Topological Learning Approach to Characterizing Biological Membranes

Andres S. Arango,[†,‡] Hyun Park,[†,‡] and Emad Tajkhorshid[*,†]

†*Theoretical and Computational Biophysics Group, NIH Resource Center for Macromolecular Modeling and Visualization, Beckman Institute for Advanced Science and Technology, Department of Biochemistry, and Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

‡*Contributed equally to this work*

E-mail: emad@illinois.edu

Abstract

Biological membranes play key roles in cellular compartmentalization, structure, and its signaling pathways. At varying temperatures, individual membrane lipids sample from different configurations, a process that frequently leads to higher-order phase behavior and phenomena. Here we present a persistent homology-based method for quantifying the structural features of individual and bulk lipids, providing local and contextual information on lipid tail organization. Our method leverages the mathematical machinery of algebraic topology and machine learning to infer temperature-dependent structural information of lipids from static coordinates. To train our model, we generated multiple molecular dynamics trajectories of DPPC membranes at varying temperatures. A fingerprint was then constructed for each set of lipid coordinates by a persistent homology filtration, in which interactions spheres were grown around the lipid atoms while tracking their intersections. The sphere filtration formed a *simplicial complex* that captures enduring key *topological features* of the configuration landscape, using homology, yielding *persistence data*. Following fingerprint extraction for physiologically relevant temperatures, the persistence data were used to train an attention-based neural network for assignment of effective temperature values to selected membrane regions. Our persistence homology-based method captures the local structural effects, via effective temperature, of lipids adjacent to other membrane constituents, e.g. sterols and proteins. This topological learning approach can predict lipid effective temperatures from static coordinates across multiple spatial resolutions. The tool, called `MembTDA`, can be accessed at `https://github.com/hyunp2/Memb-TDA`.

1 Introduction

Biological membranes are crucial for cellular compartmentalization and structural integrity, as well as act a major platform for signaling pathways that govern environmental response.[1] Membranes also serve as the primary boundary between pathogens and internal cellular compartments, thus being essential in both physiological and pathophysiological cellular response.[2,3] Lipid membranes provide a structural context for functional protein conformations, for both peripheral and transmembrane proteins.[4–6] Membrane composition is ubiquitously heterogeneous, consisting of variable lipids, sterols, and proteins.[7] Individual lipids can sample from multiple acyl tail configurations, depending on local environment, pressure, and temperature, leading to distinct membrane properties.[8–10]

In the case of homogeneous lipid compositions, membranes experience temperature-dependent, higher order phase phenomena, for example, transformation between the gel and liquid phases, as a result of shifting lipid acyl tail configurational energy basins. Lipid order parameters (S_{CH}/S_{CD}) serve as a way to capture configurational properties of lipids.[11] The order parameter for an acyl lipid tail is calculated using the angle, θ, formed between the bilayer normal and the carbon-hydrogen, or carbon-deuterium, bond vector:

$$S_{CH} = \langle 3\ \cos^2\theta - 1 \rangle / 2,$$

where the angular brackets, $\langle ... \rangle$, represent temporal/molecular ensemble averages.[11] Through molecular dynamics (MD) simulations, one can capture atomic representations of individual lipid configurations. When used in conjunction, order parameter calculations from MD trajectories can help refine force fields and provide bulk information on phase behavior.[12]

Here we present a novel method for characterizing lipid order using a topological learning approach with MD, as an alternative to S_{CD} calculations. Our approach learns the underlying temperature-dependent, potential energy surface from a wide range of lipid configurations, obeying a Boltzmann distribution sampled from equilibrium MD simulations, providing an effective temperature estimate for individual lipids (Figure 1).

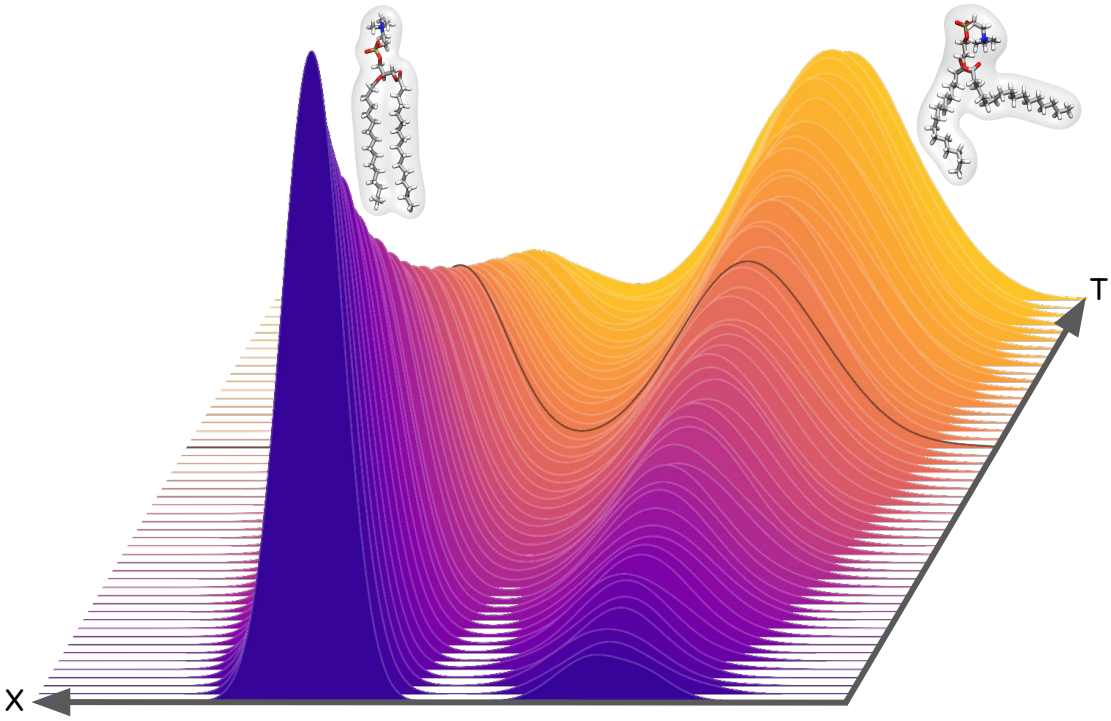

Figure 1: Schematic view of the configurational manifold mapped by `MembTDA`. Here we show configuration space X versus MD input temperatures T. The height represents a configuration probability density, with the phase transition temperature shown as a black line where two prominent configurations are of equal probability. `MembTDA` maps out this surface and infers a distribution of likely configurational temperatures, based on potential energy features derived from static coordinates, for which an expectation value yields an effective temperature.

Our method, named `MembTDA`, leverages topology, a branch of mathematics concerning sets that contain an inherent structure that is preserved under continuous deformations. Topological data analysis (TDA)[13] has rapidly become one of the main tools used by artificial intelligence researchers to agnostically extract feature information from various data sets. TDA focuses on the inherent topological features of the data, to create a topological fingerprint represented as bar codes, diagrams, or images (Figure 2). One clear benefit of TDA is the ability to cluster data consistently over other convergence-based methods, as well as being resilient to perturbations, specifically with the use of TDA techniques like persistent homology (PH).[14,15] PH is a method of extracting both geometric and topological information from a point cloud, like atomic coordinates. Major benefits of the PH method-

ology include robustness under perturbations, characterization of points clouds of varying densities, and the potential for abstract feature comparison via the Wasserstein distance.[14–16] The robustness of PH lends itself nicely to the complex nature of biological data analysis. In the case of biological membranes, lipids can have variable acyl tails as well as variable headgroups;[17] these variable factors and underlying constant features can be captured using PH even under perturbative effects, like dynamics. Capturing conserved topological features of membrane data using persistence homology, allows us to train neural networks for biophysical and thermodynamic feature prediction, like temperature.

Machine learning (ML) or deep neural network can be used to extract hidden patterns that are usually hard to detect using conventional techniques. These include, but not limited to, predicting protein structures to predicting qualitative or quantitative molecular properties.[18–23] In this work we train a deep neural network model specialized in processing images, such as Visual Transformer (`ViT`)[24,25] or `ConvNeXt`,[26] with persistent data information of lipid coordinates at varying temperatures. By marrying ML with TDA, we demonstrate our model's capability in predicting individual lipid's effective temperatures. `MembTDA` maps out the configurational manifold of lipids and allows for inferring an effective temperature from a distribution of likely configurational temperatures Figure 1.

2 Methods

Our approach is comprised of three major elements. First, we create a molecular data set for lipids at varying temperatures using molecular dynamics (MD) simulations. Then we featurize the MD data using PH (specifically, persistence images). Lastly, we train an attention based transformer using the data from PH and MD.

2.1 Molecular Dynamics Simulations

A major part of model development is data accumulation. Our data set consisted of trajectories from MD simulations of membranes in 51 different temperatures ranging 280–330 K, spaced apart by 1 K. The model lipid bilayer consisted of 117 dipalmitoyl-phosphatidylcholine

(DPPC) lipids per leaflet and was constructed using the CHARMM-GUI webserver.[27] The system was solvated and ionized with 150 mM NaCl to mimic cellular conditions. The CHARMM36m[28] force field parameters and the TIP3P water model[29] were used for all the simulations. The equilibration of the systems was performed using the NAMD.[30,31] A 100-ns production run was then performed for each temperature replica with GPU-resident NAMD to ensure optimal GPU scaling.[31] Observed gel and liquid phase transitions in the simulated bilayers were captured within the first 20 ns of each simulation replica. Only the last 200 frames (2 ns) were used for model training to minimize the effects of the degenerate starting conditions. In total, the simulation sampling amounted to 100 ns per replica, approximately an aggregate of 5 μs sampling.

As in the equilibration runs, the production simulations were performed as an NPT ensemble at varying temperatures from 280 K to 330 K with a pressure of 1.0 atm. An integration time step of 2 fs was used throughout. The Nosé-Hoover Langevin piston method[32–34] was used to maintain a constant pressure, with temperature maintained constant via Langevin dynamics with a 0.5 ps^{-1} damping coefficient.[35,36] A 12-Å cutoff was used for nonbonded interactions with a smoothing function implemented after 10 Å. The bond distances of the hydrogen atoms were constrained using the SHAKE algorithm.[37] For long-range electrostatic calculations, the particle mesh Ewald (PME) method[38] was used, with a grid density of more than 1 $Å^{-3}$. Visualization and analyses of the simulations were done using Visual Molecular Dynamics (VMD)[39] and MDAnalysis.[40]

2.2 Persistent Homology

Persistent homology (PH) is a method of extracting both geometric and topological information from a simplicial complex constructed from a point cloud, like atomic coordinates. A simplicial complex K is a collection of simplices in $\mathbb{R}^n$. Simplices are topological descriptions of connected points, where a single point is a 0-simplex, two connected points form a 1-simplex, three points a 2-simplex, and so on (Figure 2). In the case of a point cloud, like

the coordinates of a phospholipid bilayer or a protein, we have a collection of 0-simplices, i.e., isolated points.

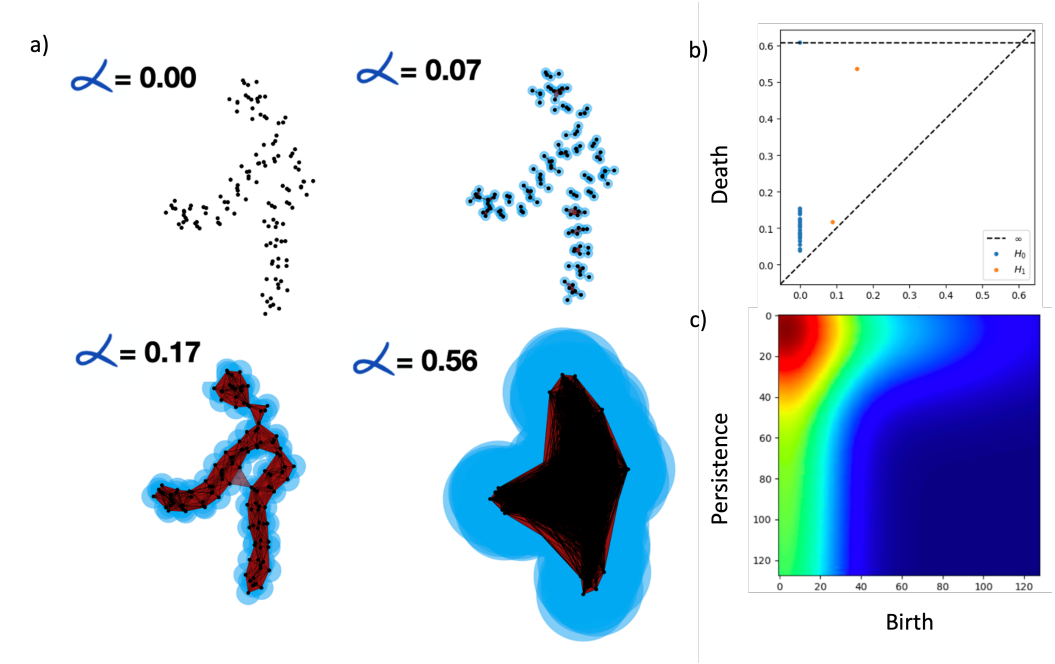

Figure 2: Overview of persistent homology (PH) calculations. a) To extract topological features via PH, we first construct a simplicial complex on the point clouds obtained from lipid atomic coordinates. A filtration value, α, is increased, which increases the radii of spheres surrounding every point in the point cloud. An example simplicial complex at varying filtration values is shown. b) The topological features in the simplicial complex can be represented as a persistence diagram depicting birth versus death of features. c) Plotting persistence (death minus birth) versus birth, and performing a Gaussian kernel approximation on the diagram yield an image that is amenable to visual transformers, called a persistence image.

In PH, we measure the unique fingerprint of a point cloud (atomic coordinates in the case of lipids) by tracking the topology of its simplicial complex at a varying filtration parameter α. To vary the filter parameter, we grow n-dimensional spheres around each vertex (individual point) of the point cloud. In our work, $n = 3$, because the points represent the atomic Cartesian coordinates of lipids. As the radius, the filtration parameter α, of the spheres increases, the ratios of n-simplices vary depending on the intrinsic structure of the data. We perform a filtration of the 0-simplicial complex by growing n-dimensional spheres around each point, starting with a radius of $\alpha = 0$ and fully disconnected 0-simplices. As we increase α, we keep track of overlaps between neighboring spheres such that two overlapping

spheres form a 1-simplex (i.e., an edge), trisecting spheres form a 2-simplex (i.e., a triangle face), and so on.

More symbolically, starting from a point cloud $\{x_1, x_2, ..., x_n\}$, as in the case of the Cartesian coordinates, we can construct a simplicial complex by growing n-dimensional spheres around each point and tracking sphere intersections. In the case of an α-complex construction,[41] we begin by creating a Voronoi partition, V_{x_i}, of our point cloud:

$$V_{x_i} = \{x : |x_i - x| \leq |x_j - x| \text{ for } i \neq j\}$$

Using the condition that our α-complex can only reside within our Voronoi partition, we then grow spheres, B_α, around each point:

$$B_\alpha(x_i) = \{x : |x_i - x| \leq \alpha \text{ and } x \in V_{x_i}\}$$

The process of growing spheres around each vertex is referred to as a filtration. Taken together, the intersection of our simplices at every radius α forms a simplicial complex, in this case an α-complex:

$$X_\alpha^k = \{(x_{i_1}, x_{i_2}, ..., x_{i_k}) : \bigcap_{j=1}^{k} B_\alpha(x_{i_j}) \neq 0\}$$

Our α-complex, X_α^k, contains information of the connectivity of our system at varying radii, α. We then identify holes in our data by applying homology to the α-complex.[42]

Ultimately, the homology of the varying filtration value α yields information on topological and geometric features. We can visualize the birth and death of the captured features obtained from PH from the varying filtration value using persistence barcodes, persistence diagrams, and persistence images, all of which are examples of persistence data Figure 2-a. Applying *homology* to the simplicial/α-complex, comprised of the simplices at all varying radii, yields topological features such as birth and death information of 0-, 1-, 2- homology groups (i.e., connected components, holes, and voids); we denote these homology groups H_0, H_1 and H_2, respectively. The topological features attained from PH are primarily represented visually as persistence diagrams and persistence images (Figure 2-b, c). In our work,

we interchangeably use terms PH, persistence data, topological fingerprints, or topological features, to refer to extracted topological data obtained from the TDA approach.

Although persistence barcodes and persistence diagrams are readily interpretable for humans, their sparse nature is less amenable to learn patterns for deep learning-based models such as visual transformers (`ViT`)[24] using the attention mechanism[43] or convolutional neural networks. To address the sparse nature of persistence diagrams, we transform them into persistence images,[42] a process called PH vectorization, which leverages Gaussian kernel approximations to create topological fingerprints that can be readily fed through a computer vision based deep neural network to learn complex patterns and predict properties. The reasons for choosing vision based deep neural network architectures are further elaborated in subsection 2.3.

Instead of birth versus death information, persistence images have persistence versus birth; where persistence is characterized by death minus birth.[42] Using every 10 ps of the last 2 ns of the membrane simulations at variable temperatures, coordinate data of individual lipids were used to form the persistence images Figure 3.

2.3 Deep Learning Model for Processing Persistence Images

In our work, we used `ViT` architecture for prediction of effective temperatures. The `ViT` architecture we used was based on window attention of `Swin Transformer version 2`[25].[44] The objective function is the temperature class prediction, $\mathcal{L}_{\text{CE}}(p_\theta(T_i), T_{\text{true}})$, and expected temperature prediction, $\mathcal{L}_{\text{MSE}}(\mathbf{E}_i, T_{\text{true}})$. By training `MembTDA` with bi-objective functions, we can optimize our neural network model for a more robust representation learning of our input. Our overall workflow is described in Figure 3 and elaborated further in Figure 5.

3 Results

Lipid bilayer phase transitions are readily characterized experimentally by metrics such as heat capacity and acyl tail order parameters. Our topological deep learning approach for characterizing lipids allows us to probe lipid phases for individual lipid molecules. We demon-

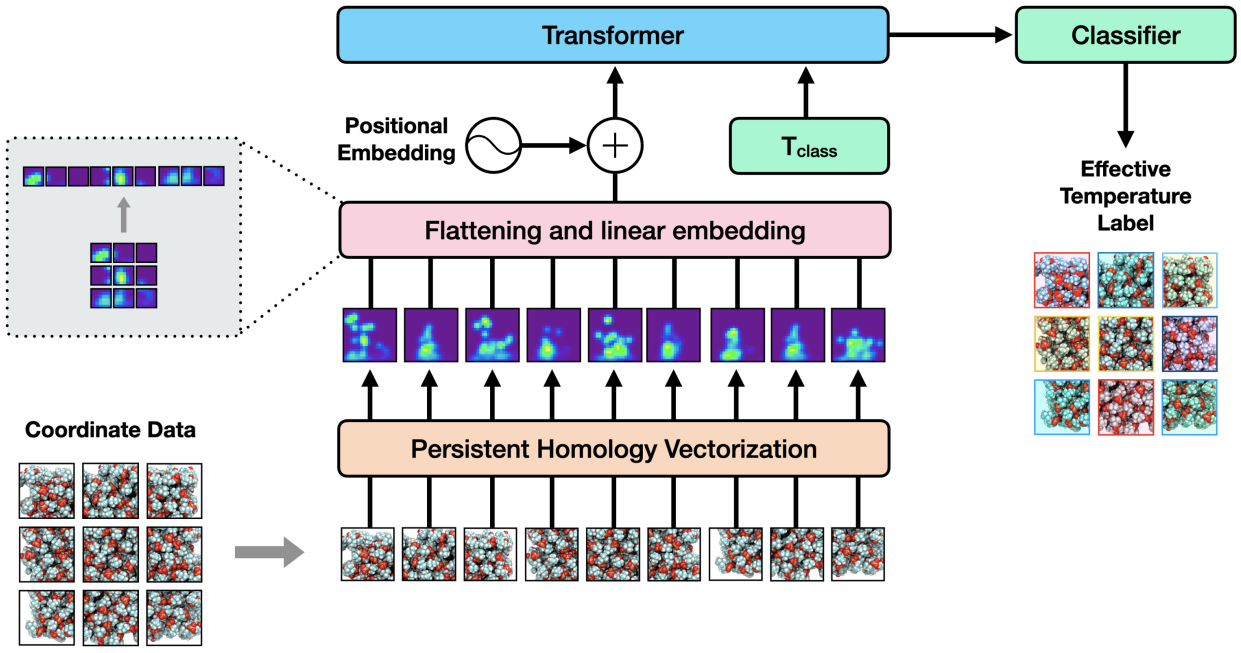

Figure 3: MembTDA overall workflow. First MD data are fed as an input to the workflow as coordinate patches, either as individual lipids or as patches. The atomic coordinates then undergo PH vectorization (diagram to image transformation), in which topological features are characterized and represented as persistence images. Each persistence image undergoes a flattening and linear embedding step, including patching and positional embedding. The embedded data are then fed through a transformer (`ViT`), with a temperature class label from the MD input temperature. The model ultimately acts as a classifier with a distribution of probability values for possible temperature classes, for which an expected value yields an effective temperature T_{E}.

strate the utility of `MembTDA` on homogeneous and heterogeneous membranes, including, lipid bilayers simulated at variable temperatures, and bilayers with transmembrane or peripheral protein systems.

3.1 Homogeneous Membranes at Variable Temperatures

To assess the ability of `MembTDA` to identify and classify lipid phases, we performed inference on the lipid training set and report the effective temperature distribution in Figure 4-b. Inference on the initial training set's effective temperatures, the distribution obtained by calculating expected values of `MembTDA` output classes, reveal a seemingly bimodal distribution with internal effective temperature minima at 297.29 K and 306.44 K.

According to Khakbaz et. al.,[45] 308.15 K is where L_α (crystalline liquid) to L_β/P_β

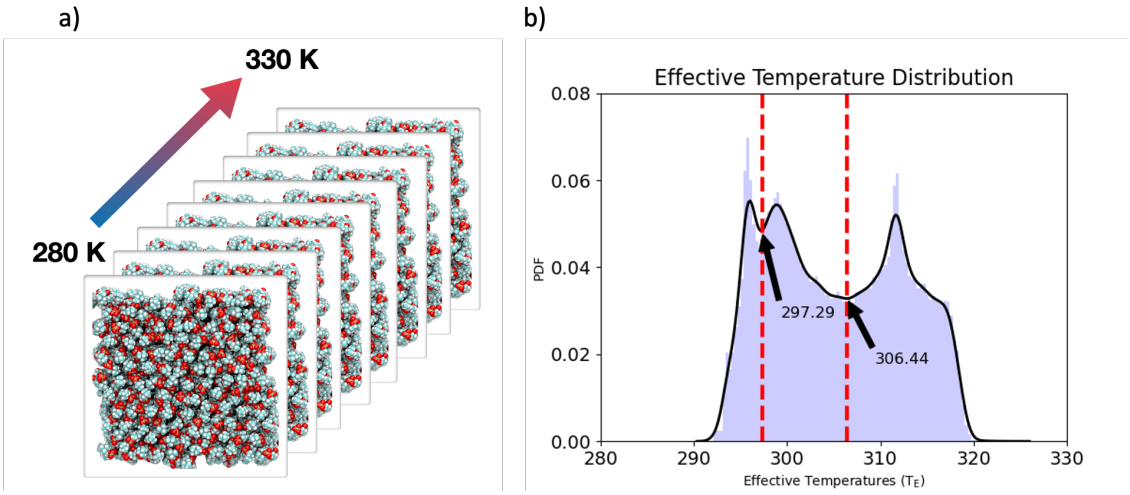

Figure 4: Inference of entire MD data set. a) A small subset of DPPC lipids from membrane simulations across 51 temperatures (280-330 K) was used to train the `ViT`-based `MembTDA`. b) Applying the trained `MembTDA` back on the training set to capture temperature class expectation values yielded a bimodal effective temperature distribution. `MembTDA` effectively classifies lipids as either gel or liquid phase. `MembTDA` also captures a L_α/L_β phase transition at 306.44 K, and acyl tail tilting at 297.29 K.

(gel/ripple, respectively) transition occurs, suggesting a sharp decrease in surface area at this temperature. In addition, according to the authors,[45] at 298.15 K, the L_β phase occurs with predominantly tilted acyl chains with respect to the membrane normal. Our MD simulation environment closely matched that used in Khakbaz et. al.[45] (i.e., NAMD,[30,31] Charmm36[46] parameters, DPPC lipids), from which `MembTDA` predicted two biophysically significant temperatures, purely from persistence data followed by neural network operation. The results demonstrate that `MembTDA` is robust at capturing this aspect of the lipid membranes, in this case phase transitions. The absolute errors between temperatures from Khakbaz et. al.[45] and `MembTDA` predictions of melting and chain tilting temperature is 1.7 K and 0.86 K, respectively.

3.2 Transmembrane protein in POPC – AQP5

Aquaporin 5 (AQP5), a protein involved in water homeostasis, was previously simulated in our lab in a 1-palmitoyl-2-oleoyl-*sn*-glycerol-3-phosphocholine (POPC) bilayer with a thermostat input temperature of 310 K. Unlike DPPC, POPC contains asymmetrical acyl tail

lengths, altering its melting temperature and membrane dynamics.[47] Inference on this system revealed an effective temperature distribution above the melting temperature of DPPC and POPC, indicative of highly disordered lipids, potentially due to protein-lipid interactions.

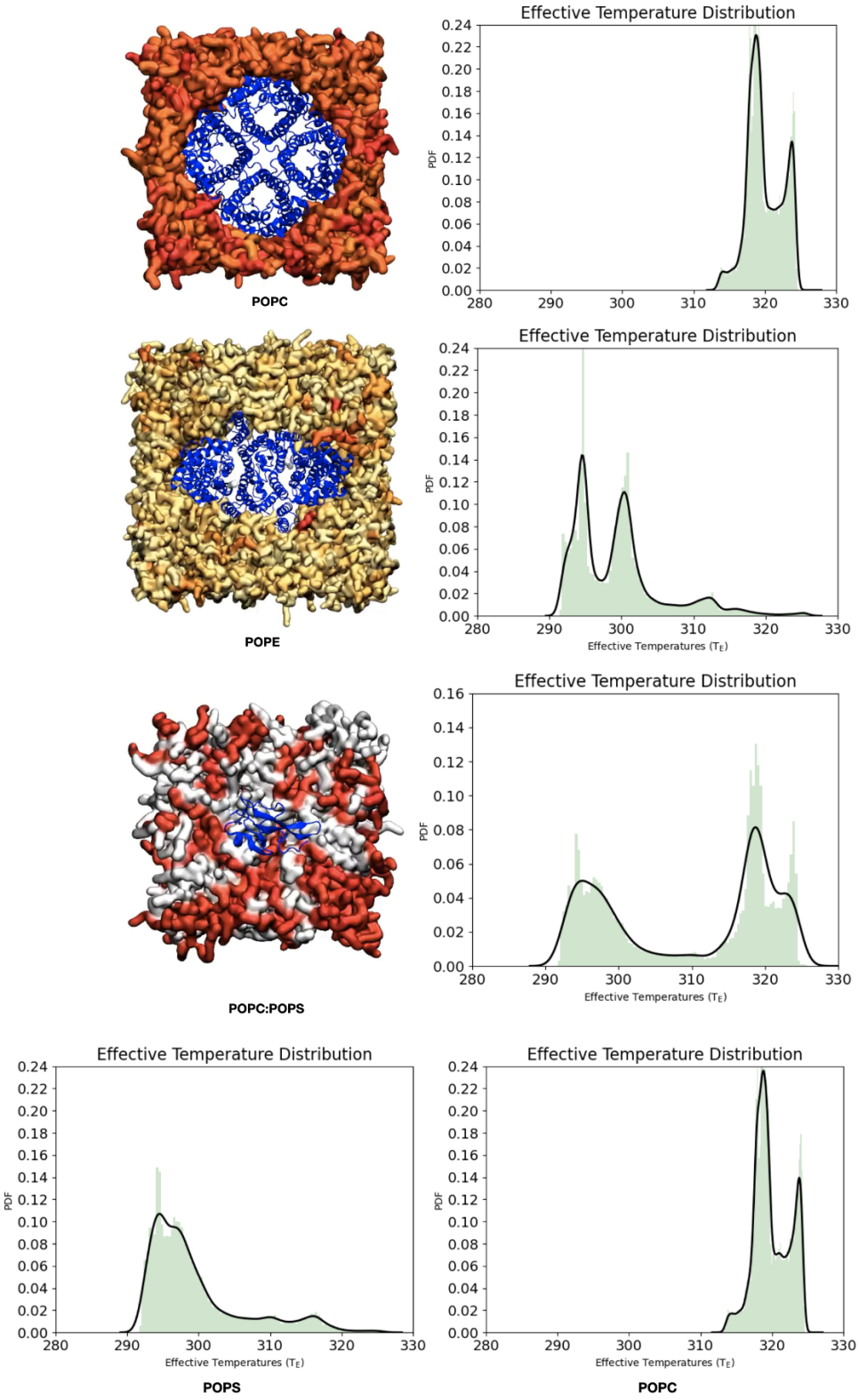

Figure 5: Inference on out of training data distribution. Effective temperature estimates for heterogeneous membranes containing varying proteins.

3.3 Transmembrane protein in POPE – LaINDY

LaINDY, a transmembrane bacterial transporter, was simulated previously in our lab in a 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine (POPE) membrane. Like POPC, POPE contains asymmetrical acyl tails, but also has a different head group (ethanolamine) than both DPPC and POPC. Inference on this system revealed two distinct peaks, potentially indicative of variable protein-lipid interactions between LaINDY and POPE. This analysis demonstrates the ability of `MembTDA` to perform inference on lipid head groups outside the training data.

3.4 Peripheral Membrane protein in POPC/POPS - β2GP1

β2GP1 a peripheral membrane protein known to bind 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-L-serine (POPS) lipids, was simulated previously in our lab in an equal ratio of POPC:POPS membrane at 310 K. Effective temperature inference on this heterogeneous membrane revealed two distinct distributions across the membrane. Upon decomposition of POPS and POPC effective temperatures, we identified that `MembTDA` is able to distinguish between POPC and POPS lipids. We find this remarkable since our model was trained purely on DPPC persistence data, and the input point cloud contained no information on charge, atom types, or connectivity. Furthermore, the low effective temperature distribution captured for POPS is interesting in the context of β2GP1 binding specificity, since β2GP1 is known to selectively bind POPS-rich membranes. Furthermore, we observe that the distribution of effective temperatures for POPC from the decomposed heterogeneous membranes, closely resembles the distribution observed for AQP5-embedded POPC.

In the Supplemental Figure S11 and Figure S12, we show the overall H_1 Wasserstein matching diagram and close-up view of the same diagram in the 0.2-0.4 birth-death range for a better demonstration. We chose 100 random lipids, each from three temperature categories, i.e. low, melting, and high temperatures, resulting in a total of 300 samples, calculated persistence diagrams and overlaid them with colors labeled in the legend (e.g.,

labeled "all lower temps" with an "o" marker). Then we calculate barycenters of each temperature categories 100 sample persistence diagram birth-death points (e.g. labeled "lower temp", with an "X" marker). Barycenters are centroids of non-linear systems such as birth-death points of persistence diagrams.[48] The Wasserstein distance is a metric to compare the similarity between two persistence diagrams. The calculation of the Wasserstein distance discerns which points of one diagram are similar to those of another diagram, hence *matching* birth-death points between two diagrams. These *matching* points are represented as an edge connecting two barycenter points in supplemental Figure S12 (i.e. blue "X" to yellow "X" and to red "X").

From the Wasserstein diagrams, we can conclude that there are prominent H_1 features (holes) appearing and persisting about 0.2 α radius. When observed closer (supplemental Figure S12), we can see that barycenters of three different temperature categories persistence diagrams show movements, indicated by edge connections. This shows that there distinct persistence data encoded across three lipid temperature categories, indicating a temperature dependent flow of topological features. This is consistent with the idea that `MembTDA` maps out the underlying configurational landscape as seen in Figure 1.

4 Discussion

We present a novel lipid characterization method, termed `MembTDA`, that estimates effective temperature of lipids from static coordinates, as a topological alternative to the commonly used S_{CH}/S_{CD} order parameter calculations. We demonstrate `MembTDA`'s effectiveness as a functional classifier by performing inference on homogeneous and heterogeneous membranes, with variable tails (e.g., DPPC, POPC), head groups (e.g., POPC, POPE, and POPS), and in the presence of proteins (e.g., AQP5, LaINDY, and β2GP1). Although our methodology functionally acts as a classifier, the reliance on coordinate data, i.e., point cloud information, necessarily implies that the method is inherently based on potential energy features, which are functions of xyz-coordinates. Furthermore, since temperature is predominately calcu-

lated as a kinetic energy parameter from the equipartition theorem,[49] ($\Sigma_i^N \frac{1}{2} m_i v_i^2 = \frac{3}{2} N k_\mathrm{B} T$), the effective temperatures of static snapshots we report are not exactly representative of traditional bulk temperature, an *ensemble* property. The effective temperature we report is a quantity obtained by taking the expected value of the `MembTDA` output distribution, which is based on a mapping of the configurational landscape with an associated input MD temperature (Figure 4).

More accurate predictions of the local temperature may be possible by taking velocity information from simulation, and using the equipartition theorem, $\langle E_K \rangle_N = \frac{3}{2} k_\mathrm{B} T$, to estimate individual lipid temperatures. Typically the average kinetic energy of the entire simulation system, including all non-lipid constituents, is used to calculate the system's temperature:

$$T = \frac{2\langle E_K \rangle_N}{k_\mathrm{B}},$$

accounting for degrees of freedom N as 3N-3 (Supplemental Figure S2, Supplemental Figure S3, Supplemental Figure S4). Due to the variable molecular degrees of freedom arising from differing intra-molecular (including the number of atoms, bonds, angles, and dihedrals) and inter-molecular interactions between lipids (contextual information such as lipid-lipid or lipid-protein association), the use of the equipartition theorem to estimate individual lipid temperatures based on their kinetic energy is non-trivial.[50] Estimates of individual lipid temperatures using kinetic energy formulations of temperature yield individual lipid temperatures, sometimes over 50 K below the global system temperature (Supplemental Figure S5, Supplemental Figure S6, Supplemental Figure S7). Alternatively to a kinetic energy-based estimate for temperature, a configurational temperature may be also calculated using the Jepps formulation:

$$k_\mathrm{B} T_\mathrm{config} = \frac{\langle \nabla U \cdot \nabla U \rangle}{\langle \nabla \cdot \nabla U \rangle},$$

where U is the system's potential energy and $\langle ... \rangle$ denotes an ensemble average.[51] Although based on the potential energy, calculation of the configurational temperature from the Jepps formulation ultimately requires computationally expensive Hessian calculations. The term

$\langle \nabla \cdot \nabla U \rangle$ necessitates calculating the divergence of the gradient of the potential energy, which requires a Hessian and degrees of freedom information, making accurate configurational estimates prohibitive.[51,52] Furthermore, in practice, MD users primarily only store atomic coordinates to save on storage and computation costs; only writing out velocity information for restart purposes.[53] `MembTDA` provides a pragmatic way to estimate effective temperatures from data typically saved by MD practitioners, allowing post-processing of extant simulation trajectories. In addition, the ability to analyze static coordinates lends to potential applications in analyzing effective temperatures in novel structural data.

`MembTDA` was originally trained on DPPC membrane structures across different temperatures. We demonstrate that the so trained `MembTDA` can capture two important temperatures of DPPC membranes, namely chain tilting and melting temperatures as mentioned in subsection 3.1. This result is particularly interesting since `MembTDA` was only given persistence data to classify into one of the 51 temperature classes, and was trained only on a partial subset. However, once trained, when given all the lipid tail dataset, `MembTDA` predicts critical unique temperatures that were never explicitly set as a training objective or part of the neural network architecture. The information captured in the persistence data and learned by the neural network has shown that the two local minima of the DPPC effective temperature distribution curve are indeed biophysically relevant temperatures as reported in Khakbaz et al.[45]

The effective temperatures predicted by `MembTDA` are dynamics of lipids, since temperature prediction inherently takes velocity information[49] into account. Also, for the configurational temperature, as in Jepps et al.,[51] force information is taken into account. Both force and velocity are vectors projected on the atomic position, responsible for propagating the MD integration. However, since there are no direct atom velocity nor complete force information (due to individual lipid force being considered without environment) in static lipid coordinates, we can speculate that persistence data carry mixed information of both velocity (a proxy for *entropic* information)[54] and force information for which the neural net-

work is effectively able to learn. Such hidden information comes in the form of persistence data, implying that effective temperature information can be retrieved by accounting for connected components and holes present in the lipid configurations. To us, this implies that `MembTDA` maps the underlying configurational manifold and its topology, which is necessary for its physical dynamics (Figure 1).

Moreover, we have shown that `MembTDA` can capture effective temperatures of lipids in membranes with different lipid compositions other than what was used in its training, and for lipids in the presence of proteins. The dynamics of lipids are shown in Figure 5 where higher effective temperature lipids, colored in red, can be observed near the periphery of proteins. Also, different lipid types (i.e., POPC, POPE, and POPS) experience different effective temperatures because different head groups have different favorability to the protein residues they are interacting with, due to polarity and charges based interactions (an *enthalpic* effect), altering allowed configurational states.

As for why `MembTDA` has accurate effective temperature predictability, we ascribe this to attention maps presented in Supplemental Figure S8, Figure S9, and Figure S10. For low-temperature lipids, we chose 16 random (hence `4 x 4` panels) lipids and extracted the attention maps `MembTDA` focuses on. An attention map (in our case, GradCAM [55]) is an explainable AI (XAI) technique to highlight which features of the input an ML model focuses on to make a prediction. We can see that there is a semi-circle on the top left part of the persistence image data, which `MembTDA` identifies as important features (Supplemental Figure S8). The semi-circle indicates prominent H_1 features (i.e., holes) in the persistence data which may be born at relatively earlier α-filtration values. Since lower temperatures render accessible lipid configurations to more rigid states, the attention map (GradCAM) can capture hole patterns more readily. On the other hand, for high-temperature lipids, the attention map (GradCAM) does not have a distinct pattern captured by our neural network (Supplemental Figure S9). This implies that a high variability of lipid configurations exists at high temperatures, and thus no consolidated patterns like those observed for low-temperature

lipids. As for the melting temperature (Supplemental Figure S10), we see mixed patterns where nearly half the attention maps (GradCAM) resemble those of low-temperature lipids (i.e., prominent attention patterns), and the rest resemble those of high-temperature lipids (i.e., no particular attention patterns). We postulate that this is due to expected equal fractions of L_β and L_α phases at the melting temperature.

At all scales, entropic contributions create system heterogeneity that result in local order.[56] In the context of biological membranes, we have shown that local order plays a role in lipid dynamics, by capturing local effective temperatures of individual lipids. The ability to recapitulate DPPC melting temperatures, demonstrated by `MembTDA`, reveals that PH is likely correlated to physical phenomena, potentially via a mapping of a manifold embedding representative of a potential energy surface with an inherent characteristic topology. We speculate that the effectiveness of our topological learning approach implies the possibility of an underlying analytical framework suitable for estimating physical properties such as melting temperatures or heat capacity, like a hybrid of the equipartition theorem and the Jepps formulation. More precisely, we posit that the use of PH on atomic data captures inherent topological features of the potential energy surface as shown in Figure 1 and Figure 4.

Acknowledgements

We would like to thank Dr. Pochao Wen, Dr. Archit Vasan, Ali Rasouli, and Matt Sinclair for their feedback and useful commentary.

The authors acknowledge support from the National Institute of General Medical Sciences of the National Institutes of Health under awards P41-GM104601, R24-GM145965, and R01-GM123455. ASA would like to acknowledge a graduate student fellowship from xxxxx. National Institutes of Health under award F31-HL136155. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

(1) Wen, P.-C.; Mahinthichaichan, P.; Trebesch, N.; Jiang, T.; Zhao, Z.; Shinn, E.; Wang, Y.; Shekhar, M.; Kapoor, K.; Chan, C. K., et al. Microscopic view of lipids and their diverse biological functions. *Current opinion in structural biology* **2018**, *51*, 177–186.

(2) Dehghani-Ghahnaviyeh, S.; Zhao, Z.; Tajkhorshid, E. Lipid-mediated prestin organization in outer hair cell membranes and its implications in sound amplification. *Nature communications* **2022**, *13*, 1–12.

(3) Ciudad, S.; Puig, E.; Botzanowski, T.; Meigooni, M.; Arango, A. S.; Do, J.; Mayzel, M.; Bayoumi, M.; Chaignepain, S.; Maglia, G., et al. Aβ (1-42) tetramer and octamer structures reveal edge conductivity pores as a mechanism for membrane damage. *Nature communications* **2020**, *11*, 1–14.

(4) Jeong, H.; Clark, S.; Goehring, A.; Dehghani-Ghahnaviyeh, S.; Rasouli, A.; Tajkhorshid, E.; Gouaux, E. Structures of the TMC-1 complex illuminate mechanosensory transduction. *Nature* **2022**, *610*, 796–803.

(5) Rasouli, A.; Yu, Q.; Dehghani-Ghahnaviyeh, S.; Wen, P.-C.; Kowal, J.; Locher, K. P.; Tajkhorshid, E. Differential dynamics and direct interaction of bound ligands with lipids in multidrug transporter ABCG2. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2213437120.

(6) Kapoor, K.; Pant, S.; Tajkhorshid, E. Active participation of membrane lipids in inhibition of multidrug transporter P-glycoprotein. *Chemical science* **2021**, *12*, 6293–6306.

(7) Muller, M. P.; Jiang, T.; Sun, C.; Lihan, M.; Pant, S.; Mahinthichaichan, P.; Trifan, A.; Tajkhorshid, E. Characterization of lipid–protein interactions and lipid-mediated modulation of membrane protein function through molecular simulation. *Chemical reviews* **2019**, *119*, 6086–6161.

(8) Pozza, A.; Giraud, F.; Cece, Q.; Casiraghi, M.; Point, E.; Damian, M.; Le Bon, C.; Moncoq, K.; Banères, J.-L.; Lescop, E., et al. Exploration of the dynamic interplay between lipids and membrane proteins by hydrostatic pressure. *Nature communications* **2022**, *13*, 1–16.

(9) Mondal, S.; Khelashvili, G.; Weinstein, H. Not just an oil slick: how the energetics of protein-membrane interactions impacts the function and organization of transmembrane proteins. *Biophysical Journal* **2014**, *106*, 2305–2316.

(10) Engelman, D. M., et al. Membranes are more mosaic than fluid. *Nature* **2005**, *438*, 578–580.

(11) Piggot, T. J.; Allison, J. R.; Sessions, R. B.; Essex, J. W. On the calculation of acyl chain order parameters from lipid simulations. *Journal of chemical theory and computation* **2017**, *13*, 5683–5696.

(12) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell Jr, A. D.; Pastor, R. W. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The journal of physical chemistry B* **2010**, *114*, 7830–7843.

(13) Chazal, F.; Michel, B. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence* **2021**, 108.

(14) Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of persistence diagrams. Proceedings of the twenty-first annual symposium on Computational geometry. 2005; pp 263–271.

(15) Otter, N.; Porter, M. A.; Tillmann, U.; Grindrod, P.; Harrington, H. A. A roadmap for the computation of persistent homology. *EPJ Data Science* **2017**, *6*, 1–38.

(16) Hamilton, W.; Borgert, J.; Hamelryck, T.; Marron, J. *Research in Computational Topology 2*; Springer, 2022; pp 223–244.

(17) Smith, P.; Owen, D. M.; Lorenz, C. D.; Makarova, M. Asymmetric glycerophospholipids impart distinctive biophysical properties to lipid bilayers. *Biophysical Journal* **2021**, *120*, 1746–1754.

(18) Nguyen, D. D.; Wei, G.-W. AGL-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling* **2019**, *59*, 3291–3304.

(19) Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018**, *58*, 287–296.

(20) Pu, L.; Govindaraj, R. G.; Lemoine, J. M.; Wu, H.-C.; Brylinski, M. DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS computational biology* **2019**, *15*, e1006718.

(21) Casadio, R.; Martelli, P. L.; Savojardo, C. Machine learning solutions for predicting protein–protein interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1618.

(22) Park, H.; Zhu, R.; Huerta, E.; Chaudhuri, S.; Tajkhorshid, E.; Cooper, D. End-to-end AI Framework for Interpretable Prediction of Molecular and Crystal Properties. *Machine Learning: Science and Technology* **2023**,

(23) Park, H.; Yan, X.; Zhu, R.; Huerta, E. A.; Chaudhuri, S.; Cooper, D.; Foster, I. T.; Tajkhorshid, E. GHP-MOFassemble: Diffusion modeling, high throughput screening, and molecular dynamics for rational discovery of novel metal-organic frameworks for carbon capture at scale. 2023.

(24) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**,

(25) Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L., et al. Swin transformer v2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp 12009–12019.

(26) Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp 11976–11986.

(27) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **2008**, *29*, 1859–1865.

(28) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **2017**, *14*, 71–73.

(29) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

(30) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **2005**, *26*, 1781–1802.

(31) Phillips, J. C.; Hardy, D. J.; Maia, J. D.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W., et al. Scalable molecular dynamics on

CPU and GPU architectures with NAMD. *The Journal of chemical physics* **2020**, *153*, 044130.

(32) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics* **1984**, *81*, 511–519.

(33) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A* **1985**, *31*, 1695.

(34) Evans, D. J.; Holian, B. L. The nose–hoover thermostat. *The Journal of chemical physics* **1985**, *83*, 4069–4074.

(35) Hoover, W. G.; Ladd, A. J.; Moran, B. High-strain-rate plastic flow studied via nonequilibrium molecular dynamics. *Physical Review Letters* **1982**, *48*, 1818.

(36) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford university press, 2017.

(37) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry* **2001**, *22*, 501–508.

(38) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N.log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993**, *98*, 10089–10092.

(39) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33–38.

(40) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry* **2011**, *32*, 2319–2327.

(41) `https://courses.cs.duke.edu/fall06/cps296.1/Lectures/sec-III-4.pdf`.

(42) Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushtanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research* **2017**, *18*.

(43) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.

(44) Tsang, S.-H. Review: Swin transformer. 2022; `https://sh-tsang.medium.com/review-swin-transformer-3438ea335585#:~:text=Swin%2DT%2C%20Swin%2DS,to%20M%3D7%20by%20default.`

(45) Khakbaz, P.; Klauda, J. B. Investigation of phase transitions of saturated phosphocholine lipid bilayers via molecular dynamics simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2018**, *1860*, 1489–1501.

(46) Huang, J.; MacKerell Jr, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, *34*, 2135–2145.

(47) Leekumjorn, S.; Sum, A. K. Molecular characterization of gel and liquid-crystalline structures of fully hydrated POPC and POPE bilayers. *The Journal of Physical Chemistry B* **2007**, *111*, 6026–6033.

(48) Turner, K.; Mileyko, Y.; Mukherjee, S.; Harer, J. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry* **2014**, *52*, 44–70.

(49) Equipartition of energy. `https://www.britannica.com/science/equipartition-of-energy`.

(50) Eastwood, M. P.; Stafford, K. A.; Lippert, R. A.; Jensen, M. Ø.; Maragakis, P.; Predescu, C.; Dror, R. O.; Shaw, D. E. Equipartition and the calculation of temperature in biomolecular simulations. *Journal of Chemical Theory and Computation* **2010**, *6*, 2045–2058.

(51) Jepps, O. G.; Ayton, G.; Evans, D. J. Microscopic expressions for the thermodynamic temperature. *Physical Review E* **2000**, *62*, 4757.

(52) Mechelke, M.; Habeck, M. Estimation of interaction potentials through the configurational temperature formalism. *Journal of Chemical Theory and Computation* **2013**, *9*, 5685–5692.

(53) Efrem, B.; Justin, G.; Heather, B. M.; David, L. M.; Jacob, I. M.; Samarjeet, P.; Daniel, M. Z. Best Practices for Foundations in Molecular Simulations [Article v1. 0]. *Living Journal of Computational Molecular Science* **2019**, *1*, 5957.

(54) `https://galileo.phys.virginia.edu/classes/152.mf1i.spring02/MolecularEntropy.htm`.

(55) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision. 2017; pp 618–626.

(56) Kauffman, S. A., et al. *The origins of order: Self-organization and selection in evolution*; Oxford University Press, USA, 1993.

(57) Zhang, Q.-L.; Yang, Y.-B. ResT V2: Simpler, Faster and Stronger. *arXiv preprint arXiv:2204.07366* **2022**,

(58) Masking and padding with Keras ;: ; Tensorflow Core. `https://www.tensorflow.org/guide/keras/masking_and_padding`.

(59) Unke, O. T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation* **2019**, *15*, 3678–3693.