependence of diffusion anisotropy on membrane thickness and area per lipid are observed in Figs. 4 f) and g). The anisotropy becomes linearly stronger with decreasing area per lipid while the dependency is stronger with increasing thickness. These results can be explained by the increasing and decreasing permeabilities with thickness and are per lipid (Figs 4 c) and d), respectively, while lateral diffusion remains approximately constant (Fig. **??** a) and c)).

3 Discussion

, 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.

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[34]. 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. 5. 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`.

[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.

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.	

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

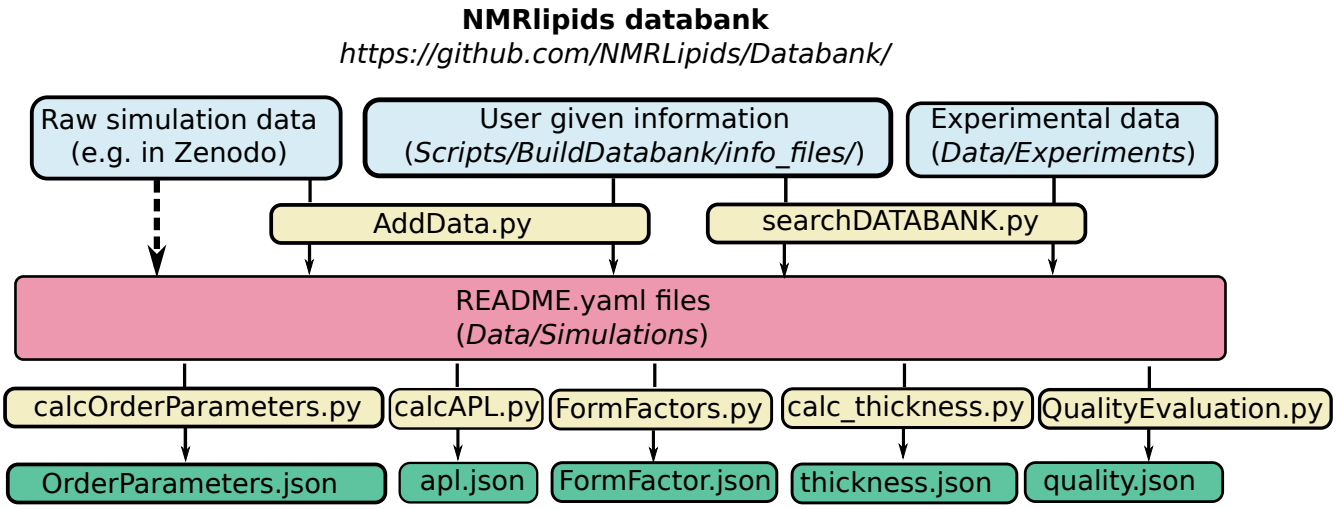

Figure 5. Structure of the NMRlipids databank. Manually added input data (blue boxes) includes basic information on the simulation, permanent links to the raw data, and experimental data if available. The databank entries (red box) and analysis results (green boxes), locating at https://github.com/NMRLipids/Databank/tree/main/Data/Simulations 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
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 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. 5) 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

4.3.1 Conformational ensembles using C-H bond order parameters

The quality of conformational ensembles of individual lipid molecules in simulations are evaluated using the C-H bond order parameters which can be directly compared with robust experimental data[21]. The order parameters are defined as

$$S_{\mathrm{CH}} = \frac{1}{2}\left\langle 3\cos^2\theta - 1\right\rangle, \tag{1}$$

where θ is the angle between the C–H bond and the membrane normal and angular brackets denote the ensemble average, i.e. average over all sampled configurations of all lipids in a simulation. Because conformational ensembles of individual lipids are independent in a simulation of a fluid lipid bilayer, the number of independent sample points for each C-H bond, n, equals the number of lipids in a simulation. Denoting the S_{CH} from Eq. 1 as the sample mean and s its variance calculated over individual lipids, $\frac{S_{\mathrm{CH}}-\mu}{s/\sqrt{n}}$ has a Student's t-distribution with $n-1$ degrees of freedom, where μ is the real mean of the order parameter. Because lipid bilayer simulations contain at least dozens of lipids, the Student's t-distribution could be safely approximated with a normal distribution. However, when calculating the probability for a simulation to locate within experimental order parameters, the normal distribution gives values below the numerical accuracy of computers for simulations values far from experiments. To avoid such numerical instability, we use the first order Student's t-distribution.Therefore, we estimate the probability for the simulation order parameter to locate within experimental error bars from equation

$$P = f\left(\frac{S_{\mathrm{CH}} - (S_{\mathrm{exp}} + \Delta S_{\mathrm{exp}})}{s/\sqrt{n}}\right) - f\left(\frac{S_{\mathrm{CH}} - (S_{\mathrm{exp}} - \Delta S_{\mathrm{exp}})}{s/\sqrt{n}}\right), \tag{2}$$

where f(t) is the first order Student's t-distribution.

The error of $\Delta S_{\mathrm{exp}} = 0.02$ is currently assumed for all experimental order parameters[21]. Because phospholipids sample their conformational ensemble within nanosecond timescale[35], 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[36]. Because large error bars widen the Student's t-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 qualities of different fragments within each lipid type in a simulation, $P^{\mathrm{frag}}[\mathrm{lipid}]$, are then estimated by averaging the probabilities for individual C-H bond order parameters to locate within experimental errors according to Eq. 2, and dividing the average with the percentage of order parameters for which the quality is available within the fragment, $p_{\mathrm{frag}}[\mathrm{lipid}]$,

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

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

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

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

4.3.2 Membrane dimensions with x-ray scattering form factors

While C-H bond order parameters relate to the conformational ensembles of individual lipids, the x-ray scattering form factors depend on membrane dimensions and density distribution[21]. Because experiments give form factors only in relative intensity scale, they are typically scaled to the simulation data. Here we use the same scaling method as in SIMtoEXP program[37] where experimental form factor intensities are scaled by a factor defined as

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

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. Quality measure based on differences in simulated and experimental form factors accross available q-range is also defined in the SIMtoEXP program. However, as shown in Fig. S2, the lobe heigths in the simulated form factors depend on the simulation box size, but locations of minima does not. Therefore, quality measure depending on lobe heights is also dependent on simulation box size. In order to define a quality measure which is independent on simulation box size, we measure the form factor quality based on the location of the first minima observed in experiments. As shown in Fig. 2, this minima correlates well with the thickness of a membrane. Although the correlation in second minima would be even stronger (Fig. S1), we have chosen to use the first minima because automatic definition of second minima is inaccurate in some experimental data. In practise, we first filter the fluctuations from the form factor data using Savitzky-Golay filter (window lenght 30 and polynomial order 1) and locate the first minima above $0.1\,AA^2$ from both simulation and experimental data. The quality of a form factor is then defined as Eucledian distance between the minima in simulated and experimental form factors, $FF_q = |FF^{\text{sim}}_{\text{min}} - FF^{\text{exp}}_{\text{min}}|$.

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.

For the permeation of water through membranes, the number of permeation events in each trajectory was first calculated using the code by Camilo et al.[38], available at https://github.com/crobertocamilo/MD-permeation. The permeation was then calculated from equation $P = r/2c_w$, where r is the rate per time and area, and c_w is the concentration of water in bulk with the value of $33.3679\,(\text{nm})^{-3}$**?**.

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. Antila, H. S. *et al.* Emerging era of biomolecular membrane simulations: Automated physically-justified force field development and quality-evaluated databanks. *The J. Phys. Chem. B* **126**, 4169–4183 (2022).

26. Hansen, S., Lehr, C.-M. & Schaefer, U. F. Improved input parameters for diffusion models of skin absorption. *Adv. Drug Deliv. Rev.* **65**, 251–264, DOI: https://doi.org/10.1016/j.addr.2012.04.011 (2013). Modeling the human skin barrier - Towards a better understanding of dermal absorption.

27. Wen, J., Koo, S. M. & Lape, N. How sensitive are transdermal transport predictions by microscopic stratum corneum models to geometric and transport parameter input? *J. Pharm. Sci.* **107**, 612–623, DOI: https://doi.org/10.1016/j.xphs.2017.09.015 (2018).

28. Nitsche, L. C., Kasting, G. B. & Nitsche, J. M. Microscopic models of drug/chemical diffusion through the skin barrier: Effects of diffusional anisotropy of the intercellular lipid. *J. Pharm. Sci.* **108**, 1692–1712, DOI: https://doi.org/10.1016/j.xphs.2018.11.014 (2019).

29. Roberts, M. S. *et al.* Topical drug delivery: History, percutaneous absorption, and product development. *Adv. Drug Deliv. Rev.* **177**, 113929, DOI: https://doi.org/10.1016/j.addr.2021.113929 (2021).

30. Nitsche, J. M. & Kasting, G. B. Permeability of fluid-phase phospholipid bilayers: Assessment and useful correlations for permeability screening and other applications. *J. Pharm. Sci.* **102**, 2005–2032, DOI: https://doi.org/10.1002/jps.23471 (2013).

31. Nitsche, J. M. & Kasting, G. B. A universal correlation predicts permeability coefficients of fluid- and gel-phase phospholipid and phospholipid-cholesterol bilayers for arbitrary solutes. *J. Pharm. Sci.* **105**, 1762–1771, DOI: https://doi.org/10.1016/j.xphs.2016.02.012 (2016).

32. Jansen, M. & Blume, A. A comparative study of diffusive and osmotic water permeation across bilayers composed of phospholipids with different head groups and fatty acyl chains. *Biophys. J.* **68**, 997–1008, DOI: https://doi.org/10.1016/S0006-3495(95)80275-4 (1995).

33. Rudakova, M., Filippov, A. & Skirda, V. Water diffusivity in model biological membranes. *Appl. Magn. Reson.* **27**, 519 (2004).

34. Khakimov, A. M., Rudakova, M. A., Doroginitskii, M. M. & Filippov, A. V. Temperature dependence of water self-diffusion through lipid bilayers assessed by NMR. *Biophysics* **53**, 147–152, DOI: 10.1134/s000635090802005x (2008).

35. 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).

36. 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).

37. 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).

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

39. Camilo, C. R. d. S., Ruggiero, J. R. & de Araujo, A. S. A method for detection of water permeation events in molecular dynamics simulations of lipid bilayers. *Braz. J. Phys.* **52**, 1–13 (2022).

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