THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB Article

Development of a PointNet for Detecting Morphologies of Self-Assembled Block Oligomers in Atomistic Simulations

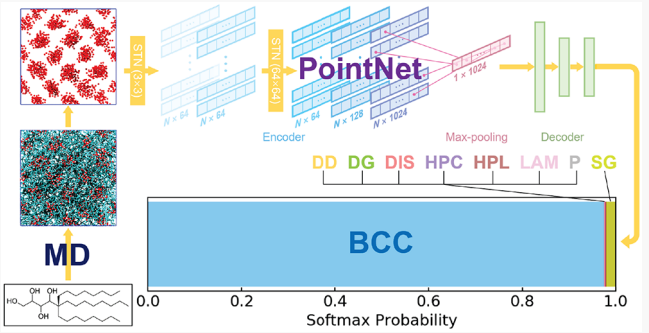
Published as part of The Journal of Physical Chemistry virtual special issue "Carol K. Hall Festschrift".

Zhengyuan Shen, Yangzesheng Sun, Timothy P. Lodge, and J. Ilja Siepmann*

Cite This: https://doi.org/10.1021/acs.jpcb.1c02389 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular simulations with atomistic or coarse-grained force fields are a powerful approach for understanding and predicting the self-assembly phase behavior of complex molecules. Amphiphiles, block oligomers, and block polymers can form mesophases with different ordered morphologies describing the spatial distribution of the blocks, but entirely amorphous nature for local packing and chain conformation. Screening block oligomer chemistry and architecture through molecular simulations to find promising candidates for functional materials is aided by effective and straightforward morphology identification techniques. Capturing 3-dimensional periodic structures, such as ordered network morphologies, is hampered by the requirement that the number of molecules in the simulated system and the shape of the periodic simulation box need to be commensurate with those of the resulting network phase. Common strategies for structure identification include structure factors and order parameters, but these fail to identify imperfect structures in simulations with incorrect system sizes. Building upon pioneering work by DeFever et al. [*Chem. Sci.* **2019**, *10*, 7503−7515] who implemented a PointNet (i.e., a neural network designed for computer vision applications using point clouds) to detect local structure in simulations of single-bead particles and water molecules, we present a PointNet for detection of nonlocal ordered morphologies of complex block oligomers. Our PointNet was trained using atomic coordinates from molecular dynamics simulation trajectories and synthetic point clouds for ordered network morphologies that were absent from previous simulations. In contrast to prior work on simple molecules, we observe that large point clouds with 1000 or more points are needed for the more complex block oligomers. The trained PointNet model achieves an accuracy as high as 0.99 for globally ordered morphologies formed by linear diblock, linear triblock, and 3-arm and 4-arm star-block oligomers, and it also allows for the discovery of emerging ordered patterns from nonequilibrium systems.



INTRODUCTION

Self-assembling amphiphiles, block oligomers, and block polymers that contain chemically distinct segments can form a wide variety of structures across length scales from a few to hundreds of nanometers. Depending on the self-assembled morphologies and domain sizes, these classes of materials can be targeted to numerous application including templates for nanopatterning,[1−4] transport membranes,[5−8] drug delivery,[9,10] and photonics.[11,12] To accelerate material design and discovery, molecular simulations can be used to efficiently screen over molecular structures and provide detailed microscopic-level insights. In our recent studies,[13−15] molecular dynamics (MD) simulations using transferable force fields[16−18] were performed to study the phase behavior of a class of block oligomers with thermotropic liquid crystallinity. Multiple mesophases were observed with domain sizes smaller than 4 nm, including lamellar (LAM), hexagonally packed cylinder (HPC), hexagonally perforated lamellar (HPL), body-centered cubic (BCC), and disordered states (DIS) (see Figure 1). Although missing from our previous simulations, 3-dimensional network structures (NET), which are often observed over narrow composition windows in self-assembling soft materials, are of increasing interest due to their interpenetrating domains that enable independent tuning of orthogonal properties in a single material.[19] Further exploring the design space of these block oligomers could facilitate the computational discovery of NET-forming materials, as well as systems with larger NET composition windows.

In molecular simulations, equilibrium mesophases can be inferred from spatial information including atomic positions

Received: March 16, 2021
Revised: April 27, 2021

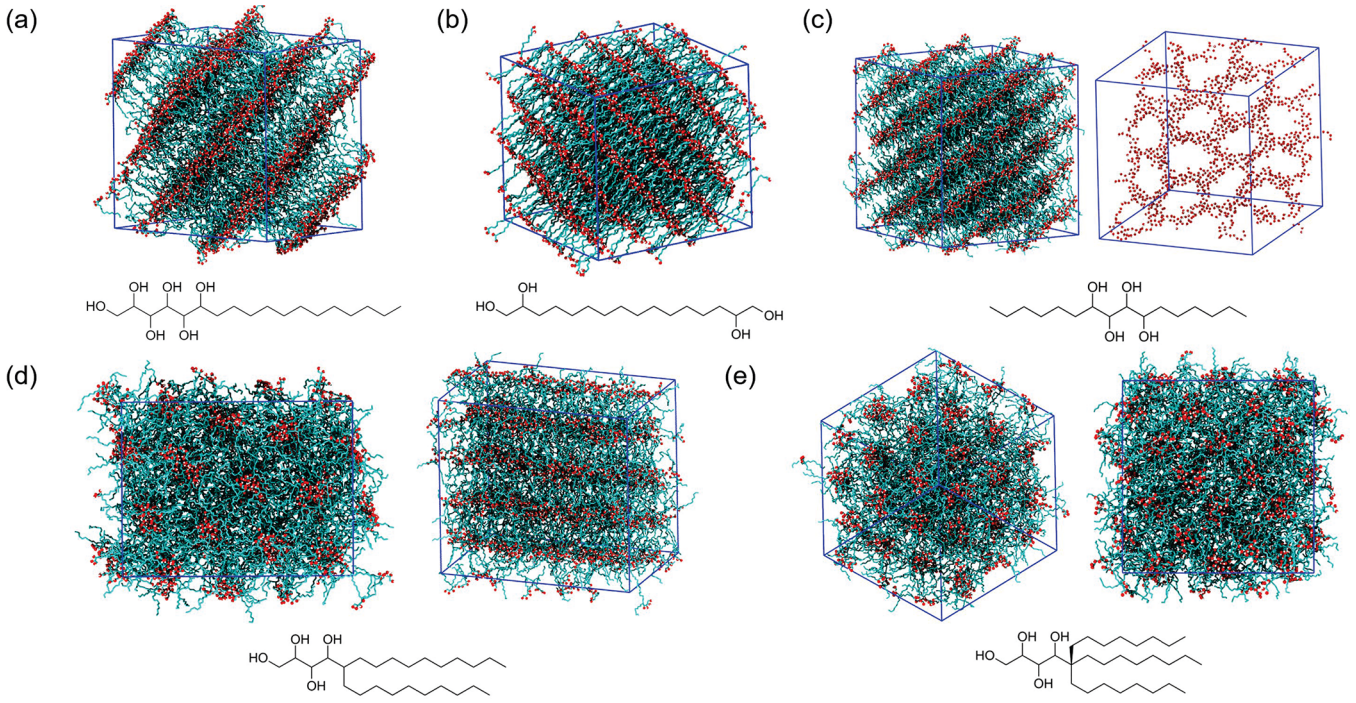


Figure 1. Snapshots showing the periodic box from previous MD simulations[13−15] for various block amphiphiles and their corresponding chemical structures: (a) A_6B_{12} and (b) $A_2B_{12}A_2$ in the LAM phase, (c) $B_6A_4B_6$ in the HPL phase (the second snapshot shows one perforated layer containing O and H atoms); (d) $A_4B(B_{10})_2$ in the HPC phase, and (e) $A_4B(B_8)_3$ in the BCC phase. Hydroxyl hydrogen and oxygen atoms are shown as white and red spheres, respectively, and the alkyl tails as cyan lines.

and identities, which allows for quantitative analysis that can bridge with macroscopic observations. One common technique is to compute the structure factor,[20−22] and the morphology can be determined from the relative peak positions and intensities. Importantly, the structure can be readily compared with experimental scattering patterns. However, in molecular simulations, the system dimensions have to be delicately selected to match integer multiples of the unit cell of the equilibrium structure.[23] When the system size is chosen incorrectly, the system may adopt a distorted, thermodynamically unstable structure in comparison to the infinite system. In principle, the incommensurability effects can be ameliorated by using a very large system size but, beyond simple coarse-grained models, such simulations are unaffordable even with current computing hardware. For morphologies that are anisotropic in one or two dimensions (e.g., LAM and HPC), the commensurability issues can be accommodated by using anisotropic orthorhombic simulation boxes that allow independent fluctuations of the three dimensions.[24,25] For 3-dimensional periodic structures, selecting the system parameters is nontrivial, since the exact stable morphology and the corresponding unit cell dimensions are not known *a priori*. In our previous work, there have been successes in achieving stable 3-dimensional periodic structures including BCC and HPL[15] by tuning the number of molecules after an initial guess of the equilibrium morphology from the imperfect structures resulting from arbitrary system sizes. Similarly, the system size for the $A_4B(B_8)_3$ miktoarm tetrablock oligomer is tuned here to yield a stable BCC morphology instead of disordered micelles.[13] However, such a human-based initial guess from emergent NET structures can be far less accurate, due to the existence of many possible NET geometries. Therefore, it is beneficial to predict the likely stable structure of a NET candidate system before fine-tuning the system size. The structure factor, in this case, can fail to detect any nonglobal features of the plausible but distorted structure. Another structure detection tool that can be used for molecular simulations is referred to as an "order parameter", which can be some mathematical quantity such as a "signature vector" as a function of atomic positions. The order parameter can tackle local structure recognition but can only distinguish among very few structures, and developing selective functional forms of order parameters can be challenging.[26−29]

Machine learning methods based on deep neural networks have been widely employed in the prediction and design of atomistic and molecular systems.[30,31] Neural networks are a class of mathematical models composed of multiple layers of neurons, where each neuron outputs a linear combination of the input from its previous layer followed by a nonlinear transformation. Although a shallow neural network with two layers is already sufficient to approximate any continuous function, increasing the number of layers introduces a hierarchy of representations of input data, which results in strong performance in various complex tasks and alleviates the need for feature engineering.[32] Deep learning methods for molecular systems are commonly based on atomic coordinates as they directly represent the structure and geometry of the system. For example, neural network potentials for MD and Monte Carlo simulations typically take a series of symmetry functions over atomic coordinates as input.[33,34] Pairwise distances have also been constructed as the features of neural networks for molecular structure generation to utilize rotational symmetry.[35] Recently, DeFever et al. reported a deep learning method for identifying local crystal structures, mesophases, and hydrophilic surfaces from MD simulations for binary mixtures of single-bead particles and for multisite water models directly from particle coordinates.[36] Their method was based on PointNet, a highly efficient and effective

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

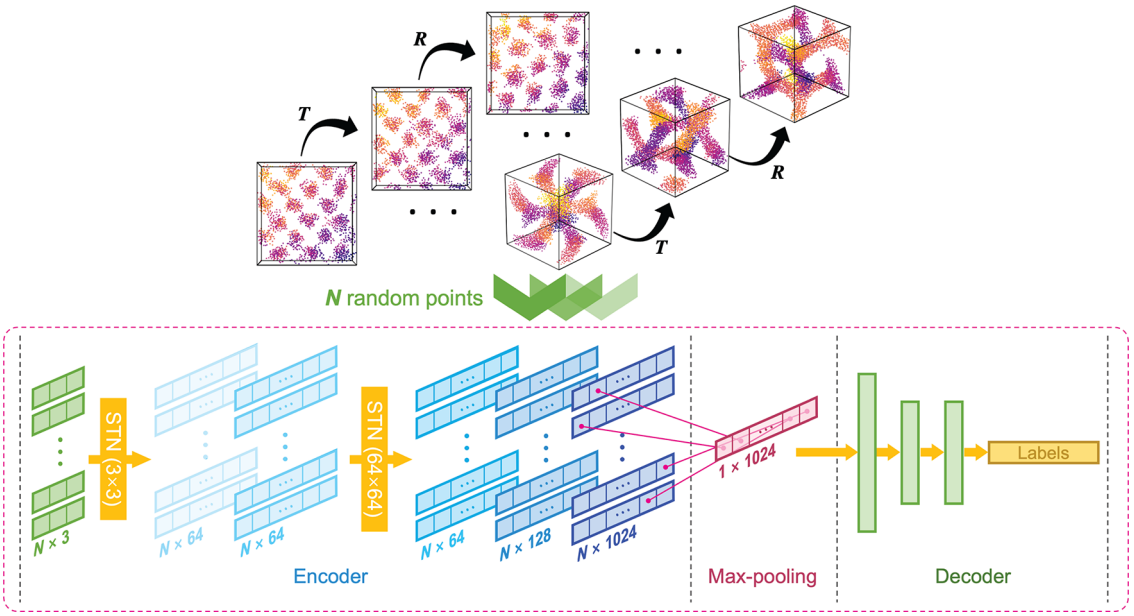


Figure 2. Data augmentation and the PointNet architecture. A random translation and a random rotation are applied on each point cloud, from which N points are randomly selected and fed into the PointNet. The N points are passed through the encoder which contains point-wise feedforward layers and spatial transformer networks (STN). The max-pooling layer performs permutation-invariant aggregations of the 1024-dimensional point features, and the following feedforward network (512, 256 neurons) with dropout outputs predicts the scores for point cloud classification.

neural network architecture in computer vision applications for point clouds,[37] and utilized the coordinates of neighboring particles within a local spherical region centered on one particle.

Here, we also adopt the PointNet architecture for morphology classification of self-assembled block oligomers investigated by MD simulations.[13−15] Apart from point clouds consisting of sets of Cartesian coordinates, meshes and voxel grids are also common representations of 3-dimensional spatial data in computer vision. Meshes are the natural choice for representing surfaces but, for the block oligomers investigated here, the surfaces are locally very rough making identification difficult. While voxel grids have been used for deep-learning-based generation of crystal structures,[38] they are computationally inefficient for larger systems and introduce discretization errors making the detected morphology potentially ambiguous. The PointNet model developed here is trained on atomic structures of different morphologies from a combination of MD simulation frames and synthetic point clouds of NET structures to address the scarcity of NET geometries encountered in our previous simulations.

METHODS

PointNet Architecture. In this work, the standard settings of the classification network in PointNet[37] are utilized (see Figure 2). The input point clouds are represented by N Cartesian coordinates without additional features, but may contain the positions of multiple beads taken from a given block oligomer. The point clouds used here represent the global structure of the system, but require a large number of points. In contrast, the PointNet introduced by DeFever et al.[36] utilizes only a local region with a small number of points for structure identification of single-site particles or water molecules. Although not guaranteeing better performance, the global point clouds used here are closer to the original computer vision application.[37]

In our approach, the input point clouds are then passed through the encoder consisting of a series of pointwise feedforward layers with 64, 64, 64, 128, and 1024 neurons with weights shared among all points. Since atomic coordinates are unordered, the network output should be invariant to the permutation of atoms, such as exchanging a pair of coordinates in the input. This is achieved in PointNet by applying a symmetric function operated on the high-dimensional point features produced from the pointwise network:

$$f(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}) \approx g(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)) \quad (1)$$

where $\mathbf{x}_1$, ..., $\mathbf{x}_n$ are input points from the point cloud, h represents the pointwise network for feature extraction, and g represents the symmetric function. The PointNet transforms the 3-dimensional Cartesian coordinates into a 1024-dimensional features space before performing max pooling. The max pooling layer then takes the largest value for all points along each dimension to give the global feature.[37] The points for which the feature coordinates contribute to the global feature are picked as the "critical points" regardless of the input order. This also agrees with the physical intuition that the morphology of the system can be identified with the most important subset of atoms in the structure. Directly after the max pooling layer, two dense feedforward layers with 512 and 256 neurons followed by a dropout layer with a 0.7 keep ratio are used to calculate class scores and to infer class labels from softmax probabilities.

In the PointNet structure used for the present work, two spatial transformer networks (STNs) are applied before the first and the third feature extraction layer. The STNs take a similar structure as the main networks and comprise the same types of modules including feature extraction, max pooling, and fully connected layers. They are aimed at learning data-dependent rigid or affine transformation matrices to align the input (3×3) point sets and higher dimensional features (64×64) into a consistent orientation to further improve the

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

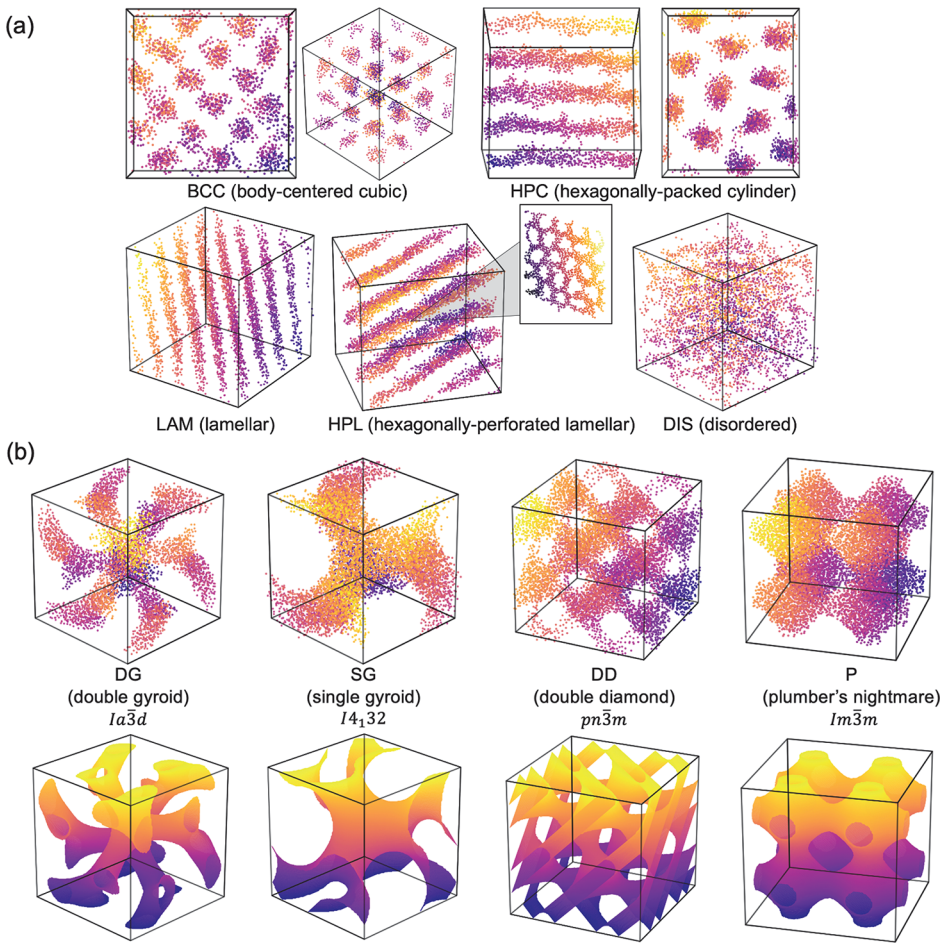


Figure 3. (a) Examples for the point clouds of the minor component of the block oligomers retrieved from MD simulation trajectories (see Table S1). (b) Generated point clouds and their corresponding surfaces for four ordered NET structures. To guide the eye, the point clouds are colored from yellow to purple according to a point's average values of its x, y, and z coordinates.

results.[37] The PointNet implementation by DeFever et al.[36] does not include such STNs. Because molecular systems are less orientation-dependent than real-life objects, we also evaluate the effectiveness of STNs in morphology detection in this work.

Batch normalization and rectified linear unit (ReLU) activation functions are applied throughout all fully connected layers. We use the Adam optimizer[39] with an initial learning rate of 0.001, which is halved every 20 training epochs. The exponential decay rates for the first and the second moment are set to values of 0.9 and 0.999, respectively. The model is implemented in PyTorch and trained on an NVIDIA Titan RTX GPU.

Data Augmentation and Learning Formalism. We use equilibrium trajectories obtained from NpT MD simulations of diblock, symmetric triblock, miktoarm triblock, and miktoarm tetrablock oligomers (for examples, see Figure 1). These amphiphiles consist of oligo-ol blocks (A_x with $2 \leq x \leq 6$ CH_rOH repeat units, where r is 1 or 2) and linear alkyl blocks (B_y with $6 \leq y \leq 12$ CH_s repeat units, where s is 2 or 3), and they assemble into various morphologies including body-centered cubic (BCC), hexagonally packed cylinders (HPC), lamellar (LAM), hexagonally perforated lamellar (HPL) and disordered (DIS).[13−15] The simulation frames used for the generation of the point clouds are selected from multiple systems (see Table S2) spanning 15 different block oligomers and, for all but one compound, two temperatures are included that represent one of the ordered morphologies and the DIS morphology. LAM, HPL, HPC, and BCC morphologies are represented by 10, 1, 2, and 1 systems, respectively, in the training set. For each morphology, 3000 point clouds for the minor component (the positions of O atoms) are extracted from frames across the different MD trajectories, with examples illustrated in Figure 3.

Due to the strict commensurability requirement for NET morphologies with crystallographic periodicity, our previous simulations on the specific class of self-assembling block oligomers did not yield any stable ordered NET structures. To train a model that can recognize common NET structures seen in other self-assembling systems, we generate synthetic data for point clouds representing minor components of double gyroid (DG), single gyroid (SG), double diamond (DD), and plumber's nightmare (P) morphologies using the following criteria, respectively:

$$|\sin(x)\cos(y) + \sin(y)\cos(z) + \sin(z)\cos(x)| > t \quad (2a)$$

$$\sin(x)\cos(y) + \sin(y)\cos(z) + \sin(z)\cos(x) > t \quad (2b)$$

$$|\sin(x)\sin(y)\sin(z) + \sin(x)\cos(y)\cos(z) + \cos(x)\sin(y)\cos(z) + \cos(x)\cos(y)\sin(z)| > t \quad (2c)$$

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

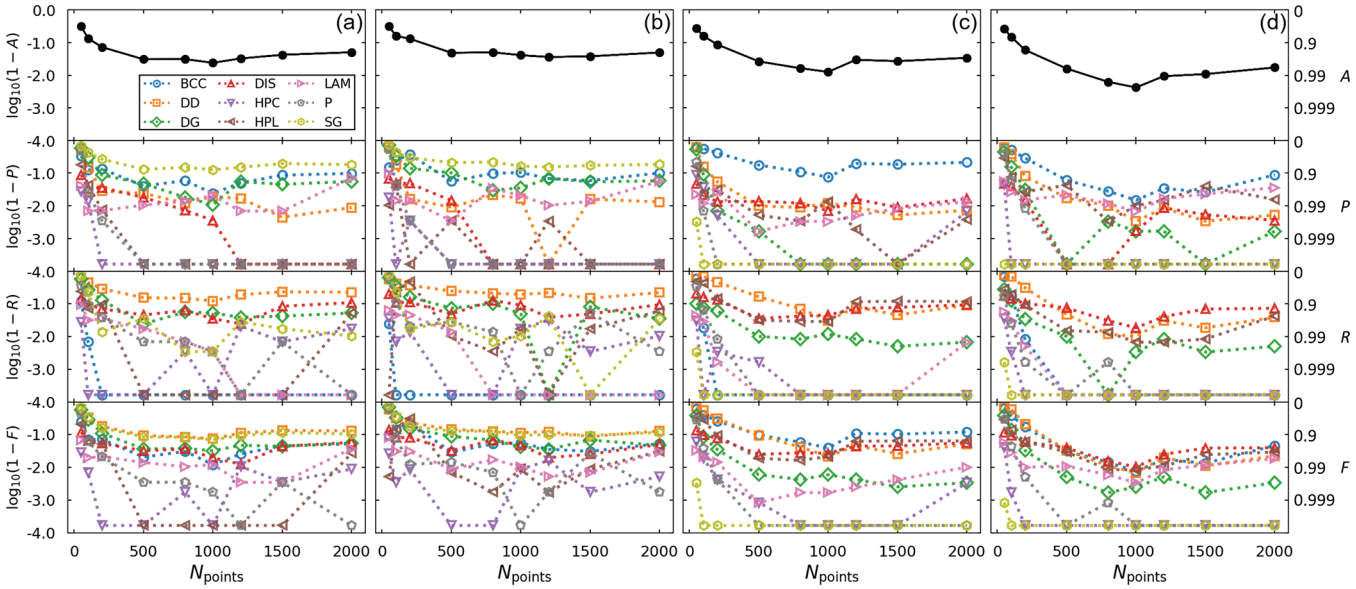


Figure 4. Overall test accuracy, *A*, and test precision, *P*, recall, *R*, and F1 score for individual morphologies as functions of input point cloud size for the models trained with (a) only random translation, (b) random translation and including STN in the PointNet, (c) random translation and rotation, (d) random translation, rotation, and including STN in the PointNet. The original point clouds of the minor components contain 2000−6000 points. Because of the finite size of the test set (600 point clouds per morphology), it is possible for recall, precision, and F1 score to reach a value of exactly 1.0000. For the logarithmic plot, this value is replaced with $\log_{10}(1/6000) = -3.778$.

$$\cos(x) + \cos(y) + \cos(z) > t \tag{2d}$$

where x, y, and z are point coordinates for the minor component, and t is an adjustable parameter to control the volume fraction of the minor component. Values of t are randomly selected within practical ranges to account for variations in volume fraction for the self-assembled systems. The t ranges are [0.9, 1.2], [0.6, 1.0], [0.6, 0.9], and [0.0, 0.4] for DG, SG, DD, and P morphologies, respectively. Points are uniformly sampled within these confined regions, and small random displacements are added to account for local composition fluctuations. Examples of the point clouds for the NET morphologies are shown in Figure 3.

Before feeding the point clouds into the PointNet, normalization and augmentation are applied on all raw point clouds. First, each point cloud is min-max scaled such that $|x|$, $|y|$, $|z| \leq 1$. In addition to the permutation invariance achieved by the PointNet structure, the model prediction should not change with translation or rotation of a point cloud. Therefore, a random translation vector $[\Delta x, \Delta y, \Delta z]^T$ is applied to all points in a given cloud, which satisfies $|\Delta x| \leq L_x/2$, $|\Delta y| \leq L_y/2$, $|\Delta z| \leq L_z/2$, where L_x, L_y, and L_z are the lengths of the x, y, and z dimensions for the orthorhombic simulation box. The transformed coordinates are then wrapped into the original bounding box using the periodic boundary conditions. A spatially uniformly random rotation matrix, $\boldsymbol{M}_{rot}$, is subsequently applied on the wrapped coordinates. To ensure a uniformly distributed rotation (see Figure S1), $\boldsymbol{M}_{rot}$ is obtained from performing a random rotation about the vertical axis followed by rotating the north pole to a random position,[40] and can be described by

$$\boldsymbol{R} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3a}$$

$$\boldsymbol{H} = \boldsymbol{I} - 2\begin{bmatrix} \cos\phi\sqrt{z} \\ \sin\phi\sqrt{z} \\ \sqrt{1-z} \end{bmatrix}[\cos\phi\sqrt{z}\sin\phi\sqrt{z}\sqrt{1-z}] \tag{3b}$$

$$\mathbf{M}_{rot} = -\mathbf{HR} \tag{3c}$$

where $\mathbf{R}$ is a simple rotation matrix around the vertical axis, $\mathbf{H}$ is the Householder matrix, $\mathbf{I}$ is the identity matrix, and θ, ϕ, and z are randomly chosen azimuthal angle ([0, π]), polar angle ([0, 2π]), and radial distance ([0, 1]), respectively. The transformed points are again wrapped using the periodic boundary conditions.

To train the PointNet, 3000 point clouds are selected for each of the nine morphologies obtained from the MD simulation trajectories and the generated NET structures. The data are split into training and test sets containing 80% and 20% of the data, respectively. A 5-fold cross-validation is performed to detect overfitting. After the data augmentation, the network is trained for 100 epochs with a batch size of 64 point clouds. Depending on the chemical compound and the number of molecules used in the simulations (see Table S2) and the volume fraction for the generated NET morphologies, each point cloud representing the minor component contains on average 4500 points with a standard deviation of 1920 points. For consistency of the input point cloud dimensions, a constant number of points ranging from 50 to 2000 is randomly drawn from each of the augmented point clouds. In addition to the standard workflow including random translation and rotation in data augmentation, and STN in the PointNet, we also compare the performance when excluding random rotation and/or STN.

RESULTS AND DISCUSSION

We examine the performance and robustness of the PointNet by training the network with the point clouds retrieved from MD simulations and the generated NET structures. These

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

point clouds represent the spatial distribution of oxygen atoms, part of the minor CH_rOH block in the self-assembled systems with a volume fraction less than 0.5. All point clouds supplied to the PointNet contain equal number of points (N_{points}) sampled from the augmented point sets. Each point cloud is labeled as one of nine different morphologies: BCC, DD, DG, DIS, HPC, HPL, LAM, P, and SG. Figure 4 shows the evaluation metrics for morphology classification including test accuracy, precision, recall, and F1 scores of the PointNet models trained by four strategies: (A) only applying random translation, (B) random translation and STN, (C) random translation and rotation, (D) random translation and rotation and STN. In each case, the model accuracy grows initially as N_{points} is increased, reaches a plateau with only small changes in the accuracy, and then drops slightly when the number of points is further increased. Confusion matrices for the four strategies are given in Figures S2 to S5. For all four models, the best performance is reached when 1000 or 1200 random points are provided as input point clouds, that is, when on average about 75% of the oxygen positions are not included in the point clouds. This peak in accuracy likely reflects the 1024 dimensions before the max pooling layer in the PointNet, but we also note that the average number of molecules in the simulated systems is close to 1000. To obtain a more reliable estimate of the accuracy of the models, we trained each model 10 times using different random seeds and randomly selected configurations with 1024 points to form the training set data. The numerical values of the average achieved accuracy and the corresponding 95% confidence interval are provided in Table 1.

Table 1. Accuracy Obtained for Different Training Strategies. The 95% Confidence Interval (CI) Is Estimated from Training Each Model 10 Times

model	strategy	accuracy	CI
A	translation	0.973	0.010
B	translation + STN	0.957	0.011
C	translation + rotation	0.983	0.009
D	translation + rotation + STN	0.990	0.005

As can be seen in Figure 4, the accuracy generally starts to exceed 90% when only 500 points are extracted, regardless of the training approach, which corresponded to a nearly 90% missing data ratio. This trend is in agreement with robustness tests of the PointNet against input corruption presented in the original paper[37] for benchmarks with the ModelNet40 data set,[41] where an accuracy of 0.75 was reported when only 256 points were used (75% missing data ratio). Interestingly, the robustness in the current study is significantly higher, which may be attributed to two factors: (i) the larger number of points in the training clouds (an average of 4500 vs 1024), and (ii) the point clouds in this study containing atomic positions that span the entire volume of the minority region instead of only representing the surface as in the ModelNet data.

The PointNet is designed to discover perceptually interesting points with the highest contribution to the max pooled features.[37] Given the max pooling dimension (see Figure 2), up to 1024 points that contribute to the max pooling layer can be selected as critical points from among the N_{points} points in a point cloud. Since a given point can take the largest value in more than one max pooling dimension, the lower bound for the number of unique critical points, N_{crit}, is the size of the input point cloud, N_{points}.[37] Figure S6 illustrates input point clouds with N_{points} = 2000 and the corresponding critical points. Although we find $N_{crit} < 1024 < N_{points}$, the global shape features are not substantially changed even when most of the noncritical points are missing. Among the critical points contributing to the max-pooled features, only a small portion of points are close to domain interfaces, which also indicates that PointNet is potentially more robust to volumetric data than surface data. Indeed, initial tests generating point clouds only from the position of the center of the bond linking the oligo-ol and alkyl regions showed less promise than utilizing the locations of the oxygen atoms. Similarly, focusing on only the local environment of this center point yielded poor performance in particular for HPL and NET structures. That is, local fluctuations and feature size for the multibead oligomers investigated here necessitate a nonlocal PointNet.

Besides the difference in the spatial distribution of points (surface vs volumetric), another major distinction between the point clouds from molecular simulations and those from real-world objects is periodicity. In most molecular simulations, 3-dimensional periodic boundary conditions are utilized to allow the finite system in the simulation box to better mimic bulk systems. Therefore, the classification model should give exactly the same label when a simulated system is shifted along, or rotated by any angles around an arbitrary vector. Examples of applying such operations to BCC and DG structures are shown in Figure 2, which is analogous to achieving equivalent morphologies from independent simulations under the same thermodynamic conditions.

The STNs are designed to disentangle part deformations of the objects by aligning input data with affine transformations.[42] While real-world objects subject to gravitational forces usually have a strongly preferred orientation in the direction parallel to the gravitational field, molecular systems in the absence of an aligning field may take any orientation. Given this difference in orientational preferences, it is interesting to compare the impact of random rotations and STNs on the model performance. As can be observed from the data presented in Figure 4 and Table 1, introducing STNs slightly reduces the highest accuracy from 0.973 to 0.957 when the point clouds are augmented only with random translations. Models A and B both incorrectly classify some of the DD structures as SG, as can be inferred from the low precision for SG and the low recall for DD. In contrast, the overall accuracy increases from model C to D (from 0.983 to 0.990 at N_{points} = 1000) when STNs are applied along with point clouds that have undergone both translations and rotations. Precision and recall are now near-perfect for the SG structures, but about 2% of DD structures are classifieded as BCC in model C.

The difference in model performance when applying random rotations and/or STNs can be understood using the orientational distributions of the simulated and generated morphologies. In the point cloud data set, all generated NET morphologies possess the same center and orientation, and multiple frames are taken from the same MD trajectory. When a stable ordered structure is formed in an MD simulation, then the overall structural orientation changes only very slowly; this results in correlation between samples in the data set, making the orientational distribution of the simulation samples highly nonuniform and multimodal due to different orientations encountered for different systems (see Figure S7). Due to the limited alignment capability of the STN originally designed for correcting slight rotations and perspective transformations,[42] it

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

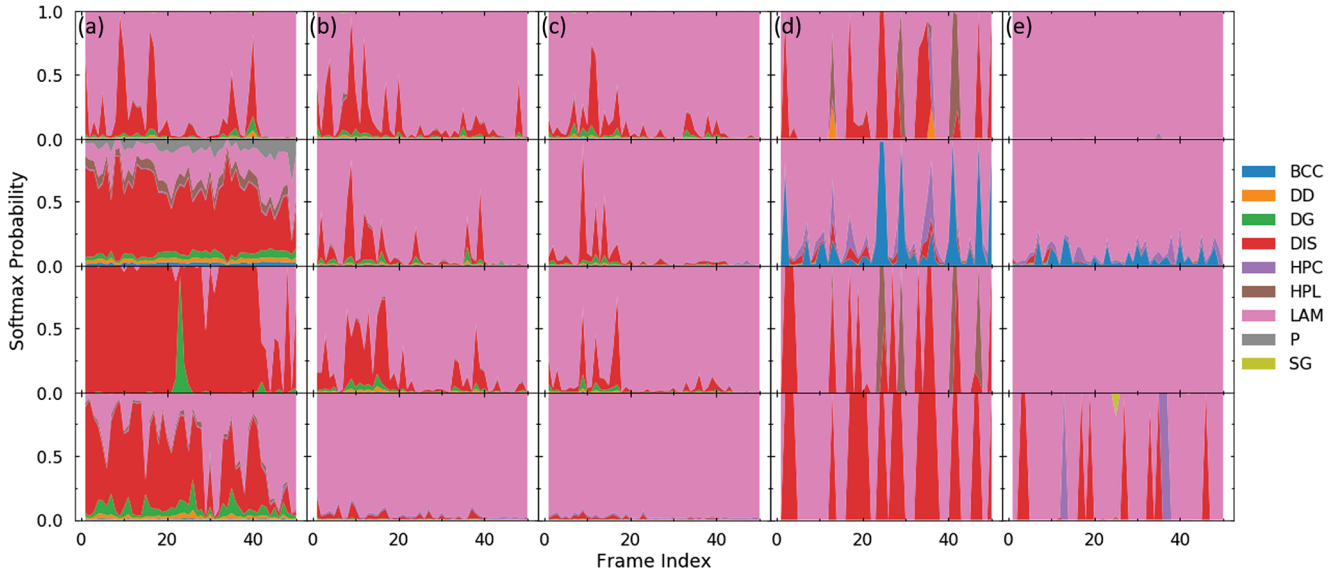


Figure 5. Stack plots of the predicted softmax classification probabilities obtained with models A, B, C, and D (top to bottom) for 50 frames taken at 10 ns intervals during a 500 ns MD trajectory. Data in columns a–c are for a 600-molecule $A_{10}B_{20}$ system with 1000 oxygen atoms for the point cloud selected at random from (a) all oxygen atoms, (b) only from oxygen atoms in positions 1, 4, 7, and 10 of the polar group, and (c) only from oxygen atoms in positions 2, 5, and 8 of the polar group. Data in columns d and e are for a 8000-molecule system of the $A_2B_8A_2$ block oligomer with 1000 oxygen atoms for the point cloud selected at random from (d) all oxygen atoms and (e) only from oxygen atoms located in a subvolume with linear dimensions of $L_x/2$, $L_y/2$, and $L_z/2$, where L_x, L_y, and L_z are the box lengths for the entire simulation box.

is only able to assist in classification of the model when the orientational distribution of input data is well behaved. Therefore, when comparing models A and B, application of the STN converts a multipeaked orientational distribution (resulting from an insufficiently diverse set of point clouds) into a more uniform distribution (see Figure S7) that, in turn, may introduce extra noise from structural distortion and lead to lower accuracy. A comparison of the performance of models B and C shows that randomly rotating the point clouds before feeding into the PointNet tends to be more effective than solely applying STNs, since the network was encouraged to capture rotational invariance from the rotated point clouds. In contrast, applying STNs for model D indeed enhances the accuracy beyond applying only translations and rotations for model C, and also leads to higher precision for classification of DD and BCC structures with model D. This can be explained by the effectiveness of STNs on a near-uniform orientational distribution after random rotations are applied. In this case, the STNs align the arbitrarily rotated point cloud onto one or a few canonical orientations (see Figure S7), which stabilizes the network and further improves the performance.

The block oligomer systems used as training sets contain between 2000 and 6016 oxygen atoms with mean and median of 4500 and 4000, respectively, and their polar blocks contain a total of four or six oxygen atoms (see Table S2). Thus, it is important to assess whether the trained PointNet models can also be applied to block oligomers with larger numbers of repeat units or to larger systems. To this extent, we performed new simulations for four systems using the same force fields and MD parameters as in prior work.[13−15] Specifically, we investigated 600-molecule systems for $A_{10}B_{20}$ and $A_8B(B_{18})_2$ (i.e., approximately doubling the number of A segments in the block oligomers), and eight times larger systems for $A_2B_8A_2$ and $A_4B(B_8)_2$ (8000 and 4000 molecules, respectively). In both cases, the linear block oligomers assemble in LAM morphology, whereas the double-tailed oligomers assemble in HPC morphology. Figure 5 and Figure S8 illustrate the performance of the four PointNet models. Despite that the total numbers of oxygen atoms for the $A_{10}B_{20}$ and $A_8B(B_{18})_2$ systems fall within the range used for the training structures, the classification performance is quite poor when the 1000 points are drawn randomly from the positions of all oxygen atoms; models A and C perform best for $A_{10}B_{20}$ and $A_8B(B_{18})_2$, respectively. That is, the PointNet models are confused when applied to larger block oligomers without pretreatment, and the number of oxygen atoms per oligomer appears to play a role. Thus, we tested pretreatments where only either four or three oxygen atoms in specific positions of the polar group are considered for selection of the point cloud. This pretreatment leads to a marked increase in performance. Model D classifies all 50 frames correctly for $A_{10}B_{20}$, and all but one frame correctly for $A_8B(B_{18})_2$. For the latter, however, models A and C classify all 50 frames correctly.

Considering the systems with larger numbers of molecules, all four models without pretreatment indicate that $A_2B_8A_2$ is likely a LAM morphology, but at least eight out of the 50 frames are incorrectly assigned. For $A_4B(B_8)_2$ with HPC morphology, models A, B, and D incorrectly indicate a preference for the DIS morphology, and model B points to either BCC or LAM. Despite that this $A_4B(B_8)_2$ system contains half the number of oxygen atoms compared to the $A_2B_8A_2$ system, the correct classification of the HPC morphology from a spatially sparse point cloud of 1000 oxygen atoms (selected from a total of 16000 oxygen atoms) appears to be more challenging. Thus, a pretreatment is also needed for the larger system sizes. In this case, our pretreatment consists of selecting the 1000 oxygen atoms for the point cloud only from a subvolume that is one-eighth of the volume of the entire system but has the same orientation. Again, pretreatment vastly improves the classification performance with models A and C giving the correct morphology for all 50 frames for both systems. Model D yields some

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

misclassifications for both systems, and model B does not assign a single frame correctly as HPC for the $A_4B(B_8)_2$ system. Since the subvolume by itself does not represent a periodic structure, application of the STN appears to be problematic. Although these pretreatment strategies allow for direct application of the PointNet models, there is clearly a limit in the number of repeat units and the system size that can be successfully classified without retraining of the PointNet models.

We also apply the models trained using the four strategies to detect morphology changes encountered in a simulation trajectory for a LAM-to-DIS transition. The simulated compound, 1,2,11,12-dodecanetetrol ($A_2B_8A_2$), forms a thermotropic liquid crystal that self-assembles into the LAM morphology at T_{SIM} = 400 and 430 K.[15] Here, an equilibrated LAM structure is simulated in the isobaric−isothermal ensemble for 10 ns at T = 430 K; at this point, a step increase is applied to raise the kinetic temperature and the thermostat temperature to 490 K (i.e., above the order−disorder temperature, T_{ODT}), followed by another 30 ns at T = 490 K. One thousand frames (spaced at 40 ps intervals) taken from the entire trajectory are used to generate point clouds for the minor component and analyzed by the four models trained only on equilibrium structures.

The time evolution of the predicted softmax scores is illustrated in Figure 6, for which each point cloud (N_{points} = 1000) is classified as belonging to a particular morphology if the softmax probability for that morphology is greater than 0.5. Models A and C consistently yield LAM softmax scores close to unity for the initial 10 ns period below T_{ODT}. For the models involving STNs, model D yields a few frames with HPC and SG false positives, but otherwise LAM softmax scores close to unity. In contrast, model B consistently shows nonzero softmax scores for other morphologies (mainly DIS and P) and also periods of false positives for the P morphology. Models A, C, and D indicate softmax scores near unity for DIS during the final 13 ns of the trajectory (t > 27 ns). In contrast, model B recognizes DIS morphologies only later (t > 35 ns) and with softmax scores significantly smaller than unity.

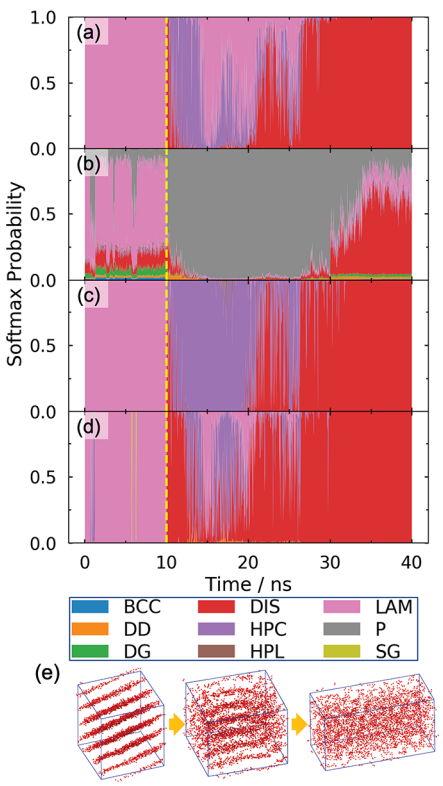


Figure 6. Stack plots of the predicted softmax classification probabilities obtained with models A, B, C, and D (subgraphs a to d) for 1000 frames taken at regular intervals during a 40 ns MD trajectory for which the kinetic temperature is increased above T_{ODT} at t = 10 ns (marked by yellow dashed line) to induce a LAM-to-DIS transition. (e) Point clouds of the minor component sampled at 5 ns (left), 20 ns (middle), and 35 ns (right).

Since the training data do not contain any point clouds reflecting a "transition" phase, it is of interest to compare the model predictions for the transition period (10 to 27 ns). All four models yield a sudden change to either DIS or P morphology immediately after the step increase in temperature, presumably because the higher temperature is almost instantaneously reflected in a change in the local structure and/or interfacial roughness (e.g., buckling of the lamellae). During the transition period, models A, C, and D indicate a mixture of LAM, HPC, and DIS structures, whereas model B predicts high probability for the P morphology with softmax scores mostly above 0.9. Models A and D show fleeting reappearance of the LAM morphology at $t \approx$ 10, 20, and 25 ns. Overall, models C and D show strong preferences for HPC and DIS, respectively, during most of the transition period, whereas model A yields more similar fractions of HPC, LAM, and DIS. These morphology classications suggest that the transition from the LAM to the DIS phase is not instantaneous, but rather involves a process of disruption and final disintegration of the lamellar planes. For the $A_2B_8A_2$ system, the DIS morphology is bicontinuous,[15] and the local packing in the DIS phase exhibits similarities to disordered cylindrical micelles and the HPC phase. Furthermore, order−order transitions from LAM to gyroid to HPC can occur before reaching T_{ODT}, and are predicted by self-consistent mean field theory for coil−coil and rod−coil block polymers within certain volume-fraction ranges.[43,44] Although PointNet classifications for the point clouds during the transition period are generalizations from models only trained on equilibrium structures, they provide additional support for the observation that STNs should not be included when the point clouds are not augmented by rotation (i.e., model B).

CONCLUSION

In this work, we train a deep neural network, PointNet, to identify morphologies of self-assembling block oligomers using points clouds taken from atomic coordinates of the minor component obtained by molecular simulations. To expand the scope of structure detection, we include synthetic point clouds of NET structures commonly observed for self-assembly of soft materials. The performances of the models trained using different strategies in performing data augmentation, building the PointNet architecture, and the number of points in the cloud are compared. A classification accuracy of 0.990 is achieved using 1000 coordinates (a missing data ratio of about 75%, but close to the number of dimensions in the max pooling layer), applying random translations and rotations under periodic boundary conditions to the training data, and including spatial transformer networks in the PointNet. With judicious pretreatment, the PointNet models can also be

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX

applied to oligomers with twice the number of repeat units and an eight times larger system than included in the training set.

The generalization ability of the trained models is tested using new point clouds from an MD trajectory sampling the lamellar-to-disorder transition of a block oligomer. We demonstrate that the PointNet models successfully predict the initial and final equilibrium structures and reflect the phase transition during intermediate time frames. The PointNet models presented in this study are generalizable and potentially transferable to discovering emerging structures of other shape-filling amphiphiles and block oligomers in molecular simulations and may guide the discovery of block oligomers forming ordered NET morphologies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c02389.

Information on PointNet model and data availability; tables providing details for the simulations of the block oligomers; figures illustrating the performance of the rotation matrix, the critical points with the highest contributions to the max-pooled features, and confusion matrices for the four model strategies (PDF)

■ AUTHOR INFORMATION

Corresponding Author

J. Ilja Siepmann − *Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455-0132, United States; Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States; Chemical Theory Center, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States;* orcid.org/0000-0003-2534-4507; Email: siepmann@umn.edu

Authors

Zhengyuan Shen − *Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455-0132, United States; Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States; Chemical Theory Center, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States*

Yangzesheng Sun − *Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States; Chemical Theory Center, University of Minnesota, Minneapolis, Minnesota 55455-0431, United States;* orcid.org/0000-0002-6505-6473

Timothy P. Lodge − *Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455-0132, United States;* orcid.org/0000-0001-5916-8834

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c02389

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation through the University of Minnesota MRSEC under Award DMR-2011401. Computer resources were provided by this NSF award and by the Minnesota Supercomputing Institute. Support through a Robert and Jill DeMaster-Excellence Fellowship (Y.S.) is gratefully acknowledged.

■ REFERENCES

(1) Feng, X.; Tousley, M. E.; Cowan, M. G.; Wiesenauer, B. R.; Nejati, S.; Choo, Y.; Noble, R. D.; Elimelech, M.; Gin, D. L.; Osuji, C. O. Scalable Fabrication of Polymer Membranes with Vertically Aligned 1 nm Pores by Magnetic Field Directed Self-Assembly. *ACS Nano* **2014**, *8*, 11977−11986.

(2) Feng, X.; Nejati, S.; Cowan, M. G.; Tousley, M. E.; Wiesenauer, B. R.; Noble, R. D.; Elimelech, M.; Gin, D. L.; Osuji, C. O. Thin Polymer Films with Continuous Vertically Aligned 1 nm Pores Fabricated by Soft Confinement. *ACS Nano* **2016**, *10*, 150−158.

(3) Nickmans, K.; Schenning, A. P. H. J. Directed Self-Assembly of Liquid-Crystalline Molecular Building Blocks for Sub-5 nm Nano-patterning. *Adv. Mater.* **2015**, *30*, 1703713.

(4) Sinturel, C.; Bates, F. S.; Hillmyer, M. A. High χ-Low N Block Polymers: How Far Can We Go. *ACS Macro Lett.* **2015**, *4*, 1044−1050.

(5) Jo, G.; Ahn, H.; Park, M. J. Simple Route for Tuning the Morphology and Conductivity of Polymer Electrolytes: One End Functional Group is Enough. *ACS Macro Lett.* **2013**, *2*, 990−995.

(6) Ichikawa, T.; Kato, T.; Ohno, H. 3D Continuous Water Nanosheet as a Gyroid Minimal Surface Formed by Bicontinuous Cubic Liquid-Crystalline Zwitterions. *J. Am. Chem. Soc.* **2012**, *134*, 11354−11357.

(7) Jackson, G. L.; Perroni, D. V.; Mahanthappa, M. K. Roles of Chemical Functionality and Pore Curvature in the Design of Nanoporous Proton Conductors. *J. Phys. Chem. B* **2017**, *121*, 9429−9436.

(8) Orilall, M. C.; Wiesner, U. Block copolymer based composition and morphology control in nanostructured hybrid materials for energy conversion and storage: solar cells, batteries, and fuel cells. *Chem. Soc. Rev.* **2011**, *40*, 520−535.

(9) Kluzek, M.; Tyler, A. I. I.; Wang, S.; Chen, R.; Marques, C. M.; Thalmann, F.; Seddon, J. M.; Schmutz, M. Influence of a pH-sensitive polymer on the structure of monoolein cubosomes. *Soft Matter* **2017**, *13*, 7571−7577.

(10) Kumar, R.; Le, N.; Tan, Z.; Brown, M. E.; Jiang, S.; Reineke, T. M. Efficient Polymer-Mediated Delivery of Gene-Editing Ribonucleoprotein Payloads through Combinatorial Design, Parallelized Experimentation, and Machine Learning. ACS. *ACS Nano* **2020**, *14*, 17626−17639.

(11) Hsueh, H.-Y.; Yao, C.-T.; Ho, R.-M. Well-ordered nanohybrids and nanoporous materials from gyroid block copolymer templates. *Chem. Soc. Rev.* **2015**, *44*, 1974−2018.

(12) Stefik, M.; Guldin, S.; Vignolini, S.; Wiesner, U.; Steiner, U. Block copolymer self-assembly for nanophotonics. *Chem. Soc. Rev.* **2015**, *44*, 5076−5091.

(13) Chen, Q. P.; Barreda, L.; Oquendo, L. E.; Hillmyer, M. A.; Lodge, T. P.; Siepmann, J. I. Computational Design of High-χ Block Oligomers for Accessing 1 nm Domains. *ACS Nano* **2018**, *12*, 4351−4361.

(14) Barreda, L.; Shen, Z.; Chen, Q. P.; Lodge, T. P.; Siepmann, J. I.; Hillmyer, M. A. Synthesis, Simulation, and Self-Assembly of a Model Amphiphile To Push the Limits of Block Polymer Nanopatterning. *Nano Lett.* **2019**, *19*, 4458−4462.

(15) Shen, Z.; Chen, J. L.; Vernadskaia, V.; Ertem, S. P.; Mahanthappa, M. K.; Hillmyer, M. A.; Reineke, T. M.; Lodge, T. P.; Siepmann, J. I. From Order to Disorder: Computational Design of Triblock Amphiphiles with 1 nm Domains. *J. Am. Chem. Soc.* **2020**, *142*, 9352−9362.

(16) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569−2577.

(17) Chen, B.; Potoff, J. J.; Siepmann, J. I. Monte Carlo Calculations for Alcohols and Their Mixtures with Alkanes. Transferable Potentials for Phase Equilibria. 5. United-Atom Description of Primary, Secondary, and Tertiary Alcohols. *J. Phys. Chem. B* **2001**, *105*, 3093−3104.
(18) Stubbs, J. M.; Potoff, J. J.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 6. United-Atom Description for Ethers, Glycols, Ketones, and Aldehydes. *J. Phys. Chem. B* **2004**, *108*, 17596−17605.
(19) Meuler, A. J.; Hillmyer, M. A.; Bates, F. S. Ordered Network Mesostructures in Block Polymer Materials. *Macromolecules* **2009**, *42*, 7221−7250.
(20) Schultz, A. J.; Hall, C. K.; Genzer, J. Obtaining Concentration Profiles from Computer Simulation Structure Factors. *Macromolecules* **2007**, *40*, 2629−2632.
(21) Overduin, S. D.; Patey, G. N. Understanding the Structure Factor and Isothermal Compressibility of Ambient Water in Terms of Local Structural Environments. *J. Phys. Chem. B* **2012**, *116*, 12014−12020.
(22) Liu, H.; Paddison, S. J. Direct calculation of the X-ray structure factor of ionic liquids. *Phys. Chem. Chem. Phys.* **2016**, *18*, 11000−11007.
(23) Schönhöfer, P. W. A.; Ellison, L. J.; Marechal, M.; Cleaver, D. J.; Schröder-Turk, G. E. Purely entropic self-assembly of the bicontinuous Ia3d gyroid phase in equilibrium hard-pear systems. *Interface Focus* **2017**, *7*, 20160161.
(24) Arora, A.; Morse, D. C.; Bates, F. S.; Dorfman, K. D. Commensurability and finite size effects in lattice simulations of diblock copolymers. *Soft Matter* **2015**, *11*, 4862−4867.
(25) Medapuram, P.; Glaser, J.; Morse, D. C. Universal Phenomenology of Symmetric Diblock Copolymers near the Order-Disorder Transition. *Macromolecules* **2015**, *48*, 819−839.
(26) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-Orientational Order in Liquids and Glasses. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1983**, *28*, 784.
(27) Larsen, P. M.; Schmidt, S.; Schiøtz, J. Robust structural identification via polyhedral template matching. *Modell. Simul. Mater. Sci. Eng.* **2016**, *24*, 055007.
(28) Mukhtyar, A. J.; Escobedo, F. A. Developing Local Order Parameters for Order-Disorder Transitions From Particles to Block Copolymers: Methodological Framework. *Macromolecules* **2018**, *51*, 9769−9780.
(29) Barnes, B. C.; Beckham, G. T.; Wu, D. T.; Sum, J. Two-component order parameter for quantifying clathrate hydrate nucleation and growth. *J. Chem. Phys.* **2014**, *140*, 164506.
(30) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.
(31) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(32) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, 2016.
(33) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(34) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(35) Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Adv. Neural Inf. Process Syst.* **2019**, *32*, 7566−7578.
(36) DeFever, R. S.; Targonski, C.; Hall, S. W.; Smith, M. C.; Sarupria, S. A generalized deep learning approach for local structure identification in molecular simulations. *Chem. Sci.* **2019**, *10*, 7503−7515.
(37) Charles, R. Q.; Su, H.; Kaichun, M.; Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *Computer Vision and Pattern Recognition. Honululu, HI* **2017**, 77.
(38) Kim, B.; Lee, S.; Kim, J. Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6*, eaax9324.
(39) Kingma, D. P.; Ba, J. A method for stochastic optimization. *arXiv:1412.6980*; https://arxiv.org/abs/1412.6980v1, **2014**.
(40) Arvo, J. In *Fast Random Rotation Matrices*; Kirk, D., Ed.; Academic Press, 1992; pp 117−120.
(41) Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D ShapeNets: A deep representation for volumetric shapes. *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* **2015**, 1912−1920.
(42) Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. *Advances in Neural Information Processing Systems 28. Montreal, Canada* **2015**, 2017−2025.
(43) Matsen, M. W.; Bates, F. S. Unifying Weak- and Strong-Segregation Block Copolymer Theories. *Macromolecules* **1996**, *29*, 1091−1098.
(44) Shao, J.; Jiang, N.; Zhang, H.; Yang, Y.; Tang, P. Target-Directed Design of Phase Transition Path for Complex Structures of Rod-Coil Block Copolymers. *ACS Omega* **2019**, *4*, 20367−20380.

https://doi.org/10.1021/acs.jpcb.1c02389
J. Phys. Chem. B XXXX, XXX, XXX−XXX