JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2019, 15, 6343−6357

pubs.acs.org/JCTC

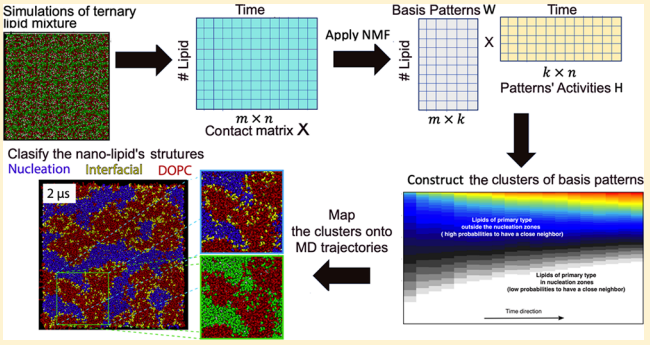
Unsupervised Machine Learning for Analysis of Phase Separation in Ternary Lipid Mixture

Cesar A. López,[†] Velimir V. Vesselinov,[‡] S. Gnanakaran,*[,†] and Boian S. Alexandrov*[,†]

[†]Theoretical Division and [‡]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States

Supporting Information

ABSTRACT: Phase separation in mixed lipid systems has been extensively studied both experimentally and theoretically because of its biological importance. A detailed description of such complex systems undoubtedly requires novel mathematical frameworks that are capable of decomposing and categorizing the evolution of thousands if not millions of lipids involved in the phenomenon. The interpretation and analysis of molecular dynamics (MD) simulations representing temporal and spatial changes in such systems are still a challenging task. Here, we present an unsupervised machine learning approach based on nonnegative matrix factorization called NMFk that successfully extracts latent (i.e., not directly observable) features from the second layer neighborhood profiles derived from coarse-grained MD simulations of a ternary lipid mixture. Our results demonstrate that NMFk extracts physically meaningful features that uniquely describe the phase separation such as locations and roles of different lipid types, formation of nanodomains, and timescales of lipid segregation.



INTRODUCTION

Cell membranes contain mixtures of different lipid types that play a key role in various mechanisms responsible for cell survival. In the past, membranes were thought to be homogeneous systems; however, new data suggests that, under a different stimulus, the lipids can segregate[1,2] into nanoscale domains. These domains are highly dynamic, varying in size and composition.[3,4] Importantly, this lateral partitioning is responsible for activation and functioning of membrane-embedded proteins.[5,6]

While it is possible to experimentally visualize the structure of segregated lipid domains,[1,7] a detailed description of such phases is inherently limited by the resolution of the experimental techniques. In this respect, molecular dynamics (MD) simulations provide a useful molecular description of the membrane's behavior.[8] In fact, the presence of such a lateral rearrangement has been studied extensively using coarse-grained (CG) MD of ternary lipid mixtures.[9] These CG simulations suggest partial segregation in large lipid systems that mimic the lipid variability of the real cell membranes.[10] MD simulations of such realistic biomolecular systems usually contain millions of particles even when simplified models are used. When the purpose of such simulations is to gain biological or physical insights, it is challenging to identify patterns in the behavior encoded in the motion of thousands of molecules, so developing analytic tools for extracting functionally relevant features from MD generated trajectories is of great importance.

Currently, machine learning (ML) methods have shown a lot of promise in many fields,[11] and recently, some of them have been applied for analysis and detection of classical and quantum phase transition data generated by simulations. Both supervised and unsupervised ML approaches have been used for this purpose,[12−19] but most of these pioneer studies used small Ising-like systems for their investigations. ML has previously been coupled with MD simulations of biomolecular systems in a limited context. ML techniques were reported to predict free-energy differences when trained with MD simulation data,[20] and unsupervised approaches such as PCA[21] as well as other techniques[22,23] have been used to reduce the dimensionality of MD-generated data.[24−28]

The unsupervised ML methods benefit ML that targets scientific understanding, because they learn relationships and similarities between elements in uncategorized data and classify the data without human's help but by revealing its hidden internal structure and latent (i.e., not directly observable) features buried in the data. Unsupervised learning methods include clustering,[29] classical neural networks,[30] and modern factorization techniques, like principle component analysis (PCA),[31] singular value decomposition (SVD),[32] independent component analysis (ICA),[33] and nonnegative matrix factorization (NMF).[34] A limitation shared by PCA, SVD, and ICA is the difficulty to relate the extracted latent features to easy

Received: January 25, 2019
Published: September 2, 2019

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

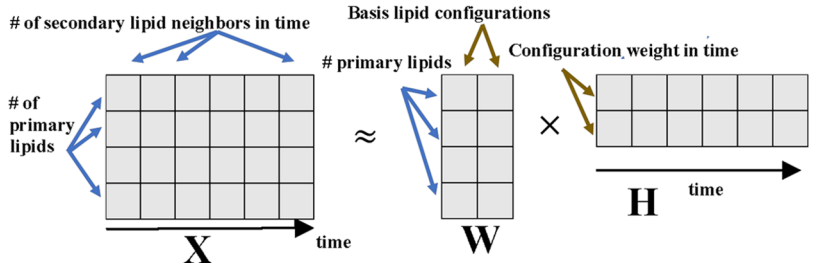


Figure 1. Illustration of a nonnegative matrix factorization. The nonnegative matrix **X** is decomposed to the product of a nonnegative matrix **W**, containing $K = 2$ basis configurations, and nonnegative matrix **H**, containing the activities of these two configurations in different time points.

interpretable quantities; NMF that we utilize here overcomes this limitation because the nonnegativity of the extracted latent features leads to a collection of strictly additive components that are sparse and parts of the data and hence are amenable to a simple and meaningful interpretation without prior assumptions.[35] Recently, the hidden Markov model (HMM),[36] in combination with the Voronoi tessellation[37] and clustering of the obtained spatial distribution via Getis–Ord local spatial autocorrelation statistics,[38] has been applied to analyze the ordered states and domains in lipid systems.[39] HMM is a statistical model with an assumed Markov process (i.e., a process where the individual states are independent and the system does not have a "memory") with latent states. Surprisingly, similar implementation of NMF has an analogous probabilistic interpretation (see below) for analysis.

Mathematically, NMF approximates a given nonnegative matrix **X**, with size (L, N), into a product of two nonnegative matrices **W**, with size (L, K), and **W**, with size (K, N), such that: $\mathbf{X}_{ij} = \sum_{s=1}^{K} \mathbf{W}_{is}\mathbf{H}_{sj}$. The factor matrices **W** and **H** are both nonnegative and have one small dimension K (Figure 1). The usual interpretation of NMF is a method for a low-rank matrix decomposition, in the sense of minimizing a given distance metrics, which was shown to be very useful for face recognition, dimension reduction, data mining, unsupervised learning, and blind source separation.[40] There is an alternative interpretation of NMF called probabilistic NMF that we utilize here: NMF is underpinned by a statistical model of superimposed components (the number of these components is equal to the size of the small dimension K) that can be treated as latent variables in Gaussian, Poisson, or other mixture models.[41] NMF minimization (with a specific distance metric) is equivalent to the expectation-minimization (EM) algorithm[42] that was developed to find the maximum likelihood estimates of parameters of statistical models when these models depend on latent variables. It is known that the probabilistic interpretation of NMF is particularly valuable when dealing with stochastic signals. Interestingly, the Baum–Welch algorithm, often utilized to estimate the parameters of HMM, is a particular specification of the EM algorithm designed to find the maximum likelihood estimate of the parameters of the hidden Markov model.[43] A mathematically rigorous formalism of the probabilistic NMF and its high-dimensional (tensor) version can be found, for example, in refs 41 44, and 45.

In the probabilistic NMF model, the observables $x_1, x_2..., x_n$, which are columns of the observational matrix **X**, are generated by the latent variables $h_1, ..., h_K$, which are columns of the matrix **H**. Specifically, each observable x_i is generated from a probability distribution with mean$\langle x_i \rangle = \sum_{s=1}^{K} \mathbf{W}_{is}h_s$, where K is the number of latent variables.[35] Thus, the influence of h_s on x_i is through the basis configurations represented by the columns of the matrix **W**, $w_1,..., w_K$. NMF has the ability to identify easy interpretable basis configurations and latent patterns that enable discoveries of new causal structures and previously unknown mechanisms hidden in the data, without prior hypotheses, in contrast, for example, to the hidden Markov models.[46]

Here, we present a new utilization of a previously proposed machine learning algorithm based on the probabilistic NMF model integrated with custom clustering, called NMFk[47−49] and demonstrate its abilities to analyze phase separation in a system of mixed lipids directly from preprocessed trajectories derived by MD simulations, without any prior hypotheses. Our observational matrix **X** is generated from data for CG MD simulations of a system that comprises a three-component lipid mixture, commonly accepted to mimic the behavior of a cellular plasma membrane.[4] We show that NMFk applied to a preprocessed data from these simulations is able to (a) identify the lipids that play distinct roles in lipid separation, (b) characterize the formation of nanolipid domains, (c) reveal timescales of interest, and (d) extract latent features that characterize the lipid phase separation.

RESULTS

Generation of Lipid Mixture Data Sets Using Coarse-Grained MD Simulations. For MD simulations of membranes and membrane-based biological systems, the Martini coarse-grained force field[50] considerably reduces the computational cost of calculations by nearly 3 orders of magnitude compared with similar MD simulations using fully atomistic force fields.[51] Particularly, the CG approach can capture relevant dynamics and fluctuations of larger membrane patches, which are prohibitive with atomistic simulations. Such access to larger spatial and temporal scales enables direct comparisons with experimental measurements.[52] Pioneering computational studies with the Martini CG force field allowed the characterization of not only lipid segregation and lipid phases but also the relative partitioning of membrane proteins between these phases.[9,53−55] More recently, Martini has been used in simulations of membranes with lipid compositions of comparable complexity to those found in specific tissues of living cells.[56,57]

Regardless of the extensive use of the Martini force field, the building-block principle of Martini along with the 4:1 atom-to-bead mapping unavoidably reduces the accuracy due to loss of detailed description of specific molecular chemical properties. Thus, despite the fact that many of the current Martini lipid parameters are sufficient to guide accurate membrane simulations,[50,58,59] global lipid properties are compromised.[60]

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

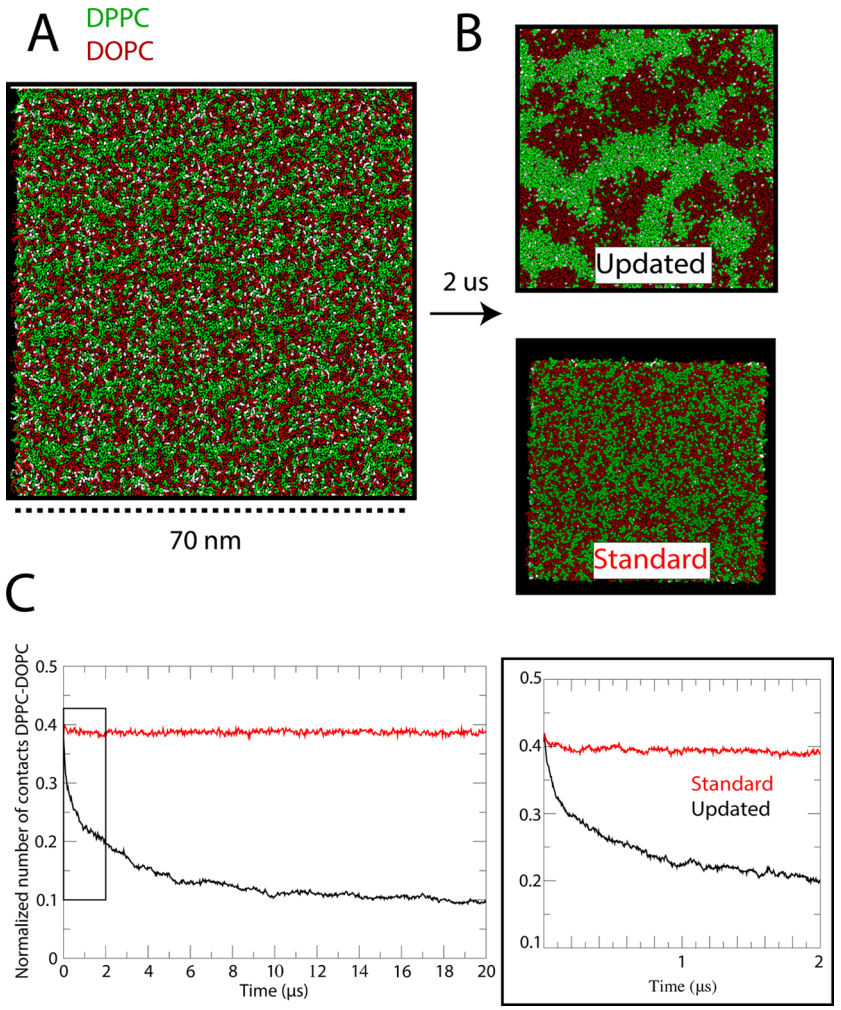


Figure 2. Coarse-grained simulations of ternary lipid mixture. (A) Initial system setup for the CG simulations. Lipids were initially randomly placed within the *XY* plane. Saturated lipid, DPPC, is colored green, while unsaturated lipid, DOPC, is colored red. Cholesterol is colored white. Water is not shown for clarity in depiction. (B) Lipid phase separation within 2 μs simulation. The Updated CG force field shows clear phase separation into Lo (green) and Ld (red). On the contrary, the Standard Martini lipid force field does not show preferential segregation in cholesterol-rich domains. (C) Time evolution of the normalized number of contacts between saturated and unsaturated lipids, showing poor separation with the Standard Martini force field. A close-up view is shown for the first 2 μs in order to enhance clarity or time frame analyzed.

Consistent with previously published works,[8,53,61] the standard Martini V2.2 parameters for DPPC, DOPC, and CHOL do not phase-separate at 298 K or even at 290 K in the physiologically relevant coexistence region of the phase diagram. Recently, we have incorporated changes in the lipid Martini force field,[62] greatly improving the simulations of lipid segregation in line with current experimental phase diagrams.[7,8]

We use the above two versions of the Martini force field, namely, one lipid parameter set that phase-separates and the other that does not, to generate two data sets for the analysis using NMFk. The first set considers the data from the coarse-grained MD simulations of the current Martini force field (called "Standard"). As mentioned above, the current Martini V2.2 lipid parameters for DPPC and DOPC are not able to properly describe segregation in a DPPC/DOPC/CHOL mixture. It serves as the prototype for the nonphase separating homogeneous lipid mixture. The second set considers the refined version of the Martini (called "Updated"), which has been optimized to reproduce the experimental phase separation and domain formation for this ternary system.[62] It serves as a prototype for phase-separating a lipid mixture. Using these two versions of force fields, we carried out 20 μs long CG MD simulations. The expectation is that we should be able to detect and analyze features associated with phase separation in one case but not in the other case.

Conventional Analysis of Domains in MD Simulations of Ternary Lipid Mixtures. The formation of lipid domains has been heavily studied both experimentally and computationally.[1−3,7,9,56,63−65] Computational observation of explicit lipid segregation at nearly atomic detail dates back to almost 10 years ago.[9] Analysis of such processes involved direct visualization of cholesterol-rich/-poor domains, as well as physical quantification of the area per lipid, cholesterol content, radial distribution functions, and membrane thickness mismatch.[9] These analyses brought enough details that membrane domains could be discriminated at the nanoscale. Regardless of these published methodologies, the analysis and prediction of such fluctuating membrane domains become inaccessible when the complexity of lipid content increases with larger size membrane patches (e.g., plasma membrane).

A typical process of molecular lipid phase separation is depicted in Figure 2. Initially, the lipid components are

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

randomized, mimicking the homogeneity at a high temperature (Figure 2A). Subsequent fast quenching of the mixture to 290 K, well below the melting temperature of the fully saturated DPPC lipid, leads to the rapid formation of nanoscale domains on a sub-microsecond timescale with the Updated Martini force field (Figure 2B, top panel). These nanodomains are eventually formed over the entire surface of the membrane but in different regions. After 0.5 μs, the nanodomains start to interconnect, leading to the formation of larger cholesterol-rich regions. In agreement with the general raft hypothesis[63] and previous computational studies,[9] the "ordered" nanodomains contain most of the saturated lipids together with cholesterol forming a Liquid-ordered (Lo) domain, whereas the "disordered" (Ld) nanodomain is mainly composed of the polyunsaturated DOPC lipid. Contrary to the Updated force field, the lipid mixture based on the Standard Martini force field does not show any tendency for phase separation or domain formation (Figure 2B, bottom panel), in agreement with previously published data.[62]

Following the conventional approaches to quantify the segregation tendency, we compute the normalized total contacts between DPPC and DOPC as a function of simulation time (Figure 2C). Initially, these contacts are featured by larger values, meaning that these two lipid types are indeed in close contact, highlighting the initial homogeneous lipid mixing of the system. However, with the Updated Martini, lipid segregation leads to a decrease in the total number of contacts between DPPC and DOPC (Figure 2C, black line). Meanwhile, the contacts remain unchanged in the simulations with the Standard Martini (Figure 2C, red line). In the case of Updated Martini, contacts decay during the simulations and begin to plateau with increasing simulation time. We should note here that the proper convergence to a stationary state may not be achievable within the simulation timescales considered here, as already published,[64] while significant transition toward segregation occurs within the first 2 μs (Figure 2C, inset). We consider this time regime suitable for NMFk analysis to extract latent features associated with the phase separation.

A more sophisticated approach can be found in the work of Baoukina et al.,[61] where a Voronoi tessellation methodology was applied in order to delineate the boundaries between ordered/disordered domains in monolayers. This approach can be also directly combined with automated predictive tools like Markovian based methodologies.[64] For instance, Sodt et al.[66] provided a straightforward description of lipid membrane partitioning using lipid-neighbor counts coupled with an HMM. The assumption is that the first shell composition around each lipid is able to provide enough description of Lo/Ld behavior. In a similar effort, Park and Im[39] introduced a two-step protocol for analyzing lipid order states and domains from the trajectories generated by simulations. Under the assumption that both surface area and chain thickness are good descriptors of domain formation, the output of their protocol allows mapping of the states of the observables to recognize the hidden end-states in a small lipid patch and mapping it to the spatial states of the lipids in the simulations. Importantly, such a method brings improvements that can be applied to membrane systems of poor segregation properties.

NMFk Framework for Analysis of Ternary Lipid Mixture Simulations. Our approach for description of segregations in a lipid bilayer is based on the following physical hypothesis: The number of lipids up to the secondary shell of a given lipid, although stochastic by nature, on average, is different if (a) this lipid is located in a nucleated domain in comparison with if (b) this lipid is situated out of any nucleated domain. Therefore, this number can be used as a descriptor of lipids' domains.

Hence, our natural observables are the packing of up to the second shell of a given lipid, that is, the number of specific neighbors of each lipid that are located up to its second shell, at time t. Specifically, we tagged all lipids of a given type (we call them lipids of Type I or primary lipids) and count the number of their neighbors of a given type (we call these lipids of Type II or secondary lipids) up to the second shell of each tagged primary lipid.

To extract and characterize the lipids' domains based on the above descriptor, we introduce a new unsupervised approach based on the previously reported NMFk algorithm.[47−49,67] After the MD simulations, our approach contains the following steps:

(1) Creating the contact matrix $\mathbf{X}_L(t)$

To create our observational data, we first define the Type I and Type II lipids (e.g., DPPC/DOPC), and then, based on the trajectories of the MD simulations, we calculate the contact matrix at each given time t, $\mathbf{X}_L(t)$. The contact matrix $\mathbf{X}_L(t)$ contains the number of lipids of Type II surrounding the lipid of Type I (each lipid of Type I is tagged by the index L), up to the second shell at each moment t.

Most of the details below are based on the consideration of DPPC/DOPC as the primary (Type I) and secondary (Type II) lipids, respectively. However, this selection is arbitrary, and different combinations of primary and secondary lipid types can be chosen. To demonstrate that, we also describe the phase separation, considering CHOL/DPPC, DPPC/DPPC, DOPC/DPPC, and DOPC/DOPC as primary/secondary lipids, respectively.

(2) Applying NMFk to $\mathbf{X}_L(t)$: extraction of the basis lipid configurations

NMFk approximates the contact matrix $\mathbf{X}_L(t)$ by a product of two matrices $\mathbf{W}_s(L)$ and $\mathbf{H}_s(t)$, where $\mathbf{X}_L(t) = \sum_{s=1}^{K} \mathbf{W}_s(L)\mathbf{H}_s(t)$. The columns of the matrix $\mathbf{W}$ are the basis configurations (as in the probabilistic NMF model), and the columns of matrix $\mathbf{H}$ are the hidden variables. An example of such decomposition for three hidden variables, namely, h_1, h_2, and h_3,is shown in Figure 3. The latent variables, which are the columns of the matrix $\mathbf{H}$, can be seen also as coefficients of the linear combinations of the basis lipid patterns, namely, w_1, w_2, and w_3, representing each observed lipid configuration x_i at time t_i (Figure 3).

In our approach, each basis lipid configuration w_k contains the probabilities of the tagged (by L) lipids of Type I to have a neighbor from Type II, situated up to the secondary shell. The combinations of the basis lipid configurations, namely, w_1, w_2, and w_3, with different activities h_1, h_2, and h_3, reconstruct all observed contact configurations, as they are obtained by the MD simulations of the investigated lipid bilayer system.

Importantly, NMHk, based on the stability of the decomposition (see the Materials and Methods section), is capable of determining the optimal number K of latent

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

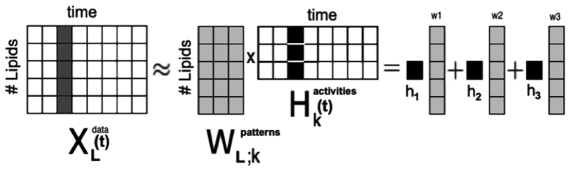


Figure 3. Illustration of the probabilistic hidden variable model underlying the nonnegative matrix factorization for lipid bilayer. The nonnegative contact matrix $\mathbf{X}_L(t)$, containing the contact configurations of the primary lipids L_1, L_2, ..., L_5 at consecutive time points t_1, t_2, ..., t_8, is decomposed into the product of two factor matrices $\mathbf{W}_k(L)$, containing $k = 3$ basis lipid configurations, and matrix $\mathbf{H}_k(t)$, containing activities (hidden variables) of these three configurations that generate the observables $x_1(L)$, $x_2(L)$, ..., $x_8(L)$, at every given time point t. In this figure, we represent the configuration x_3 at time t_1, expressed as a linear combination of the columns of $\mathbf{W}$.

features,[48] which here is the number of basis lipid patterns.

(3) Extraction of the lipids' domains and corresponding timescales

To characterize and extract the lipids' domains based on the obtained K basis lipid configurations, we applied standard k-means clustering on each of them separately. The idea here is to determine the groups of primary lipids in each basis configuration that have similar probabilities to have a neighbor of secondary type up to the secondary shell. Apparently, these groups of lipids are forming different domains. To find the most probable number of clusters, in each basis lipid configuration, we combined the k-means clustering with Silhouette statistics.[68] Finally, the time interval where each basis lipid configuration is active is determined by the corresponding row of matrix $\mathbf{H}_k(t)$.

(4) Mapping

Finally, we map the extracted feature by k-means clustering of each basis lipid configuration group of lipids on the MD trajectories to validate and rationalize the obtained results.

A flowchart of our approach is presented in Figure 4.

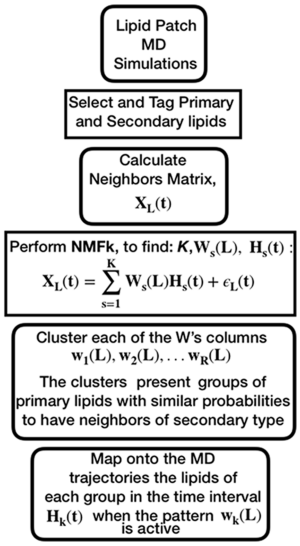


Figure 4. Flowchart of the NMFk approach for data analysis of MD simulations of phase separation in lipid systems.

NMFk Implementation for Analysis of Ternary Lipid Mixture Simulations. We first define the Type I and Type II lipids to be DPPC/DOPC tagging each DPPC. Using the MD trajectories, we compute the number of DOPC neighbors around every DPPC lipid, within the distance corresponding to the second peak of the DOPC/DPPC radial distribution function (Figure 5, blue dashed line). The number of

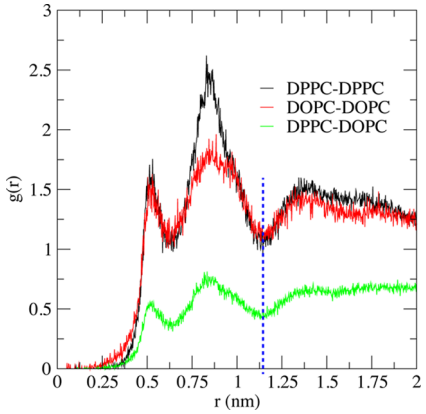


Figure 5. Lateral radial distribution function for the different lipid combinations. RDF was computed considering the center of mass of the molecules. The dashed blue line indicates the chosen cutoff distance for profiling neighbor list needed for the NMFk matrix construction.

neighbors within this distance serves as the order parameter for description of the lipid phase separation and create the contact matrix $\mathbf{X}_L(t)$. Then NMFk decomposes the $\mathbf{X}_L(t)$ as a product of two matrices: $\mathbf{W}_s(L)$ and $\mathbf{H}_s(t)$, where the columns of $\mathbf{W}_s(L)$ are the K basis lipid configurations, describing the state of the lipid system, and $\mathbf{H}_s(t)$ contains the activities of each one of these configurations at time t_i. Thus, for a given arrangement of the primary lipids and contact matrix $\mathbf{X}_L(t)$, we have

$$\mathbf{X}_L(t) = \sum_{s=1}^{K} \mathbf{W}_s(L)\mathbf{H}_s(t) + \varepsilon_L(t) \quad (1)$$

where $\varepsilon_L(t)$ denotes the presence of a noise or unbiased error of the decomposition (see Figure 1, where $K = 2$).

The accuracy of the reconstruction of $\mathbf{X}_L(t)$, provided by the product of the two matrices $\mathbf{W}_s(L)$ and $\mathbf{H}_s(t)$, serves as a measure for the significance and quality of the extracted latent features. At the beginning of the Updated Martini simulations, the lipid mixture is homogeneous and there are no distinct features, that is, there is no specific structure in $\mathbf{X}_L(t)$. Hence, reconstruction of $\mathbf{X}_L(t)$ did not reproduce well the simulation data. However, by 2 μs, the primary lipid type, DPPC, segregates into the Lo region, while the secondary lipid type, DOPC, segregates into the Ld domain. This process leads to distinct latent features. Table 1 presents the Pearson correlation coefficient between the reconstructed lipid arrangements at each time point t (obtained by NMFk) and the original arrangements at the same time points (i.e., the corresponding column of the neighbor matrix $\mathbf{X}_L(t)$).

The K unique basis lipid configurations (encoded in the matrix $\mathbf{W}_s(L)$) reproduced accurately $N - s$ lipid arrangements forming the matrix $\mathbf{X}_L(t > t_s)$. With the Standard Martini, the NMFk was not able to provide a set of lipid configurations that can reconstruct the simulations accurately. Indeed, Table 1

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

Table 1. Quality of Reconstruction of X Estimated by the Mean Pearson Correlation between the Columns of X and Columns of W × H

coarse-grained simulations of ternary lipid mixture	time (μs)	mean correlation coefficient	standard deviation
Standard Martini	0.45−2	0.54	0.040
Updated Martini	0.45−2	0.90	0.020
	16−18	0.96	0.004

shows that the NMFk analysis does not reproduce well the simulation data obtained via Standard Martini.

NMFk Extracts Basis Lipid Configurations Associated with Phase Separation. A typical outcome of an NMFk analysis is presented in Figure 6. According to the Silhouette-Reconstruction criterion (see the Materials and Methods section), NMFk determines that the optimal number of stable basis lipid configurations represented by the columns w_i of $\mathbf{W}_s(L)$ is 20 (Figure 6A). Next, the rows of the matrix $\mathbf{H}_s(t)$ that encode the contribution/activity of the corresponding basis lipid configuration w_i at time t are presented in Figure 6B. It is clear that, for each basis lipid configuration w_i, there is a well-defined time interval containing a number of consecutive frames where w_i is dominant and only a few other basis configurations are active in this time interval. From Table 1, it can be seen that, after 0.45 μs, NMFk reproduces the neighbor matrix $\mathbf{X}_L(t > t_s)$ very well: the cross-correlation between the contact matrix and the reconstructed matrix for each time frame is above 0.95. The much lower cross-correlation for reconstruction of the contact matrix $\mathbf{X}_L(t)$ at early times suggests a prevalence of a relatively homogeneous lipid mixture in the first 0.45 μs. The 15 basis lipid configurations that are active at consecutive time intervals corresponding to the rows of $\mathbf{H}_s(t > t_s)$ after the first 0.45 μs are #6, #10, #12, #15, #17, #19, #20, #18, #16, #14, #13, #11, #7, #2, and #1. Each one of these basis lipid configurations contains a set of probabilities for each lipid of primary type to have a neighbor lipid of secondary type up to the secondary shell. The consecutive basis lipid configurations in time represent the evolution of the phase separation in the lipid system.

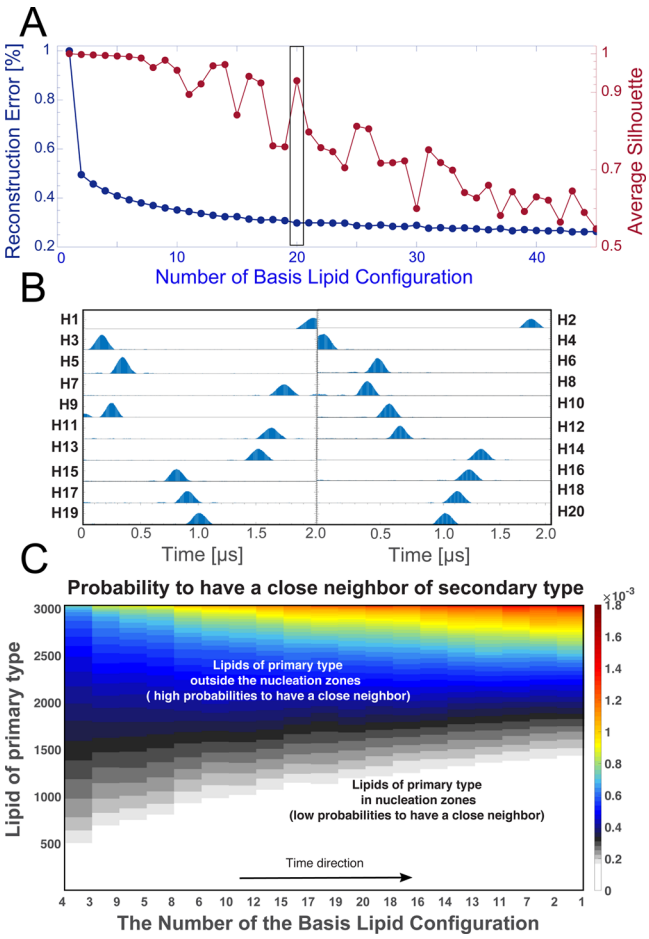


Figure 6. Outcome from NMFk analysis. (A) Silhouette-Reconstruction criterion (see the Materials and Methods Section). The x axis is denoted the number of basis lipid configurations, and the y axis is the average Silhouettes (right y axis, red marking) as well as the Reconstruction (left y axis, the blue marking). The box denotes that NMFk estimates the number of basis lipid configurations to be 20. (B) Presentation of the rows h_i of the matrix $\mathbf{H}_s(t)$ that encodes the activities of the basis lipid configurations in time t. (C) Heatmap representing the basis lipid configurations ordered along the x axis according to the time interval they are active, from early time to late, as gathered from the corresponding activities h_i. The color gradient from white to blue to red represents the increase in probability for a specific primary lipid to have a secondary lipid-neighbor within each basis lipid configuration. The increase in white and gray colors with time is indicative of the evolution of the phase separation.

Structure of the Basis Lipid Configurations Extracted by NMFk. Each one of the 20 basis lipid configurations extracted by NMFk contains a set of probabilities for the lipids of a primary type to have a neighbor lipid of a secondary type up to the secondary shell. Each basis lipid configuration is active at a given time interval defined by the activity of this basis configuration, quantified by the corresponding row of the matrix $\mathbf{H}_s(t)$. Importantly, the set of probabilities in each basis lipid configuration has a clear structure interrelated with the phase separation and formation of lipids' domains in the system.

By applying k-means clustering combined with Silhouette statistics, we found that each basis lipid configuration possesses well defined groups/clusters with different average probabilities. In each basis lipid configuration, we found at least two groups of probabilities: (a) the group of probabilities of the primary lipids situated in domains that have relatively small number of neighbor-lipid of secondary type and therefore, on average, approach zero and (b) the probabilities of the primary lipids that are outside any nucleated domain whose number of neighbor-lipids of secondary type is much higher and hence on average bigger.

Figure 6C shows the basis lipid configurations ordered according to the sequence of the frames corresponding to consecutive time intervals determined by their respective activities. The k-means clustering procedure determined that each of the extracted 20 lipid basis configurations can be separated into two clusters: The first cluster contains the primary lipids with a low probability to have a DOPC lipid-neighbor, and the second cluster contains the primary lipids with more than 4 times higher probability to have a DOPC lipid-neighbor. Next, we colored differently the lipids in each of the 20 lipid basis configurations, with two clusters each, at the time intervals where the respective lipid basis configuration is active. The color gradient in Figure 6C captures these two groups of primary lipids: white to gray for the primary lipids within the nucleated domains and blue to red for the primary lipids at the interface or outside any nucleated domain.

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

When the lipids' domains are forming, there is always fast changeover of the locations of the primary type of lipid with time. Lipids move in and out of the nucleated domains or at the interface of the nucleated domains: at a certain moment, a given primary lipid can be situated outside a nucleated domain formed by primary lipids, but after a while, it can reach interfacial region and eventually get absorbed into the nucleated domain. Alternatively, primary lipids inside the nucleated domain or at the interfacial region may venture out into the liquid-disordered region enriched with secondary lipids. These exchanges continuously alter the probability of a given lipid of the primary type to have a neighbor-lipid of secondary type, as the phase separation proceeds, which results in different basis lipid configurations at different time points as the system goes to a phase separation. At a long timescale, when the phase separation has reached equilibrium, basis lipid configurations capture slower exchanges of the primary lipids governed only by the stochasticity and diffusion.

Mapping the Structure of the Basis Lipid Configurations onto the MD Trajectories and Validation of NMFk Approach. We use the MD simulations trajectories to visualize, rationalize, and validate the sets of two primary lipid clusters, as extracted by the clustering of each basis lipid configuration.

We considered lipid basis configurations #19 and #1 corresponding to 1 and 2 μs time points, respectively. All lipids contributing to those lipid basis configurations are mapped into the trajectories at the corresponding time points in Figure 7. At 2 μs, approximately 70% of primary lipids (DPPC, colored blue) are situated in large nucleated domains, which are well packed and condensed by the high cholesterol concentration, leading to the Ld phase. These primary lipids are predominantly shielded from the secondary (DOPC, colored red) lipids, as they are localized and form the Lo phase. Hence, the primary lipids corresponding to the first cluster identified by NMFk are localized in nucleated domains and form the Ld phase.

On the other hand, at 2 μs, approximately 30% of the primary lipids (DPPC, colored yellow) in the same lipid basis configuration #1 are located near the interfacial regions or deep into the Ld regions formed by secondary lipid types. Unlike the previous set of primary lipids, these lipids are in continuous contact with the secondary lipids. The inset in Figure 7 shows that NMFk distinguishes these lipids from all DPPC lipids available. These lipids correspond to the second cluster identified by NMFk.

Thus, each extracted lipid basis configuration contains physical and easily interpretable features that enable us to make a distinction on primary lipids (of the same type) depending on their location and their contribution to the lipid segregation.

The strength of NMFk analysis is the ability to extract lipid basis configurations as a function of time. These configurations enable us to determine the time dependence of the latent features (the columns of matrix $\mathbf{H}_s(t)$) related to the kinetics of phase separation without carrying out tedious multiple analyses. A conventional analysis that seeks to probe the temporal profile would have considered the normalized number of contacts among DPPC lipids and between DPPC and DOPC to capture the nucleation process (Figure 2C). NMFk decomposes the nucleation process into two components, as shown in Figure 8 where the distinction is made on the temporal profiles of primary lipids from the

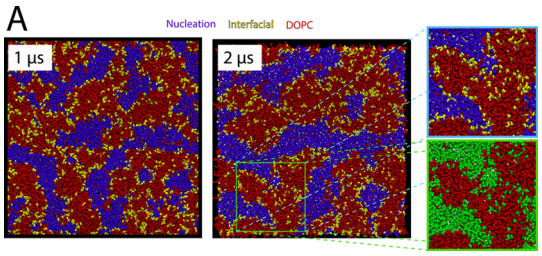


Figure 7. Visual inspection of simulations trajectories for evaluation of NMFk categorization primary lipids according to their localization. Primary (DPPC) lipids are colored according to the NMFk output as purple (in nucleated domains) and yellow (out of nucleated domains) in MD configurations corresponding to two time points. Secondary (DOPC) lipids are colored red, and cholesterol is colored white. Insets highlight a particular region where lipids can be differentiated as nucleating (purple) and boundary (yellow) lipids. As a comparison, the same region is represented using the conventional way of addressing the distinction between saturated (green) and unsaturated (red) lipids.

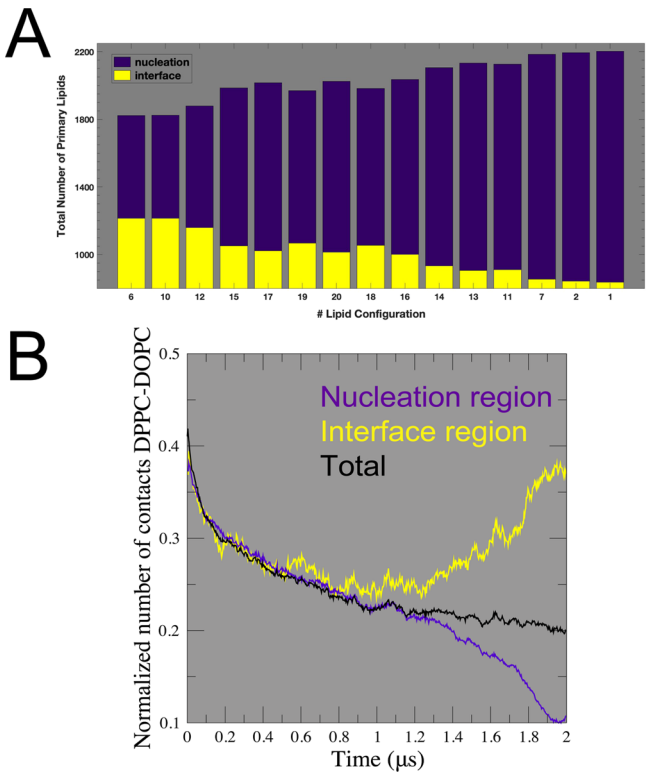


Figure 8. Time-dependent variation of the total number of nucleating primary (DPPC) lipids. (A) The purple bars represent the total number of primary lipids in nucleated domains, as concluded from their membership in different clusters of the corresponding basis lipid configuration, while the primary lipids outside any nucleated domain are presented by the yellow bars. The labels on the x axis (the numbers) correspond to consecutive (in time) processes extracted from 0.45 to 2 μs simulation time. (B) Same distinction of nucleating primary lipids as in panel (A) obtained using the normalized number of contacts of the lipids identified by NMFk. The same color scheme is used. The black line corresponds to the normalized number of contacts using the total fraction of saturated lipids (i.e., without making the distinction within the nucleating primary lipids).

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

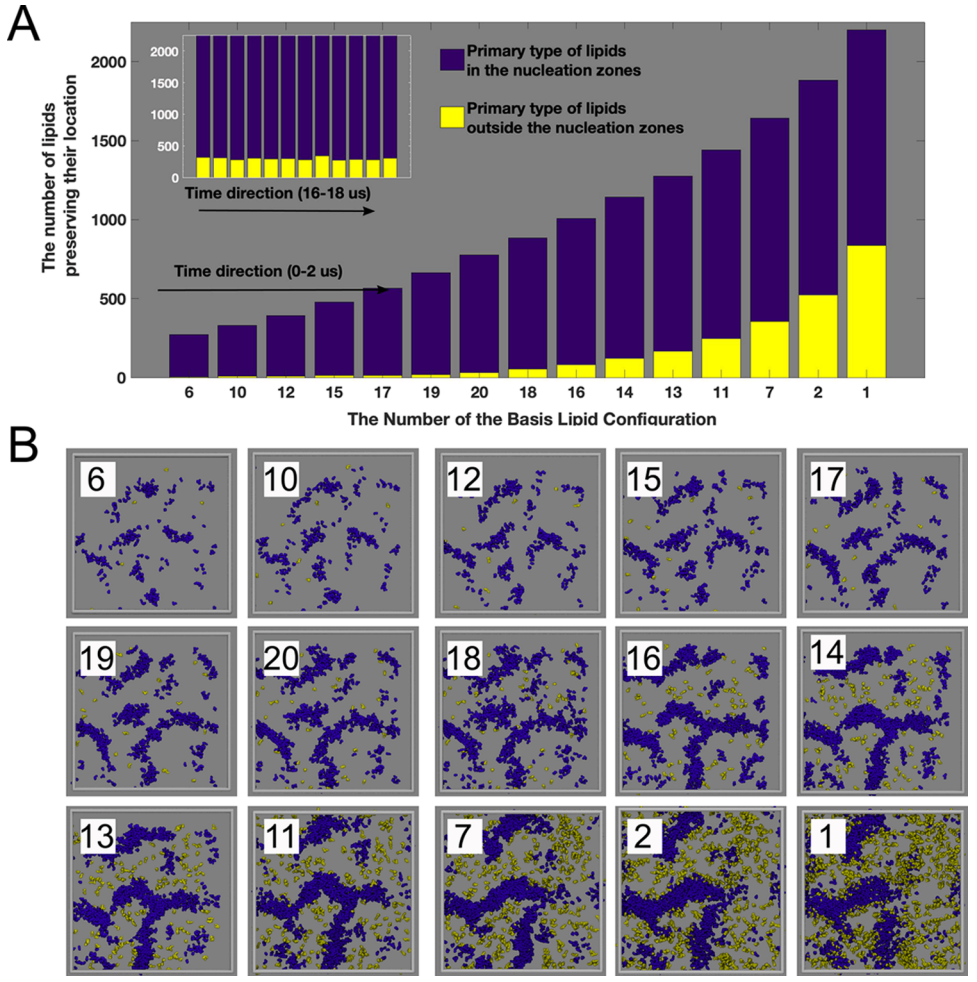


Figure 9. Temporal description of membership profile of primary lipids as captured by NMFk analysis. (A) The clear growth of nucleated domains with time, formed by primary lipids, is extracted from the clusters containing primary lipids with low probabilities to have lipid-neighbors of secondary type (refer to Figure 5C). The numbers of primary lipids that belong to nucleated domains are colored purple, whereas those that appear at the interface (edges of the nucleated domain) or outside the nucleated domain are colored yellow. The exponential growth of the number of primary lipids that continue to be in the same cluster demonstrated the evolution of the steady state of the phase separation. The inset demonstrates the same process but at much later time (~20 μs) when the phase separation is in equilibrium. Here, although the minor changes still exist, primary lipid membership numbers have stabilized. (B) Spatial visualization of the primary lipid memberships extracted by NMFk analysis (between 0.45 and 2 μs). These primary lipids were tracked using MD trajectories and rendered with the same color scheme as in panel (A) to distinguish the primary lipids in and out of the nucleated domains. Other lipids are not rendered (silver background). The MD simulation box is represented as solid gray lines.

nucleated domains from those that are still outside the nucleated domains. In Figure 8A, the purple bars represent the total number of primary lipids in nucleated domains at consecutive time intervals (ordered on the x axis), while the primary lipids outside any nucleated domain are presented by yellow bars. At early times, rapid growth of domains is seen, which is directly correlated with the increase in nucleated domains of primary lipids. After this rapid growth, a steadier behavior is observed, which continues until the end of the simulation. In Figure 8B, we represent the normalized number of contacts with the secondary lipids for the primary lipids as identified from NMFk. At early times, these contacts are high due to random encounters between lipids. This behavior begins to change around 0.6 μs, and then the contacts between primary and secondary lipids are indicative of steady growth of the number of primary lipids in the nucleated domains and a decrease in the total number of lipids in the nonnucleating regions as the lipid mixture system phase-separates.

Furthermore, NMFk analysis extracts the temporal evolution of the primary lipids' membership: in or out of a nucleated domain. Specifically, the membership is defined by identifying primary lipids that join a nucleated domain and remain in that domain until the phase separation. We ordered the 15 significant basis lipid configurations that are acting in $t > t_s$ extracted by NMFk according to the time intervals when the specific configurations are active (Figure 9). We identify the primary lipids that participate in the nucleation by keeping track of the lipids in basis lipid configurations with low probabilities to have a secondary lipid-neighbor. In Figure 9A, we present the primary lipids that remain in the same cluster with a small (purple color) or high (yellow color) probability to have a secondary lipid-neighbor at consecutive time intervals, which represents the evolution of the membership and hence stability of the nucleation. The inset in Figure 9A demonstrates the system at much later times (~20 μs) after the initial nucleation processes when the phase separation has reached equilibrium and the primary lipids that are located in

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

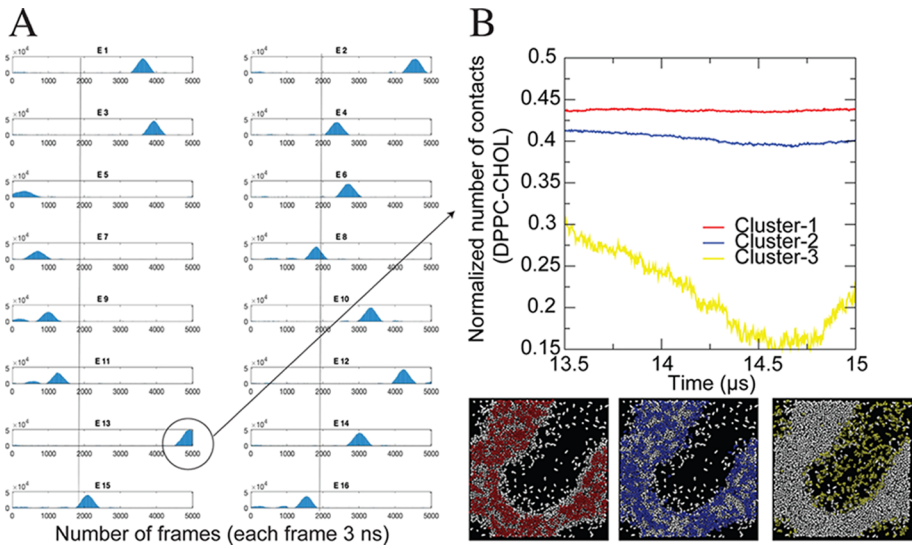


Figure 10. Lipids' domains extracted by NMFk applied to DPPC/CHOL contact matrix and mapped onto the MD trajectories. (A) Groups of lipids at given time frame. (B) Normalized number of contacts between DPPC and cholesterol, within the range of 13.5−15 μs.

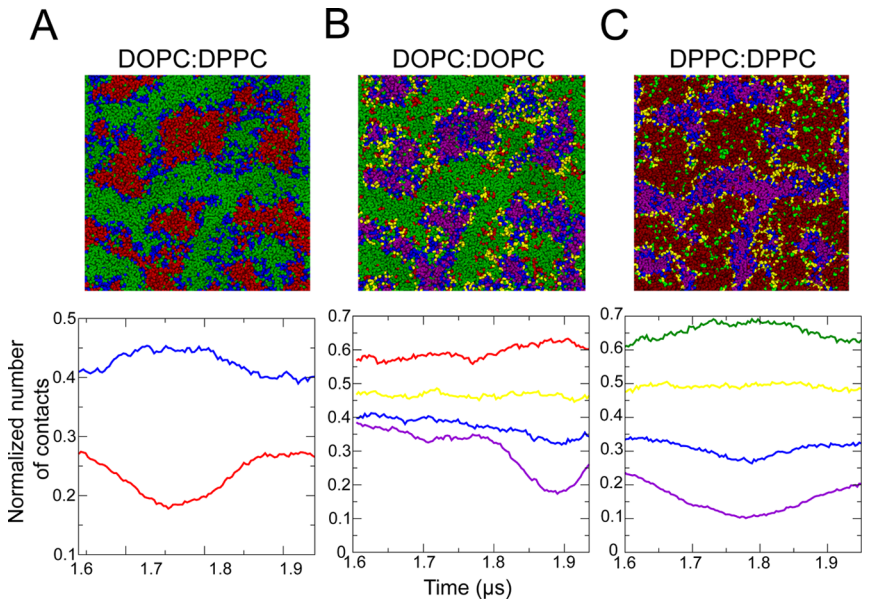


Figure 11. Mapping of the groups of lipids extracted via NMF analysis for a diverse set of contact matrices: DOPC/DPPC, DOPC/DOPC, and DPPC/DPPC.

nucleated domains mostly preserve their membership in time. Importantly, this evolution of the primary lipids can be easily mapped at their spatial coordinates. In Figure 9B, the patterns extracted by NMFk are mapped via MD trajectories to their spatial coordinates (at each time interval) to visualize the evolution of the different groups of primary lipids based on their membership: in nucleated domains (purple color) or outside those domains (yellow color). This again highlights the power of the NMFk to extract physical properties and details that can be easily visually tracked as the system undergoes localization and lipid segregation.

NMFk Analysis of DPPC/CHOL, DOPC/DPPC, DOPC/DOPC, and DPPC/DPPC Contact Matrices. The NMFk-based analysis of the DPPC/DOPC contact matrix demonstrated that NMFk extracts features of lipid phase separation that can be easily recognized after mapping onto MD trajectories. Here, we demonstrate that our NMFk approach does not require tagging of a particular type of lipids as "primary" or "secondary" since the underlying analysis uses only the set of contact profiles as an input. In particular, we show that the previous analysis of the DPPC/DOPC contact matrix can be performed on alternative contact matrices associated with different types of primary and secondary lipid types. To demonstrate that, we performed NMFk analyses on DPPC/CHOL, DOPC/DPPC, DPPC/DPPC, and DOPC/DOPC contact matrices.

First, we decomposed the same MD trajectories in terms of DPPC/CHOL contacts. In order to avoid the effect of the cholesterol flip-flop, we analyzed the contacts in the context of a full membrane and not per leaflet. The results are provided in Figure 10. For the DPPC/CHOL contact matrix, NMFk extracts 16 basis lipid configurations that reproduce well the contact matrix and are summarized in Figure 10A. The clustering analysis of the basis lipid configurations extracts a set

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

of lipids' groups that we mapped onto the MD trajectories. In Figure 10B, we show the groups of lipids obtained via *k*-means clustering of the basis lipid configuration #13 (indicated by the arrow) within the range of activity of this basis lipid configuration (13.5−15 μs). *k*-means clustering extracts three lipid groups, colored red, blue, and yellow, for better visualization. The red group corresponds to lipids that are in constant contact with the cholesterol-enriched region domain and are featured by a nonchanged normalized number of contacts with cholesterol. The second group of DPPC lipids (blue) is also part of the cholesterol-enriched region; however, at certain times along the trajectory, this particular group of lipids diffuses toward the boundaries of the cholesterol-enriched domains. A direct analysis of the number of contacts pertaining to this process would be hard to asses given the narrow differences between the red and blue groups. However, the fact that the lipids can be directly separated into these two groups confirms the strength of NMFk analysis to distinguishing subtle differences in lipids' domains. A third group of lipids is easier to distinguish as they diffuse within the poor cholesterol domain (yellow color). Thus, analyzing the DPPC/CHOL contact matrix, NMFk successfully identifies distinct groups of lipids that clearly belong to different domains of the membrane. In this case, cholesterol enrichment is directly linked to the formation of the Lo domain, which is fluid and continuously changing, giving rise to various nanostructures.

Next, we perform NMFk analyses on another contact matrix, namely, DOPC/DPPC, as shown in Figure 11A. The presented images highlight the extracted feature via the *k*-means clustering groups of lipids within the timescales presented underneath. Each extracted group of lipids is accompanied by the calculated normalized number of contacts. Thus, the NMFk analysis based on the DOPC/DPPC contact matrix has DOPC as the primary lipids, which complements the outcome introduced previously where DPPC was considered as the primary lipid type using the DPPC/DOPC contact matrix. In the opposite DOPC/DPPC case, NMFk distinguishes the presence of two different lipids' groups: red that is fully segregated from the DPPC domains and blue that is localized within the boundaries of the red and green (DPPC) regions. Such a distinction and formulation of lipids' groups resemble the detailed description provided previously for the DPPC/DOPC contact matrix.

Finally, we considered contact matrices obtained by decomposing the same MD trajectories in terms of DOPC/DOPC and DPPC/DPPC contacts. In these self-contact matrices, we are using the same type of lipid as the primary and secondary type. The results are shown in Figure 11B,C. Applied to these contact matrices, NMFk analysis, by applying *k*-means clustering to the last (in time) basis lipid configuration, extracts two more groups. These groups share specific structural properties that can be clearly seen once mapped into the MD trajectories. One of the groups of lipids (red and green lines in the lipid−lipid contact profile) displays a high-normalized number of contacts with the lipids pertaining to the domain that we are not considering here (e.g., if analysis is performed on DPPC, computed contacts are with respect to DOPC). In the case of the DOPC/DOPC contact matrix (Figure 11B), the red lipids are surrounded by the domain colored green (forming DPPC regions). For the DPPC/DPPC contact matrix (Figure 11C), the domains are extracted again from the last basis lipid configuration. In Figure 11C, the green lipids are deeply localized within the DOPC-enriched region. Similar properties can be observed for the other extracted lipids' groups: The yellow-colored lipids are constantly localized within the boundaries of the two membrane domains, while the blue-colored lipids are localized within the boundary and the nucleated domains, and the purple-colored lipids are clearly forming the deeper parts of the nucleated domains. According to the contact profile plots, the blue subset can be idealized as a buffer region between the boundary and nucleated regions.

DISCUSSION

In this paper, we introduce a new unsupervised machine learning approach for extraction of latent variables based on NMFk, which combines NMF with custom clustering, for analysis of MD simulations. Specifically, we implement NMFk to detect and describe the lateral lipid phase separation in a simplistic lipid raft model composed of a well-characterized ternary lipid mixture. Based on this study, we believe that the NMFk approach can be also implemented for extracting relevant features from more complex biological membranes.

The DOPC/DPPC/CHOL ternary lipid mixture considered here can exist as a homogeneously mixed mixture or exhibit two distinct phases of Lo and Ld depending on temperature. NMFk is utilized to detect and analyze this phase separation behavior of this well-studied system. The distinction between the two phases, Lo and Ld, is sensitive to the spatial localization of DPPC lipids and the number of DOPC neighbors. We designate DPPC and DOPC as primary and secondary lipids in the NMFk formalism, respectively. At the beginning of the simulations, there are no distinct phases, and the lipid mixture is homogeneous. As simulation time progresses, lipids begin to segregate. By 0.6 μs, the primary lipid type, DPPC, begins to segregate into the Lo regions, while the secondary lipid type, DOPC, begins to segregate into the Ld domains. Thus, the number of secondary lipid-neighbors around a given primary lipid should reflect that phase separation.

Given that a neighborhood profile of lipids can track the DOPC/DPPC/CHOL mixture, we first built a specific time-dependent nonnegative data matrix $\mathbf{X}N(t)$XNt (see the Materials and Methods section) whose elements represent the number of DOPC neighbors to each one of the DPPC lipids in the system within a specific radius r_{cutoff} extracted from careful analysis of the radial distribution function. NMFk decomposed the matrix $\mathbf{X}N(t)$ into a product of two matrices: (i) the matrix of the basis lipid configurations $\mathbf{W}_s(L)$WNK whose columns contain the probability of each of the DPPC lipids to have a DOPC neighbor and (ii) the matrix $\mathbf{H}_s(t)$ in which rows contain the activity of each one of these basis lipid configurations in time *t*. NMFk determines the number of basis lipid configurations *K* based on the robustness of the decomposition. Each one of the *K* basis lipid configurations in $\mathbf{W}_s(L)$WNK contains the probabilities of the tagged DPPC lipids to have DOPC neighbors up to the secondary shell. The NMFk Silhouette-Reconstruction criterion was used to estimate the optimal number of basis lipid configurations to be 20. Further, *k*-means clustering of these 20 basis lipid configurations demonstrates the tendency of increasing the number of primary lipids with a neglecting probability to have a neighbor of type DOPC when the time advances, which correspond to an increased total number of DPPC lipids located in nucleated domains.

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

By relating the basis lipid configurations to MD trajectories, we were able to show details of phase separation extracted by NMFk analysis and to validate our results. The NMFk discriminates lipids, depending on whether they belong to Lo or Ld phases or interfacial regions, as they undergo phase separation. Unlike other analyses, basis lipid configurations provide details of lipids that take part in the nucleation versus those that establish line tension. NMFk tracks the complicated features of the lipid segregation process leading to Lo and Ld phases. We identify lipids within the boundary of the Lo phase with lipid configuration basis corresponding to a signature of the Lo domain where DPPC is well packed and condensed by the high cholesterol concentration. Separately, another basis lipid configuration captures interfacial lipids that shield the DOPC lipids from such Lo domains during phase separation. Importantly, we demonstrate that the evolution of the nucleation process is captured by NMFk in terms of lipid membership to a different basis lipid configuration active in consecutive time intervals. NMFk identifies the lipids that take part in the initial nucleation and remains as part of the domain toward equilibrium. The method also identifies the lipids that join the initially nucleated regions and remains in them until the phase separation reaches equilibrium.

Finally, it is clear that the definition of contacts between lipids specifies structural and dynamical information that NMFk can extract in order to understand membrane properties. We also demonstrate that a direct selection of primary versus secondary types of lipids in our approach does include bias toward specific latent features. The NMFk analysis of different contact matrices demonstrates that the different contact matrices contain various different membrane physical and structural properties and latent features, which cannot be covered by a single contact matrix. For instance, the presence of a buffer region between the boundary and the nucleated domain can have a physical and structural implication that cannot be traced by solely comparing the contacts between lipids but uncovering new hidden properties using NMFk.

CONCLUSIONS

The high variability and complexity of plasma membranes are still poorly understood. Higher-resolution spectroscopy in combination with atomic detailed computer simulations provides new insights. However, we are not close yet to fully understanding or describing the membrane processes regulating cellular functions. A detailed description of such complex systems undoubtedly requires novel mathematical frameworks that are able to decompose and categorize the evolution of thousands if not millions of lipids involved in the phenomenon.

Here, we show the power of our new NMFk approach on analyzing lipid phase separation and providing categorization of the lipids according to their localization in the membrane as well elucidating time dependencies along the nucleation process. The NMFk discriminates the different types of lipids, part of Lo, Ld, or interface, due to their particular behavior along the trajectory and the resulting probability to have a neighbor up to their second shell. If there is no clear pattern in the behavior of the lipids, for example, when MD simulations do not produce any distinct behavior associated with phase separation, NMFk analysis does not produce false features. This is the first demonstration of NMFk serving as a tool in detecting time-dependent domains formation and lipid separation in MD simulations of a complex lipid mixture system. Even though we have exhibited the usefulness of NMFk in the context of a well-studied ternary lipid mixture, an extension to more complicated mixtures is feasible with a tensor formalism,[69] and it is currently under consideration.

MATERIALS AND METHODS

Membrane Patch. An initial configuration of a CG membrane patch was obtained using the script tools provided in the Martini force field website (http://cgmartini.nl/). Our CG lipid system contains DPPC/DOPC/CHOL lipids in a 37:36:27 ratio, which initially were randomly placed within an *XY* plane. This lipid ratio has been experimentally and computationally observed to transitioning toward a phase-separated Liquid-ordered/Liquid-disordered state.[64,70] The lipids were represented using the Martini V2.2 force field[50] with the "optimal" set of parameters, which has shown an improved phase separation behavior.[62] Similarly, simulation with the standard Martini lipid model was also carried out. The total system was composed of 16366 lipids, 718830 Martini water beads (175 atomistic water molecules per lipid), and 150 mM NaCl to preserve an overall constant ionic strength. In order to avoid spontaneous freezing of the Martini water beads (a well-known artifact previously reported in the original model[50]), 0.1% M water beads were replaced by antifreeze particles.

Molecular Dynamics Protocol. We followed a current update in parameters setup for performing the CG simulations.[71] The equations of motion were integrated every 30 fs time step. A reaction-field electrostatics algorithm was used with a Coulomb cutoff of 1.1 nm and dielectric constant of 15 or 0 within or beyond this cutoff. Lennard-Jones interactions were cut off at 1.1 nm where the potential was shifted to zero. In order to accelerate the lipid phase demixing, constant temperature was maintained at 290 K via separate coupling of the solvent (water and ions) and membrane components using a velocity-rescaling thermostat[72] with a relaxation time of 1.0 ps. During equilibration, the Martini beads representing the phosphate groups of the lipid head regions were positionally (*xyz* components) restrained in order to preserve the initial random positions. In this stage, the solvent molecules (water and ions) were allowed to diffuse, and the box pressure was maintained semi-isotropically coupled at 1 bar using the Berendsen barostat[73] with a relaxation time of 12 ps and compressibility of 3×10^{-4} bar^{-1}. After that, production runs were performed using a Parrinello–Rahman barostat.[74] Simulations were run for 20 μs using the GROMACS Version 5.2.1,[75] and the trajectories were saved every 3 ns providing the frames for the construction of the NMFk matrix (see later). The simulated MD data containing the lipids' trajectories (~100 GB) is available freely, but because of its size, it is upon request to gnana@lanl.gov.

Generation of the Contact Matrix $X_L(t)$. Every frame stored within 2 μs (667 frames in total) was used for generating the corresponding matrix for NMFk analysis. We rely on the implemented GROMACS tool gmx select to output the number of DOPC lipids around every DPPC molecule within 1.1 nm. This cutoff radius structurally corresponds to the second layer of neighbors, as estimated by the second maximum peak of the radial distribution function $g(r)$ (Figure 3). Thus, each column of the contact matrix $\mathbf{X}_L(t))$XNt corresponds to a variable number of DOPC neighbors of a given DPPC lipid per frame, while the rows correspond to the number of 3038 DPPC lipids in the system. Similarly, matrix

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

reconstruction was carried out for the 20 μs collection. An example of the matrix can be found in the Supporting Information.

Nonnegative Matrix Factorization and NMFk Algorithm. Nonnegative matrix factorization (NMF) is a well-known unsupervised method created for parts-based representation.[34] In the low-rank decomposition via NMF, only the observational matrix is known initially. In the probabilistic interpretation of NMF, the hypothesis is that the observational matrix $\mathbf{X}$ is formed by a mixing of K Gaussian (or other) components.[41] Since both factor matrices $\mathbf{W}$ and $\mathbf{H}$ are unknown and even their size K (i.e., the number of hidden patterns) is unknown, the problem is typically underdetermined. NMF can solve that kind of problem by leveraging, for example, the multiplicative update algorithm[35] to minimize the Frobenius norm of the objective function: $\mathbf{O} = \|\mathbf{X} - \mathbf{W} \times \mathbf{H}\|_F^2$ or Kullback–Leibler divergence of $\mathbf{O} = \|\mathbf{X} - \mathbf{W} \times \mathbf{H}\|_{KL}$ that correspond to Gaussian and Poisson mixed models, respectively. An additional advantage of the NMF method is that it can extract hidden patterns that are not independent but partially correlated.[40]

One of the difficulties of the NMF algorithm is that it requires prior knowledge of the number of latent features K. Recently, a new algorithm called NMFk addressing this limitation has been reported.[48,49,67] NMFk complements classical NMF with custom k-means clustering and Silhouette[68] statistics, which allows identification of the optimal number of unknown latent features (see the next paragraph). The NMFk was utilized for successful decomposition of the largest available data set of human cancer[47] genomes, for extraction of physical pressure transients[49] and contaminants[76] originating from an unknown number of sources that may propagate with a finite speed in nondispersive[77] or dispersive media,[78] and for extraction of the original crystal structures and phase diagram of X-ray spectra of material combinatorial libraries.[79]

Silhouette-Reconstruction Criterion for Estimation the Number of Latent Features. NMFk determines the number of the unknown number of hidden features based on the robustness of the NMF solutions and accuracy of the reconstruction of the original data-matrix, which is described in detail elsewhere.[67] Specifically, NMFk first constructs an ensemble of contact matrices where each element $x_{ij}^p \in \mathbf{X}_L^p(t)$ is defined by the original element $x_{ij} \in \mathbf{X}_L(t)$ "shuffled" by a small random deviation $x_{ij}^p = x_{ij} + \Delta_{ij}^p$.

In this way, the ensemble of the initial matrices $\{\mathbf{X}_L{}^1(t), \mathbf{X}_L{}^2(t), ..., \mathbf{X}_L^M(t)\}$ is constructed with Poisson resampling and $\mathbf{X}_L^p(t) \in \mathrm{Pois}(\mathbf{X}_L(t))$, if the Kullback–Leibler divergence[80] is used for the NMF's minimizations, or with Gaussian resampling, if the Frobenius norm is used for the NMF's minimizations (in this case, $\mathbf{X}_L^p(t)$ is a Gaussian distributed variable). In the Gaussian case, we use a small deviation of ~5% of the original value and preserve the nonnegativity of all elements. Further, NMF explores consecutively all possible numbers of latent features ($\tilde{K}$ can go from 1 to $N - 1$, where N is the total number of frames) by obtaining sets of NMF-minimization solutions for each explored $\tilde{K}$. Thus, for each explored $\tilde{K}$, we derived a set of solutions obtained with different (random) initial conditions in the minimization of the set of the contact matrices $\{\mathbf{W}_{\tilde{K}}{}^1; \mathbf{H}_{\tilde{K}}{}^1, \mathbf{W}_{\tilde{K}}{}^2; \mathbf{H}_{\tilde{K}}{}^2, ..., \mathbf{W}_{\tilde{K}}^M; \mathbf{H}_{\tilde{K}}^M\}$. Further, NMFk leverages a custom clustering on the columns of a set of $\mathbf{W}$'s $\{\mathbf{W}_{\tilde{K}}{}^1, \mathbf{W}_{\tilde{K}}{}^2, ..., \mathbf{W}_{\tilde{K}}^M\}$, utilizing cosine similarity as a metrics, which is the natural choice for similarity between nonnegative vectors.[81] Comparing the quality of the derived $\tilde{K}$ clusters, that is, the cohesion and separability of the clusters, measured via the standard Silhouette statistics,[68] together with the accuracy of the minimization (among the sets with various $\tilde{K}$), NMFk determines the optimal numbers of unknown hidden patterns. Thus, NMFk utilizes average Silhouette width S to measure how good the clusters are and hence how stable the solutions are under the influence of small perturbations on the initial contact matrix, for a particular choice of $\tilde{K}$. Specifically, the optimal number of patterns is picked by selecting the value of $\tilde{K}$ that leads to both (a) an acceptable mean of the reconstruction errors R^p where

$$R^p = \frac{\|\mathbf{X}^p - \mathbf{W}^p \times \mathbf{H}^p\|_{KL}}{\|\mathbf{X}^p\|_{KL}} \tag{2}$$

and (b) a large average Silhouette width (i.e., S close to 1, which means stable clusters). The combination of these two criteria, which we call a Silhouette-Reconstruction criterion, is easy to understand intuitively. For solutions with $\tilde{K}$ less than the actual number of patterns ($\tilde{K} < K_{\mathrm{true}}$), we expect the $\tilde{K}$ clusters to be well-separated and with a relatively good cohesion (i.e., with an average Silhouette width close to 1), because few of the actual clusters could be combined to produce one "super-cluster"; however, the reconstruction error will be high, due to the model being too constrained (with too few degrees of freedom) and thus on the under-fitting side. In the opposite limit of overfitting, that is, when $\tilde{K} > K_{\mathrm{true}}$ ($\tilde{K}$ exceeds the actual number of patterns), the average reconstruction error could be quite small—each solution reconstructs the contact matrix very well—but the solutions will not be well-clustered (e.g., with S less than 0.8) since there is no unique way to reconstruct with more than the actual number of clusters and hence no well-separated clusters will be formed.

Thus, the best estimate for the number of unknown hidden patterns is given by the value of $\tilde{K}$ that optimizes the above described Silhouette-Reconstruction criterion. Finally, after finding the optimal $\tilde{K}$, we use the medoids of the $\tilde{K}$ clusters as a final robust representation of the latent features.

NMFk Minimization Algorithm. Here, we leveraged the multiplicative algorithm[35] based on the Kullback–Leibler divergence as well as the block coordinate descent algorithm[82] based on the Frobenius norm. We did not observe any significant differences between the results obtained via these two algorithms.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.9b00074.

COORD.gro with the coordinates; INPUT.mdp with the parameters set to start the simulation with Gromacs; PARAMETERS.top with the Martini force field based topology for the membrane system; DPPC-DOPC-1.1 nm.xvg with an example matrix generated from the MD trajectories used for NMFk calculation. NMFk is based on the SigProfile software created for identification of mutational signatures in human cancer.[47] The SigProfile code is freely available at https://www.mathworks.com/matlabcentral. To use SigProfile, an input file should be at place. In our case, the input file is the contact matrix $\mathbf{X}_L(t)$ with size $(L \times M)$, where L is the number of lipids

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343–6357

in the MD simulations and M is the number of frames. A detailed description of NMFk is available elsewhere.[49] Input data file, containing the contact matrix $\mathbf{X}_L(t)$ as well as a script needed to run the SigProfile; README.docx file with a list of included files and links to publicly available repositories along with their brief description and instructions (ZIP)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: gnana@lanl.gov (S.G.).
*E-mail: boian@lanl.gov (B.S.A.).

ORCID

Cesar A. López: 0000-0003-4684-3364
Boian S. Alexandrov: 0000-0001-8636-4603

Author Contributions

All authors designed the research. C.A.L. performed the MD simulations. B.S.A. performed the NMFk analyses. All authors analyzed the data and wrote the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was performed at the Los Alamos National Laboratory and carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy under Contract No. 89233218CNA000001. We like to acknowledge Tyler Reddy for generating synthetic data to test preliminary results. The work was also supported by LANL LDRD grant 20180060DR and LDRD grant 20190020DR. This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) Program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, Oak Ridge National Laboratory under Contract DE-AC05-00OR22725, and Frederick National Laboratory for Cancer Research under Contract HHSN261200800001E. We thank the LANL Institutional Computing for the computing resources.

REFERENCES

(1) Nickels, J. D.; Chatterjee, S.; Stanley, C. B.; Qian, S.; Cheng, X.; Myles, D. A. A.; Standaert, R. F.; Elkins, J. G.; Katsaras, J. The in vivo structure of biological membranes and evidence for lipid domains. *PLoS Biol.* **2017**, *15*, No. e2002214.

(2) Kraft, M. L. Sphingolipid Organization in the Plasma Membrane and the Mechanisms That Influence It. *Front. Cell Dev. Biol.* **2017**, *4*, 154.

(3) Rosetti, C. M.; Mangiarotti, A.; Wilke, N. Sizes of lipid domains: What do we know from artificial lipid membranes? What are the possible shared features with membrane rafts in cells? *Biochim. Biophys. Acta* **2017**, *1859*, 789−802.

(4) Levental, K. R.; Lorent, J. H.; Lin, X.; Skinkle, A. D.; Surma, M. A.; Stockenbojer, E. A.; Gorfe, A. A.; Levental, I. Polyunsaturated Lipids Regulate Membrane Domain Stability by Tuning Membrane Order. *Biophys. J.* **2016**, *110*, 1800−1810.

(5) Dietrich, C.; Volovyk, Z. N.; Levi, M.; Thompson, N. L.; Jacobson, K. Partitioning of Thy-1, GM1, and cross-linked phospholipid analogs into lipid rafts reconstituted in supported model membrane monolayers. *Proc. Natl. Acad. Sci.* **2001**, *98*, 10642−10647.

(6) Van Meer, G.; Voelker, D. R.; Feigenson, G. W. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 112.

(7) Veatch, S. L.; Keller, S. L. Separation of liquid phases in giant vesicles of ternary mixtures of phospholipids and cholesterol. *Biophys. J.* **2003**, *85*, 3074−83.

(8) Davis, R. S.; Sunil Kumar, P. B.; Sperotto, M. M.; Laradji, M. Predictions of phase separation in three-component lipid membranes by the MARTINI force field. *J. Phys. Chem. B* **2013**, *117*, 4072−80.

(9) Risselada, H. J.; Marrink, S. J. The molecular face of lipid rafts in model membranes. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17367−72.

(10) Ingólfsson, H. I.; Lopez, C. A.; Uusitalo, J. J.; de Jong, D. H.; Gopal, S. M.; Periole, X.; Marrink, S. J. The power of coarse graining in biomolecular simulations. *WIREs. Comput. Mol. Sci.* **2014**, *4*, 225−248.

(11) Michalski, R. S.; Carbonell, J. G.; Mitchell, T. M. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media: 2013.

(12) Broecker, P.; Carrasquilla, J.; Melko, R. G.; Trebst, S. Machine learning quantum phases of matter beyond the fermion sign problem. *Sci. Rep.* **2017**, *7*, 8823.

(13) Tanaka, A.; Tomiya, A. Detection of phase transition via convolutional neural networks. *J. Phys. Soc. Jpn.* **2017**, *86*, No. 063001.

(14) Zhang, Y.; Kim, E.-A. Quantum loop topography for machine learning. *Phys. Rev. Lett.* **2017**, *118*, 216401.

(15) Wang, L. Discovering phase transitions with unsupervised learning. *Phys. Rev. B* **2016**, *94*, 195105.

(16) Carrasquilla, J.; Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **2017**, *13*, 431.

(17) Zhang, P.; Shen, H.; Zhai, H. Machine learning topological invariants with neural networks. *Phys. Rev. Lett.* **2018**, *120*, No. 066401.

(18) Hu, W.; Singh, R. R. P.; Scalettar, R. T. Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination. *Phys. Rev. E* **2017**, *95*, No. 062122.

(19) Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* **2017**, *96*, No. 022140.

(20) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726−741.

(21) García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.

(22) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7426−7431.

(23) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−23.

(24) Glazer, D. S.; Radmer, R. J.; Altman, R. B., Combining molecular dynamics and machine learning to improve protein function recognition. In *Biocomputing 2008*; World Scientific: 2008, 332−343.

(25) Karamzadeh, R.; Karimi-Jafari, M. H.; Sharifi-Zarchi, A.; Chitsaz, H.; Salekdeh, G. H.; Moosavi-Movahedi, A. A. Machine Learning and Network Analysis of Molecular Dynamics Trajectories Reveal Two Chains of Red/Ox-specific Residue Interactions in Human Protein Disulfide Isomerase. *Sci. Rep.* **2017**, *7*, 3666.

(26) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

(27) Ash, J.; Fourches, D. Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *J. Chem. Inf. Model.* **2017**, *57*, 1286−1299.

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

(28) Sgourakis, N. G.; Merced-Serrano, M.; Boutsidis, C.; Drineas, P.; Du, Z.; Wang, C.; Garcia, A. E. Atomic-Level Characterization of the Ensemble of the Aβ(1−42) Monomer in Water Using Unbiased Molecular Dynamics Simulations and Spectral Algorithms. *J. Mol. Biol.* **2011**, *405*, 570−583.
(29) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100−108.
(30) Kohonen, T. The self-organizing map. *Neurocomputing* **1998**, *21*, 1−6.
(31) Jolliffe, I. *Principal component analysis*. Wiley Online Library: 2002.
(32) Golub, G. H.; Reinsch, C. Singular value decomposition and least squares solutions. *Numerische mathematik* **1970**, *14*, 403−420.
(33) Cichocki, A.; Yang, H. H. A new learning algorithm for blind signal separation. *Advances in neural information processing systems* **1996**, *8*, 757−763.
(34) Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, *5*, 111−126.
(35) Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788−791.
(36) Baum, L. E.; Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **1966**, *37*, 1554−1563.
(37) Voronoi, G Nouvelles applications des paramètres continus á la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die reine und angewandte Mathematik* **1908**, 97−178.
(38) Ord, J. K.; Getis, A. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis* **1995**, *27*, 286−306.
(39) Park, S.; Im, W. Analysis of Lipid Order States and Domains in Lipid Bilayer Simulations. *J. Chem. Theory Comput.* **2018**, *15*, 688−697.
(40) Cichocki, A.; Zdunek, R.; Phan, A. H.; Amari, S.-i. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons: 2009.
(41) Févotte, C.; Cemgil, A. T. Nonnegative matrix factorizations as probabilistic inference in composite models. In *2009 17th European Signal Processing Conference*; IEEE: 2009, 1913−1917.
(42) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via theEMAlgorithm. *J. R. Stat. Soc.* **1977**, *39*, 1−22.
(43) Cappé, O.; Moulines, E.; Rydén, T. *Inference in Hidden Markov Models*; Springer: New York, 2005.
(44) Yılmaz, Y. K.; Cemgil, A. T. Algorithms for probabilistic latent tensor factorization. *Signal Process.* **2012**, *92*, 1853−1863.
(45) Kubjas, K.; Robeva, E.; Sturmfels, B. Fixed points of the EM algorithm and nonnegative rank *boundaries*. *Ann, Stat.* **2015**, 422−461.
(46) Chakraborty, C.; Talukdar, P. H. Issues and limitations of HMM in speech processing: a survey. *Int. J. Comput. Appl.* **2016**, *141*, 13−17.
(47) Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Aparicio, S. A.; Behjati, S.; Biankin, A. V.; Bignell, G. R.; Bolli, N.; Borg, A.; Børresen-Dale, A.-L.; Boyault, S.; Burkhardt, B.; Butler, A. P.; Caldas, C.; Davies, H. R.; Desmedt, C.; Eils, R.; Eyfjörd, J. E.; Foekens, J. A.; Greaves, M.; Hosoda, F.; Hutter, B.; Ilicic, T.; Imbeaud, S.; Imielinski, M.; Jäger, N.; Jones, D. T. W.; Jones, D.; Knappskog, S.; Kool, M.; Lakhani, S. R.; Löpez-Otín, C.; Martin, S.; Munshi, N. C.; Nakamura, H.; Northcott, P. A.; Pajic, M.; Papaemmanuil, E.; Paradiso, A.; Pearson, J. V.; Puente, X. S.; Raine, K.; Ramakrishna, M.; Richardson, A. L.; Richter, J.; Rosenstiel, P.; Schlesner, M.; Schumacher, T. N.; Span, P. N.; Teague, J. W.; Totoki, Y.; Tutt, A. N. J.; Valdés-Mas, R.; van Buuren, M. M.; van't Veer, L.; Vincent-Salomon, A.; Waddell, N.; Yates, R. L.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain; Zucman-Rossi, J.; Futreal, P. A.; McDermott, U.; Lichter, P.; Meyerson, M.; Grimmond, S. M.; Siebert, R.; Campo, E.; Shibata, T.; Pfister, S. M.; Campbell, P. J.; Stratton, M. R. Signatures of mutational processes in human cancer. *Nature* **2013**, *500*, 415.
(48) Alexandrov, B. S.; Alexandrov, L. B.; Iliev, F. L.; Stanev, V. G.; Vesselinov, V. V. Source identification by non-negative matrix factorization combined with semi-supervised clustering. US20180060758A1, 2018.
(49) Alexandrov, B. S.; Vesselinov, V. V. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. *Water Resour. Res.* **2014**, *50*, 7332−7347.
(50) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812−24.
(51) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42*, 6801−6822.
(52) Vögele, M.; Köfinger, J.; Hummer, G. Hydrodynamics of Diffusion in Lipid Membrane Simulations. *Phys. Rev. Lett.* **2018**, *120*, 268104.
(53) Domański, J.; Marrink, S. J.; Schäfer, L. V. Transmembrane helices can induce domain formation in crowded model membranes. *Biochim. Biophys. Acta* **2012**, *1818*, 984−994.
(54) Schafer, L. V.; de Jong, D. H.; Holt, A.; Rzepiela, A. J.; de Vries, A. H.; Poolman, B.; Killian, J. A.; Marrink, S. J. Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1343−8.
(55) Klingelhoefer, J. W.; Carpenter, T.; Sansom, M. S. P. Peptide nanopores and lipid bilayers: interactions by coarse-grained molecular-dynamics simulations. *Biophys. J.* **2009**, *96*, 3519−3528.
(56) Ingólfsson, H. I.; Melo, M. N.; van Eerden, F. J.; Arnarez, C.; Lopez, C. A.; Wassenaar, T. A.; Periole, X.; de Vries, A. H.; Tieleman, D. P.; Marrink, S. J. Lipid organization of the plasma membrane. *J. Am. Chem. Soc.* **2014**, *136*, 14554−14559.
(57) Ingólfsson, H. I.; Carpenter, T. S.; Bhatia, H.; Bremer, P.-T.; Marrink, S. J.; Lightstone, F. C. Computational Lipidomics of the Neuronal Plasma Membrane. *Biophys. J.* **2017**, *113*, 2271−2280.
(58) Löpez, C. A.; Sovova, Z.; van Eerden, F. J.; de Vries, A. H.; Marrink, S. J. Martini Force Field Parameters for Glycolipids. *J. Chem. Theory Comput.* **2013**, *9*, 1694−708.
(59) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108*, 750−760.
(60) Khelashvili, G.; Kollmitzer, B.; Heftberger, P.; Pabst, G.; Harries, D. Calculating the Bending Modulus for Multicomponent Lipid Membranes in Different Thermodynamic Phases. *J. Chem. Theory Comput.* **2013**, *9*, 3866−3871.
(61) Baoukina, S.; Mendez-Villuendas, E.; Tieleman, D. P. Molecular view of phase coexistence in lipid monolayers. *J. Am. Chem. Soc.* **2012**, *134*, 17543−53.
(62) Carpenter, T. S.; Löpez, C. A.; Neale, C.; Montour, C.; Ingólfsson, H. I.; di Natale, F.; Lightstone, F. C.; Gnanakaran, S. Capturing Phase Behavior of Ternary Lipid Mixtures with a Refined Martini Coarse-Grained Force Field. *J. Chem. Theory Comput.* **2018**, *14*, 6050.
(63) Heberle, F. A.; Feigenson, G. W. Phase separation in lipid membranes. *Cold Spring Harbor Perspect. Biol.* **2011**, *3*, a004630.
(64) Sodt, A. J.; Pastor, R. W.; Lyman, E. Hexagonal Substructure and Hydrogen Bonding in Liquid-Ordered Phases Containing Palmitoyl Sphingomyelin. *Biophys. J.* **2015**, *109*, 948−55.
(65) Simons, K.; Sampaio, J. L. Membrane organization and lipid rafts. *Cold Spring Harbor Perspect. Biol.* **2011**, *3*, a004697.
(66) Sodt, A. J.; Sandar, M. L.; Gawrisch, K.; Pastor, R. W.; Lyman, E. The molecular structure of the liquid-ordered phase of lipid bilayers. *J. Am. Chem. Soc.* **2014**, *136*, 725−732.
(67) Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Campbell, P. J.; Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **2013**, *3*, 246−259.
(68) Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357

(69) Alexandrov, B. S.; Stanev, V. G.; Vesselinov, V. V.; Rasmussen, K. Ø. Nonnegative tensor decomposition with custom clustering for microphase separation of block copolymers. *Stat. Anal. Data Min.* **2019**, *12*, 11407.
(70) García-Sáez, A. J.; Chiantia, S.; Schwille, P. Effect of line tension on the lateral organization of lipid membranes. *J. Biol. Chem.* **2007**, *282*, 33537−33544.
(71) de Jong, D. H.; Baoukina, S.; Ingólfsson, H. I.; Marrink, S. J. Martini straight: Boosting performance using a shorter cutoff and GPUs. *Comput. Phys. Commun.* **2016**, *199*, 1−7.
(72) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, No. 014101.
(73) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(74) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(75) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19−25.
(76) Vesselinov, V. V.; Alexandrov, B. S.; O'Malley, D. Contaminant source identification using semi-supervised machine learning. *J. Contam. Hydrol.* **2018**, *212*, 134.
(77) Iliev, F. L.; Stanev, V. G.; Vesselinov, V. V.; Alexandrov, B. S. Nonnegative Matrix Factorization for identification of unknown number of sources emitting delayed signals. *PLoS One* **2018**, *13*, No. e0193974.
(78) Stanev, V. G.; Iliev, F. L.; Hansen, S.; Vesselinov, V. V.; Alexandrov, B. S. Identification of release sources in advection-diffusion system by machine learning combined with Green's function inverse method. *Appl. Math. Modell.* **2018**, *60*, 64.
(79) Stanev, V.; Vesselinov, V. V.; Kusne, A. G.; Antoszewski, G.; Takeuchi, I.; Alexandrov, B. S. Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering. *npj Comput. Mater.* **2018**, *4*, 43.
(80) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. stat.* **1951**, *22*, 79−86.
(81) Korenius, T.; Laurikkala, J.; Juhola, M. On principal component analysis, cosine and Euclidean measures in information retrieval. *Inf. Sci.* **2007**, *177*, 4893−4905.
(82) Xu, Y.; Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* **2013**, *6*, 1758−1789.

DOI: 10.1021/acs.jctc.9b00074
J. Chem. Theory Comput. 2019, 15, 6343−6357