# NMRlipids Databank makes data-driven analysis of biomembrane properties accessible for all

**Anne M. Kiirikki[1], Hanne S. Antila[2], Lara Bort[3], Pavel Buslaev[4], Fernando Favela[5], Tiago Mendes Ferreira[6], Patrick F.J. Fuchs[7,8], Rebeca Garcia-Fandino[9], Ivan Gushchin[10], Batuhan Kav[11,12], Patrik Kula[13], Milla Kurki[14], Alexander Kuzmin[10], Jesper J. Madsen[15,16], Markus S. Miettinen[17,18], Ricky Nencini[1], Thomas Piggot[19], Ángel Piñeiro[20], Suman Samantray[10], Fabián Suarez-Leston[9,20,21], and O. H. Samuli Ollila[1,*]**

[1]University of Helsinki, Institute of Biotechnology, Helsinki, Finland
[2]Department of Theory & Bio-Systems and Department of Biomaterials, Max Planck Institute of Colloids and Interfaces, Germany
[3]University of Potsdam, Institute of Physics and Astronomy, Potsdam-Golm, 14476, Germany
[4]Nanoscience Center and Department of Chemistry, University of Jyväskylä, P.O. Box 35, Jyväskylä, 40014 , Finland
[5]Departamento de Ciencias Básicas, Tecnológico Nacional de México - ITS Zacatecas Occidente, Sombrerete, Zacatecas, 99102, México
[6]NMR group - Institute for Physics, Martin Luther University Halle-Wittenberg, Halle (Saale), 06120, Germany
[7]Sorbonne Université, Ecole Normale Supérieure, PSL University, CNRS, Laboratoire des Biomolécules (LBM), Paris, 75005, France
[8]Université Paris Cité, UFR Sciences du Vivant, Paris, 75013, France
[9]Center for Research in Biological Chemistry and Molecular Materials (CiQUS), Universidade de Santiago de Compostela, Santiago de Compostela, E-15782, Spain
[10]no affiliation
[11]Institute of Biological Information Processing: Structural Biochemistry (IBI-7), Forschungszentrum Jülich, Jülich 52428, Germany
[12]ariadne.ai GmbH (Germany), Häusserstraße 3 Heidelberg 69115, Germany
[13]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nám. 542/2, Prague, CZ-16610, Czech Republic
[14]School of Pharmacy, University of Eastern Finland, 70211 Kuopio, Finland
[15]Global and Planetary Health, College of Public Health, University of South Florida, Tampa, Florida, 33612, United States of America
[16]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, Florida, 33612, United States of America
[17]Department of Chemistry, University of Bergen, Norway
[18]Computational Biology Unit, Department of Informatics, University of Bergen, Norway
[19]Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom
[20]Department of Applied Physics, Faculty of Physics, University of Santiago de Compostela, Santiago de Compostela, E-15782, Spain
[21]MD.USE Innovations S.L., Edificio Emprendia, 15782 Santiago de Compostela, Spain
[*]samuli.ollila@helsinki.fi

## ABSTRACT

Cellular membrane lipid composition is implicated in diseases and major biological functions in cells, but membranes are difficult to study experimentally due to their intrinsic disorder and complex phase behaviour. Molecular dynamics (MD) simulations have been useful in understanding membrane systems, but they require significant computational resources and often suffer inaccuracies in model parameters. Applications of data-driven and machine learning methods, currently revolutionising many fields, are limited to membrane systems due to the lack of suitable training sets. Here we present the NMRlipids Databank—a community-driven, open-for-all database featuring programmatic access to quality evaluated atom-resolution MD simulations of lipid bilayers. The NMRlipids databank will benefit scientists in different disciplines by providing automatic ranking of simulations based on their quality against experiments, flexible implementations of data-driven and machine learning applications, and rapid access to simulation data via graphical user interface. To demonstrate the unlocked possibilities beyond current MD simulation studies, we analyzed how anisotropic diffusion of water and cholesterol flip–flop rates depend on membrane properties.

## 1 Introduction

Cellular membranes contain hundreds of different types of lipid molecules that regulate the membrane properties, morphology, and biological functions[?,?,?]. Membrane lipid composition is implicated in diseases, such as cancer and neurodegenerative disorders, and therapeutics that affect membrane compositions are emerging[?]. However, biomembranes are often difficult to study experimentally because they are complex mitures of lipids and proteins that are in disordered fluid state with complex phase behaviour at biological conditions. For those reasons, the detailed connections between complex lipid interactions and biological functions taking place in or around membranes remain poorly understood. Molecular dynamics (MD) simulations have been particularly useful in understanding membrane systems, although their accuracy has often been compromised by artefacts such as the quality of model parameters[?,?]. Presently, the accuracy of models is becoming increasingly important as researches are progressing from simulations of individual molecules to simulating whole organelles or even cells using interdisciplinary approaches[?,?,?]. Such systems exhibit complex emergent behaviour making inaccuracies more difficult to detect and accumulation of even modest errors may have major impact on the conclusions.

In contrast to experimental structural biology, where standard protocols to share and quality-evaluate the resolved structures are firmly established *via* the Protein Data Bank (PDB)[?], the equivalent best practices are yet to be defined for MD simulations. The importance of such approaches is widely recognized[?,?,?,?,?,?,?,?,?] and data-sharing solutions are emerging for proteins in solution[?,?], proteins in membranes[?,?,?], nucleic acids[?], nucleic acids and proteins[?], cyclodextrins[?], COVID-19–involved macromolecules (`bioexcel-cv19.bsc.es`). However, automatic quality evaluation of simulation data[?,?] and programmatic access are still rare. In particular, tools for automatic quality evaluation of membrane simulations, or training sets for machine learning models of membrane-containing systems, are not yet available. Recent advances using machine learning approaches utilizing publicly available databanks, for example to solve protein structures[?], emphasize the increasing importance of such resources.

Here we present the NMRlipids Databank—a community-driven, open-for-all database featuring programmatic access to atom-resolution MD simulations of lipid bilayers. The programmatic access enables users to apply data-driven approaches on the Databank, thus facilitating the creation of new tools for (and by) researchers in a wide range of fields covering academia and industry, from cell membrane biology and lipid nanoparticle formulations to computational chemistry and machine learning. As two usage examples of what is already possible, we demonstrate here (i) how data-driven analysis of water anisotropic diffusion in all the membrane systems available in the Databank can extend the scope of MD simulations to magnetic resonance imaging (MRI) and pharmacokinetics, and (ii) how the Databank allows its users to analyse rare phenomena that are beyond the scope of standard MD simulation investigations. Wider adaptation of the NMRlipids Databank will open even more possibities. Furthermore, the Databank performs automatic quality evaluation of membrane simulations, which facilitates the selection of best models for specific applications and accelerates the development of simulation parameters and methodology.

While the NMRlipids Databank currently contains only lipid bilayer systems, its key elements that enable the programmatic access to large scale MD simulation data can be applied also for other molecules, such as disordered proteins or sugars. The powerful combination of the overlay databank structure and a community-wide open collaboration can be used to build databases that enable data-driven and machine learning applications also in other fields where storage of raw data requires significant resources, best practices in the field are not defined, and incentives to share data do not exist, such as assignment of NMR spectra[?].

## 2 Results

### 2.1 MD simulations of membranes composed of most common biologically abundant lipids

NMRlipids Databank is a community-driven database containing atomistic MD simulations of biologically relevant lipid membranes emerging from the NMRlipids open collaboration[?,?,?,?,?]. It has been designed to improve data Findability,

Accessibility, Interoperability, and Reuse (FAIR)[?] for MD simulation trajectories. The NMRlipids databank is constructed using the NMRlipids project protocol where all the content is openly accessible throughout the project[?]. Currently the NMRlipids databank contains 726 simulation trajectories with the total length of approximately 0.4 ms. The distribution of available simulations containing a specific lipid, shown in Fig. 1B, roughly resembles the biological abundance of different lipid types with phosphatidylcholine (PC) being the most common followed by cholesterol, phosphatidylethanolamine (PE), phosphatidylserine (PS), phosphatidylglycerol (PG), phosphatidylinositol (PI), and other lipids depending on organism and organelle[?]. Abbreviations and full names of all lipids present in the databank are listed in Table **??**. Single component lipid membranes and binary mixtures, depicted in Fig. 1F, are currently most abundant in the NMRlipids databank, yet mixtures with up to five lipid types are available. Force fields used in simulations cover all the essential parameters commonly used in lipid simulations, see Fig. 1C and table **??**, including also united atom and polarizable force fields. Therefore, the averages calculated over the databank can be considered as mean predictions from available lipid models for an average cell membrane.

The overlay structure of the NMRlipids databank, illustrated in Fig. 1A, is designed to enable efficient upcycling of MD simulations for data-driven and machine learning applications with mimimal investment on new infrastructure. Raw simulation data in the *data layer* can be stored in any publicly available location with long term stability, such as Zenodo (`www.zenodo.org`), and with permanent links to the data, such as digital object identifiers (DOIs). The *Databank layer* (`www.github.com/NMRLipids/Databank`) is the core of the databank containing all the relevant information about the simulations, including links to the raw data, relevant metadata describing the systems, universal naming conventions for lipids and their atoms, quality evaluation of simulations against experimental data, and computer programs to create the entries and analyse basic properties calculated from all simulations (area per lipid, C–H bond order parameters, X-ray scattering form factors, membrane thickness and equilibration times of principal components). Also the results for these basic properties are stored in the *Databank layer*.

The *application layer* is composed of repositories and tools that read information from the *databank layer* for further analyses. Because *application layer* does not interfere with the *databank layer*, it can be freely extended by anyone for wide range of purposes. This is demonstrated here with two examples: NMRlipids databank graphical user interface (NMRlipids databank-GUI) tool at `www.databank.nmrlipids.fi/` and a repository exemplifying novel analyses utilizing NMRlipids databank as discussed below (`www.github.com/NMRLipids/DataBankManuscript`). More detailed description of the NMRlipids databank structure is available in the supplementary information.

### 2.2 NMRlipids databank-GUI: graphical access to the MD simulation data

NMRlipids databank-GUI, available at `www.databank.nmrlipids.fi/`, provides easy access to the NMRlipids databank content via graphical user interface (GUI). Simulations can be searched based on their composition, used force field, temperature, properties and quality. The search results are ranked based on simulation quality evaluated against experimental data. NMRlipids databank-GUI enables also visualization of simulations and comparison of properties between different simulations and experiments. The NMRlipids databank-GUI enables rapid surveys of available simulation data, selection of best available simulations for specific systems based on ranking lists, and comparisons of basic properties between different types of membranes. Notably, GUI enables these operations to be performed by scientists with wide ranges of backgrounds who do not necessarily have expertise in programming or other means to access to MD simulation data.

### 2.3 NMRlipids databank-API: programmatic access to the MD simulation data

The NMRlipids databank-API, available at `www.github.com/NMRLipids/Databank`, provides programmatic access to all simulation data in the NMRlipids databank. This enables wide range of novel data-driven applications from construction of machine learning models that predict membrane properties to automatic analyses of virtually any property from all simulations in the databank. Practical implementation of such analyses is illustrated in the flowchart in Fig. 2D. First, raw data of each simulation can be accessed by iterating through the README.yaml files in the *databank layer* that contain links pointing to the locations of simulation trajectories and other relevant files. Simulations can be then automatically analyzed with the help of README.yaml and mapping files that affiliate specific naming conventions in each simulation to the universal molecule and atom names. Finally, the analysis results are stored using the same structure as in the *databank layer*. Examples of codes that utilize NMRlipids databank-API and a template for new analyses can be found from locations listed in table **??**.

While analysis codes and results for basic membrane properties are included in the *Databank layer*, unlimited amount of further analyses can be implemented by anyone in separate repositories in the *application layer*. When *Application layer* repositories are organized by mimicking the *Databank layer* structure, they can be accessed programmatically and further analyzed using the tools in NMRlipids databank-API by implementing the flowchart demonstrated in Fig. 2E. Novel analyses demonstrating the power of the NMRlipids databank in selecting best simulation models, analyses of rare phenomena and extending MD simulations to new fields are implemented in an *application layer* repository located at `www.github.com/NMRLipids/DataBankManuscript`. The related codes are listed in table **??**.

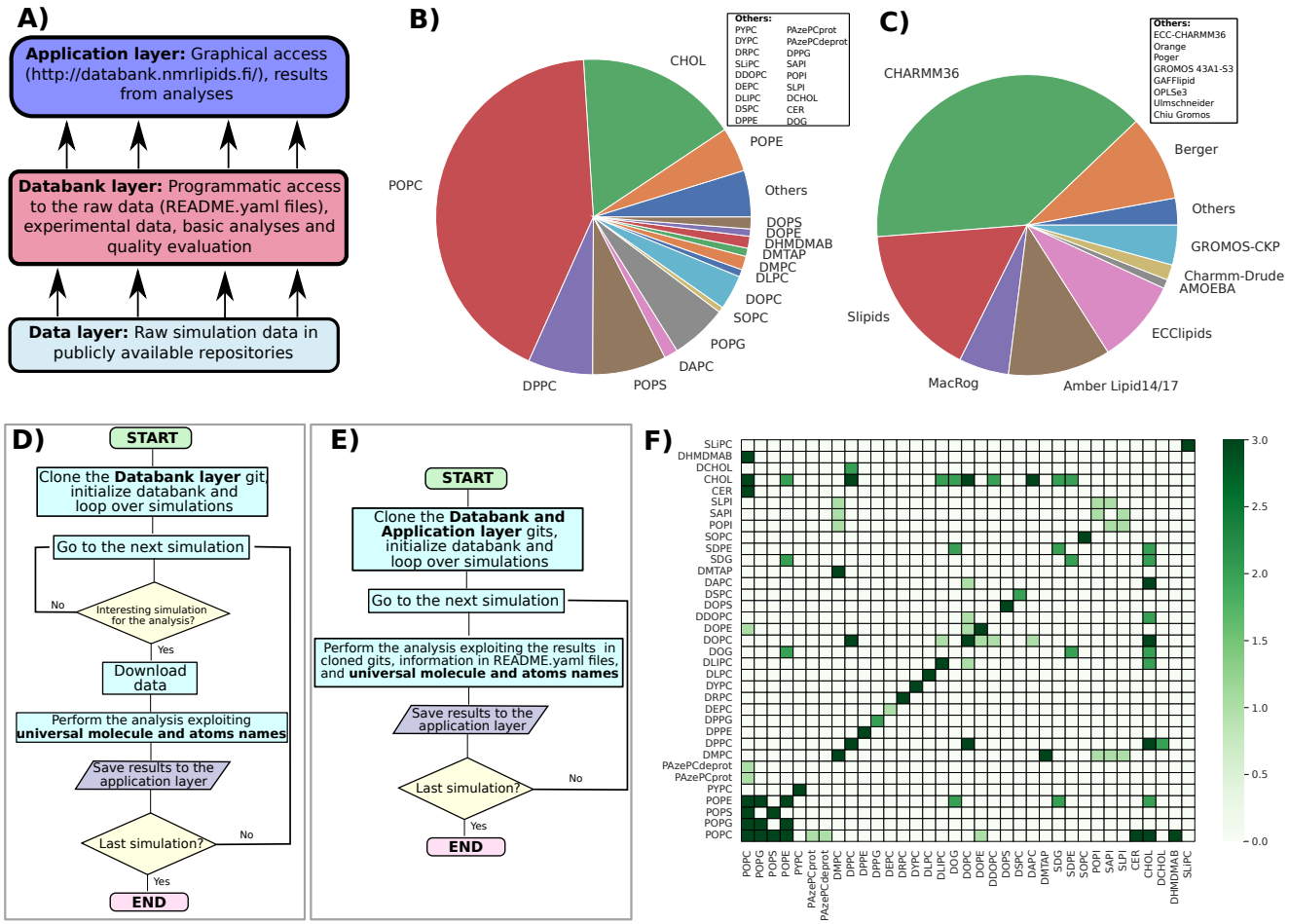



**Figure 1.** A) Schematic presentation of the overlay structure used in the NMRlipids databank. A more detailed structure of the databank layer is shown in Fig. **??** in the SI. B) Distribution of lipids present in the trajectories in the NMRlipids databank. Lipids occuring in five or fewer simulations ('others') are listed on the right. C) Distribution of force fields in the simulations in the NMRlipids databank. References for each force field are given in table **??**. D) Flowchart for performing an analysis of properties through all MD simulations in the NMRlipids databank using the API. E) Flowchart for accessing results calculated from the NMRlipids databank and stored to the application layer. F) Currently available binary mixtures in the NMRlipids databank.

## 2.4 Applications of the NMRlipids databank

### *2.4.1 Selection of simulation parameters using the NMRlipids databank: Best models for most abundant neutral membrane lipids*

The quality of lipid bilayer MD simulations has to be carefully assessed in their applications to minize the detrimental consequences of artificial MD simulation results for lipid bilayers[?]. This is possible, for example, by using C–H bond order parameters from NMR and X-ray scattering form factors[?], although it requires comparisons between large number of simulations which is laborious even with collaborative approaches[?,?,?,?]. To streamline this process we have defined quantitative quality measures for lipid bilayer simulations that enable automatic ranking of lipid bilayer simulations based on their quality against experiments. Conformational ensembles of individual lipid molecules are evaluated in the NMRlipids databank by calculating the probabilities for each C-H bond order parameter to locate within experimental error. Furthermore, the quality against X-ray scattering experiments is estimated from the distance of the first form factor minimum between the experiment and simulation. These measures can be used to evaluate membrane dimensions as the form factor minima locations and acyl chain order parameters correlate with the membrane lateral packing and thickness as shown in Figs. 2G, **??**, and **??**. In addition, the ergodicity of conformational sampling of lipids is estimated by calculating the convergence time of the slowest principal component divided by the length of a simulation. Here we demonstrate how automatic simulation quality evaluation and NMRlipids databank-API enable rapid selection of the best models for simulations of membranes with POPC and POPE lipids

that are the two most biologically abundant neutral membrane lipids[?].

Predictions for the lateral packing of POPC and POPE membranes in terms of area per lipid, one of the most important parameter used to characterize cellular membranes, diverge between different force fields as illustrated in Fig. 2A. To select the most realistic models among these, we first ranked all simulations in the databank according to the quality of C-H bond order parameters evaluated against experimental data in Fig. **??**. From these results, we then picked the force fields occurring in Fig. 2A and ranked them according to the quality of *sn*-1 chain in Figs. 2B and C. Simulations with $\tau_{rel}$ clearly above one were discarded. Also the highest ranked simulation (POPC bilayer with OPLS3e parameters) is included for the reference. Because the membrane packing correlates with the average order parameter of the *sn*-1 chain (Fig. 2G), rankings in Figs. 2B and C were used to select the simulations giving most realistic results in Fig. 2A. Also direct comparisons with the experimental data for the most relevant simulations are shown in Figs. 2C–E and **??**. Based on these rankings, Lipid17 and Slipids simulations give the most realistic predictions for POPC membrane, while simulations with CHARMM36 and GROMOS-CKP parameters predict overly packed bilayers (overestimated order in Fig. **??**). On the other hand, GROMOS-CKP and Slipids give the most realistic results for POPE, while CHARMM36 and Lipid17 predict membranes that are too packed. In conclusion, the quality evaluation based on the NMRlipids databank suggests that Slipids parameters are the best available choice for simulations with PC and PE lipids, at least for applications where membrane packing is relevant.

#### *2.4.2 Detecting rare phenomena using NMRlipids databank: Cholesterol flip–flops*

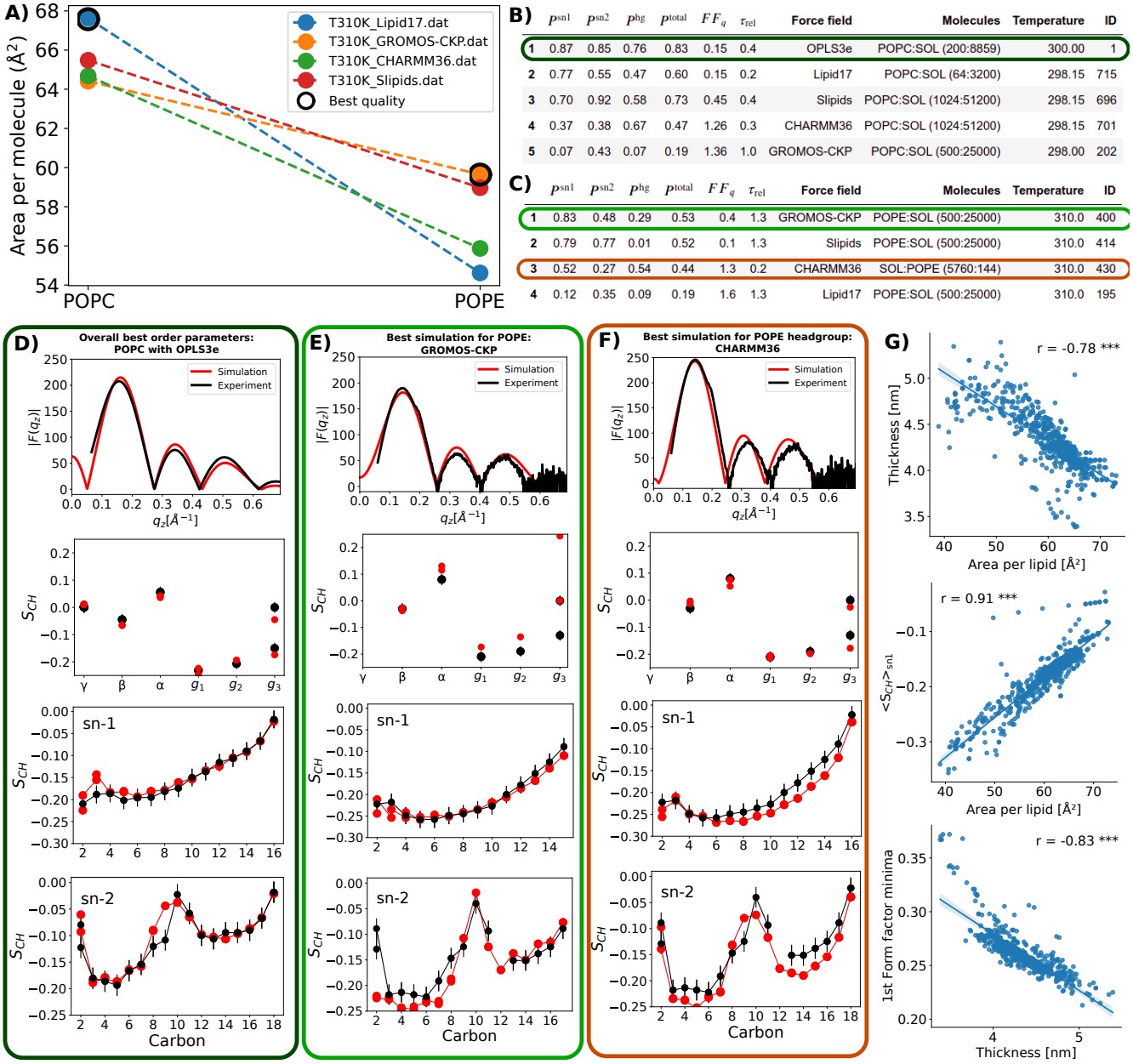
Lipid flip–flops from one bilayer leaflet to another play an important role in lipid trafficking and regulating membrane properties[?]. Phospholipid flip–flops are slow when not facilitated by proteins, occurring spontaneously on the timescale of hours or days, while cholesterol, diacylglycerol and ceramide flip-flops are faster. Still, the reported timescales range from minutes to sub-millisecods[?,?,?,?]. These timescales were previously accessible only by coarse-grained simulations or free energy calculations[?], yet atomistic simulations reporting cholesterol flip–flop events have been published only recently[?,?,?]. These studies report an increase in cholesterol flip–flop rates with increasing acyl chain unsaturation level and decreasing cholesterol concentration[?,?], but the amount of data in these individual studies was not sufficient to systematically assess correlations between cholesterol flip–flop rates and membrane properties. Here, we demonstrate that the NMRlipids databank API makes analyses of such rare phenomena accessible to everyone by enabling the access to large amount of MD simulation data as described in section 2.2. This is particularly useful for scientists in different fields of science and industry who do not have the access to required computational resources and expertise to produce the large amounts of MD simulation data.

Using workflow depicted in Fig. 1D, we first calculated the cholesterol flip–flop rates from all the simulations available in the NMRlipids databank. Flip–flops were observed for cholesterol, DCHOL (18,19-di-nor-cholesterol), DOG (1,2-dioleoyl-*sn*-glycerol), and SDG (1-stearoyl-2-docosahexaenoyl-*sn*-glycerol). The observed cholesterol flip–flop rates, ranging between 0.001–1.6 $\mu s^{-1}$ with the mean of 0.16 $\mu s^{-1}$ and median of 0.07 $\mu s^{-1}$, are in line with the previously reported values from atomistic MD simulations[?,?,?]. The flip–flop rate of DCHOL, 0.2 $\mu s^{-1}$, was close to the average value of cholesterol, while average rates for diacylglycerols DOG and SDG were higher than for cholesterol, 0.4 $\mu s^{-1}$ and 0.5 $\mu s^{-1}$, respectively. Flip–flops were not observed for other lipids, giving the upper limits for PC lipid flip–flop rate of $9 \cdot 10^{-6}$ $\mu s^{-1}$ and for ceramide (*N*-palmitoyl-D-*erythro*-sphingosine) of 0.002 $\mu s^{-1}$. Thus, available data in the NMRlipids databank suggests that the lipid flip–flip rate decreases in the order of diacylglycerols > cholesterol > other lipids including ceramides. However, the amount of data for diacylgycerols (8 simulations with Lipid17 force field) and ceramide (3 simulations with CHARMM36 force field) is less than for cholesterol (83 simulations), thus we cannot fully exclude the effect of force field or composition on this comparison.

Nevertheless, we used the workflow depicted in Fig. 1E to analyse how flip–flop rates calculated from the NMRlipids databank depend on membrane properties. Cholesterol flip–flop rates and their histograms are plotted as a function of membrane thickness, lateral density and order in Figs. 3 B–D. The results reveal a non-linear correlation between cholesterol flip–flop rate and membrane packing which is depicted as area per lipid. Flip–flop rates increase by an order of magnitude when membrane packing density decreases and a major jump is observed at low membrane packing. The order of magnitude changes in cholesterol flip–flop rate with the membrane composition may have major implications in understanding lipid trafficking and membrane biochemistry[?,?]. Because the results from the NMRlipids databank are averaged over large range of membrane compositions and force fields, they show that the strong dependence of cholesterol flip–flop rate on membrane properties is not limited to certain lipid compositions or force fields used in previous studies[?,?,?].

#### *2.4.3 Extending the scope of MD simulations to new fields using the NMRlipids databank: Water diffusion anisotropy in membrane systems*

The anisotropy in water diffusion in parallel and perpendicular directions with respect to membranes can be related to the translocation of drugs through biological material, particularly in skin[?,?,?,?], and to the formation of signals in diffusion tensor MRI imaging[?]. MD simulations are rarely used to analyze the anisotropic diffusion of water since only few membrane permeation events for water are typically observed in a single MD simulation trajectory[?,?], thereby making the collection of sufficient amount of data challenging. Here, we show that the API access to the data in NMRlipids databank enables systematic




B)

| | $P^{sn1}$ | $P^{sn2}$ | $P^{hg}$ | $P^{total}$ | $FF_q$ | $\tau_{rel}$ | Force field | Molecules | Temperature | ID |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.87 | 0.85 | 0.76 | 0.83 | 0.15 | 0.4 | OPLS3e | POPC:SOL (200:8859) | 300.00 | 1 |
| 2 | 0.77 | 0.55 | 0.47 | 0.60 | 0.15 | 0.2 | Lipid17 | POPC:SOL (64:3200) | 298.15 | 715 |
| 3 | 0.70 | 0.92 | 0.58 | 0.73 | 0.45 | 0.4 | Slipids | POPC:SOL (1024:51200) | 298.15 | 696 |
| 4 | 0.37 | 0.38 | 0.67 | 0.47 | 1.26 | 0.3 | CHARMM36 | POPC:SOL (1024:51200) | 298.15 | 701 |
| 5 | 0.07 | 0.43 | 0.07 | 0.19 | 1.36 | 1.0 | GROMOS-CKP | POPC:SOL (500:25000) | 298.00 | 202 |

C)

| | $P^{sn1}$ | $P^{sn2}$ | $P^{hg}$ | $P^{total}$ | $FF_q$ | $\tau_{rel}$ | Force field | Molecules | Temperature | ID |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.83 | 0.48 | 0.29 | 0.53 | 0.4 | 1.3 | GROMOS-CKP | POPE:SOL (500:25000) | 310.0 | 400 |
| 2 | 0.79 | 0.77 | 0.01 | 0.52 | 0.1 | 1.3 | Slipids | POPE:SOL (500:25000) | 310.0 | 414 |
| 3 | 0.52 | 0.27 | 0.54 | 0.44 | 1.3 | 0.2 | CHARMM36 | SOL:POPE (5760:144) | 310.0 | 430 |
| 4 | 0.12 | 0.35 | 0.09 | 0.19 | 1.6 | 1.3 | Lipid17 | POPE:SOL (500:25000) | 310.0 | 195 |

**Figure 2.** A) Area per lipids of POPC and POPE lipid bilayers predicted by different force fields at 310 K in simulations that are available in the NMRlipids databank. The points from the best performing simulations based on rankings in B and C are surrounded by black circles. B) Best POPC simulations ranked based on *sn*-1 acyl chain order parameter quality. C) Best POPE simulations ranked based on *sn*-1 acyl chain order parameter quality. D)–F) Direct comparison against experimental (NMR order parameters and X-ray scattering ) 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). G) Scatter plots and Pearson correlation coefficients, *r*, for the membrane area per lipid, thickness, first minimum of X-ray scattering form factor and average order parameter of the *sn*-1 acyl chain extracted from the NMRlipids databank. All correlation coefficients have p-value below 0.001 as indicated by ***. For more correlations see Fig. **??**.

analyses on how the anisotropic diffusion of water depends on membrane properties in multilamellar membrane systems, thereby extending the applications of MD simulations to new fields.

To this end, we first calculated the water permeability through membranes from all simulations in the NMRlipids databank using the workflow depicted in Fig. 1D. The resulting non-zero values range between 0.3 and 322 µm/s with the mean and median of 14 µm/s and 8 µm/s, respectively. These values agree with the previously reported simulation results[?,?], but are on

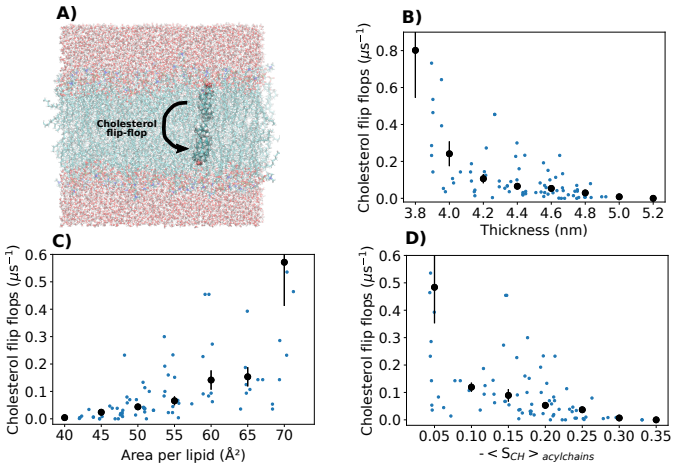



**Figure 3.** **A** Illustration of cholesterol flip–flop. **B–D** Cholesterol flip–flops analyzed from the databank as a function of membrane thickness, area per lipid, and acyl chain order. Values from simulations with non-zero flip–flop rates are shown with blue dots. Histogrammed values are shown with black dots.

average slightly larger than experimental values reported for PC lipids in the liquid crystalline phase, 0.19–0.33 µm/s[?]. Using the workflow depicted in Fig. 1E, we then plotted the observed permeabilities and their histogrammed values in Figs. 4B–E as a function of temperature, membrane thickness, area per lipid, and acyl chain order. As expected, the permeability increases with the temperature, giving an average energy barrier of $17 \pm 4\, k_B T$ for the water permeation from the Arrhenius plot in Fig. 4B. On the other hand, the permeability of water decreases on average when membranes become more packed, *i.e.*, with decreasing area per lipid and increasing thickness and acyl chain order (Figs. 4 C–E). Permeation of water through bilayers depends on membrane properties also according to previous studies, but there is no established consensus on whether the area per lipid[?] or bilayer thickness[?] is the main parameter determining the permeability. Our analysis over the NMRlipids databank, containing significantly more data than what was available in previous studies, suggest non-linear dependencies on both of these parameters, yet the linear correlation would be a good approximation for thicknesses above $\sim 3.9$ nm and area per lipids below $\sim 69$ Å$^2$ (insets in Figs. 4 C–D). Clear dependencies of permeability on charged lipids, cholesterol, POPE, or hydration level were not observed in Fig. **??** in the supplementary information.

To analyze how water diffusion anisotropy depends on membrane properties in a multi-lamellar lipid bilayer system, we calculated the water diffusion parallel to the membrane surface from all simulations in the NMRlipids databank using the workflow depicted in Fig. 1D. The parallel diffusion coefficient of water, $D_{||}$, decreases with reduced hydration and increases with the temperature, but dependencies on membrane area per lipid, thickness, or fraction of charged lipids were not observed in Figs. 4 and **??** when the workflow depicted in Fig. 1E was used. Simulation results are close to the experimental values with low hydration levels in Fig. 4 F, but increase to approximately two times higher than the experimental value for bulk water diffusion value ($3.1 \cdot 10^{-9}$ m$^2$/s at 313 K[?]) with high hydration levels. This is not surprising as the most common water model used in membrane simulations, TIP3P, overestimates the bulk water diffusion[?]. To estimate the diffusion anisotropy of water, $D_\perp/D_{||}$, in multilamellar membrane system, the permeability coefficients of water through membranes were translated to perpendicular diffusion coefficients, $D_\perp$, using the Tanner equation[?,?]. The resulting perpendicular diffusion coefficients are approximately five orders of magnitude smaller than lateral diffusion coefficients of water (Figs. 4 G and H), which is at the upper limit of the anisotropy estimated from the experimental data[?]. Significant increase in the diffusion anisotropy with membrane packing is observed, as $D_\perp/D_{||}$ drifts away from one with decreasing area per lipid and increasing thickness in Figs. 4 G and H. This follows from decreasing water permeability with membrane packing (Figs. 4 C and D), while lateral diffusion remains approximately constant (Figs. **??** A and C).

In summary, our results suggest that the bilayer packing has a substantial effect on anisotropic water diffusion in multi-membrane lipid systems. Several folds larger anisotropy in membranes with higher lateral density is expected to play a role in pharmacokinetic models not only for water but also for other hydrophilic molecules[?]. Furthermore, the enhanced understanding of this anisotropy may help in developing new diffusion tensor based MRI imaging methods where signals originate from anisotropic diffusion of water in biological material[?].

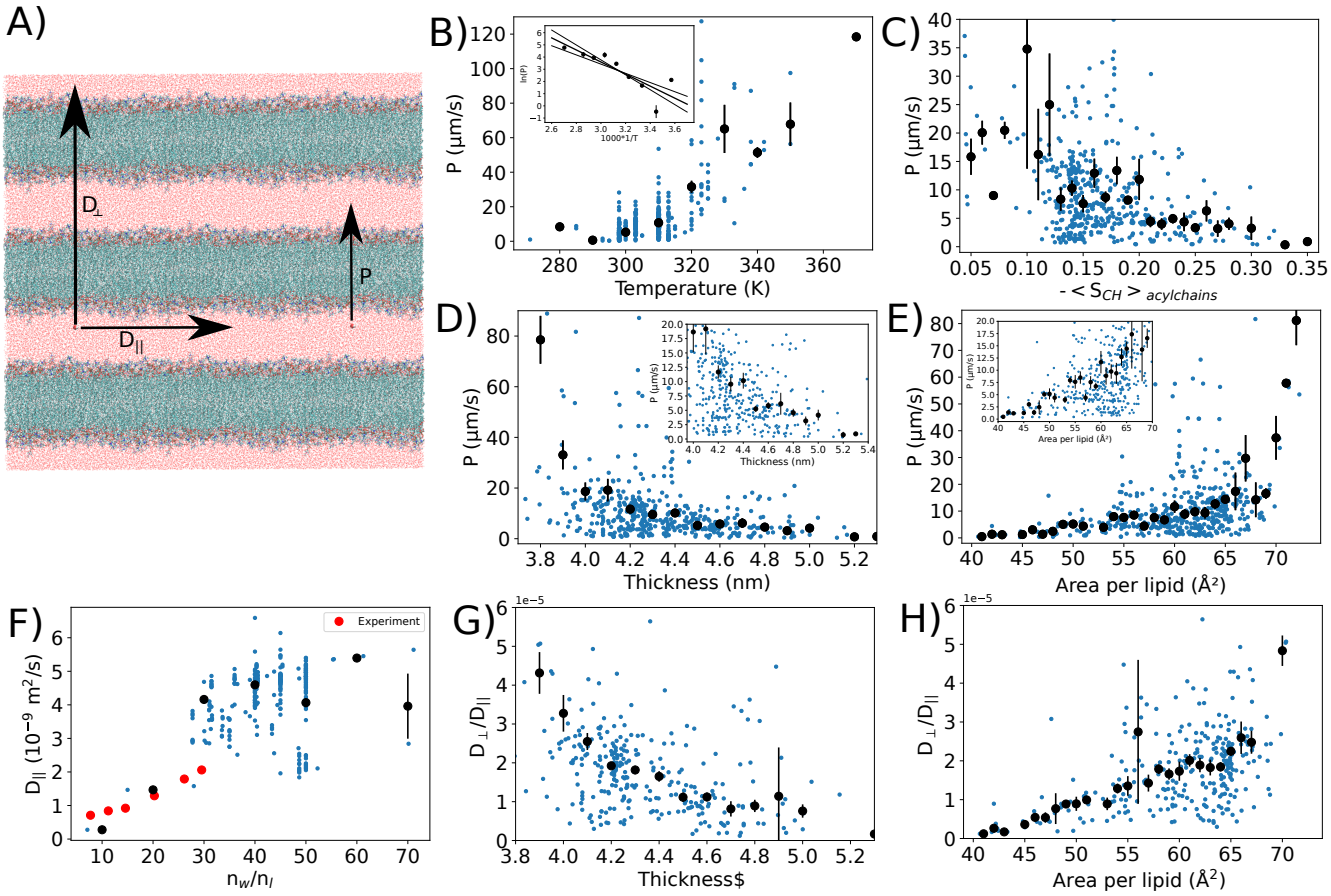



**Figure 4.** **A** Water diffusion, $D_{\text{perp}}$, and permeability, $P$, through membranes, and lateral diffusion along the membrane, $D_{||}$, illustrated in a multilamellar stack of lipid bilayers. **B–E** Water permeation through membranes analyzed from the databank as a function of temperature, thickness, area per lipid, and acyl chain order. Values from simulations with non-zero permeation values are shown with blue dots. Histogrammed values are shown with black dots. Inset in B) shows the Arrhenius plot of permeation ($\ln(P)$ vs. $1/T$) that gives $17 \pm 4\, k_{\text{B}}T$ for the average activation energy for water permeation through lipid bilayer. Insets in C) and D) show the region where the dependence could be considered approximately linear. **F** Lateral diffusion of water as a function of hydration level. Experimental points for DMPC bilayers at 313 K at different hydration levels are shown[?]. **G–H** Diffusion anistoropy of water as a function of thickness and area per lipid.

## 3 Discussion

The focus of biomolecular simulations is moving from studies of individual molecules to larger complexes and even whole cells and organelles[?,?,?]. Simultaneously, machine learning based models predicting behaviour of biomolecules and automatic approaches to parametrize models are emerging[?,?]. The resources delivered by the NMRlipids databank will support the development in both of these directions. Automatic quality evaluation and ranking of simulations against experimental data enables the selection of best simulations for specific applications without laborious manual force field evaluation. This also streamlines automatic parametrization procedures for atomistic and coarse grained simulations. Such practises fostering the accuracy of simulations are becoming increasingly important due the accumulation of small errors when complexity and size of simulated systems are increasing.

The open programmatic access to unprecedented amount of MD simulation data via NMRlipids databank-API enables wide range of novel data-driven analyses and machine learning applications. The analyses of cholesterol flip–flop events (Fig. 3) and water permeation through membranes (Fig. 4) demonstrate how large amount of accessible simulation data in terms of quantity (*e.g.*, simulation length and number of conformations) and content (*e.g.*, lipid compositions and ion concentrations) enable analyses of rare phenomena that are beyond the current possibilities for a single research group. Such analyses also pave the way for applications of MD simulations in new fields as demonstrated here by analysing the anisotropic diffusion of water in membrane systems (Fig. 4) that is essential parameter in pharmacokinetic modeling and MRI imaging[?,?]. Furthermore, the NMRlipids databank-API delivers access to a training set data that paves the way for diverse machine learning applications to

| Type of application | Practical examples | Target group |
|---|---|---|
| Analyses of rare phenomena | Lipid flip–flops, water permeation (Figs. 3 and 4) | Membrane scientists |
| Correlations between membrane properties | Membrane structural properties, water dynamics (Figs. 2 and 4) | Membrane scientists |
| Applications that are outside typical scope of MD simulations | Anisotropic water diffusion for pharmacokinetics and MRI imaging applications (Fig. 4) | Scientist in fields where MD simulations are not usually applied |
| Selection of the best simulation model for a specific application | Best model for membranes with PC and PE lipids (Fig. 2), lipid headgroup conformations[?], packing of PS[?] and PE (Fig. 2) containing membranes. | Scientists using MD simulations |
| Guidance for force field development | Improvements in ion binding to lipids[?,?] and lipid headgroup conformational ensembles[?,?,?] | Scientists developing parameters for MD simulations |
| Training and target data for the development of coarse grained models | Optimizing parameters of coarse grained models against NMRlipids databank, extracting continuum parameters for membranes. | Scientists developing and using coarse grained MD simulations |
| Training set for machine learning applicatons | Programmatic access to the data and results enables training of machine learning type of models for various applications, such as predictions of membrane properties from composition | Scientists building and using machine learning applications for biomolecules. |

**Table 1.** Examples on applications of the NMRlipids databank.

predict membrane properties. Such applications are analogous to AlphaFold[?] and other tools that predict proteins structures from their sequence using artificial intelligence. These possibilities are particularly valuable for scientist who do not typically have access to large scale MD simulation data. Different types of expected applications of the NMRlipids databank in wide range of fields are listed in Table 1, yet the scope of such applications is expected to further widen with increasing amount of data in the NMRlipids databank.

Potential benefits of sharing MD simulation data are widely recognized by different stakeholders from scientists[?,?,?,?,?,?,?,?,?] to funders and journal staff, but data sharing with efficient tools for upcycling have been rare. The main barriers for such solutions have been the required commitment for the long term support of hardware and software, and the lack of incentives for researchers to share the data. The first issue is solved in the NMRlipids databank using the overlay design where raw data is distributed to already publicly available decentralized locations while the core of the databank is composed only of metadata stored in a version controlled git repository with an open access licence. On the other hand, the open collaboration approach developed in the NMRlipids project[?] creates incentives for sharing the data by offering authorship in published articles to the contributors. Advantages of such approach are demonstrated here for membrane simulations, but the concept can be applied also not only to other biomolecules, such as disordered proteins and membrane-protein systems, but also in other fields where similar barriers hinder the machine learning revolution, such as assignment of NMR spectra[?].

# 4 Methods

## 4.1 Structure of the databank

The overlay structure designed for the NMRlipids databank is composed of three layers illustrated in Fig. 1A. The *data layer* contains raw data that can be distributed to publicly available servers such as Zenodo (`www.zenodo.org`). The core content of the databank locates in the *databank layer* which is a git repository at `https://github.com/NMRLipids/Databank/` and is also permanently stored in Zenodo repository (`www.zenodo.org`)[?]. The Essential information of each simulation is stored in a human and machine readable README.yaml file located at subfolders of `/Data/Simulations` folder in the *databank layer* repository. Each subfolder has a unique name constructed based on hash code of trajectory and topology files of each simulation. The README.yaml files in these folders contain access to all information that are needed for further analyses of simulations, such as links to the raw data and affiliations of universal molecule and atom names. The content of these files is described in detail in Table **??** in the supplementary information. Results from analyses of basic membrane properties (area per lipid, thickness, C-H bond order parameters, x-ray scattering form factors and relaxation of principal components) are stored in the same folders with README.yaml files. Experimental data used for ranking is stored in `/Data/experiments` folder, results from such ranking in `/Data/Ranking` and relevant scripts in `/Scripts/` folder in the *databank layer* repository. The scripts in the NMRlipids databank are mainly written in Python and many of them utilize the MDAnalysis module[?,?]. The databank structure is illustrated more detailed in Fig. **??** in the supplementary information. Whenever specific files or folders are referred here, they locate at *datbank layer* repository unless stated otherwise.

## 4.2 Molecule and atom naming convention

When analysing simulations, atoms and molecules needs to be often called by their names defined in the simulation trajectory. However, these names typically vary between force fields because the universal naming convention has not been defined for lipids. To enable automatic analyses over simulations with different atom and molecule names in the NMRlipids databank, we have defined unique naming conventions for molecules and atoms therein. Unique abbreviations used in the NMRlipids databank for each molecule are listed in table **??** in the supplementary information. Atom names used in simulation trajectories are connected to unique atom names using mapping files that are defined in the NMRlipids project (`https://nmrlipids.blogspot.com/2022/04/new-yaml-format-of-mapping-files.html`). These files are located at `/Scripts/BuildDatabank/mapping_files` in the NMRlipids databank repository. These files also define whether an atom belongs to headgroup, glycerol backbone, or acyl chain region in a lipid. In practise, force field specific molecule names and mapping files names are given in the COMPOSITION dictionary in README.yaml files for each molecule in each simulation in the NMRlipids databank.

## 4.3 Adding data into the NMRlipids databank

The NMRlipids databank is open for additions of simulation data by anyone. The list information that a contributor has to deliver is given in Table **??**. The rest of the information to be stored in README.yaml files, also listed in Table **??**, will be automatically extracted using the `/Scripts/BuildDatabank/AddData.py` script. In practise, the manually entered data is first stored into an info.yaml file that is then added into `/Scripts/BuildDatabank/info_files` folder via pull request. To avoid ineligible entries and minimize human errors, the pull requests are monitored before the acceptance and generation of README.yaml files. Currently, the NMRlipids databank is composed of simulations that are found from the Zenodo repository with an appropriate licence. Most, but not all, of these trajectories originate from previous NMRlipids projects[?,?,?,?].

## 4.4 Experimental data

Experimental data used in the quality evaluation, currently composed of C–H bond order parameters and X-ray scattering form factors, are stored in `/Data/experiments` in the NMRlipids databank repository. Similarly to simulations, each experimental data set has a README.yaml file containing all the relevant information about the experiment. The keys and their descriptions for the experimental data are given in Table **??**. NMR data currently in the NMRlipids databank are taken from Refs. **?**, **?**, **?**, **?**, **?** and X-ray scattering data from Refs. **?**, **?**, **?**, **?**, **?**, **?**, **?**. In addition, previously unpublished NMR data for POPE, POPG and DOPC was acquired as described in the supplementary information (Figs. **??**-**??**) and contributed to the databank.

## 4.5 Analysing simulations

In practise, simulation data in the NMRlipids databank can be currently analyzed by first downloading the data using the information in README.yaml and then performing the analysis on local computer utilising the universal naming conventions for molecules and atoms therein. This procedure is illustrated in Fig. 1D and tables **??** and **??** list examples of codes that perform such analyses in practise.

### 4.6 Calculation of C–H bond order parameters

The C–H bond order parameters were calculated directly from the carbon and hydrogen positions using the definition

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

where angular brackets denote the ensemble average, *i.e.*, average over all sampled configurations of all lipids in a simulation, and $\theta$ is the angle between the C–H bond and the membrane normal. As in previous NMRlipids publications, the order parameters were first calculated separately for each lipid and the standard error of the mean over different lipids was used as the error estimate[?]. The script that calculates C–H bond order parameters from all simulations in the NMRlipids databank is available at `/Scripts/AnalyzeDatabank/calcOrderParameters.py` in the NMRlipids databank repository. The resulting order parameters are stored for all simulations in files named `[lipid_name]OrderParameters.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.7 Calculation of X-ray scattering form factors

X-ray scattering form factors were calculated with the standard equation for lipid bilayers[?] that allows also for unsymetry in membranes,

$$F(q) = \left| \int_{-D/2}^{D/2} \Delta\rho_e(z) \exp(izq_z) \mathrm{d}z \right|, \tag{2}$$

where $\Delta\rho_e(z)$ is the difference between total and solvent electron densities, and $D$ is the simulation box size in the $z$-direction (normal to the membrane). For the calculation of density profiles, atom coordinates were first centred around the centre of mass of lipid molecules for every time frame, and a histogram of these centred positions, weighted with the number of electrons in each atom, was then calculated with the bin width of $1/3$ Å. Electron density profiles were then calculated as an average of these histograms over the time frames in simulations. The script to calculate form factors for all simulations in the NMRlipids databank is available at `Scripts/AnalyzeDatabank/calc_FormFactors.py`. The resulting form factors are stored for all simulations in files named `FormFactor.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.8 Calculation of a bilayer area per lipid and thickness

Area per lipids of bilayers were calculated by dividing the time-averaged area of the simulation box with the total number of lipid and surfactant molecules (see the list in table **??**) in the simulation. The script that calculates area per lipids from all simulations in the NMRlipids databank repository is available at `Scripts/AnalyzeDatabank/calcAPL.py` in the NMRlipids databank repository. The resulting area per lipids are stored for all simulations in files named `apl.json` at folders in `/Data/Simulations`.

Thicknesses of lipid bilayers were calculated from the intersections of lipid and water electron densities. The script that calculates thickness of all simulations in the NMRlipids databank is available at `Scripts/AnalyzeDatabank/calc_thickness.py` in the NMRlipids databank repository. The resulting thicknesses are stored in files named `thickness.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.9 Principal Components Analysis of equilibration of simulations

To estimate how well the provided trajectories are converged, the Principal Components Analysis (PCA) was used[?,?]. Following the PCALipids[?] protocol, the lipids of a specific type were first selected, and trajectories of individual lipids were concatenated. Next, each lipid configuration was aligned to the average lipid structure. Afterwards, the PCA was applied on the cartesian coordinates of all heavy atoms of the lipid. Since it was shown before, that the motions along the first, major, principal component (PC) are the slowest ones[?], the equilibration of individual lipids was estimated based on the dynamics of the first PC. The suggested way to estimate the characteristic timescales needed to adequately sample the configurational space of individual lipids is to calculate the distribution convergence of the trajectories projected on the first PC[?]. However, due to the linear dependence between autocorrelation decay times and distribution convergence times, faster calculation times and higher computational stability of autocorrelation decay times[?,?], the distribution convergence timescales were calculated from autocorrelation decay times:

$$\tau_{\mathrm{convergence}} = k * \tau_{\mathrm{autocorrelation}} \tag{3}$$

The empirical coefficient $k$ is equal to $k = 49$ and was calculated based on the analysis of 8 trajectories from the Databank and included simulations of POPC, POPS, POPE, POPG, and DPPC with CHARMM36 force fields. All simulations, used for $k$

calculation, were at least 200 ns long. As was shown previously[?], the coefficient does not depend on the force field, thus only CHARMM36 simulations, as the most frequent in the Databank, were used.

As a measure of equilibration quality for each trajectory, the ratio between the convergence time of first PC and the trajectory length, $\tau_{rel} = \tau_{convergence}/\tau_{sim}$, was calculated for each lipid type individually. Sterols were excluded from the analysis. The script that calculates the equilibration of lipids is available at `Scripts/BuildDatabank/NMRPCA_timerelax.py` in the NMRlipids databank repository. The resulting values are stored in files named `eq_times.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.10 Quality evaluation of C–H bond order parameters

As the first step to evaluate simulation qualities against experimental data, each simulation is connected to an experimental data if molar concentrations of all molecules are within $\pm 3$ percentage units, charged lipids have the same counterions, and temperatures are within $\pm 2$ K. For molar concentrations of water, the exact hydration level is considered only for systems with molar water to lipid ratio below 25, otherwise the systems are considered as fully hydrated. In practise, this is implemented by adding the path of the experimental data into the simulation README.yaml file using the `/Scripts/BuildDatabank/searchDATABANK.py` script in the NMRlipids databank repository.

The quality of each C–H bond order parameter is estimated by calculating the probability for a simulated value to locate within the error bars of the experimental value. Because conformational ensembles of individual lipids are assumed to be independent in a fluid lipid bilayer, $\frac{S_{CH}-\mu}{s/\sqrt{n}}$ has a Student's t-distribution with $n-1$ degrees of freedom and $\mu$ representing the real mean of the order parameter. The probability for an order parameter from simulation to locate within experimental error bars can be estimated from equation

$$P = f\left(\frac{S_{CH} - (S_{exp} + \Delta S_{exp})}{s/\sqrt{n}}\right) - f\left(\frac{S_{CH} - (S_{exp} - \Delta S_{exp})}{s/\sqrt{n}}\right), \tag{4}$$

where $f(t)$ is the Student's t-distribution, $n$ is the number of independent sample points for each C–H bond which equals the number of lipids in a simulation, $S_{CH}$ is the sample mean from Eq. 1, $s$ is the variance of $S_{CH}$ calculated over individual lipids, $S_{exp}$ is the experimental value, and $\Delta S_{exp}$ its error. The error of $\Delta S_{exp} = 0.02$ is currently assumed for all experimental order parameters[?], yet more accurate ones may be available in the future[?]. Because a lipid bilayer simulation contains at least dozens of lipids, the Student's t-distribution could be safely approximated with a normal distribution. However, the normal distribution gives probability values that are below the numerical accuracy of computers when simulation values are far from experiments. To avoid such numerical instabilities, we use the first order Student's t-distribution having slightly higher probabilities for values far away from the mean. On the other hand, some force fields exhibit too slow dynamics which leads to large error bars in order parameter values[?]. Such artificially slow dynamics widens the Student's t-distribution in Eq. 4, thereby increasing the probability to find the simulated value within experimental error bars. Therefore, the order parameters with simulation error bars above the experimental error 0.02 are not included in the quality evaluation.

To streamline the comparison between simulations, we define the qualities for different fragments (frag = *sn*-1, *sn*-2, headgroup or total referring to all order parameters within a molecule) within each lipid type in a simulation as

$$P^{frag}[\text{lipid}] = \langle P[\text{lipid}]\rangle_{frag} F_{frag}[\text{lipid}], \tag{5}$$

where $\langle P[\text{lipid}]\rangle_{frag}$ is the average of individual C-H bond order parameters qualities within the fragment, and $F_{frag}[\text{lipid}]$ is the percentage of order parameters for which the quality is available within the fragment. The overall quality of different fragments in a simulation are then defined as a molar fraction weighted average over different lipid components

$$P^{frag} = \sum_{lipid} \chi_{lipid} P^{frag}[\text{lipid}], \tag{6}$$

where $\chi_{lipid}$ is the molar fraction of a lipid in the bilayer.

The quality evaluation of order parameters is implemented in `/Scripts/BuildDatabank/QualityEvaluation.py` in the NMRlipids databank repository. The resulting qualities for each order parameters are stored in files named `[lipid_name]_OrderParameters_quality.json`, for individual lipids in files named `[lipid_name]_FragmentQuality.json`, and for overall quality for fragments in files named `system_quality.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.11 Quality evaluation of X-ray scattering form factors

Because experiments give form factors only in relative intensity scale, they should be scaled before comparing with the simulation data. Here we use the scaling coefficient for experimental intensities defined in the SIMtoEXP program[?]

$$k_e = \frac{\sum_{i=1}^{N_q} \frac{|F_s(q_i)||F_e(q_i)|}{(\Delta F_e(q_i))^2}}{\sum_{i=1}^{N_q} \frac{|F_e(q_i)|^2}{(\Delta F_e(q_i))^2}}, \tag{7}$$

where $F_s(q)$ and $F_e(q)$ are form factors from a simulation and experiment, respectively, $\Delta F_e(q)$ is the error of the experimental form factor, and summation goes over the experimentally available $N_q$ points.

Also, a quality measure based on differences in simulated and experimental form factors across available $q$-range is defined in the SIMtoEXP program[?]. However, the lobe heights in simulated form factors depend on the simulation box size as shown in Fig. **??**. Therefore, the quality measure defined in SIMtoEXP would also depend on the simulation box size. In contrast, locations of form factor minima are independent on simulation box size in Fig. **??**. Here we use the location of the first form factor minima for the quality evaluation because automatic detection of the location of second minima is inaccurate for some experimental data due to fluctuations, such as for the POPE data in Figs. 2 D and E. The first minimum correlates well with the thickness of a membrane (Fig. 2 F), although the correlation of the second minima would be even stronger (Fig. **??**). In practise, we first filter the fluctuations from the form factor data using Savitzky–Golay filter (window length 30 and polynomial order 1) and locate the first minima above 0.1 $Å^{-1}$ from both simulation and experimental data. The quality of a form factor is then defined as Euclidean distance between the minima in simulated and experimental form factors, $FF_q = |FF_{\text{min}}^{\text{sim}} - FF_{\text{min}}^{\text{exp}}|$.

The quality evaluation of form factors is implemented in `/Scripts/BuildDatabank/QualityEvaluation.py` in the NMRlipids databank repository. The resulting form factor qualities are stored in files named `FormFactorQuality.json` at folders in `/Data/Simulations` in the NMRlipids databank repository.

### 4.12 Calculation of lipid flip–flops

Flip–flop rates were calculated using `AssignLeaflets` and `FlipFlop` tools from LiPyphilic package[?]. Headgroup atoms of each molecule as defined in the mapping file were used to determine in which leaflet the molecule locates. The midplane cut-off defining the region between leaflets was 1 nm and frame cutoff was 100. This means that if the headgroup of a molecule entered within the distance of 1 nm from the bilayer midplane and was found from the opposing leaflet after 100 steps, this event was considered as a successful flip–flop event. The code that finds flip–flop events from all simulations in the NMRlipids databank is available at `scripts/FlipFlop.py` and the results at `Data/Flipflops/` in the repository at https://github.com/NMRLipids/DataBankManuscript/.

### 4.13 Analysing anisotropic diffusion of water in membrane environment from the NMRlipids databank

Water permeability through membranes was calculated from equation $P = r/2c_w$, where $r$ is the rate of permeation events per time and area, and $c_w$=33.3679 $(\text{nm})^{-3}$ is the concentration of water in bulk[?]. The number of permeation events in each trajectory was calculated using the code by Camilo et al.[?], available at https://github.com/crobertocamilo/MD-permeation. The code that calculates permeabilities for all simulations in the NMRlipids databank is available at `/scripts/calcMD-PERMEATION.py` and the resulting permeabilities are stored at `/Data/MD-PERMEATION` in the repository containing all analyses specific for this publication at https://github.com/NMRLipids/DataBankManuscript/. This repository is organized similarly to the NMRlipids databank repository, enabling the upcycling of also the analyzed data without overloading the main NMRlipids databank repository.

The lateral diffusion of water along the membrane surface, $D_{||}$, was calculated with Einstein's equation using `-lateral` option in `gmx msd` program within the Gromacs software package[?]. The code that calculates $D_{||}$ for water from all simulations in the NMRlipids databank is available at `/scripts/calcWATERdiffusion.py`, and the resulting diffusion coefficients are stored at `/Data/WATERdiffusion` in the repository at https://github.com/NMRLipids/DataBankManuscript/.

Water diffusion in the perpendicular direction of lipid bilayers in a multilamellar stack was estimated from the Tanner equation $D_{\text{perp}} = \frac{D_{||} P z_w}{D_{||} + P z_w}$[?,?], where the water layer thickness, $z_w$, was estimated by subtracting bilayer thickness from the size of the simulation box in membrane normal direction.

## Acknowledgements

P.B. was supported by the Academy of Finland (Grants 311031 and 342908). R.G.-F, A.P. and F.S.-L. thank Centro de Supercomputación de Galicia for computational support. R.G.-F. thanks Ministerio de Ciencia, Innovación y Universidades for

a ”Ramón y Cajal” contract (RYC-2016-20335), and Spanish Agencia Estatal de Investigación (AEI) and the ERDF (RTI2018-098795-A-I00, PDC2022-133402-I00), Xunta de Galicia and the ERDF (ED431F 2020/05, 02_IN606D_2022_2667887 and Centro singular de investigación de Galicia accreditation 2016-2019, ED431G/09). AP thanks Spanish Agencia Estatal de Investigación (AEI) and the ERDF (PID2019-111327GB-I00, PDC2022-133402-I00), Xunta de Galicia and the ERDF (ED431B 2022/36, 02_IN606D_2022_2667887). F. S.-L. thanks Axencia Galega de Innovación for his predoctoral contract (02_IN606D_2022_2667887), and Spanish Agencia Estatal de Investigación (AEI) and the ERDF (PID2019-111327GB-I00). M.S.M. was supported by the Trond Mohn Foundation BFS2017TMT01. O.H.S.O, A.M.K, and R.N. acknowledge CSC — IT Center for Science for computational resources and Academy of Finland (grants 315596 and 319902) for financial support.

## Author contributions statement

P.B. implemented the PCA-based equilibration metric, supervised the work of A.K., and participated in code refactoring. R.G.-F was responsible for the design, validation, supervision and funding of the back-end and front-end for the graphical user interface (GUI) (http://www.databank.nmrlipids.fi/). A.P. was responsible for the design, validation, supervision and funding of the back-end and front-end for the graphical user interface (GUI) (http://www.databank.nmrlipids.fi/). F.S.-L. developed the back-end and participated in the design and development and validation of the graphical user interface (GUI) (http://www.databank.nmrlipids.fi/). P.F.J.F. contributed to changes in the buildH software so that the Databank can handle united atom simulation data.

## Additional information