PCCP

Physical Chemistry Chemical Physics

Accepted Manuscript

Volume 19
Number 1
7 January 2017
Pages 1-896
PCCP
Physical Chemistry Chemical Physics
rsc.li/pccp
ISSN 1463-9076
ROYAL SOCIETY OF CHEMISTRY
PAPER
H.-P. Loock et al.
Determination of the thermal, oxidative and photochemical degradation rates of scintillator liquid by fluorescence EEM spectroscopy

View Article Online
DOI: 10.1039/D0CP02058C

PCCP

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

A machine learning study of the two states model for lipid bilayer phase transitions†

Vivien Walter,*[a] Céline Ruscher,[b] Olivier Benzerara,[c] Carlos M. Marques,[c] and Fabrice Thalmann**[c]

We have adapted a set of classification algorithms, also known as Machine Learning, to the identification of fluid and gel domains close to the main transition of dipalmitoyl-phosphatidylcholine (DPPC) bilayers. Using atomistic molecular dynamics conformations in the low and high temperature phases as learning sets, the algorithm was trained to categorise individual lipid configurations as fluid or gel, in relation with the usual two-states phenomenological description of the lipid melting transition. We demonstrate that our machine can learn and sort lipids according to their most likely state without prior assumption regarding the nature of the order parameter of the transition. Results from our machine learning approach provides strong support in favour of a two-states model approach of membrane fluidity.

1 Introduction

Phospholipid molecules play a major structural role in biological membranes[1,2] where a deep understanding of the physical properties can only be acquired from a detailed knowledge of the lipid assemblies. Thanks to their amphiphilic nature and geometrical characteristics, most phospholipid molecules spontaneously self-assemble in water as bilayers. Supported or free-standing lipid bilayers and vesicles can easily be made, controlled and studied, and have now become standard tools in membrane biophysics studies, referred as model lipid bilayer systems[3]. Early studies on pure phospholipid bilayers indicated that lipids were subject to thermodynamic transitions[2,4–6], with in particular a sharp transition associated to a significant change in enthalpy called *main*, or *melting* transition. This transition separates a low temperature well-packed assembly from a high temperature expanded, disordered lipid tail organisation. This transition is considered to be weakly first order, with significant pretransitional effects and an almost continuous variation of many of the membrane structural, thermodynamic and kinetic properties[7]. The low temperature region is commonly referred as the *gel* phase, while the high temperature region is the *fluid* phase. Lipid mixtures also display melting transitions spreading along a finite temperature range, usually accompanied by gel-fluid domain coexistence. It is usually assumed that most biological lipid membranes are found in a fluid phase, and many scenarii aiming at explaining the lateral lipid and protein segregation observed in biological membranes involve ordering of the lipid tails.

The consensual description of the single component lipid melting transition assumes that dominant molecular conformations evolve from all-*trans* extended, well oriented hydrocarbon chain conformations in the low temperature phase, to disordered chains melted by rotation isomerism, as proposed in the earliest theoretical proposals[8–11]. Lipids in the fluid phase have more configuration entropy and more enthalpy than those in the gel phase, due to lower density and cohesive energy, and higher chain torsion energy. Balance between entropy and enthalpy holds at the melting temperature T_m. Despite some asymmetry between the two phases, a large number of experimental facts related to melting transition of pure and mixed lipid compositions have been successfully interpreted by means of a phenomenological two-states model, originally proposed by Doniach[5,12–19]. This model can be expressed as an Ising model, each lipid taking binary discrete values (say $s = 0$ for *gel* and $s = 1$ for *fluid*) with neighbouring lipids being coupled[5]. In the framework of the two-states model, the Ising variables stand for a coarse-grained description of the lipid tail conformations, assuming that lipids can be classified into two classes, according to their molecular conformations. Within this description, an effective temperature dependent "magnetic field" $h(T)$ biases the odds in favour of one or the other state, while

[a] Department of Chemistry, King's College London, Britannia House, 7 Trinity Street, SE1 1DB, London, United Kingdom
* E-mail: vivien.walter@kcl.ac.uk
[b] Stewart Blusson Quantum Matter Institute, University of British Columbia, Vancouver BC V6T 1Z1, Canada
[c] Institut Charles Sadron, CNRS and University of Strasbourg, 23 rue du Loess, F-67034 Strasbourg, France
** E-mail: fabrice.thalmann@ics-cnrs.unistra.fr

† Electronic Supplementary Information (ESI) available: one *pdf* file, see DOI: 10.1039/cXCP00000x/. Sample MD trajectory files are available from the Zenodo repository, see DOI:10.5281/zenodo.3950029. Machine Learning analysis scripts can be obtained from the Git server github.com/vivien-walter/mllpa.

View Article Online
DOI: 10.1039/D0CP02058C

cooperativity results from nearest neighbour state coupling.

Usually, the determination of the lipid bilayer phase relies on a structural scalar order parameter, such as the membrane thickness, given for instance by the head to tail lipid extension or the tail molecular order parameter. We show in the present approach that Machine Learning approaches can also elegantly distinguish the gel and fluid lipid bilayer structures. Machine Learning has already been applied successfully to a number of situations in statistical thermodynamics and phase transitions. For instance, Cubuk *et al.* used support vector machines to localise plastic flow regions in amorphous structures[20], Carrasquilla and Melko revealed the strong aptitude of neural networks models to recognise various spin ordering regimes in condensed matter systems[21], Le and Tran succeeded in predicting the polymorphism of complex lipid mixtures given a set of structural, chemical and composition parameters, by means of an artificial neural network approach[22]. Very recently, Iyer et al. adapted support vector machines (SVM) to the determination of lipid domains in raft forming mixtures[23]. We address in this work the validity of the two-states description for the transition of pure DPPC (1,2-dipalmitoyl-*sn*-glycero-3-phosphatidylcholine) bilayers, using atomistic molecular dynamics (MD) simulations and supervised Machine Learning (ML) classification algorithms. Assessing the two states model requires ones to analyse single lipids and sort them into their respective states. Our ML classification works without reference to an existing or newly defined scalar order parameter. It only relies on a procedure for lifting the orientation degeneracy of each lipid configuration. In addition, the ML approach may serve as testing the relevance of a given scalar order parameter *a posteriori*.

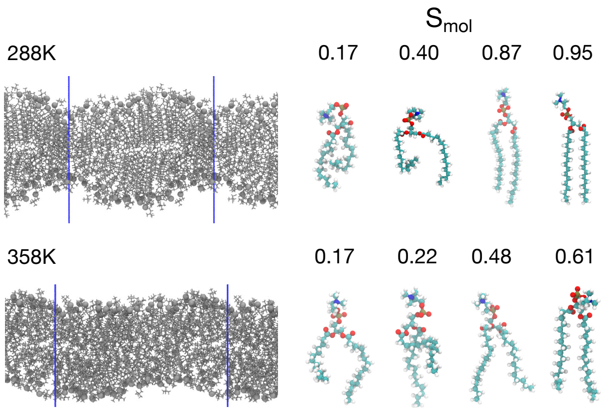


Fig. 1 Snapshots of a DPPC bilayer simulated at (top-left) 288 K and at (bottom-left) 358 K, respectively below and above the experimental T_m of the lipid. Lipid molecules are shown in gray, with their phosphorus atom displayed as a plain big sphere to distinguish the lipid orientation in the bilayer. Blue lines delimit the simulation box beyond which periodic boundary conditions are applied to the system. Right: lipids extracted from the bilayers displaying an important diversity in their conformations, with their associated molecular order parameter S_{mol}. Top-right: conformations at 288K, bottom right: conformations at 358K.

2 Simulated systems and Machine Learning analysis

DPPC molecules are among the best known phospholipids[6]. They display experimentally a melting transition at 314 K, which is well reproduced by the CHARMM-36/TIP3P force-field for atomistic simulations of lipid bilayers in aqueous solutions[24,25]. In our simulations, the bilayer was found in a L_α fluid phase above the main transition, and in a disordered or "ripple" gel phase (Figure 1), *i.e* a spatially modulated gel phase with peristaltic variations in the bilayer thickness and normal direction[26,27] at lower temperatures.

It is known experimentally that pure DPPC bilayers undergo a premelting transition at temperature $T_p = 307$ K. The pretransition calorimetric signature is an order of magnitude smaller than the one associated to the main transition. In the temperature interval between the pre- and the main transitions, the stable thermodynamic phase is expected to be a spatially modulated $P_{\beta'}$ ripple phase[2,6,26,28,29]. Below premelting, the thermodynamically stable phase is a tilted lamellar $L_{\beta'}$ gel phase. Such low temperature states are difficult to investigate using molecular dynamics, as they behave in practice as solid phases and equilibrate very slowly. Using the Charmm36 lipid-SPC water force field, Khakbaz et Klauda simulated small systems (64 lipids) and found a low temperature structure consistent with a $L_{\beta'}$ organisation[27]. Our simulations of larger systems (212 lipids) did not show evidence of such $L_{\beta'}$ state but display instead a noticeable corrugation of the membrane better consistent with a $P_{\beta'}$ lipid arrangement. In agreement with[27], a $L_{\beta'}$-like structure was indeed obtained for smaller systems (64 lipids, ESI†). The amplitude and period of the corrugation turns out to strongly depend on the simulation box size, and all the more visible than the simulated system becomes large. A precise study of the structure of membranes at low temperature is out of the scope of the current study and is devoted to an upcoming paper[30].

As the ML analysis shows below, our numerical low temperature gel structure display some amount of chain disorder, consistent with recent experimental findings on related phosphocholine bilayer systems[31]. We therefore assume that our structure is closer to the ripple than to the tilted gel phase, and the former appears to be numerically favoured in all the MD simulations performed in the present work. In what follows, we refer to this low temperature organisation as a **disordered gel state**, or simply gel state†.

The principle of the analysis is as follows. A 212 lipid molecules system was thermalised at low (288 K) and high temperature (358 K) and pressurised with a semi-isotropic barostat. At such temperatures, we assume that lipid conformations are predominantly gel and fluid respectively. Our training set was therefore composed of an equal number of conformations coming from the 288 K MD trajectory (all labelled as gel) and the 358 K trajectory (labelled as fluid). Details regarding MD simulations are given as ESI†.

Raw molecular conformations (spatial coordinates) were processed to remove all translation and rotation degeneracy, and the dimension of the initial conformation space was slightly reduced, resulting respectively in a 100-dimensional reduced coordinate space $\mathscr{X}$, and a 61-dimensional mutual distances space $\mathscr{D}$ (Figure 2 and Appendix). The coordinate space $\mathscr{X}$ and the distance space $\mathscr{D}$ provide a very detailed description of individual lipid conformations, and constitute the starting point of the ML ap-

View Article Online
DOI: 10.1039/D0CP02058C

proach. They are referred as "feature spaces" in the context of automated classification. The processed conformations were then fed to the ML algorithms.

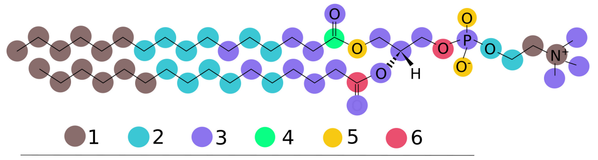


Fig. 2 DPPC molecular structure after removal of the hydrogen atoms. It contains 50 centers of forces associated to the "heavy" atom species C,O,N,P. Enumerating all atom pairs separated by 6 bonds along the molecular structure graph gives 61 different pairs. The number of pairs to which an atom belongs ranges between 1 and 6 (color code).

Three different algorithms were selected for the purpose of classifying the molecular conformations: Naive Bayes (NB), K-Nearest neighbours (KNN) and Support Vector Machines (SVM). They were all used as implemented in the *Python/Scikit-learn* package[32]. The different ML algorithms were tested and found to perform moderately well when used separately, some methods performing better for gel lipids, other methods for fluid lipids. The resulting scores shown in Figure 3 shows that the NB predictive capacity is independent from the phase of the lipid considered. KNN performs better for fluids than for gels. At the opposite, ”coordinates” $\mathscr{X}$-SVM and ”distances” $\mathscr{D}$-SVM seems to perform better in the gel phase.

We therefore decided to combine the predictions of the above models, retaining those who perform best in each phase. The following decision chain was implemented:

(1) if the 4 models agree on the same prediction for the state, this prediction is retained;

(2) if the $\mathscr{D}$-SVM algorithm predicts a gel state, the lipid configuration is assumed to be *gel*;

(3) if the $\mathscr{X}$-SVM algorithm predicts a fluid state, the lipid configuration is assumed to be *fluid*;

(4) if the NB algorithm predicts a gel state, the lipid configuration is assumed to be *gel*;

(5) if none of the above conditions have been met, the KNN algorithm makes the final decision on the configuration classification.

Combining the approaches together, we managed to get a success rate of 88% upon validation, *i.e.* using 80% of the training set configurations for learning, our best ML algorithm was able to assign 86% of the low temperature configurations to a gel state, and 91% of the high temperature configurations to a fluid state. After training, the ML model was used to analyse simulations at arbitrary temperatures. MD trajectories and ML scripts used in this work are respectively available in the Zenodo and GitHub repositories[33].

3 Results and discussion

One important prediction of the two states model is the presence of minor components within the majority state, under the form of "thermal excitations". A finite fractions of molecules tend to adopt a conformation different from their immediate environment, in spite of the presence of a "local field" biasing the statistics in favour of the dominant state. In that respect, the training set cannot be considered as containing only pure gel and fluid conformations. The presence of minor components is inherent to the presence of thermal excitations, and it does not seem possible to curate the training set without introducing further unwanted biases into the analysis. Our results show, however, that the training algorithm is not sensitive to the presence of a small fraction of non representative lipid conformations. In other words, the learning procedure was found to be robust so long as the training sets temperatures were chosen far apart from the melting transition.

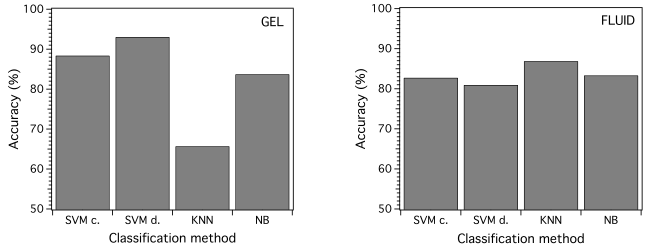


Fig. 3 Prediction scores of the Machine Learning classification methods for lipids in the gel phase (Left) and in the fluid phase (Right). Besides the Naive Bayes method (NB), all methods have an important asymmetry in their accuracy between each phase.

Let us first analyze how ML predictions differ from those based on standard scalar structural order parameters. Figure 4 shows the distribution of the molecular segmental order parameter S_{mol}, below and above the transition. This scalar observable was defined, for a given lipid and an instantaneous configuration (coming from a MD trajectory frame), by averaging over all the CC bonds in the two aliphatic tails a nematic parameter $(3\cos(\theta)^2-1)/2$, θ being the bond orientation with respect to the bilayer normal direction z. Both histograms overlap significantly. Using for instance a threshold value $S_{\mathrm{th}} = 0.508$, it was found that altogether 13% of lipids (1 in 8) were assigned to the opposite state, either gel at 358 K, or fluid at 288 K. The gel S_{mol} distribution appears to be strongly skewed, as 23% (1 in 4) of the lipid molecules ended up in the fluid state at 288 K. The relatively large fraction of lipids with $S_{\mathrm{mol}} \leq 0.508$ found at 288 K is in part due to the ripple structure of the bilayer. This clearly show that using S_{mol} for categorising individual lipid conformations gives noticeably different results. Figure 6 compares the ML and S_{mol} classifiers as a function of temperature. It was found that S_{mol} assigns 77% of the 288 K lipid configurations to the *gel* state (86% for ML) and 87% of 358 K lipid configurations to the *fluid* state (91% for ML). Based on a criterion consisting in minimizing the fraction of minor component in each state, ML outperforms S_{mol} by a small margin. Though their predictions differ at the level of single molecules, both methods seem consistent in terms of molecular averages.

A similar analysis was conducted for two other scalar determinants: the lipid head-to-tail extension L, and the area per lipid A_l computed from a Voronoi tessellation of the 2d xy projection of lipid center of mass positions (assigning to each lipid a polygonal cell of given area). In both cases, the low and high temperature

View Article Online
DOI: 10.1039/D0CP02058C

distributions were found to overlap significantly (ESI†).

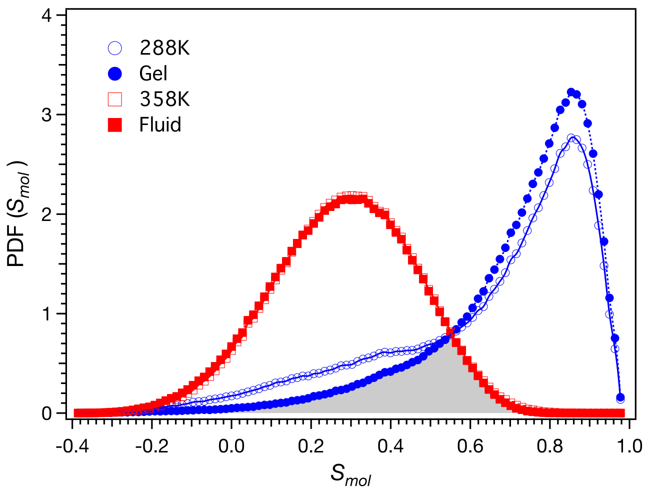


Fig. 4 Distribution of the molecular order parameters S_{mol} obtained from MD trajectories at 288 K (blue open circles) and 358 K (red open squares) respectively (using 10^6 lipid conformations). Distribution of molecular order parameters S_{mol} for lipids which have been classified as *gel* (blue full circles) and as *fluid* (red full squares) irrespective of the simulation temperature. The dashed black line shows the threshold value that best determine the internal lipid state based on S_{mol} values (according to maximum likelihood without *a priori* bias). The fraction of the population which would be incorrectly classified by using this scalar order parameter as compared with our Machine Learning approach is represented by the shaded area in gray.

To provide further evidence in favour of our approach, we compared the ML predictions to the membrane structure as a function of temperature. Figure 5 shows on the same graph the average area per lipid A_l, the average volume per lipid V_l, the molecular order parameter S_{mol} and the ML prediction for the fraction of lipid assigned to the fluid state. The area per lipid is defined as the projected membrane area divided by the number of lipids per leaflet, neglecting out-of-plane membrane undulations. The volume per lipid results from a Voronoi analysis of the lipid centre of masses. The structural values evolve monotonically and reversibly with temperature, from 288 to 358 K. The three curves superimposes very well, which shows that structural data supports the finding of our classification tool. Two snapshots of the membrane upper leaflet are also provided, at low and intermediate temperature. More snapshots are provided in Figure 7. The low temperature membrane shows a small number of isolated small clusters of fluid lipids. A one to one fluid/gel ratio is observed at 318 K (51 ± 3% of lipids in the fluid state), a temperature close to the experimental melting temperature (314 K). At high temperature (358 K) a few small dispersed clusters of lipid in gel state remain visible.

Snapshots of the upper leaflet at 318 K show two large distinct domains, separated by a smooth boundary. The fluid state ratio evolves smoothly and reversibly from an estimated value of 14 ± 2% at 288 K to 95 ± 1% at 358 K. The strong correlation between fluid ratio x_f and average area per lipid A_l can be naturally interpreted in the framework of the two-states model by assigning constant values A_f and A_g to the fluid and gel states, resulting in $A_l = x_f A_f + (1 - x_f) A_g$. A similar conclusion could be reached when considering the average volume per lipid V_l and the average molecular order parameter S_{mol}.

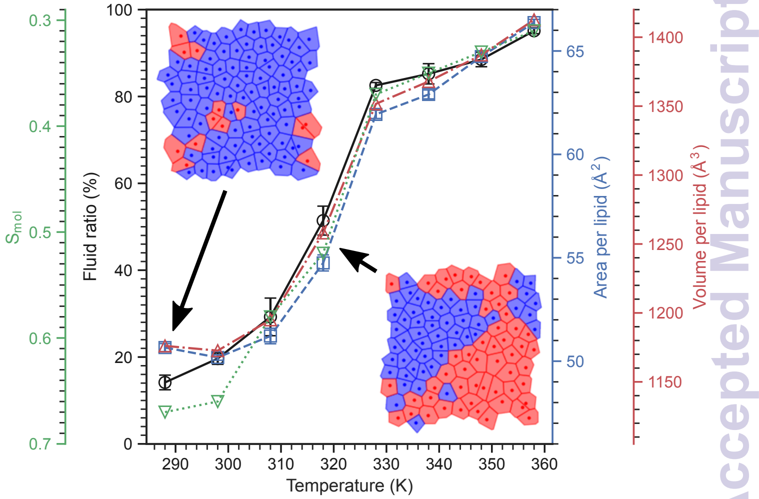


Fig. 5 Ratio of lipids in fluid state (black circles curve, vertical axis on the left), average area per lipid (blue squares curve, vertical axis on the right), average volume per lipid (red triangles up curve, vertical axis on the right) and average molecular order parameter (green triangles down curve, inverted vertical axis on the left) as functions of temperature. Two snapshots of the upper membrane leaflet, respectively taken at 288 and 318 K, are shown in the inset. The dark spots correspond to the 2d projection of the lipid center of masses. Boundaries between lipids result from the associated 2d Voronoi tessellation. Each cell colour corresponds to the assignment of the ML, blue for the gel state and red for the fluid state.

The ML predictions agree well with the global structural properties of the membrane and we now consider the local correlation properties. The two-states model combines the internal spontaneous dynamics of each lipid with local interactions promoting cooperativity between neighbours. The interactions, known as J coupling in the Ising model context, are responsible for the sharp structural and thermodynamic changes with temperature, the emergence of local correlations, clustering and domain formation, such as seen in Figures 5 and 7. Such couplings create a local field whose effect is to bias the lipid state in favour of the dominant local state, according to the majority rule. Internal reversible gel ↔ fluid state transitions occur spontaneously, according to a non conserved parameter, or Glauber dynamics[34]. In order to get insight into the local correlations and flip rates, we performed a systematic statistical neighbour analysis, and estimated the conditional conversion states.

Voronoi tessellations provide an operational method for deciding unambiguously which lipid pairs are nearest neighbours. Following a protocol described in the ESI†, we consider two consecutive MD frames and divide the lipid population into four subsets: (a) lipids categorised as fluid, which stay in the fluid state, (b) lipids categorised as fluid, switching to the gel state, (c) lipids in the gel state, remaining in the gel state and (d) lipids in the gel state switching to a fluid state. At $T = 318$ K the fluid and

View Article Online
DOI: 10.1039/D0CP02058C

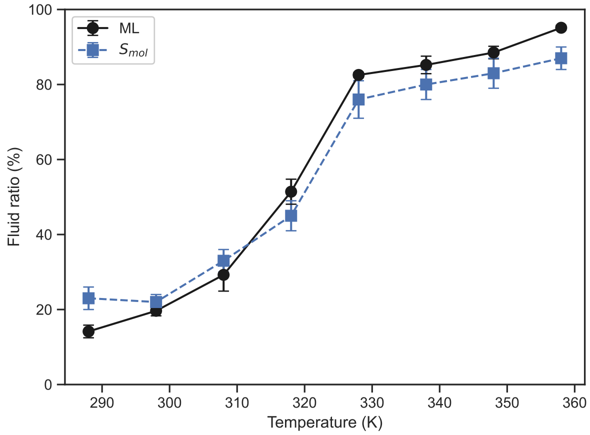


Fig. 6 Fluid ratio predictions (%) *vs* temperature for ML and S_{mol} based binary classifiers. The ML classifier discriminates better than S_{mol} both at high and low temperatures.

gel states compete evenly and the fluid and gel populations are roughly equal. We count for each molecule, the number n_g of gel neighbours and n_f of fluid neighbours. We find in these conditions that the most typical environment of a lipid in fluid state (a and b) is $n_g = 2 \pm 1$ and $n_f = 7 \pm 1$, while for a lipid in gel state (c and d) $n_g = 7 \pm 1$ and $n_f = 1 \pm 1$. However, the internal lipid state fluctuate spontaneously, and clearly the dynamics is strongly influenced by the local environment, as demonstrated in Figure 8. Indeed, for gel to fluid (case d) and fluid to gel (case b) transitions, we notice that all lipids that are just about to switch are more likely to have an equal number of gel and fluid neighbours (typically $n_g = 3 \pm 1$ and $n_f = 5 \pm 1$). We conclude that lipids subject to internal state transitions are mostly located at the border between domains. The results of our neighbour analysis clearly support the idea that internal state dynamics is under the control of some local field.

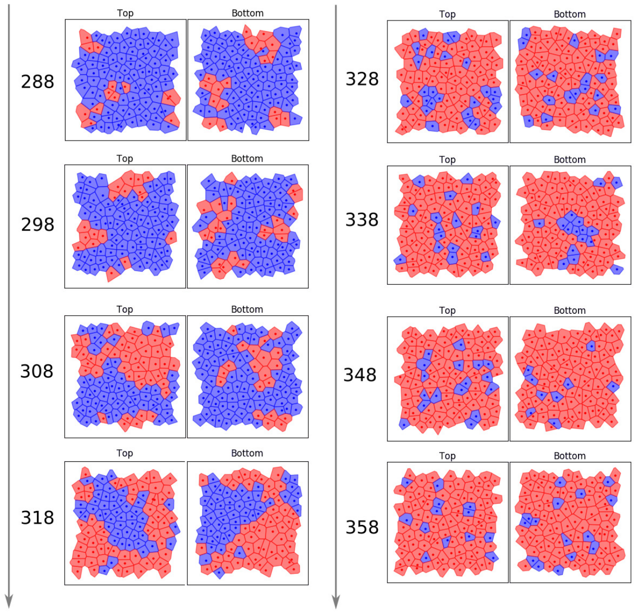


Fig. 7 Snapshots of the top and bottom bilayer leaflets for increasing 288, 298, 308, 318, 328, 338, 348 and 358 K temperatures. Each frame represents the 2d Voronoi cells of lipids in gel state (dark blue) and fluid state (light red) such as predicted by the ML approach.

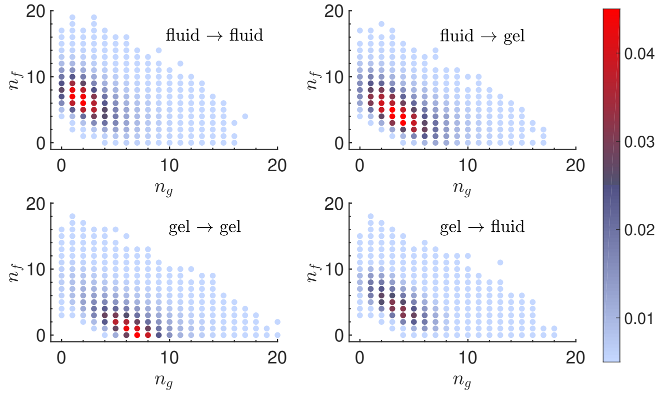


Fig. 8 Color coded histograms of the (n_l, n_g) distribution for the 4 subsets described in the text, at $T = 318$ K. Probability at $T = 318K$ for a lipid to have its local environment composed of n_g neighbours in gel state and n_f neighbours in fluid state. Left column informs about the local environment probability of lipids in the fluid state (Top) and gel state (Bottom) when no change in the state is observed. Right column shows the probability of the local environment just before lipids experience a transition to the other state.

4 Conclusion

As a summary, we trained a Machine Learning algorithm to classify phospholipid molecular conformations obtained by atomistic molecular dynamics simulations. Lipids were sorted into two classes, gel and fluid, according to their similarity with a reference (training) set of conformations originating from low and high temperature trajectories respectively. The efficiency of the ML approach is superior to simple schemes based on scalar order parameters as far as minimizing the fraction of minor component in each phase is concerned, and deals successfully with the ripple structure at low temperature. The measured fraction of fluid/gel conformations correlates very well with the observed structural changes as temperature evolves. A finite fraction of the minor phase is always present which can be associated to thermal excitations in the framework of a two-states Ising model interpretation. Lipids with similar state tend to cluster into large domains, while spontaneous internal state conversions are more likely to occur at the boundary between domains. The local distribution of neighbours supports the concept of local field and nearest neighbour coupling. We overall conclude that our ML approach provides convincing evidence in favour of the two-states phenomenological model.

We foresee numerous applications of the present approach. A first straightforward extension concerns lipid binary mixtures[35] and liquid ordered phases caused by the presence of cholesterol[36–39], as well as membranes of more complex composition[40,41]. We also anticipate that our ML approach will be useful to study the influence of membrane solutes that

View Article Online
DOI: 10.1039/D0CP02058C

are known to influence the thermodynamics of melting in model membranes. This includes hydrophobic pollutants, *e.g.* pyrene[42], carbohydrates[19,43], anaesthetic molecules[5,44] and synthetic oligomers[45,46]. A ML approach could then quantify the lipid state alteration induced by these compounds. Importantly, this analysis is also well-suited to study the local lipid environment of membrane proteins, for which the existence of lipid mediated interactions and minor phase nucleation is speculated[47–49]. For all these cases, significant improvements over approaches relying on scalar order parameters can be anticipated.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

V.W. warmly thanks Tiago Espinosa de Oliveira for helping with simulations set-up. The authors gratefully acknowledge support from the high performance cluster (HPC) Equip@Meso from the University of Strasbourg, through grant n° G2018A53.

Appendix: Machine Learning assignment of lipid states and training

Lipid classification procedure

The lipid classification process involves three steps:

1. the molecular conformation of a lipid is recorded as a list of atom positions. Lipids are then shifted and rotated in order to remove all rotation and translation degeneracy.

2. the lipid molecular conformations are further simplified, reducing each single lipid conformation to a set of 100 (r,z) coordinates (configuration space $\mathscr{X}$), or 61 mutual distances (configuration space $\mathscr{D}$).

3. 424 conformations were selected with the purpose of training the algorithms. A number of *Machine Learning* procedures were tested and compared. A combination of 4 algorithms was found to maximize the training success rate. We therefore combine the 4 algorithms in our final classification procedure.

Definition of the configuration spaces

Machine Learning algorithms share the ability to discriminate well multidimensional data. Our purpose is to feed the algorithms with single lipid conformations and have them sorted into two classes. We observe that the raw chemical formula of DPPC (CAS 2644-64-6) is $PNO_8C_{40}H_{80}$. Disregarding hydrogens, each lipid molecule comprises 50 "heavy atoms", and therefore the associated single lipid configurations belong to a 150 dimensions vector space. As some configurations can be mapped onto each other by means of a spatial displacement (translation and/or rotation), the set of configurations has only 144 independent degrees of freedom.

For simplicity, we decided to work with slightly smaller configuration spaces, by projecting further the 144 dimensional molecular conformations onto two configuration spaces, $\mathscr{X}$ and $\mathscr{D}$, that we now define. Within the first approach, one determines the proper inertial frame of each lipid configuration, locating the center of mass, fitting the position of all atoms with a 3D line to find the longest axis. Lipid configurations can be recast into the inertial frame, using the center of mass as origin, and the longest axis (smaller moment of inertia) as vertical axis. This almost always results in having the lipid directed along the bilayer normal, with two possible orientations. When necessary, the lipid is flipped in order to place by convention the phosphocholine group into the positive z upper half space. This combination of reorientations lifts entirely the translation and orientation degeneracy of the original coordinate space. New cylindrical coordinates $\{r_i, \theta_i, z_i\}$, $i = 1 \dots 50$ can be associated to the resulting lipid conformation. We decided then to disregard entirely the angles θ_i, keeping only the coordinate subset $\{X_i = r_i, z_i\}$, $i = 1 \dots 50$. This defines a $N_{\mathscr{X}} = 100$ dimensional space $\mathscr{X}$ that will subsequently be referred as "Coordinate space".

For the second approach, starting from the original 150 lipid coordinates, we calculate a set of mutual euclidean distances between pairs of atoms. Two atoms participate in a pair when they are separated by 6 positions along the chemical graph defining the molecule (see Figure 2). Enumerating the possibilities of the tree-like graph, one finds 61 non equivalent pairs of atoms, associated to 61 distances $\{D_j\}, j = 1 \dots 61$. This defines a $N_{\mathscr{D}} = 61$ dimensional "Distance space" $\mathscr{D}$, deprived of translation and rotation degeneracy.

In both cases, the original dimension of the problem is reduced (from 144 to 100 and 61 respectively) but yet the configuration spaces $\mathscr{X}$ and $\mathscr{D}$ are large and preserve to a large extent the complexity of the original conformations. As our comparisons show, the classification of lipid states is efficient, whether one starts from $\mathscr{X}$ or $\mathscr{D}$.

Three different algorithms commonly used for Machine Learning classifications were used in this Letter. A short description of these algorithms is given below.

Naive Bayes

Given a configuration space $\mathscr{D} = \{D_j\}, j = 1 \dots N_{\mathscr{D}}$, and two classes $s = \{\text{gel}, \text{fluid}\}$, the Bayesian approach assumes that it exists a joint probability distribution $P(s, \{D_j\})$, that weighs the respective likelihood of each state s once a configuration $\{D_j\}$ is provided. The Bayesian model makes a decision regarding the state s by comparing the conditional probability densities $P(\text{gel}|\{D_j\})$ and $P(\text{fluid}|\{D_j\})$. To be precise, the Bayesian approach attributes a *fluid* label to a configuration if the sign of

$$\ln\left(\frac{P(\text{fluid}|\{D_j\})}{P(\text{gel}|\{D_j\})}\right) \quad (1)$$

is positive, and a *gel* label otherwise.

Each Bayesian approach provides a mathematical model $P(\{D_j\}|s)$ describing the expected configuration distribution for a given class s: *fluid* and *gel*. There is in principle entire freedom in choosing the model, but the efficiency and the optimization requirements limit such choices in practice. Training a Bayesian model therefore amounts to finding the most realistic function $P(\{D_j\}|s)$ as far as classifying a given training set of data is concerned.

The Bayes theorem provides the connection between the con-

View Article Online
DOI: 10.1039/D0CP02058C

ditional probabilities entering in the choice function (1) and the model:

$$P(s|\{D_j\}) = \frac{P(\{D_j\}|s)P(s)}{P(\{D_j\})}, \qquad (2)$$

where $P(s)$ represents the *prior* distribution of s, *i.e.* the statistical distribution of s in the absence of any configuration related information, and $P(\{D_j\}) = P(\text{gel}, \{D_j\}) + P(\text{fluid}, \{D_j\})$ a normalization factor which cancels out in eq (1). The Naive Bayes (NB) gaussian model assumes that $P(\{D_j\}|s)$ factorizes as a product of independent gaussian distributions of D_j.

$$P(\{D_j\}|s) \prod_{j=1}^{N_{\mathscr{D}}} \mathrm{d}D_j = \prod_{j=1}^{N_{\mathscr{D}}} \frac{\mathrm{d}D_j}{\sqrt{2\pi}\sigma_{s,j}} \exp\left(-\frac{(D_j - D_{s,j})^2}{2\sigma_{s,j}^2}\right) \qquad (3)$$

Training the algorithm means finding the best mean value $D_{s,j}$ and standard deviation $\sigma_{s,j}$ for every parameter D_j in the distance space $\mathscr{D}$ and each class s. The number of parameters to compute turns out to be equal to twice the dimension of $\mathscr{D}$. In our case, the training set contains an equal number of gel and fluid conformations, and there is no *a priori* bias between classes, meaning that $P(\text{gel}) = P(\text{fluid}) = 1/2$. Therefore, the number of parameters to determine during training is $2 \times 61 = 122$.

To sum up, the Naive Bayes approach classifies a lipid conformation by computing a quadratic score function in the conformation space,

$$\sum_{j=1}^{N_{\mathscr{D}}} \frac{(D_j - D_{\text{fluid},j})^2}{2\sigma_{\text{fluid},j}^2} - \sum_{j=1}^{N_{\mathscr{D}}} \frac{(D_j - D_{\text{gel},j})}{2\sigma_{\text{gel},j}^2} + \text{Const} \qquad (4)$$

and deciding whether a $\{D_j\}$ lies closer to $\{D_{\text{fluid},j}\}$ or to $\{D_{\text{gel},j}\}$ according to this generalized distance.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classification method is based on defining a distance between any arbitrary pairs of objects to discriminate. A natural choice is the Euclidean norm of the feature space, here the $N_{\mathscr{X}}$ dimensional coordinate space $\mathscr{X}$.

The KNN algorithm finds the K nearest neighbors within the training set, of each new configuration to classify. Decision is taken based on majority rule, *i.e.* the most abundant class found among the K closest neighbors. The optimal K is determined during the training and validation process, and was set equal to 5, a typical value for this method.

Support Vector Machines

Support Vector Machines (SVM) classify data by means of linear separation in high dimensional representation spaces. Denoting $\boldsymbol{\phi}$ an arbitrary element in a given representation space $\mathscr{R}$, the binary decision is given by the sign of the affine expression $\boldsymbol{w} \cdot \boldsymbol{\phi} + b$, with b a numerical constant and $\boldsymbol{w}$ the hyperplane of separation normal vector. In a few favorable cases, it is possible to use directly the data definition space as representation space. However, in many practical situations, efficient classification can only be achieved by mapping the data (*e.g.* $\mathscr{X}$ or $\mathscr{D}$) onto a larger representation space, namely $\boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})$. Training a SVM corresponds to choosing a suitable representation space $\mathscr{R}$, and finding the optimal b and $\boldsymbol{w}$. As shown in[50,51], given a training set $\{\boldsymbol{x}_i\}, i = 1 \dots n$, the optimal $\boldsymbol{w}$ can always be expressed as a linear combination $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$ with either positive or vanishing α_i coefficients, and $y_i = \pm 1$, depending on the class (*fluid* 1, *gel* -1) of the corresponding vector data $\boldsymbol{x}_i$.

The subset of vectors $\{\boldsymbol{x}_i\}$ participating in the definition of $\boldsymbol{w}$ with non vanishing coefficients $\alpha_i > 0$ forms the so-called *support vectors*. Denoting $\mathscr{J}$ the sequence of indices of support vectors, and $\mathscr{K}(\boldsymbol{x}, \boldsymbol{x}')$ the product $\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$ in $\mathscr{R}$, called kernel function, the SVM decision function for any vector data $\boldsymbol{x}$ reads:

$$\text{sign}\left(b + \sum_{j \in \mathscr{J}} \alpha_j y_j \mathscr{K}(\boldsymbol{x}_j, \boldsymbol{x})\right). \qquad (5)$$

The SVM training optimization problem can therefore be formulated without any explicit reference to the representation space $\mathscr{R}$, nor the mapping $\boldsymbol{\phi}(x)$. It only requires an explicit positive kernel function $\mathscr{K}(\boldsymbol{x}, \boldsymbol{x}')$. As explained in[50,51], there are efficient quadratic optimization algorithms for determining the support vectors $\boldsymbol{x}_j$, the non-vanishing coefficients α_j and the shift constant b.

In this study, we used the standard radial basis kernel function

$$\mathscr{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2\right), \qquad (6)$$

with a default value γ equal to the inverse of the dimension of the configuration space. This choice assumes that each component of $\boldsymbol{x}$ is of order 1. When using the coordinate space $\mathscr{X}$, the trained SVM ends up using 297 non vanishing support vectors and coefficients α_j, $\gamma = 1/100$, $b = -0.416$ (167 *fluid*, 130 *gel* support vectors). When considering the distance space $\mathscr{D}$, the trained SVM used 161 non vanishing coefficients α_j, $\gamma = 1/61$ and $b = 0.134$ (100 *fluid*, 61 *gel* support vectors).

Training

The Machine Learning analysis performed in this Letter were conducted using the `Scikit-Learn` module (version *0.19*) for Python 3[32]. An unbiased selection of lipid conformations extracted from a trajectory at low temperature (288 K) was part of the training set, with a label *gel*. Similarly, an unbiased selection of conformations from a trajectory at high temperature (358 K) was added to the training set with a label *fluid*. The number of gel and fluid conformations were in equal number in the training set. As customary, 20 % of conformations were removed from the training sets, and used for verification and scoring purposes. The training set therefore consists of a sequence of 370 vectors (elements of the configuration spaces $\mathscr{X}$ or $\mathscr{D}$), used for building the prediction model.

Asserting the predictive capacity of each model

Once defined the training set, with properly labelled gel and fluid states, 20% of the lipids in each phase were taken apart for forming a validation set. After training the models (*i.e.* optimizing the parameters with respect to the 80% remaining conformations) a prediction score was separately calculated for the *gel* and *fluid* conformations in the validation set. The overall procedure was repeated 10 times, each time with the same training set, but independently drawn validation subsets. The average prediction scores are those shown in Figure 3.

Notes and references

1 O. G. Mouritsen, *Life-as a matter of fat: the emerging science of lipidomics*, Springer, 2005.

2 G. Cevc and D. Marsh, *Phospholipid Bilayers. Physical Principles and Models*, John Wiley & Sons, New-York, 1987.

3 *The Giant Vesicle Book*, ed. R. Dimova and C. M. Marques, CRC Press, Taylor and Francis, 1st edn, 2019.

4 S. Mabrey and J. M. Sturtevant, *Proceedings of the Natural Academy of Sciences USA*, 1976, **73**, 3862–3866.

5 T. Heimburg, *Thermal Biophysics of Membranes*, Wiley-VCH, 2007.

6 D. Marsh, *Handbook of Lipid Bilayers*, CRC Press, Boca Raton, 2nd edn, 2013.

7 O. Mouritsen, *Chemistry and Physics of Lipids*, 1991, **57**, 179 – 194.

8 J. Nagle, *Journal of Chemical Physics*, 1973, **58**, 252.

9 D. Marsh, *Journal of Membrane Biology*, 1974, **18**, 145–162.

10 S. Marcelјà, *Biochim. Biophys. Acta*, 1974, **367**, 165–176.

11 D. A. Pink and D. Chapman, *Proceedings of the National Academy of Sciences of the United States of America*, 1979, **76**, 1542–1546.

12 S. Doniach, *The Journal of Chemical Physics*, 1978, **68**, 4912–4916.

13 I. P. Sugàr and G. Monticelli, *Biophysical Chemistry*, 1983, **18**, 281–289.

14 T. Heimburg and R. Biltonen, *Biophysical Journal*, 1996, **70**, 84 – 96.

15 R. Jerala, P. F. Almeida and R. Biltonen, *Biophysical Journal*, 1996, **71**, 609–615.

16 I. P. Sugàr, T. E. Thompson and R. L. Biltonen, *Biophysical Journal*, 1999, **76**, 2099–2110.

17 V. P. Ivanova and T. Heimburg, *Physical Review E*, 2001, **63**, 041914.

18 J. Wolff, C. M. Marques and F. Thalmann, *Physical Review Letters*, 2011, **106**, 128104.

19 M. I. Morandi, M. Sommer, M. Kluzek, F. Thalmann, A. P. Schroder and C. M. Marques, *Biophysical Journal*, 2018, **114**, 2165 – 2173.

20 E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras and A. J. Liu, *Physical Review Letters*, 2015, **114**, 108001.

21 J. Carrasquilla and R. G. Melko, *Nature Physics*, 2017, **13**, 431.

22 T. C. Le and N. Tran, *ACS Applied Nano Materials*, 2019, **2**, 1637–1647.

23 S. S. Iyer, A. Negi and A. Srivastava, *Journal of Chemical Theory and Computation*, 2020, **16**, 2736–2750.

24 J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell and R. W. Pastor, *The Journal of Physical Chemistry B*, 2010, **114**, 7830–7843.

25 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell Jr., *Journal of Chemical Theory and Computation*, 2012, **8**, 3257–3273.

26 A. H. de Vries, S. Yefimov, A. E. Mark and S. J. Marrink, *Proceedings of the National Academy of Sciences*, 2005, **102**, 5392–5396.

27 P. Khakbaz and J. B. Klauda, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2018, **1860**, 1489 – 1501.

28 T. Heimburg, *Biophysical Journal*, 2000, **78**, 1154–1165.

29 O. Lenz and F. Schmid, *Phys. Rev. Lett.*, 2007, **98**, 058104.

30 V. Walter, C. Ruscher, O. Benzerara, C. M. Marques and F. Thalmann, *In preparation*.

31 K. Akabori and J. F. Nagle, *Soft Matter*, 2015, **11**, 918–926.

32 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

33 *Zenodo repository* doi:10.5281/zenodo.3950029
GitHub repository https://github.com/vivien-walter/mllpa.

34 R. J. Glauber, *Journal of Mathematical Physics*, 1963, **4**, 294–307.

35 J. H. Ipsen and O. G. Mouritsen, *Biochim. Biophys. Acta*, 1988, **944**, 121–134.

36 J. H. Ipsen, G. Karlström, O. Mouritsen, H. Wennerström and M. Zuckermann, *Biochimica Biophysica Acta*, 1987, **905**, 162–172.

37 D. Marsh, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2010, **1798**, 688–699.

38 A. J. Sodt, M. L. Sandar, K. Gawrisch, R. W. Pastor and E. Lyman, *Journal of the American Chemical Society*, 2014, **136**, 725–732.

39 M. Javanainen, H. Martinez-Seara and I. Vattulainen, *Scientific Reports*, 2017, **7**, year.

40 D. Marsh, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2009, **1788**, 2114–2123.

41 H. I. Ingòlfsson, M. N. Melo, F. J. van Eerden, C. Arnarez, C. A. Lopez, T. A. Wassenaar, X. Periole, A. H. de Vries, D. P. Tieleman and S. J. Marrink, *Journal of the American Chemical Society (JACS)*, 2014, **136**, 14554–14559.

42 M. D. Fraňovà, I. Vattulainen and O. S. Ollila, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2014, **1838**, 1406 – 1411.

43 L. M. Crowe and J. H. Crowe, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1991, **1064**, 267 – 274.

44 R. S. Cantor, *Biochemistry*, 1997, **36**, 2339.

45 G. Rossi, J. Barnoud and L. Monticelli, *Journal of Physical Chemistry Letters*, 2014, **5**, 241–246.

46 G. Rossi and L. Monticelli, *Journal of Physics: Condensed Matter*, 2014, **26**, 503101.

47 S. Marčelja, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1976, **455**, 1–7.

48 T. Gil, M. Sabra, J. Ipsen and O. Mouritsen, *Biophysical Jour-*

View Article Online
DOI: 10.1039/D0CP02058C

nal, 1997, **73**, 1728 – 1741.

49 T. Gil, J. H. Ipsen, O. G, M. C. Sabra, M. M. Sperotto and M. J. Zuckermann, *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes*, 1998, **1376**, 245 – 266.

50 B. E. Boser, I. M. Guyon and V. N. Vapnik, COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

51 C. Cortes and V. Vapnik, *Machine Learning*, 1995, **20**, 273–297.

A machine learning study of the two states model for lipid bilayer phase transitions
- Electronic Supplementary Informations -

Vivien WALTER
Department of Chemistry
King's College London
Britannia House, 7 Trinity Street, London, United Kingdom

Céline RUSCHER
Stewart Blusson Quantum Matter Institute
University of British Columbia
Vancouver BC V6T 1Z1, Canada

Olivier BENZERARA, Carlos M MARQUES, and Fabrice THALMANN*
Institut Charles Sadron
CNRS and University of Strasbourg
23 rue du Loess, F-67034 Strasbourg, France

(Dated: June 30, 2020)

* fabrice.thalmann@ics-cnrs.unistra.fr

I. SIMULATIONS

A. Simulation conditions

All simulations were performed using GROMACS 2016.4 [1, 2] along with the CHARMM-36 all-atom force-field [3] (June 2015 version). A lipid bilayer made of 106 lipid molecules per leaflet, each containing 130 explicit atoms, was created using CHARMM-GUI [4–7]. It was hydrated with two 8 nm thick water layers on each side (connected through periodic boundary conditions), using the TIP3P water model, for a total of 29826 solvent molecules. The force field parameters for DPPC molecules were provided directly by CHARMM-GUI [8, 9].

The above system was first subject to energy relaxation using steepest descent energy minimization, followed by a 10 ps NVT thermalization stage at 288 K. Then, the bilayer was subject to a 1 ns NPT equilibration run coupled to a semi-isotropic barostat (1 bar in all directions). The system was then further equilibrated at the desired temperature with the same semi-isotropic barostat during a second NPT equilibration step of 10 ns. Molecular dynamics production runs of 50 ns were finally generated at the same temperature and with the same semi-isotropic barostat. The analysis were performed on the last 25 ns of simulations. All time steps were set to 2 fs.

All the molecular dynamics simulations used the leap-frog integration algorithm [10]. Temperature and pressure were kept constant using respectively a Nosé-Hoover thermostat [11, 12] (correlation time $\tau_T = 0.4$ ps) and a Parrinello-Rahman semi-isotropic barostat [13, 14] (correlation time $\tau_P = 2.0$ ps, compressibility 4.5×10^{-5} bar^{-1}).

Lipid and water molecules were separately coupled to the thermostat. Following GROMACS recommendations for the CHARMM-36 all-atom force field, a Verlet cut-off scheme on grid cells was used with a distance of 1.2 nm, and non-bonded interactions cut-offs (Van der Waals and Coulombic) were also set to 1.2 nm. Fast smooth Particle-Mesh Ewald electrostatics was selected for handling the Coulombic interactions, with a grid spacing of 4 nm. A standard cut-off scheme with a force-switch smooth modifier at 1.0 nm was applied to the Van der Waals interactions. We did not account for long range energy and pressure corrections, and constrained all the hydrogen bonds of the system using the LINCS algorithm.

B. Nature of the gel phase

Whenever a large system (212 lipids or more) was simulated at low temperature, either starting from a molecular builder configuration (Charmm-gui), or resulting from the annealing of a high temperature configuration, a structure showing a longitudinal corrugation in the x and y direction was obtained (Fig. S1 right). As many other authors before, we think that this structure could actually be reminiscent of the experimental P$_{\beta'}$ DPPC ripple phase [15–17], see also [18]. In this work we refer to this corrugated phase as a **disordered gel phase** to distinguish it from the flat, tilted chains, $L_{\beta'}$ structure.

On the other hand, simulations of small systems made of 64 lipids each show much less corrugation, and looks closer to a flat tilted L$_{\beta'}$ [19] gel phase. The same holds if the 64 lipids system originates from a slow cooling of the high temperature phase. Figure S2 summarizes the various pathways leading to either a disordered, or a flat tilted structure.

The stability of the disordered gel phase was challenged by putting the system in contact with an anisotropic barostat (3 independent axis, same pressure) for 50 ns at 288 K in order to remove any box induced residual stress. The ripple phase was not perturbed except but a 6% change in the box lateral size. For all practical purposes, the bilayer behaves mechanically as a solid (fluid bilayers display large box size fluctuations when subject to an anisotropic barostat).

In addition, the ripple phase was put under tension, both under anisotropic and semi-isotropic barostat conditions, imposing a 10 mN/m stretching condition during 50 ns at 288 K. The longitudinal instability persisted and no tilted L$_{\beta'}$ emerged. The stresses and box sizes obtained at low temperature (288 K) are given in Table I.

The outcome of these "stress-tests" was that the disordered gel phase shows robustness and metastability (*e.g.* apparent stability without evidence of thermodynamic stability). Meanwhile, it is possible to duplicate a 64 lipids tilted flat configuration, and simulate it at low temperature (Fig. S2). The resulting 256 lipids solid phase also showed (meta)stability within accessible simulation times. However, once this system was heated and melted, the flat tilted configuration could not be recovered under quenching, or cooling. Only the disordered gel structure seems to be spontaneously favored upon system cooling, and reversible temperature cycling.

We therefore considered that the disordered "ripple-like" gel phase was indeed our low temperature reference phase, and performed the training and analysis on it.

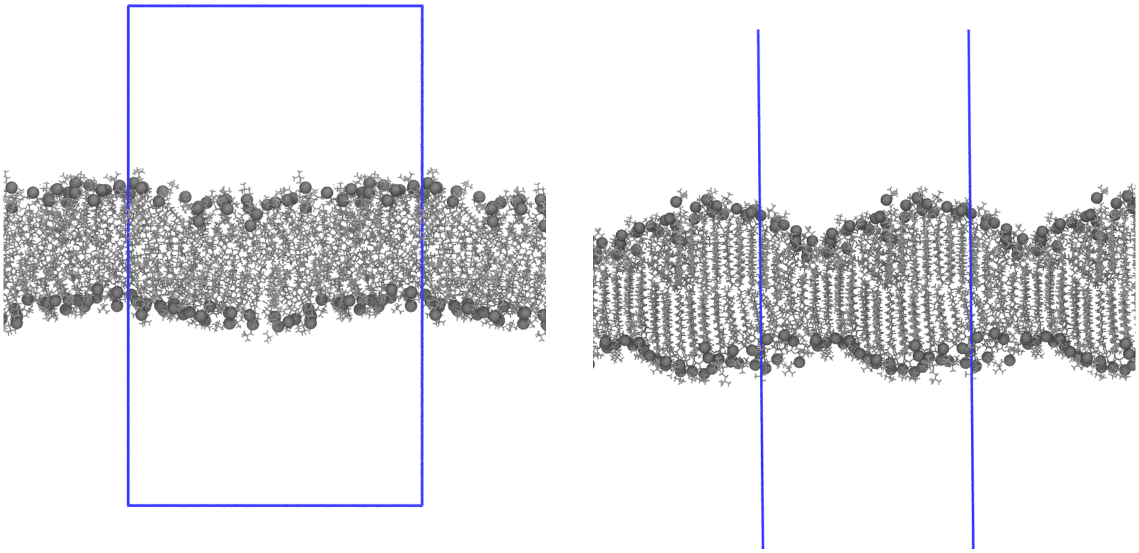


FIG. S1. Snapshots of bilayer configurations in the fluid (left) and disordered gel (right) phase, with periodic boundary conditions (box).

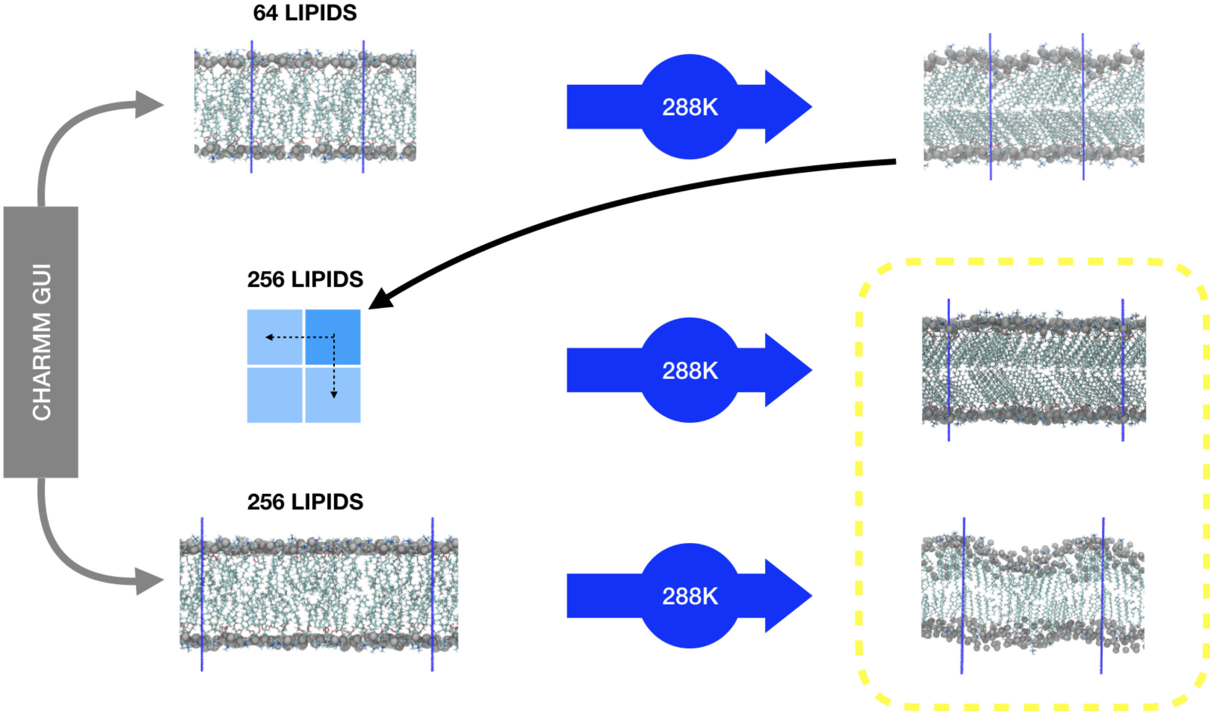


FIG. S2. Synopsis of the conditions allowing for the emergence of a disordered gel (large system, bottom row) or tilted gel (small system 64 lipids, top row and replication of the small system, middle row). The tilted lipid structures does not show up spontaneously if the number of lipids is larger than 64, and can only be found by replicating a small system.

II. SYSTEM ANALYSIS

A. Determination of structural parameters

Values of the average area per lipid A_l and order parameter S_{mol} of the bilayer were respectively obtained using the GROMACS built-in commands `gmx energy` and `gmx order`.

For measurements of individual lipid properties (area, order parameter, elongation), atom positions were collected from trajectories using the Python 3 `MDAnalysis` module [20, 21]. The individual areas per lipid were obtained from Voronoi tessellations of the two-dimensional projections of the lipid center of masses, computed using the `Voro++` library [22]. The individual volumes per lipid were derived from three-dimensional tessellations of the lipid centers of mass, again using the `Voro++` library. Note that the bilayer geometry requires a specific tessellation procedure: this was done by introducing *ghosts* lipids in the water regions. Without these ghost lipids, the tessellation cells cannot be correctly defined and are unbounded across the membrane-water interface, thus overestimating significantly the individual volume per lipid. Ghosts lipids are mirror images of bilayer lipids across the local lipid-water interface (*cf.* Fig. S3). After the tessellation was made, ghost lipids described in the previous section and their corresponding cells were removed the lists, and only the volumes of physical lipids were collected and analyzed.

The molecular order parameter S_{mol} of individual lipids was calculated by measuring, for every $N_C - 2 = 14$ non-terminal carbon atoms $k = 2 \dots 15$ of the 2 tails of the lipids, the angle formed between the z-axis of the system directed along $\vec{u}_z$ and the vector $\overrightarrow{C_{(j,k-1)}C_{(j,k+1)}}$ defined by the carbon atoms surrounding atom k within the same tail $j = 1, 2$. The order parameter $S_{\mathrm{mol},(j,k)}$ of the atom k is obtained from the 2^{nd} Legendre polynomial P_2 using $\cos(\theta_{(j,k)}) = \vec{u}_z \cdot \overrightarrow{C_{(j,k-1)}C_{(j,k+1)}}$, and averaging over j and k:

$$S_{\mathrm{mol}} = \frac{1}{2(N_C - 2)} \sum_{j=1}^{2} \sum_{k=2}^{N_C - 1} \frac{1}{2} \left(3\cos(\theta_{(j,k)})^2 - 1\right) \quad (1)$$

B. Next-nearest neighbors statistics

After completion of the 3d Voronoi tessellation using `Voro++`, a list of next-nearest neighbors was established for each lipid center of mass. We also collected the areas of the polygonal surfaces separating each pair of neighboring Voronoi cells. The ghost lipids and their corresponding faces were removed from the lists. The neighbor lists were further curated by removing all the neighbor pairs for which the corresponding face area accounted for less than 1% of the total surface area of each Voronoi cell in contact. The number of next-nearest neighbors were finally counted to build the coordination statistics (n_g, n_f), where each lipid molecule has n_g gel and n_f fluid neighbors.

III. COMPARISON BETWEEN MACHINE LEARNING DECISIONS AND STRUCTURAL CHARACTERIZATIONS OF THE LIPID CONFIGURATIONS

Machine Learning predictions were compared to two typical lipid structural properties: the carbon carbon (CC) order parameter S_{mol} along the chains and the area per lipid A in the 2d Voronoi tessellation of the lipid projected centers of mass. The corresponding results are presented in Fig. S4(a) and (b).

The order parameter curves (Fig. S4(a)) clearly discriminate among the low temperature (288 K) and high temperature (358 K) fluid phases, in agreement with published results on these systems [23].

Simulation	Stress xx	Stress yy	Box x	Box y
(1)	-6	8	7.2	7.2
(2)	0.8	1	7.0	7.4
(3)	-21	-18	7.5	7.5
(4)	-20	-20	7.4	7.6

TABLE I. Mechanical resistance of the low temperature phase (288 K). Simulation conditions: (1) tensionless semi-isotropic barostat, (2) tensionless anisotropic barostat, (3) under tension semi-isotropic barostat, (4) under tension anisotropic barostat. Stress xx: virial stress in the x direction in bars. Stress yy: virial stress in the y direction. Box x: lateral x box size in nm. Box y: lateral y box size.

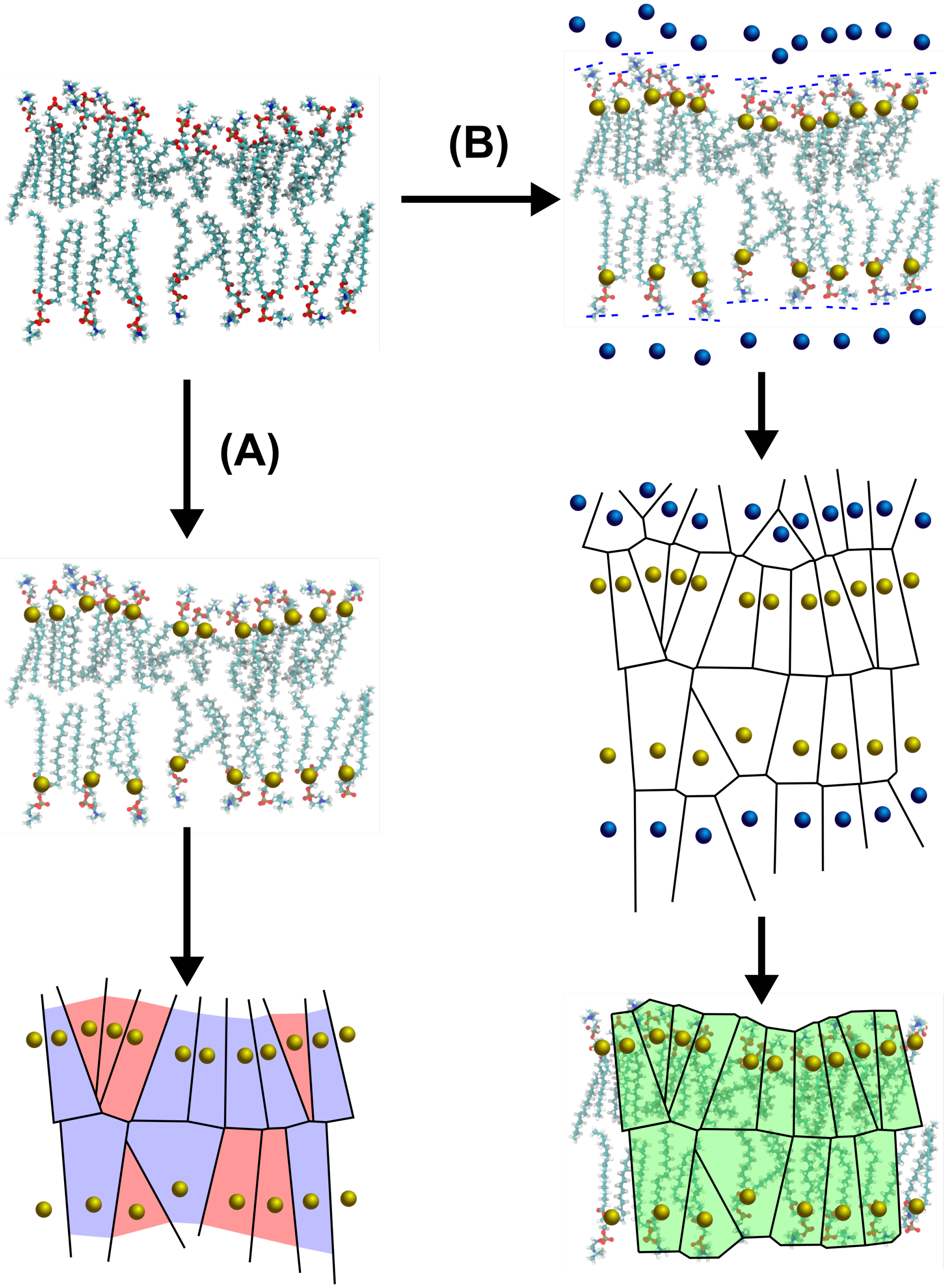
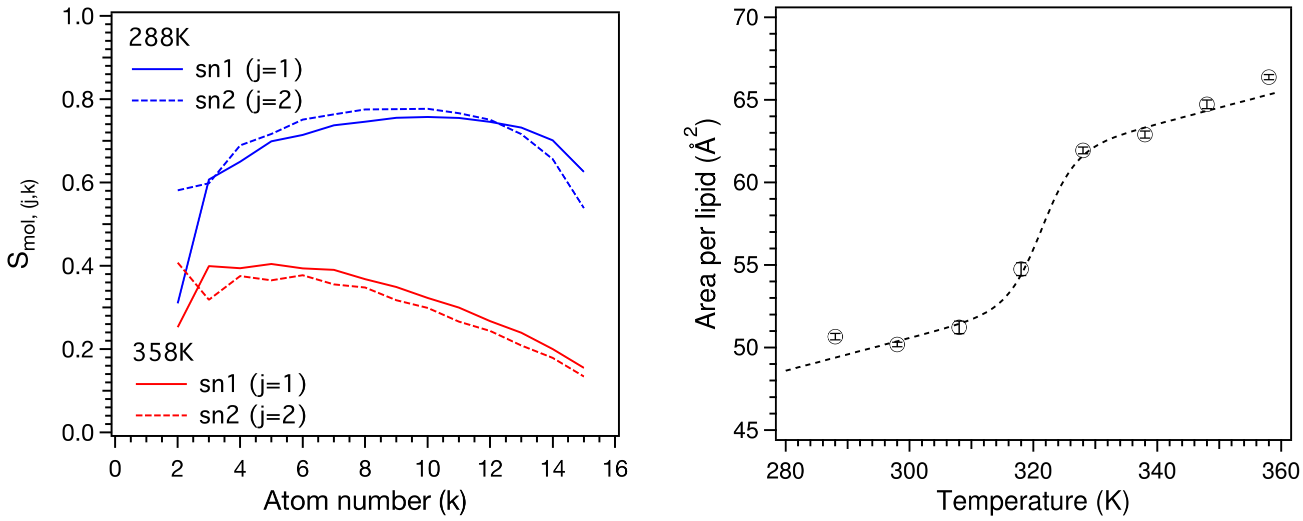


FIG. S3. Comparison of 3-dimensional Voronoi tessellations of a lipid bilayer configuration (A) without and (B) with ghost lipids. Without ghost lipids, most cells are unbounded, with infinite volume, due to the absence of particle on the opposite side of the water-membrane interface. As a practical solution of this problem, ghost lipids are added to the data set, mirroring the lipid center of mass positions. The resulting cells for lipids inside the bilayer display a realistic volume and shape, accounting for the water interface in a natural way.

The order parameter of the atoms in the lipid tails at low (288 K) and high (358 K) temperature is characteristic from membranes in the gel and fluid phases respectively [19]. The phase transition can be clearly seen in Fig. S4(b) as a significant variation in the evolution of the area per lipid A_l around 321 K.

Experimental structural values are available at 323 K [24–27]. Nagle *et al.* obtained for DPPC an area per lipid equals to 64 ± 1 Å^2 significantly close to the value *circa* 60 Å^2 we obtained in our simulations. Using the average DPPC bilayer thickness reported by Nagle *et al.*, we could estimate an experimental volume per lipid of 1220 ± 50 Å^3,



(a)

(b)

FIG. S4. Confirmation of the thermodynamic phase of the bilayer using two common structural parameters: (a) the order parameter $S_{\text{mol},(j,k)}$ with k the atom number, $j = 1$ (tail sn1) or $j = 2$ (tail sn2), and (b) the area per lipid A_l. (Left) the average order parameter is shown as a function of the carbon atom index along the chain (from glycerol to terminal end), for each *sn1* and *sn2* chain. (Right) the phase transition can be clearly seen in the evolution of the area per lipid as a function of temperature. A sigmoid fit points to a transition temperature T_m equal to 321 K in our system.

which agrees fairly with our Voronoi value of 1300 Å^3

IV. NAIVE CLASSIFICATIONS

The distributions of the areas per lipid and molecular elongations at low and high temperatures are shown in Fig. S5. Using a naive classification scheme based on a single threshold value for either of the two previous scalar parameters would at best result in a prediction accuracy of respectively 69% and 67%.

Fig. S6 compares the histogram of molecular order parameters S_{mol} as a function of the temperature of the lipid bilayer from which the configurations are extracted (288 K or 358 K), and as a function of the result of the Machine Learning classification procedure. The difference between the distribution at 288 K and the distribution in the *a posteriori* gel state ensemble indicates that a small fraction of lipids in the fluid state are already present at 288 K.

[1] H Berendsen, D van der Spoel, and R van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 1995.

[2] M J Abraham, T Murtola, R Schulz, S Páll, J C Smith, B Hess, and E Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 2015.

[3] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain khi1 and khi2 dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

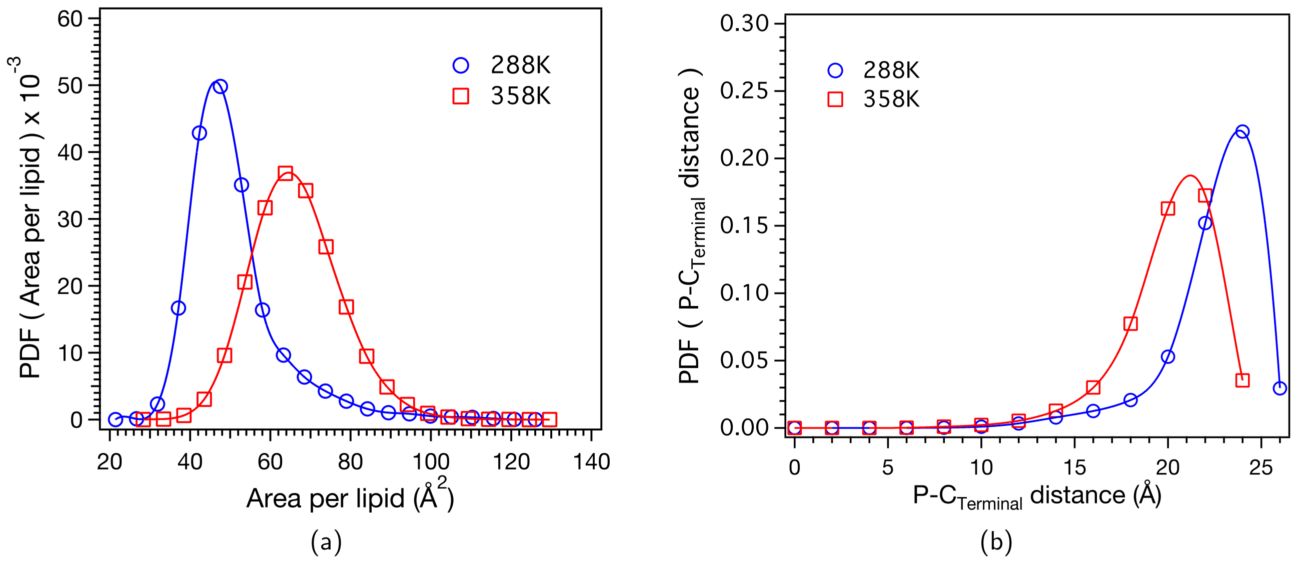
[4] S. Jo, T. Kim, V. G. Iyer, and W. Im. Charmm-gui: A web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29:1859–1865, 2008.

[5] E. L. Wu, X. Cheng, S. Jo, H. Rui, H. K. Song, E. M. Davila-Contreras, Y. Qi, J. Lee, V. Monje-Galvan, R. M. Venable, J. B. Klauda, and W. Im. Charmm-gui membrane builder toward realistic biological membrane simulations. *Journal of Chemical Theory and Computation*, 35:1997–2004, 2014.

[6] S. Jo, J. B. Lim, J. B. Klauda, and W. Im. Charmm-gui membrane builder for mixed bilayers and its application to yeast membranes. *Biophysical Journal*, 97:50–58, 2009.

[7] S. Jo, T. Kim, and W. Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS ONE*, 2(9):880, 2007.

[8] B. R. Brooks, C. L. Brooks III, A. D. MacKerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels,



(a)

(b)

FIG. S5. Distributions of the area per lipid (Left) and the average elongation between phosphorus and *sn1* terminal carbon atoms (Right) from lipids conformations at 288 and 358 K.

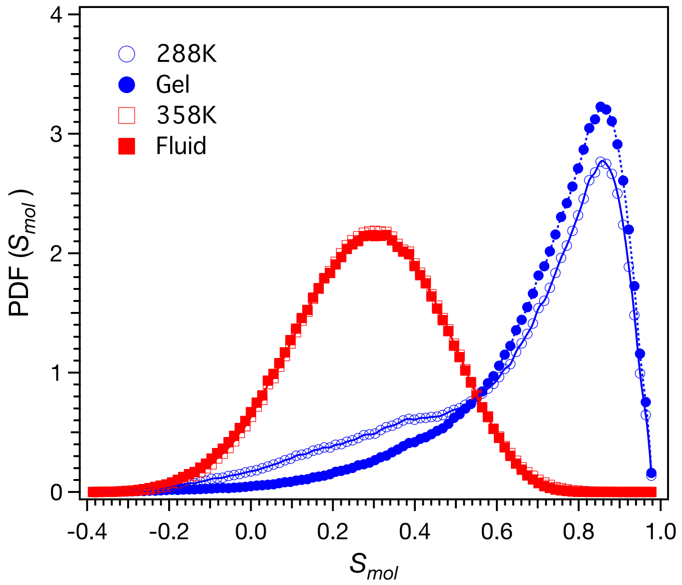


FIG. S6. Histograms of the molecular order parameter $S_{\rm mol}$ of lipids sorted by temperature (blue circles) and state ML classification (red squares).

S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30:1545–1614, 2009.

[9] J. Lee, X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi, S. Jo, V. S. Pande, D. A. Case, C. L. Brooks III, A. D. MacKerell Jr, J. B. Klauda, and W. Im. Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *Journal of Chemical Theory and Computation*, 12(1):405–413, 2016.

[10] R. W. Hockney, S. P. Goel, and J. W. Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.

[11] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.

[12] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31:1695, 1985.

[13] S. Nose and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50:1055–1076, 1983.

[14] M. Parinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52:7182, 1998.

[15] Alex H. de Vries, Serge Yefimov, Alan E. Mark, and Siewert J. Marrink. Molecular structure of the lecithin ripple phase. *Proceedings of the National Academy of Sciences*, 102(15):5392–5396, 2005.

[16] Olaf Lenz and Friederike Schmid. Structure of symmetric and asymmetric “ripple” phases in lipid bilayers. *Phys. Rev. Lett.*, 98(5):058104, 2007.

[17] Kiyotaka Akabori and John F. Nagle. Structure of the DMPC lipid bilayer ripple phase. *Soft Matter*, 11(5):918–926, 2015.

[18] Ananya Debnath, Foram M. Thakkar, Prabal K. Maiti, V. Kumaran, and K. G. Ayappa. Laterally structured ripple and square phases with one and two dimensional thickness modulations in a model bilayer system. *Soft Matter*, 10(38):7630–7637, 2014.

[19] Gregor Cevc and Derek Marsh. *Phospholipid Bilayers. Physical Principles and Models.* John Wiley & Sons, New-York, 1987.

[20] R J Gowers, M Linke, J Barnoud, T J E Reddy, M N Melo, S L Seyler, D L Dotson, J Domanski, S Buchoux, I M Kenney, and O Beckstein. Mdanalysis: A python package for the rapid analysis of molecular dynamics simulations. *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.

[21] N MichaudAgrawal, E J Denning, T B Woolf, and O Beckstein. Mdanalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, 2011.

[22] C R Rycroft. Voro++: A three-dimensional voronoi cell library in c++. *Chaos*, 19:041111, 2009.

[23] Jeffery B. Klauda, Richard M. Venable, J. Alfredo Freites, Joseph W. OConnor, Douglas J. Tobias, Carlos Mondragon-Ramirez, Igor Vorobyov, Alexander D. MacKerell, and Richard W. Pastor. Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *Journal of Physical Chemistry B*, 114(23):7830–7843, 2010.

[24] J F Nagle, R Zhang, S Tristram-Nagle, W Sun, H I Petrache, and R M Suter. X-ray structure determination of fully hydrated l alpha phase dipalmitoylphosphatidylcholine bilayers. *Biophysical Journal*, 70(3):1419–1431, 1996.

[25] J F Nagle and S Tristram-Nagle. Structure of lipid bilayers. *Biochimica and Biophysica Acta*, 1469(3):159195, 2000.

[26] N Kuçerka, S Tristram-Nagle, and J F Nagle. Closer look at structure of fully hydrated fluid phase dppc bilayers. *Biophysical Journal*, 90(11):L83–L85, 2006.

[27] N Kuçerka, J F Nagle, J N Sachs, S E Feller, J Pencer, A Jackson, and J Katsaras. Lipid bilayer structure determined by the simultaneous analysis of neutron and x-ray scattering data. *Biophysical Journal*, 95(5):23562367, 2008.