THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Cite This: J. Phys. Chem. B 2020, 124, 69−78

pubs.acs.org/JPCB

Analysis of Self-Assembly Pathways with Unsupervised Machine Learning Algorithms

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

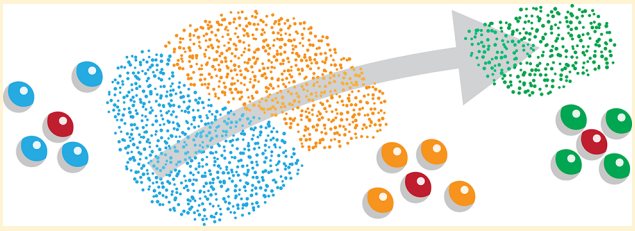
Carl S. Adorf,[†] Timothy C. Moore,[†] Yannah J. U. Melle,[†] and Sharon C. Glotzer*[,†,‡,§]

[†]Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109, United States

[‡]Department of Materials Science and Engineering, University of Michigan, Ann Arbor, Michigan 48109, United States

[§]Biointerfaces Institute, University of Michigan, Ann Arbor, Michigan 48109, United States

ABSTRACT: Colloidal and nanoparticle systems display a rich and exciting phase behavior including the self-assembly of highly complex crystal structures. Nucleation and growth pathways toward crystallization have been studied both computationally and experimentally, but the mechanisms for the formation of the precritical nucleus and consequent crystal growth are yet to be fully understood. Recent advances in the application of machine learning algorithms applied to many-particle systems have led to significant breakthroughs in the ability for high-throughput analysis of phase transitions and the identification of crystal structures. We build upon these techniques to identify and analyze pathways for nucleation and growth in supercooled liquids of colloidal systems modeled with isotropic pair potentials. Our study involves the development of unsupervised machine learning models trained on spherical-harmonics-based descriptors. These models allow us to determine clusters of local environments that are present prior to and during crystallization. We analyze these environments to identify prevalent motifs and local order within the supercooled liquid prior to formation of the critical nucleus.



INTRODUCTION

A full and detailed mechanistic understanding of crystallization pathways would have enormous implications on our ability to develop novel materials with unique properties. The remaining discrepancy between predicted nucleation and growth rates and those measured in experiment suggests that our models of the crystallization process are still insufficient in capturing all relevant effects.[1,2] One particular ambiguity in studying crystallization is the identification of local motifs that particles participate in as they transform from fluid to solid. The advent of machine learning (ML) in the field of chemical physics provides an avenue to reduce such ambiguities and autonomously identify relevant particle environments, as we will show in this paper.

Here we present the analysis of crystallization pathways of colloidal systems, which are especially useful, because simple model systems such as hard spheres and pairwise interacting potentials are amenable to study both theoretically[3−5] and experimentally.[6,7] The fact that colloidal building blocks on the nano- and microscale can be artificially designed and fabricated and, thus, in principle, be optimized for the design of self-assembly routes has sparked the imagination of the scientific community[8] and led to a plethora of computational and experimental studies.[9−20]

One of the main challenges in studying the exact pathways of self-assembly is the ability to access relevant time and length scales. That holds true for both physical and computational experiments but is especially problematic for the study of rare events including nucleation.[21] Rare events like these may have such a minuscule probability to occur on the time scales accessible via brute-force simulation that even increases in orders of magnitude of available computational resources would not be sufficient. Rare event sampling techniques that fall into the general category of advanced sampling techniques, such as umbrella sampling,[22] transition interface sampling,[23] and forward flux sampling,[24,25] provide ways to study rare events but require more consideration and a better understanding of the underlying process compared to simple brute-force approaches. They also might have restrictions such as being only suitable to equilibrium or quasi-equilibrium processes.

Even with sufficient sampling, describing and identifying particle environments in a systematic, unbiased, and computationally feasible manner represents its own challenge. One particularly successful approach is to use Steinhardt bond orientational order parameters[26] based on spherical harmonics, which we will from here on refer to as Steinhardt order parameters, or Q_l in short. Steinhardt order parameters capture

Received: October 13, 2019
Revised: December 7, 2019
Published: December 8, 2019

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78

symmetries that are broken by the particle's local environment in a rotationally invariant manner, which is especially important when we are concerned with the characterization and identification of local motifs before and during the crystallization process as opposed to characterizing bulk crystalline phases.

While useful, Steinhardt order parameters are often insufficient: For one, they require some heuristic on what neighbors to include in their computation. A simple heuristic is to select all points within a certain radius as neighbors. Another heuristic may be based on the number of *n*-nearest neighbors regardless of their distance. A more advanced approach is the determination of neighbors based on the topology of the local environment: The Polyhedral Template Matching (PTM) structure identification algorithm selects the number of neighbors based on the Voronoi cell surrounding a particular particle.[27] Similarly, the solid-angle-based nearest-neighbor algorithm (SANN) also adjusts the number of neighbors for each particle individually on the basis of reaching a threshold of the sum of the solid angle.[28]

However, no matter the environment descriptor and environment selection heuristic applied, the use of order parameters to unambiguously identify structures and phase changes still requires the evaluation and comparison of distributions and arbitrary cutoffs. The design of suitable order parameter (OPs) could, in many instances, be better characterized as an art than a science.[29] However, in the past few years the field has seen a significant shift toward the use of ML-based OPs,[30−34] an approach that has the potential to greatly simplify and accelerate the search for suitable OPs for a new system or process.

In this work, we advance techniques for the autonomous identification of local environments present within a given system, and apply them to characterize the crystallization pathways of colloidal systems, including hard spheres, the Lennard-Jones fluid, and systems of point particles interacting via isotropic pair potentials (IPPs) designed to self-assemble specifically targeted complex crystal structures.[35] We use a descriptor composed of the bispectrum of the three-dimensional rotation group (SO(3)) that is a high-dimensional rotationally invariant representation of a local particle environment. The bispectrum descriptor is similar to the Steinhardt order parameter based on spherical harmonics but preserves more information about the environment than just its symmetry. We show that the high-dimensional descriptor is effective in devising an environment detection model through dimensionality reduction and a clustering-based unsupervised ML workflow.

THEORETICAL METHODS

Models. We studied the crystallization behavior of Weeks−Chandler−Anderson (WCA) spheres and systems of point particles interacting via the Lennard-Jones (LJ) potential and potentials optimized for the self-assembly of A15-type *cP*8−Cr_3Si (*cP*8) and *β*-tin *tI*4-Sn (*tI*4) crystal structures, all of which are plotted in Figure 1. The models for *cP*8 and *tI*4 were obtained with an algorithm designed for the optimization of simple, short-ranged IPPs that will self-assemble complex crystal structures.[35] To our knowledge, detailed self-assembly pathways of these crystal structures using computational models have not been reported.

Unsupervised Machine Learning Workflow. To study the self-assembly pathways of models for colloidal self-assembly, we employ a workflow, illustrated in Figure 2, that uses a generalized descriptor space as input and produces an environment classification as output. In this way we are able to assign to each particle at each sampled time step a label corresponding to exactly one environment group. The descriptor is in principle independent of the studied systems but must be general enough to incorporate all important features that sufficiently define a local particle environment at least up to its first, but optionally also to its second, neighbor shell. At a minimum, for a system that assembles a specific target structure, the descriptor space must have features that capture the broken symmetries of that structure.

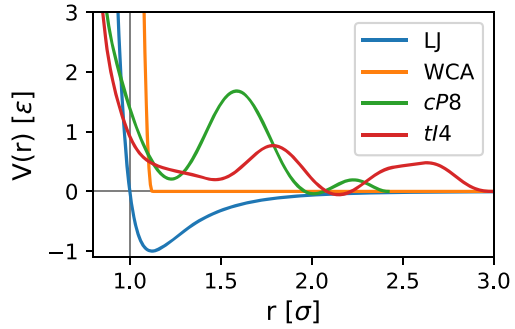


Figure 1. We analyzed the nucleation and growth crystallization pathways for systems of point particles interacting via IPPs. We studied two benchmark models (Lennard-Jones (LJ) and Weeks−Chandler−Anderson (WCA)) as well as two models that were specifically designed to assemble complex crystal structures (*cP*8 and *tI*4).

The "complete feature vector" has length D and is a concatenation of multiple descriptors with the assumption that the unsupervised ML model is going to be able to discern the relevant features automatically and will discard all features that are irrelevant for a given system and pathway. The complete descriptor space is then an array with shape $N \times D$, where N corresponds to our sample size, which is spanned by the number of sampled particles and time steps.

After computing the descriptor space, we apply two dimensionality reduction steps to reduce the overall data size and thus computational demand and to prepare the data for the crucial clustering step. We first reduce the dimensionality from D to 20 by transforming the descriptor space with incremental Principal Component Analysis (PCA)[36,37] and selecting the first 20 components with the largest variance and hence the most information. Our preliminary studies showed that the relative variance typically drops drastically after 10−20 components, which means that including more than that would not lead to more accurate results.

A subset corresponding to 1000 randomly selected particles of the resulting data space embedding is then further reduced in dimensionality using the density-based and nonlinear Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm.[38] The UMAP algorithm has natural advantages for our problem space. It reduces the dimensionality of the descriptor space computationally more efficiently than comparable algorithms such as as t-SNE.[39] Furthermore, UMAP optimizes the resulting manifold as much as possible to preserve distances between points within the higher-dimensional space. This means that the resulting topology and distances between points is related to the actual high-dimensional representation of the system and thus

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78

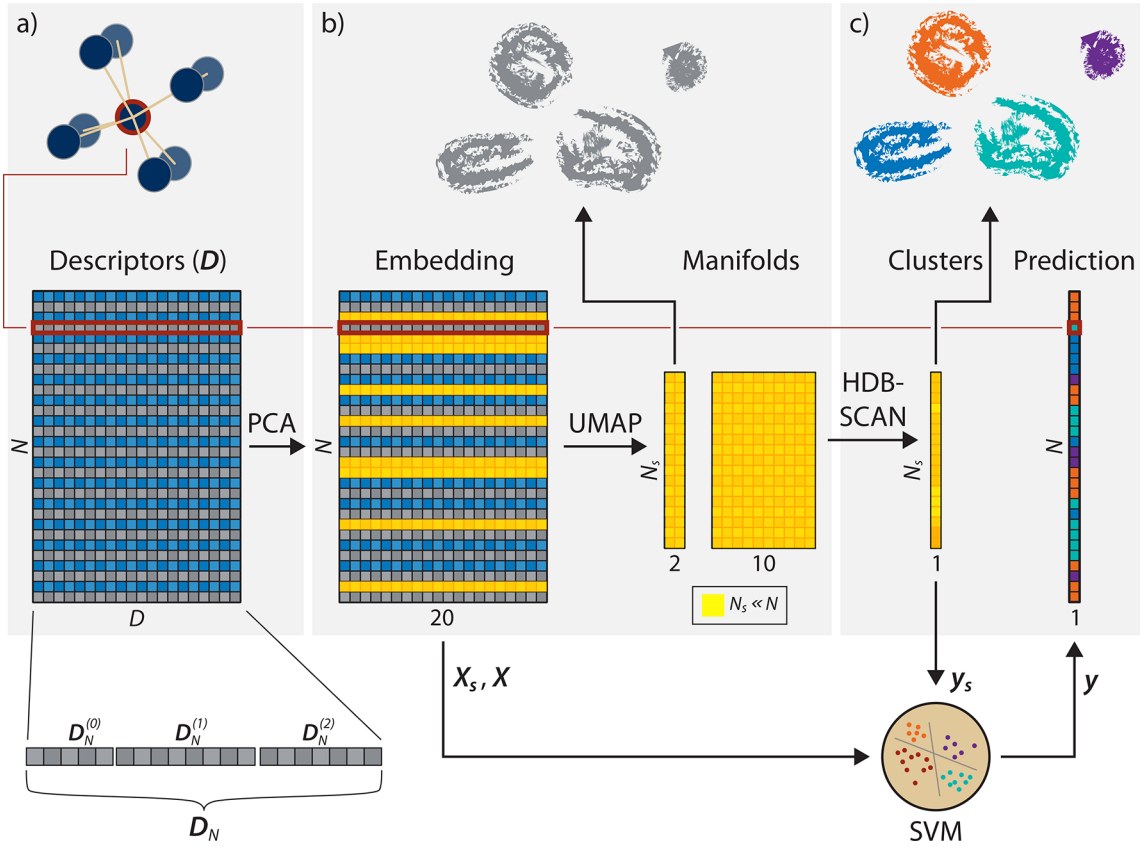


Figure 2. We employ an unsupervised ML workflow for the automated detection of environments within the crystallization pathway. Particle environments are first mapped to a descriptor space, which should capture the local environment of each particle in a rotationally invariant manner (a). This potentially high-dimensional descriptor space is then reduced to its first 20 principal components using a Principal Component Analysis (PCA) decomposition and subsequently reduced to 2 and 10 dimensions using the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm (b). Environments are detected from the UMAP manifold using the HDBSCAN* clustering algorithm (c).

potentially informative and interpretable. The UMAP algorithm was applied with the number of neighbors parameter set to 30, the minimum distance set to zero, and with point distances computed using the "Manhattan" norm. The "Manhattan" metric is advantageous compared to the Euclidean metric in preserving distances in higher dimensional spaces.[40]

We reduced the descriptor space embedding to two dimensions for visualization and validation purposes and to 10 dimensions for the actual unsupervised learning via a clustering algorithm. The lower-dimensional manifolds constitute a representation of the self-assembly process as described by the descriptor vectors, which naturally cluster into groups that may be associated with phase changes and particle environments present in the studied system. Concretely, we expect environments that are found within the same cluster to be more similar to each other than to environments in a different cluster.

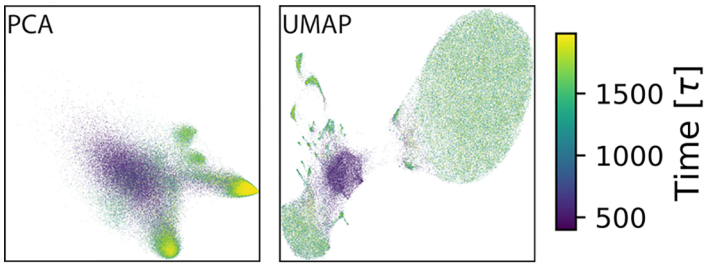
We apply the HDBSCAN* clustering algorithm[41] to automatically identify clusters within the 10-dimensional descriptor space manifold. The HDBSCAN* algorithm represents an extension of the popular DBSCAN algorithm that is more robust against fluctuations in density and cluster size within the clustered space compared to standard DBSCAN.

Like DBSCAN, HDBSCAN* clusters points by their local neighborhood density, building up a graph of connected points referred to as a cluster hierarchy. However, unlike DBSCAN, which selects clusters within the hierarchy based on a global cutoff distance, HDBSCAN* merges connected nodes until the cluster associated with each node reaches a specified minimum cluster size. That means the effective cutoff range will be different for each cluster. We determined that a minimum cluster size within the range of 1−4% of the system size in combination with a minimum sample size of 1% of the system size is suitable for our application. We further used the "leaf" cluster selection mode, which means that HDBSCAN* selects all leafs within the merged cluster hierarchy, resulting in smaller and more homogeneous clusters.

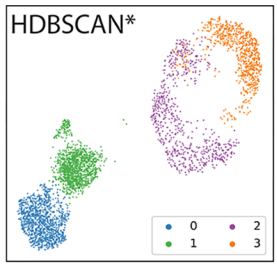
As aforementioned, the manifold generation and its subsequent clustering was applied to a subset of 1000 randomly selected particles for all or a subset of all sampled time steps. To predict labels for *all* particles at *all* time steps, we trained a linear Support Vector Machine (SVM) model on the embedded PCA descriptor space (excluding all points classified as noise) with a 3-to-1 training-validation split. The validation accuracy obtained for these models was typically close to or even above 98%. Training a separate model has the distinct advantage of avoiding the need to compute the manifold for the complete descriptor space as well as simplifying the labeling of new data from either the same system or possibly even a different system.

Descriptors. Within the scope of this study, we employed various descriptors based on the local neighborhood including

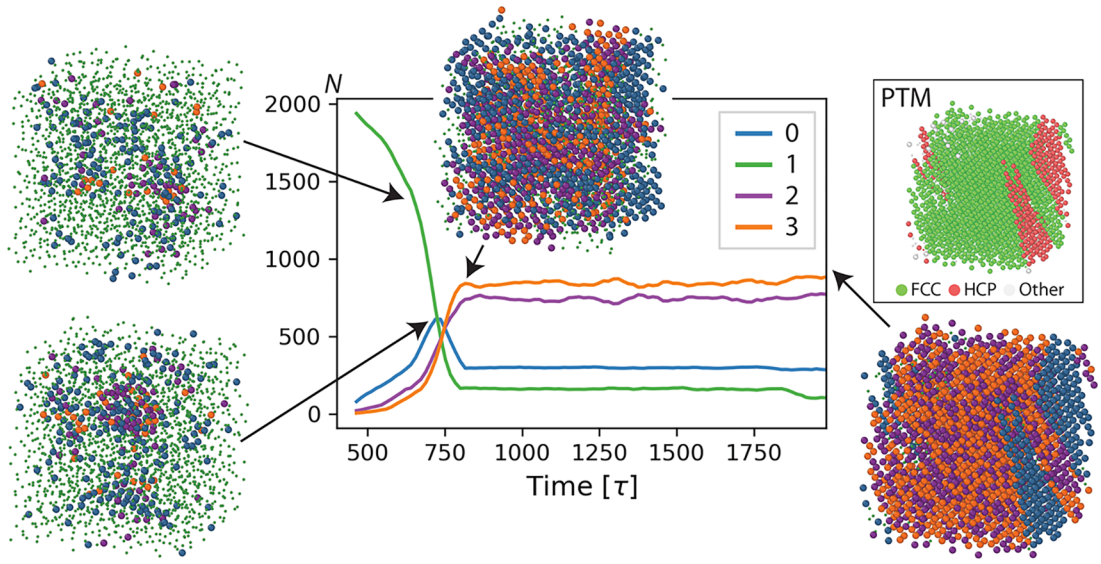
DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78



(a) PCA embedding (left) and 2D-UMAP manifold (right)



(b) UMAP colored by environments



(c) Number of environments over time; fluid environment 1 is visualized with reduced size for clarity

Figure 3. The first two principal components of the descriptor space (a left), where each point represents one particle and its environment colored by time, have a broad initial basin with two smaller competing basins at a later stage. In contrast to the PCA embedding, the two-dimensional UMAP manifold is optimized to represent the complete topology of the high-dimensional descriptor space and for this system displays better separation between unrelated environments (a right). We used the HDBSCAN* algorithm to cluster the 10-dimensional manifold and identify four environments (b). The supersaturated fluid is dominated by a liquid environment (1), with fluctuations of random hexagonal close packing (RHCP) environments (0). The RHCP environments present a precursor for the nucleation of the crystal nucleus, which is primarily made up of FCC environments (2 and 3). We compared our identification of environments to those obtained with PTM (c).

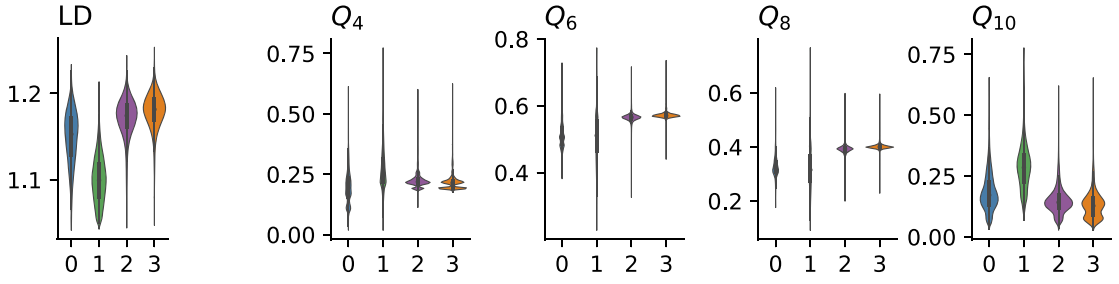
bond distances and bond angles, as well as spherical-harmonics-based descriptors such as Steinhardt order parameters and the bispectrum. These order parameters are implemented within the open-source pythia ML package.[33,42] We found the spherical-harmonics-based bispectrum descriptor[43] to generally be most robust to universally describe the self-assembly pathway characteristics across all studied systems compared to Steinhardt-order-parameter- and bond- or angle-based descriptors.

We also compared our results against the traditionally employed Steinhardt order parameters for comparison and validation purposes. While the bispectrum descriptor preserves much more information compared to the Steinhardt order parameters, the latter require only one dimension per symmetry variant and neighborhood size. That is significantly less than the bispectrum descriptor, whose dimension is proportional to the cube of l_{max}, which is not a huge problem for the initial dimensionality reduction using PCA but vastly increases the required storage size if the descriptor is to be preserved. We find it encouraging that our ML algorithm easily handles descriptors with up to at least 5368 dimensions, as is the case for the *cP*8 system presented later.

Environment Characterization. The environments identified by our unsupervised ML workflow can be characterized through visual inspection as well as various order parameters, including the Local Density (LD, the Steinhardt order parameter for different values of l, and their root-mean-squared deviation. For this we collect a sample of particles that are all classified to belong to the same environment and then compute the given order parameter such as the LD for that sample. This postclassification analysis serves as additional validation, since the original classification via the descriptor-based unsupervised ML model is independent of this step.

Data Management and Workflow Implementation. The computational workflow in general and data management in particular for this publication was primarily supported by the signac data management framework.[44] Simulations were performed with HOOMD-blue.[45] Trajectories were analyzed with freud.[46] The descriptors were calculated with pythia.[42] The rendered visualizations of particle systems were generated with OVITO.[47] Computations were carried out on the Oak

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78



(a) Local Density (b) Steinhardt order parameters Q_l

Figure 4. The Local Density (LD) order parameter is significantly lower for the fluid environment (2) compared to all other environments, with the HCP-like environment (0) between the fluid and FCC (3 and 4) (a). The Steinhardt order parameter Q_l, here evaluated for l = 4, 6, 8, and 10, shows narrow distributions of Q_6 for both HCP (0) and FCC (3 and 4). We see that there is significant overlap for all environments, which supports that using Q_l alone in combination with hard cutoffs would make it difficult to reliably distinguish between the identified environments (b). The plots show the interquartile range of the data as a vertical bar, the median as a dot, and whiskers that extend to the full range of the data excluding all data points that are considered outliers, overlaid by a kernel density estimate of the distribution. Please see the appendix for details.

Ridge National Laboratory Titan and Summit super computers as well as the University of Michigan Flux cluster.

RESULTS AND DISCUSSION

Weeks−Chandler−Anderson (WCA). The self-assembly of colloidal hard spheres was simulated with the WCA model

$$V_{\mathrm{WCA}}(r)\begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12}-\left(\frac{\sigma}{r}\right)^{6}+\frac{1}{4}\right] & r \le 2^{1/6}\sigma \\ 0 & r > 2^{1/6}\sigma \end{cases} \quad (1)$$

which approximates the interaction of physical hard spheres via a purely repulsive and short-ranged energy function. The self-assembly behavior of WCA spheres has been studied extensively[2,4,48] and was recently analyzed with a similar unsupervised ML approach by Boattini et al.[34] The system represents a suitable benchmark for the validation of novel colloidal crystallization pathway analysis methods. We sampled the system with Molecular Dynamics in the *NPT* ensemble with a constant system size N, pressure P, and temperature T. The integration was carried out via a MTK barostat−thermostat,[49−51] a time step of $\mathrm{d}t = 0.001\tau$, a thermal coupling constant of $\tau_T = 0.01$, and a pressure coupling constant of $\tau_P = 0.1$. We initialized the system with a largely expanded box and at an elevated temperature of $\beta\epsilon = 40/3$ and then equilibrated at the same temperature to a pressure of $\beta P\sigma^3 = 30$, where β is the inverse temperature $1/kT$. At time $t = 400\tau$ the temperature was abruptly reduced to $\beta\epsilon = 40$.

We computed the bispectrum descriptor for 12 neighbors and, as described in the previous section, mapped the descriptor space with PCA and the UMAP algorithm from $D = 1342$ dimensions to 20, 10, and 2 dimensions (Figure 3a). Although we visualize only the first two principal components, we apply the UMAP algorithm twice, once to generate a two-dimensional manifold for visualization and once to generate a higher-dimensional representation used for clustering. Both UMAP manifolds are inferred from the 20-dimensional PCA reduced embedding.

Each point in the plots in Figure 3a represents one particle at a specific time step, which is signified in terms of its color from dark to bright. The location of each point on the two-dimensional graph represents its environment in terms of the descriptor vector mapped into two dimensions.

Simply by examining this lower-dimensional representation of the descriptor space, we can already infer some characteristics of the self-assembly pathway and the final phase:

1. There are two distinct environment groups within the system.
2. One group is present at both an early and a late stage of the simulation while the other one is only present at a later stage.
3. The environment present at both the early and late stages is more similar to the initial (disordered) fluid phase than it is to the second environment group.

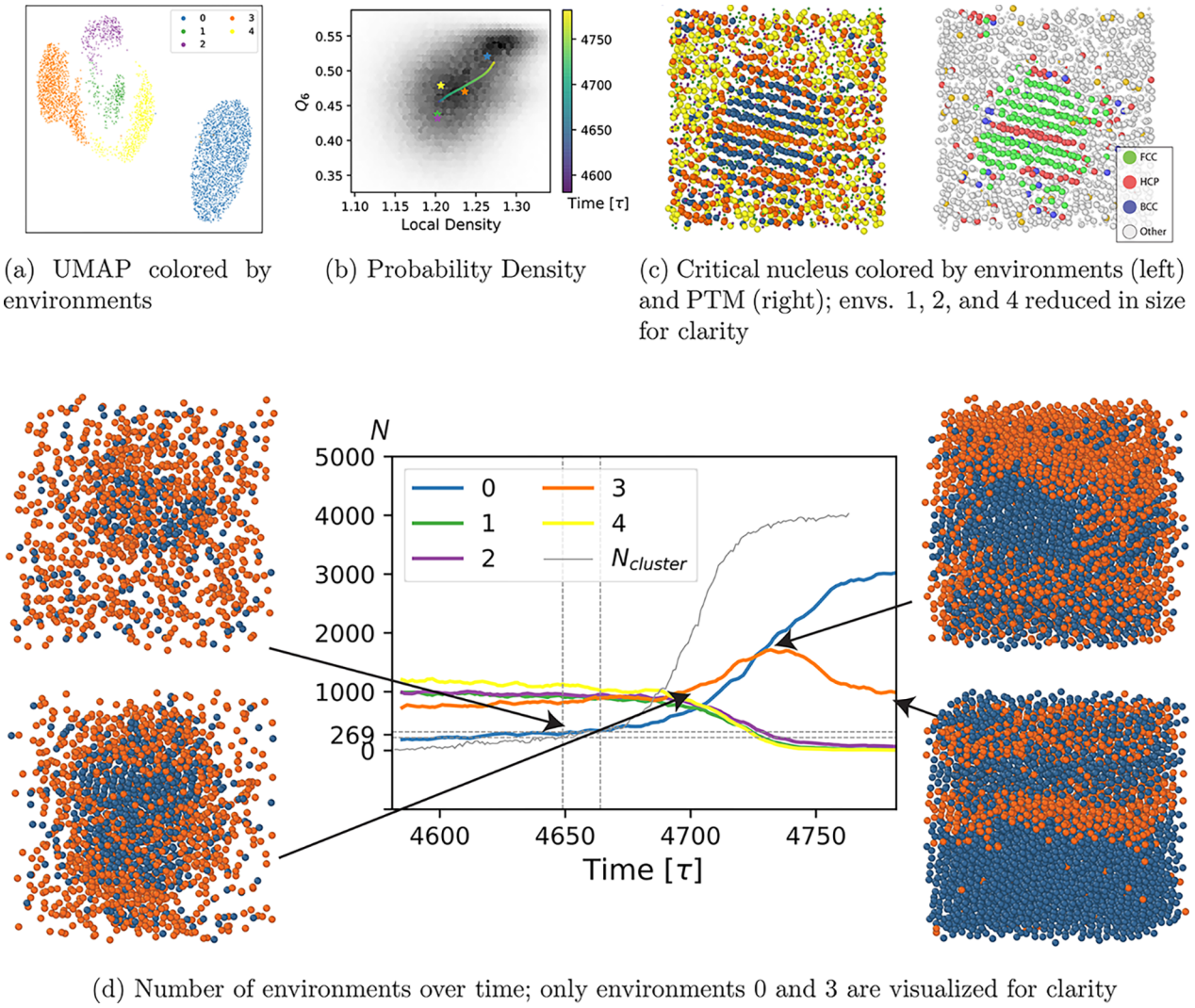
We then applied the HDBSCAN* clustering algorithm to automatically cluster the indiscriminate point cloud shown in Figure 3a into the four clusters shown in Figure 3b. The clustering algorithm clearly delineates between the two primary clusters; however, it also splits these further into two smaller clusters that might not have been immediately obvious through simple visual inspection in two dimensions. As stated before, the clustering algorithm operates on the 10-dimensional reduced descriptor space, taking in more information about distances between points than visible to the eye.

Through visual inspection in combination with PTM, we identify the four principal environments to correspond to two phases, as shown in Figure 3c. We identify the initially dominant environment to correspond to a disordered fluid (environment 1). The second most dominant phase consists of random fluctuations of RHCP (0), which find a local maximum around 750τ, playing a strong role in the formation of the crystal nucleus (Figure 3c). Shortly after, the two FCC-like environments (2 and 3) become dominant. The remaining 15% of all particles identified within environment 0 persist in the form of a hexagonal close packing (HCP)-stacking fault.

The fluid-like environment (1) has a significantly lower median LD compared to all other environments; see Figure 4a. The HCP-like environment (0) has also a slightly lower median LD compared to the two FCC-like environments (2 and 3), which are both very similar to each other. Both environments 3 and 4 are highly similar; however, environment 4 has a slightly higher LD and lower Q_{10} Steinhardt order parameter (Figure 4b), which means its local order is slightly higher.

Our simulation and analysis of the WCA system confirms observations made previously by Auer and Frenkel,[52] Kawasaki and Tanaka,[4] and recently by Ren et al.,[2] that fluctuations of

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78



(a) UMAP colored by environments

(b) Probability Density

(c) Critical nucleus colored by environments (left) and PTM (right); envs. 1, 2, and 4 reduced in size for clarity

(d) Number of environments over time; only environments 0 and 3 are visualized for clarity

Figure 5. The UMAP representation of the Lennard-Jones (LJ) assembly pathway descriptor space (a) is divided into two major cluster groups with a clear separation of the FCC environment (0) on the right. The probability density diagram (b) features two wells with the deepest well located at the top coinciding with environment 0. We observe that the two HCP-like environments (3 and 4) are located halfway between the fluid environments (1 and 2). Environment 3 has a larger LD compared to environment 4. The visualization of the critical nucleus (c) colored by environments on the left and with PTM on the right shows that the HCP-like environments with higher density (3) are incorporated into the critical nucleus, while those with lower density (4) are primarily found in the bulk. The critical nucleus has a significantly higher concentration of FCC environments compared to the bulk; however, there is a large number of spontaneous fluctuations of HCP order within the fluid at all times, consituted as HCP environments, that then increase in density and finally transform into FCC (d).

RHCP provide the necessary precursory conditions for the formation of the nucleus, which predominantly consists of FCC-like environments with higher Q_6 compared to HCP. Local body-centered cubic (BCC)-like environments do not play a role in the formation of the critical nucleus. Having now validated our methodology, we study increasingly complex systems in the following subsections.

Lennard-Jones (LJ). The nucleation and growth characteristics of the LJ system have been extensively studied;[53−55] however, there are a few remaining questions. The nucleation rates determined from simulations are still orders of magnitude away from experimental measurements,[56] and the exact composition of the nucleus in its early stage is still somewhat dependent on the chosen order parameter.[21]

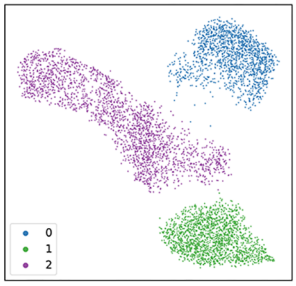
We sampled the crystallization behavior of 4096 point particles interacting via the LJ potential

$$V_{LJ}(r)\begin{cases} V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] & r \le r_{cut} \\ 0 & r > r_{cut} \end{cases} \quad (2)$$

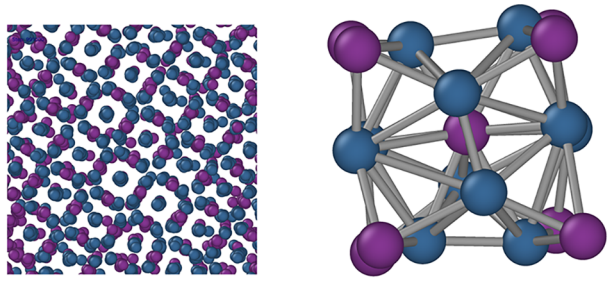
with a cutoff range of $r_{cut} = 3\sigma$ without shift at the cutoff using Forward Flux Sampling (FFS) at moderate supercooling with a temperature of $kT = 0.85\epsilon$ and a pressure of $P = 5.76\epsilon\sigma^{-3}$ in the *NPT* ensemble. The integration time step was set to $dt = 0.01\tau$, and the system was coupled to the environment with $\tau_T = 0.1$ and $\tau_P = 1$. Details on the implementation of the advanced sampling with FFS are described in the appendix.

Similar to the WCA system, the pathway analysis of the LJ system reveals two dominant environment groups: Group one represents the environment of the final crystal structure FCC (0), and group two is made up of fluid- (1 and 2) and HCP-like (3 and 4) environments. The two fluid-like environments

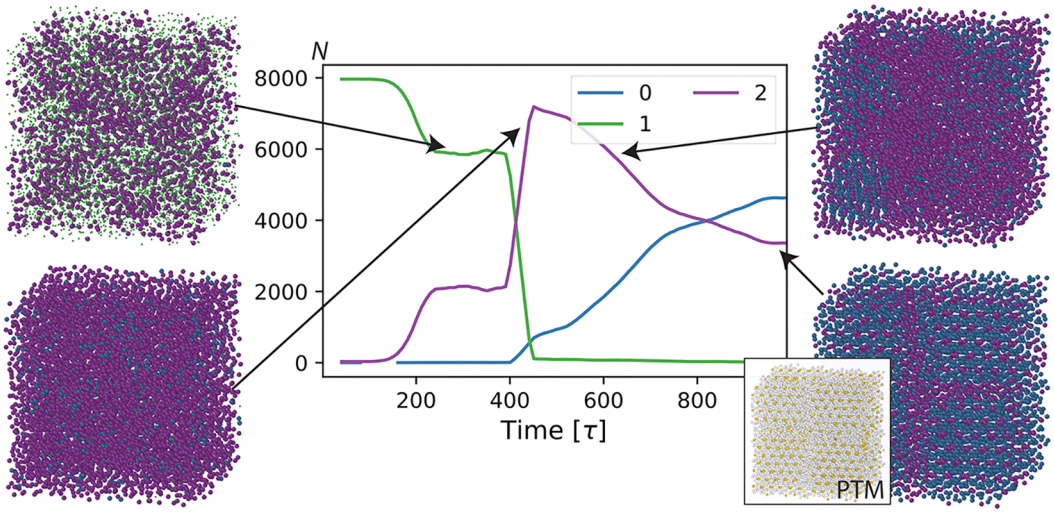
DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78



(a) UMAP colored by environments



(b) A cut of the final crystal structure (left) and one unit cell (right) colored by environments



(c) Number of environments over time; fluid environment 1 is visualized with reduced size for clarity

Figure 6. The *cP*8 UMAP descriptor space representation (a) has three primary clusters, a disordered fluid (1), an icosahedral environment (2) that corresponds to *cP*8 Wyckoff site a, and a third environment only present in the ordered crystal structure (0), which corresponds to Wyckoff site c (b). The manifold topology suggests that environments 2 and 1, i.e., the icosahedral environment and the fluid, are more closely related. Environment 2 reaches a brief local maximum right after the temperature quench at 400τ and then partially transitions to environment 0 (c). The PTM representation of the final crystal structure shows that environment 2 corresponds to an icosahedral motif (gold), while environment 0 is not recognized (white) (see inset).

have low density and order, whereas the two HCP-like environments have similar densities, but higher local order; see Figure 5b. We can infer from the manifold that RHCP/HCP is overall more similar to the fluid than FCC (Figure 5a).

The pathway as visualized in the probability density plot of LD vs Q_6 (Figure 5b) proceeds from the fluid (1 and 2) to RHCP/HCP (3 and 4) to FCC (0) and is characterized by simultaneous densification and ordering. There are two primary wells with a clear separation between FCC and all other environments, with the latter being shallower and broader.

The critical crystal nucleus (see Figure 5c) is primarily made up of FCC with one HCP stacking fault and wetted by HCP environments (3). Fluctuations of RHCP (environment 4) are present within the bulk, meaning that the system has overall very high HCP-like order at all times with a significant density gradient at the surface of the crystal nucleus.

Looking at the crystallization over time, fluctuations of RHCP and HCP environments (3 and 4) are present at all times. However, the increase in the number of particles with FCC environment (0) coincides with the nucleus reaching a critical size at 4650τ, which makes sense considering that the nucleus is primarily made up of particles with the FCC environment. The crystal growth phase is characterized by a simultaneous increase of both HCP (3) and FCC (0) environments with HCP being initially dominant and reaching a maximum at 4730τ at which point FCC becomes the dominant environment.

A15-type *cP*8-Cr_3Si. The crystallization of the *cP*8 structure was simulated with the pairwise interaction potential presented in previous work.[35] The system was equilibrated with the Langevin integrator (time step dt = 0.001τ) at an elevated temperature of kT = 3.0ϵ and a slightly expanded box for 100τ and then compressed to the final density for another 100τ. The temperature was then reduced abruptly to kT = 1.0ϵ at 400τ.

The descriptor for the *cP*8 structure was computed for 12, 13, 24, and 26 neighbors to include the first and second neighbor shells, which results in the identification of the three environments shown in Figure 6a. These environments can be attributed to the disordered fluid (1), an intermittently dominant icosahedral environment (2) corresponding to Wyckoff site a, and an environment that is only present in the final crystal structure (0) and corresponds to Wyckoff site c (see Figure 6b).

Roughly one-fourth of the particles were found to be within the icosahedral environment (2) prior to the quench, with a significant uptake to almost 100% immediately after the

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78

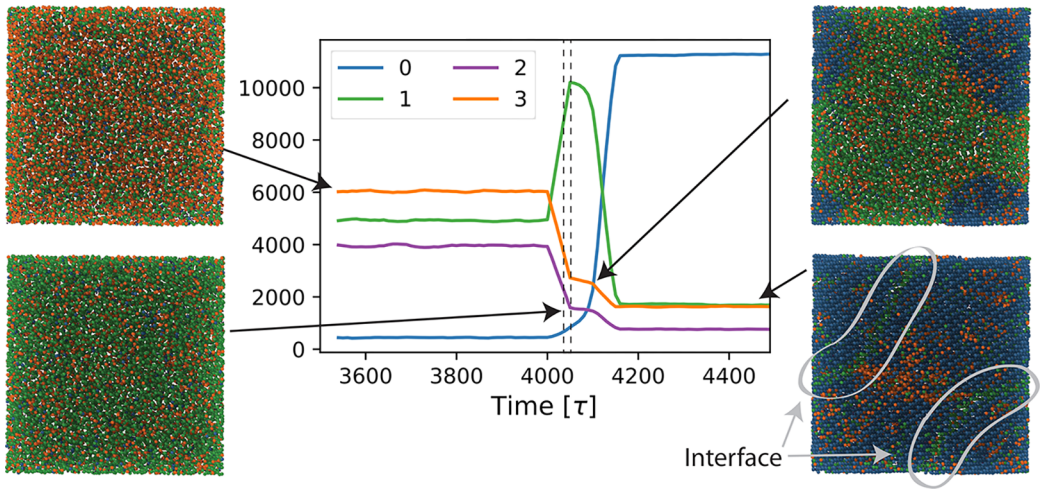


Figure 7. The *tI*4 crystallization pathway is characterized by a precursor environment (1), which becomes dominant shortly after the temperature quench at 4000τ and is then consumed by the crystalline environment (0). The precursor environment persists in the form of a spherical defect with interfaces circled in the bottom right visualization.

temperature quench (see Figure 6c). These environments became incorporated into the crystal nucleus and the final crystal structure over a prolonged growth period. The final crystal structure has two defect planes, which are visually apparent by a reduced concentration of environment 0.

***β*-tin *tI*4-Sn.** The crystallization of the *tI*4 structure was simulated with the Langevin integrator (dt = 0.001) with the potential presented in previous work.[35] The system, composed of 14400 point particles, was equilibrated at $kT = 3.0\epsilon$ for 4000τ, and then quenched to $kT = 0.1\epsilon$.

There are three environments present in significant amounts prior to the temperature quench: environments 1, 2, and 3. Environment 1 becomes dominant immediately after the temperature quench at 4000τ only to be replaced shortly after by the crystalline environment 0. Environment 1 remains present in significant concentration localized at a defect in the form of a spherical shell (see Figure 7).

The UMAP representation of the *tI*4 self-assembly descriptor space is interpretable in terms of both the relative distance between clusters and their relative location within the two-dimensional coordinate system. The clustering algorithm identifies four clusters within the descriptor space, with two fluid environments (2 and 3), a precursor environment (1), and a crystalline environment (0) visualized in Figure 8.

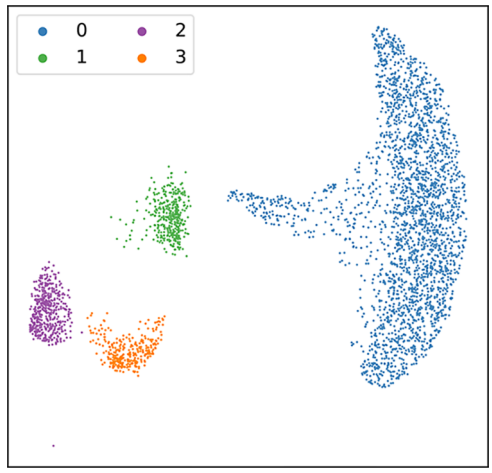


Figure 8. The UMAP representation of the *tI*4 descriptor space separates into two main cluster groups, which we can identify as pre- and postcrystallization from the time coding alone (not shown). The clustering algorithm identifies three distinct environments within the left group, which we later associate with the disordered fluid state (2 and 3) and a precursor environment (1).

CONCLUSIONS

Using a general spherical-harmonics-based descriptor in combination with an unsupervised ML workflow, we studied the self-assembly of four interaction models that assemble a variety of crystal structures. The applied clustering algorithm autonomously identified groups of local structural motifs, here referred to as environments, that helped us to discriminate between a fluid and a solid phase as well as specific crystal structures, e.g., FCC and HCP and related structures such as RHCP. The autonomous environment classification performed as well as established methods such as PTM in discriminating between fluid and solid phases and in identifying the final structure, interfaces, and defects.

The employed bispectrum descriptor takes multiple symmetries into account and was only adjusted with respect to the number of neighbors for a particular system. We used the coordination number of particles within the final crystal structure as a guideline on the number of neighbors selection; however, future studies should use a more advanced neighbor selection heuristic, for example, based on the SANN algorithm[28] or by employing Voronoi cell division.[27]

All pathways show significant fluctuations of local structural order within the disordered bulk fluid phase prior to nucleation. This suggests that nucleation events are triggered by a combination of random fluctuations in both local order and density, providing further evidence for the hypothesis by Russo and Tanaka[57] that the supercooled bulk liquid is highly structured to lower its free energy compared to a completely isotropic bulk phase. We can further confirm their insight on the importance of the structure of the embryonic crystal nucleus on the final macro state, evidenced by our observation that stacking faults that are present even at a very early stage, for example in the LJ nucleus, persist even after substantial crystal growth and opportunity for defect healing (see Figure 5c,d).

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78

Finally, the relationship between the detected environments, as manifested in their placement on the low-dimensional manifold representation of the descriptor space, provides further insight into the role of precursory environments for the crystallization process. All studied systems have precursory environments that are more similar to those predominantly found in the liquid compared to those predominantly found in the solid. We allege that this mechanism, akin to Ostwald's step rule, is more obvious and easier to infer from the manifold representation as opposed to distributions of more conventional order parameters, for example, those of Steinhardt and local density.

Violin Plots

Violin plots show the median of the data (dot), the interquartile range in form of a vertical box, and the complete range of the data including the minimum and maximum except for points that are considered outliers as the gray line. Outliers are defined as any point that extends 1.5 times the interquartile range below the first and above the third quartile. In addition, the graph also features a kernel density estimate (KDE) of the distribution, which makes it easier to visually asses the overall distribution of the data.

Forward Flux Sampling (FFS)

Forward Flux Sampling (FFS) is an advanced sampling technique originally presented by Allen et al.[25] that allows the sampling of pathways that have rare events in a way that is computational feasible, is largely unbiased with respect to the chosen order parameter, and allows the study of non-equilibrium systems.[24]

We studied the crystallization of the LJ system at moderate supercooling, which made it necessary to use advanced sampling techniques to observe nucleation events on a reasonable computational time scale. The nucleation rate of the WCA spheres as well as of the systems of point particles interacting via the *tI*4 and *cP*8 pairwise interactive potentials is comparatively higher at the chosen state points and could therefore be observed with brute-force sampling.

We employed the solid−liquid order parameter introduced by ten Wolde et al.[53] as implemented in the freud analysis package.[46] Using this order parameter, a particle is considered solid-like if and only if it has more than 10 nearest neighbors where the dot product

$$q_l = \bar{Q}^*_{lm}(i)\cdot\bar{Q}_{lm}(j) \tag{3}$$

exceeds the threshold of $q_l > 0.6$. The value for $Q_{lm}(i)$ is determined by evaluating

$$\bar{Q}_{lm}(i) = \frac{1}{N_b(i)}\sum_{j=1}^{N_b(i)} Y_{lm}(\vec{r}_{ij}) \tag{4}$$

over the N_b nearest neighbors of particle i, where $\vec{r}_{ij}$ denotes the bond vector between particle i and particle j and $Y_{lm}(\vec{r})$ are spherical harmonics. The number of nearest neighbors was determined using a ball-metric with a cutoff radius of $r_{cut} = 1.35\sigma$.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sglotzer@umich.edu.

ORCID

Sharon C. Glotzer: 0000-0002-7197-0085

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a grant from the Simons Foundation (256297, SCG). This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. This work was also supported in part through computational resources and services supported by Advanced Research Computing at the University of Michigan, Ann Arbor.

REFERENCES

(1) Whitelam, S.; Jack, R. L. The Statistical Mechanics of Dynamic Pathways to Self-Assembly. *Annu. Rev. Phys. Chem.* **2015**, *66*, 143−163.

(2) Ren, S.; Sun, Y.; Zhang, F.; Travesset, A.; Wang, C.-Z.; Ho, K.-M. Calculation of Critical Nucleation Rates by the Persistent Embryo Method: Application to Quasi Hard Sphere Models. *Soft Matter* **2018**, *14*, 9185−9193.

(3) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *J. Chem. Phys.* **1971**, *54*, 5237−5247.

(4) Kawasaki, T.; Tanaka, H. Formation of a Crystal Nucleus from Liquid. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 14036−14041.

(5) Russo, J.; Tanaka, H. The Microscopic Pathway to Crystallization in Supercooled Liquids. *Sci. Rep.* **2012**, *2*, 505.

(6) Gasser, U.; Weeks, E. R.; Schofield, A.; Pusey, P. N.; Weitz, D. A. Real-Space Imaging of Nucleation and Growth in Colloidal Crystallization. *Science* **2001**, *292*, 258−262.

(7) Tan, P.; Xu, N.; Xu, L. Visualizing Kinetic Pathways of Homogeneous Nucleation in Colloidal Crystallization. *Nat. Phys.* **2014**, *10*, 73−79.

(8) Glotzer, S. C.; Solomon, M. J. Anisotropy of Building Blocks and Their Assembly into Complex Structures. *Nat. Mater.* **2007**, *6*, 557−562.

(9) Tang, Z.; Zhang, Z.; Wang, Y.; Glotzer, S. C.; Kotov, N. A. Self-Assembly of CdTe Nanocrystals into Free-Floating Sheets. *Science* **2006**, *314*, 274−278.

(10) Iacovella, C. R.; Keys, A. S.; Horsch, M. A.; Glotzer, S. C. Icosahedral Packing of Polymer-Tethered Nanospheres and Stabilization of the Gyroid Phase. *Phys. Rev. E* **2007**, *75*, 040801.

(11) Zhang, Z.; Tang, Z.; Kotov, N. A.; Glotzer, S. C. Simulations and Analysis of Self-Assembly of CdTe Nanoparticles into Wires and Sheets. *Nano Lett.* **2007**, *7*, 1670−1675.

(12) Leunissen, M. E.; Dreyfus, R.; Cheong, F. C.; Grier, D. G.; Sha, R.; Seeman, N. C.; Chaikin, P. M. Switchable Self-Protected Attractions in DNA-Functionalized Colloids. *Nat. Mater.* **2009**, *8*, 590−595.

(13) Henzie, J.; Grünwald, M.; Widmer-Cooper, A.; Geissler, P. L.; Yang, P. Self-Assembly of Uniform Polyhedral Silver Nanocrystals into Densest Packings and Exotic Superlattices. *Nat. Mater.* **2012**, *11*, 131−137.

(14) Knorowski, C.; Burleigh, S.; Travesset, A. Dynamics and Statics of DNA-Programmable Nanoparticle Self-Assembly and Crystallization. *Phys. Rev. Lett.* **2011**, *106*, 215501.

(15) Macfarlane, R. J.; Lee, B.; Jones, M. R.; Harris, N.; Schatz, G. C.; Mirkin, C. A. Nanoparticle Superlattice Engineering with DNA. *Science* **2011**, *334*, 204−208.

(16) Damasceno, P. F.; Engel, M.; Glotzer, S. C. Predictive Self-Assembly of Polyhedra into Complex Structures. *Science* **2012**, *337*, 453−457.

(17) Ye, X.; Chen, J.; Engel, M.; Millan, J. A.; Li, W.; Qi, L.; Xing, G.; Collins, J. E.; Kagan, C. R.; Li, J.; et al. Competition of Shape and Interaction Patchiness for Self-Assembling Nanoplates. *Nat. Chem.* **2013**, *5*, 466−473.

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78

(18) de Nijs, B.; Dussi, S.; Smallenburg, F.; Meeldijk, J. D.; Groenendijk, D. J.; Filion, L.; Imhof, A.; van Blaaderen, A.; Dijkstra, M. Entropy-Driven Formation of Large Icosahedral Colloidal Clusters by Spherical Confinement. *Nat. Mater.* **2015**, *14*, 56−60.
(19) Grünwald, M.; Geissler, P. L. Patterns without Patches: Hierarchical Self-Assembly of Complex Structures from Simple Building Blocks. *ACS Nano* **2014**, *8*, 5891−5897.
(20) Wang, M. X.; Brodin, J. D.; Millan, J. A.; Seo, S. E.; Girard, M.; Olvera De La Cruz, M.; Lee, B.; Mirkin, C. A. Altering DNA-Programmable Colloidal Crystallization Paths by Modulating Particle Repulsion. *Nano Lett.* **2017**, *17*, 5126−5132.
(21) Sosso, G. C.; Chen, J.; Cox, S. J.; Fitzner, M.; Pedevilla, P.; Zen, A.; Michaelides, A. Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations. *Chem. Rev.* **2016**, *116*, 7078−7116.
(22) Torrie, G. M.; Valleau, J. P. Monte Carlo Free Energy Estimates Using Non-Boltzmann Sampling: Application to the Sub-Critical Lennard-Jones Fluid. *Chem. Phys. Lett.* **1974**, *28*, 578−581.
(23) Van Erp, T. S.; Moroni, D.; Bolhuis, P. G. A Novel Path Sampling Method for the Calculation of Rate Constants. *J. Chem. Phys.* **2003**, *118*, 7762−7774.
(24) Allen, R. J.; Valeriani, C.; Rein Ten Wolde, P. Forward Flux Sampling for Rare Event Simulations. *J. Phys.: Condens. Matter* **2009**, *21*, 463102.
(25) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Forward Flux Sampling-Type Schemes for Simulating Rare Events: Efficiency Analysis. *J. Chem. Phys.* **2006**, *124*, 194111.
(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*, 784−805.
(27) Larsen, P. M.; Schmidt, S.; Schiøtz, J. Robust Structural Identification via Polyhedral Template Matching. *Modell. Simul. Mater. Sci. Eng.* **2016**, *24*, 055007.
(28) van Meel, J. A.; Filion, L.; Valeriani, C.; Frenkel, D. A Parameter-Free, Solid-Angle Based, Nearest-Neighbor Algorithm. *J. Chem. Phys.* **2012**, *136*, 234107.
(29) Sethna, J. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*; OUP Oxford, 2006.
(30) Wang, L. Discovering Phase Transitions with Unsupervised Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *94*, 195105.
(31) Wang, J.; Ferguson, A. L. Nonlinear Machine Learning in Simulations of Soft and Biological Materials. *Mol. Simul.* **2018**, *44*, 1090−1107.
(32) Jadrich, R. B.; Lindquist, B. A.; Truskett, T. M. Unsupervised Machine Learning for Detection of Phase Transitions in Off-Lattice Systems. I. Foundations. *J. Chem. Phys.* **2018**, *149*, 194109.
(33) Spellings, M.; Glotzer, S. C. Machine Learning for Crystal Identification and Discovery. *AIChE J.* **2018**, *64*, 2198−2206.
(34) Boattini, E.; Dijkstra, M.; Filion, L. Unsupervised Learning for Local Structure Detection in Colloidal Systems. *J. Chem. Phys.* **2019**, *151*, 154901.
(35) Adorf, C. S.; Antonaglia, J.; Dshemuchadse, J.; Glotzer, S. C. Inverse Design of Simple Pair Potentials for the Self-Assembly of Complex Structures. *J. Chem. Phys.* **2018**, *149*, 204102−204102.
(36) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2*, 559−572.
(37) Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* **1933**, *24*, 417−441.
(38) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* **2018**.
(39) Maaten, L. v. d.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(40) Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. In *Database Theory — ICDT 2001*; Goos, G., Hartmanis, J., van Leeuwen, J., Van den Bussche, J., Vianu, V., Eds.; Springer: Berlin, Heidelberg, 2001; Vol. *1973*, pp 420−434.
(41) McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2*, 205.
(42) Spellings, M. P. A Library to Generate Numerical Descriptions of Particle Systems: glotzerlab/pythia. 2019; https://github.com/glotzerlab/pythia, Accessed on: 2019-09-18.
(43) Kondor, R. A. Novel Set of Rotationally and Translationally Invariant Features for Images Based on the Non-Commutative Bispectrum. *arXiv:cs/0701127 [cs.CV]* **2007**.
(44) Adorf, C. S.; Dodd, P. M.; Ramasubramani, V.; Glotzer, S. C. Simple Data and Workflow Management with the signac Framework. *Comput. Mater. Sci.* **2018**, *146*, 220−229.
(45) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General Purpose Molecular Dynamics Simulations Fully Implemented on Graphics Processing Units. *J. Comput. Phys.* **2008**, *227*, 5342−5359.
(46) Ramasubramani, V.; Dice, B. D.; Harper, E. S.; Spellings, M. P.; Anderson, J. A.; Glotzer, S. C. freud: A Software Suite for High Throughput Analysis of Particle Simulation Data. *arXiv:1906.06317 [cond-mat, physics:physics]* **2019**.
(47) Stukowski, A. Visualization and Analysis of Atomistic Simulation Data with OVITO−the Open Visualization Tool. *Modell. Simul. Mater. Sci. Eng.* **2010**, *18*, 015012.
(48) Filion, L.; Ni, R.; Frenkel, D.; Dijkstra, M. Simulation of Nucleation in Almost Hard-Sphere Colloids: The Discrepancy Between Experiment and Simulation Persists. *J. Chem. Phys.* **2011**, *134*, 134901.
(49) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101*, 4177−4189.
(50) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-Operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in the Isothermal−Isobaric Ensemble. *J. Phys. A: Math. Gen.* **2006**, *39*, 5629−5651.
(51) Yu, T.-Q.; Alejandre, J.; López-Rendón, R.; Martyna, G. J.; Tuckerman, M. E. Measure-Preserving Integrators for Molecular Dynamics in the Isothermal−Isobaric Ensemble Derived from the Liouville Operator. *Chem. Phys.* **2010**, *370*, 294−305.
(52) Auer, S.; Frenkel, D. Prediction of Absolute Crystal-Nucleation Rate in Hard-Sphere Colloids. *Nature* **2001**, *409*, 1020.
(53) ten Wolde, P. R.; Ruiz-Montero, M. J.; Frenkel, D. Numerical Evidence for bcc Ordering at the Surface of a Critical fcc Nucleus. *Phys. Rev. Lett.* **1995**, *75*, 2714−2717.
(54) Rein ten Wolde, P.; Ruiz-Montero, M. J.; Frenkel, D. Numerical Calculation of the Rate of Crystal Nucleation in a Lennard-Jones System at Moderate Undercooling. *J. Chem. Phys.* **1996**, *104*, 9932−9947.
(55) Turci, F.; Schilling, T.; Yamani, M.; Oettel, M. Solid Phase Properties and Crystallization in Simple Model Systems. *Eur. Phys. J.: Spec. Top.* **2014**, *223*, 421−438.
(56) Kalikmanov, V. I.; Wölk, J.; Kraska, T. Argon Nucleation: Bringing Together Theory, Simulations, and Experiment. *J. Chem. Phys.* **2008**, *128*, 124506.
(57) Russo, J.; Tanaka, H. Crystal Nucleation as the Ordering of Multiple Order Parameters. *J. Chem. Phys.* **2016**, *145*, 211801.

DOI: 10.1021/acs.jpcb.9b09621
J. Phys. Chem. B 2020, 124, 69−78