Soft Matter

Deep learning for characterizing the self-assembly of three-dimensional colloidal systems

Journal:	*Soft Matter*
Manuscript ID	SM-ART-10-2020-001853.R1
Article Type:	Paper
Date Submitted by the Author:	20-Nov-2020
Complete List of Authors:	O'Leary, Jared; University of California Berkeley Mao, Runfang; Lehigh University, Chemical and Biomolecular Engineering Pretti, Evan; Lehigh University Paulson, Joel; The Ohio State University Mittal, Jeetain; Lehigh University, Mesbah, Ali; University of California Berkeley,

Soft Matter

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Deep learning for characterizing the self-assembly of three-dimensional colloidal systems

Jared O'Leary,[a] Runfang Mao,[b] Evan J. Pretti,[b] Joel A. Paulson,[c] Jeetain Mittal,*[b] and Ali Mesbah*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Creating a systematic framework to characterize the structural states of colloidal self-assembly systems is crucial for unraveling the fundamental understanding of these systems' stochastic and nonlinear behavior. The most accurate characterization methods create high-dimensional neighborhood graphs that may not provide useful information about structures unless these are well-defined reference crystalline structures. Dimensionality reduction methods are thus required to translate the neighborhood graphs into a low-dimensional space that can be easily interpreted and used to characterize non-reference structures. We investigate a framework for colloidal system state characterization that employs deep learning methods to reduce the dimensionality of neighborhood graphs. The framework next uses agglomerative hierarchical clustering techniques to partition the low-dimensional space and assign physically meaningful classifications to the resulting partitions. We first demonstrate the proposed colloidal self-assembly state characterization framework on a three-dimensional *in-silico* system of 500 multi-flavored colloids that self-assemble under isothermal conditions. We next investigate the generalizability of the characterization framework by applying the framework to several independent self-assembly trajectories, including a three-dimensional *in-silico* system of 2052 colloidal particles that undergo evaporation-induced self-assembly.

1 Introduction

Colloidal self-assembly (SA) is the process by which particles in solution spontaneously organize into an ordered structure[1]. The spontaneous self-organization central to SA enables "bottom-up" materials synthesis, which would allow for manufacturing advanced, highly ordered crystalline structures with up to sub-nanometer precision in an inherently parallelizable and cost-effective manner. Thus, colloidal SA can create new avenues for highly scalable, economical manufacturing of novel metamaterials with unique optical, electrical, or mechanical properties[1–6]. Creating a systematic framework to characterize the states of colloidal SA systems is crucial for unraveling our fundamental understanding of the stochastic and nonlinear behavior of these systems.

The most common method to characterize the colloidal SA system state is Steinhardt bond order parameters[7,8], which use spherical harmonic basis functions to evaluate the symmetry of a particle's neighbors. However, Steinhardt bond order parameters are extremely sensitive to thermal fluctuations that smear local bond order into broad overlapping distributions and interfere with the ability to resolve the character of small domains. Other commonly used methods include Common Neighbor Analysis (CNA)[9,10], Polyhedral Template Matching (PTM)[11], and Bond Angle Analysis (BAA)[12]. CNA and PTM evaluate the topology of each particle's nearest neighbors to generate neighborhood graphs that describe a given particle's local structure, while BAA evaluates the symmetry of each particle's nearest neighbors to create neighborhood graphs. These methods, however, fail to provide quantitative information about particles whose topologies or symmetries do not correspond to well-defined reference crystalline structures[13]. The high-dimensional, discrete nature of these neighborhood graphs prevents intuitive understanding of how these graphs are related and dimensionality reduction methods are thus required to translate the neighborhood graphs into a (continuous) low-dimensional space that can be easily interpreted and used to characterize non-reference structures.

The current state-of-the-art method for colloidal system state characterization accomplishes dimensionality reduction by implementing diffusion maps[13–23]. However, diffusion maps are computationally expensive to implement. In fact, the most recent implementations of diffusion maps require the choice of (arbi-

[a] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720, USA. Tel: +1-510-642-7998; E-mail: mesbah@berkeley.edu
[b] Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, PA 18015, USA. Tel: +1-610-758-4791; E-mail: jem309@lehigh.edu
[c] Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, OH 43210, USA.

trarily chosen) "landmark points" to reduce the size of the high-dimensional space before dimensionality reduction takes place. Diffusion maps further do not provide an explicit functional mapping between the high and low-dimensional spaces, thereby, limiting physical interpretation of the low-dimensional space.

Several other researchers have implemented different varieties of "machine learning" for dimensionality reduction and/or classification of colloidal structures. For example, refs. 24, 25 implement principal component analysis (PCA) to detect phase transitions in off-lattice systems. PCA, however, is not designed to reduce the dimensionality of variables with highly nonlinear relationships among one another[26]. On the other hand, ref. 27 uses a combination of Gaussian Mixture Models and shallow artificial neural networks to identify the overall crystal structures of bulk self-assembled systems. However, this work does not explicitly employ machine learning techniques for dimensionality reduction and instead investigates the learning techniques' ability to create and interpret large neighborhood graphs.

The overarching goal of this work is to develop a characterization framework for investigating the stochastic and nonlinear dynamics of entire SA trajectories (as opposed to merely characterizing individual lattices). We thus propose an alternative approach to dimensionality reduction based on a deep neural network called an autoencoder[28,29]. Autoencoders are easy to implement with available tools and cheap to evaluate. The computational efficiency allows autoencoders to simultaneously reduce the dimensionality of the thousands of neighborhood graphs that can appear during SA, an operation which would likely be intractable for diffusion maps. The nonlinear activation functions within deep neural networks also allow the autoencoder to explicitly account for the nonlinear relationships among the diverse neighborhood graphs that may appear during SA. Autoencoders further provide an explicit mapping between the low- and high-dimensional spaces, elucidating which of the high-dimensional inputs are the most "important" for the system under analysis.

We note that Boattini *et al.* previously applied autoencoders based on shallow neural networks for dimensionality reduction and subsequent classification of colloidal systems[30]. However, their approach creates neighborhood graphs using a vector of only 8 Steinhardt order parameters. We instead create neighborhood graphs via a well-established methodology based on Delaunay triangulation and graphlet decomposition[15,31–34]. This methodology is much less sensitive to thermal fluctuations and has also been shown to quantify detailed colloidal lattice configurations by Reinhart *et al.*. Our deep neural network-based autoencoders further employ dropout regularization to prevent model overfitting and achieve continuity in the low-dimensional space. We finally note that Boattini *et al.* primarily focused on classifying individual lattices whereas the focus of this work is to study entire SA trajectories.

We propose a three-step framework for colloidal system state characterization (see Fig. 1). The first step establishes neighborhood graphs with a precise methodology that has been shown to be robust to thermal fluctuations and capable of describing complex topologies[15,31–34]. The second step uses deep learning techniques to reduce the dimensionality of the neighborhood graphs. The third step employs agglomerative hierarchical clustering to partition the low-dimensional space and assign physically meaningful classifications to the resulting partitions.

We demonstrate the proposed three-step colloidal system state classification framework on a three-dimensional *in-silico* system of 500 DNA-functionalized multiflavored colloidal particles (i.e., silica colloids that are coated with blends of complementary single strands of DNA)[35–38] that self-assemble into a variety of FCC, HCP, and BCC-like lattices. We also examine the generalizability of the characterization framework by applying the framework to several independent colloidal SA trajectories (i.e., trajectories that were not used to train the autoencoder), including a system consisting of 2052 *in-silico* colloidal particles that undergo three-dimensional evaporation-induced SA[23]. We have placed the entire dimensionality reduction framework in an easily accessible GitHub format that is explicitly designed for people to use and modify[39]. More in-depth descriptions of the "multi-flavored" and "evaporation-induced" /textitin-silico self-assembly systems can be found in Section 2.4.

2 Methods

2.1 Neighborhood Graph Construction

The first step in classifying the structure of a given colloidal particle is to generate a "neighbor list" that consists of a list of particles that are considered topologically or symmetrically adjacent to the particle of interest. This neighbor list is then used to construct a neighborhood graph that quantifies the local structure of the given particle. Two of the most common local structure classification methods, Common Neighbor Analysis (CNA)[9,10] and Steinhardt order parameters[7,8], heavily rely on the concept of particles being "bonded" to establish neighbor lists. These methods thus require a strict definition of a bond, where two particles are considered bonded if they fall within a predefined cutoff radius. However, such a cut-off radius is, by necessity, somewhat arbitrary. In addition, thermal vibrations, the coexistence of various phases, and fluctuations in the local density will introduce noise into the analysis and can even make finding a suitable cut-off radius impossible. This problem is partially mitigated by adaptive CNA[40], where the cutoff radius is determined by the average distance to a heuristically chosen number of particles. Despite the use of averaging, radii for low-density and vapor phase particles can be extremely large and inhibit classification accuracy. The approach further assumes that a given particle's neighborhood is isotropic, which is often not the case for open lattices[41]. We employ the methodology described in refs. 15, 31 to obtain the neighbor list of topologically adjacent particles and subsequent neighborhood graph. Because this method avoids the concept of bonds between particles and instead uses a geometry-based, fixed number of particles to establish the neighborhood, it is less sensitive to thermal fluctuations, density gradients, and anisotropy mentioned above.

2.2 Dimensionality Reduction

The high dimensionality of the neighborhood graphs and non-uniformity in the distances among them indicate that dimension-

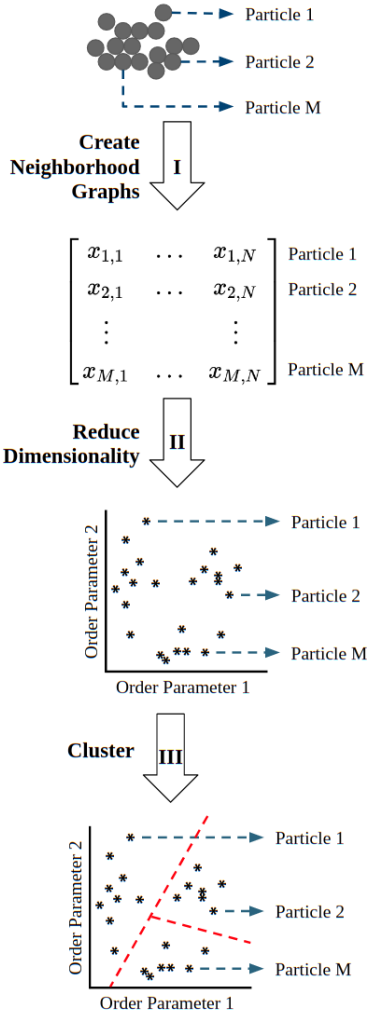


Fig. 1 Colloidal self-assembly state characterization framework summary. The particle positions are recorded and translated into neighborhood graphs. The dimensionality of the neighborhood graphs is next reduced using deep learning techniques. Agglomerative hierarchical clustering is finally used to partition the low-dimensional space and assign discrete classifications to each particle.

ality reduction must be performed to produce a low-dimensional manifold from which relationships among neighborhood graphs can be more easily inferred. We reduce the dimensionality of our neighborhood graphs using a deep neural network called an autoencoder. An autoencoder is comprised of an *encoder* that constructs a low-dimensional representation of its input (i.e., the neighborhood graph in this case) and a *decoder* that reconstructs the input from the low-dimensional representation[28,29]. The encoding process is often lossy, meaning that part of the information is lost during the encoding process and cannot be recovered during decoding. Dimensionality reduction is thus accomplished by finding the encoder/decoder pair that keeps the maximum information when encoding and correspondingly has the minimum reconstruction error (e.g., mean-squared error or MSE) when decoding. Note that only the encoder is used to reduce dimensionality, while the decoder is used to find the encoder model that creates the best low-dimensional representation of the input data. As discussed in detail in the supplementary information (SI) in Section S4, the "optimal" encoder/decoder scheme is found through an iterative training process. Finally note that training the autoencoder can be thought of as a "self-supervised" learning process, as training determines a (nonlinear) function that maps the neural network's inputs (i.e., the neighborhood graphs) to themselves (i.e., neighborhood graphs that are reconstructed from their low-dimensional representation).

To train the autoencoder, we first collected particle position data for 11 different isothermal trajectories of an *in-silico* three-dimensional system of 500 multi-flavored colloidal particles[35–38]. The inter-particle interactions in each trajectory were varied such that a variety of vapor, low-density, defective, and FCC, HCP, and BCC-like lattices appear during assembly. Neighborhood graphs for each particle in each simulation frame were recorded according to the Delaunay triangulation and graphlet decomposition methodology of Section 2.1 (527,500 total neighborhood graphs). We then used only the unique neighborhood graphs (4153 unique neighborhood graphs) to train the autoencoder.

One of the main advantages of autoencoders over diffusion maps is that autoencoders provide an exact analytical mapping from the high to low-dimensional spaces. This mapping allows us to assess the *relative importance* of each entry in the neighborhood graph via input perturbation and stepwise methods[30,42–45]. Relative importance is measured by the variation in MSE caused by perturbing samples in the training data set (see SI Section S5).

2.3 Classification and Interpretation of the Low-Dimensional Space

The key challenge of colloidal SA state classification is then partitioning this low-dimensional space into discrete regions to make final decisions regarding structural identity. We use agglomerative hierarchical clustering with a Ward's minimum variance linkage metric to partition the low-dimensional space[46,47]. Agglomerative hierarchical clustering via Ward's linkage operates by initially placing each data point in its own cluster. In each iteration, two clusters are combined into one by finding the pair of clusters that leads to the minimum increase in total intra-cluster variance after merging. This variance increase is a weighted squared distance between cluster centers and these iterations continue until all data points are grouped into one cluster. The method creates clusters of various shapes, sizes, intra-cluster variances, and membership populations. By iteratively minimizing the increase in total intra-cluster variance, the method can naturally discover both clusters that are adjacent in the low-dimensional manifold with small intra-cluster variances and clusters with high intra-cluster variances that span larger, less-populated sections of the coordinate space. Moreover, the clustering strategy makes no assumptions regarding the distribution of the low-dimensional space as it only assesses similarities between pairs of objects.

Most importantly, agglomerative hierarchical clustering establishes a "cluster tree" that reveals the underlying hierarchical structure of the data. The branches within this tree allow us to extract informed descriptions of the discrete regions of the low-dimensional space and choose a number of clusters that is appro-

priate for our specific application. We next visualize the results of the classification by assigning a unique color to each identified class and using OVITO visualization software[48] to create a color-coordinated image of each simulation frame. A qualitative analysis of these simulation frames provides a general idea of what each class physically represents.

2.4 Self-Assembly System Descriptions

In this work, two separate three-dimensional *in-silico* SA systems are used to demonstrate the proposed characterization framework. The first consists of a system of 500 multi-flavored colloids that self-assemble under isothermal conditions. The second consists of a system 2052 silica colloids that undergo evaporation-induced self-assembly[23]. These systems are described in more detail below.

2.4.1 Self-Assembly of Multi-Flavored Colloids

One way to promote the SA of colloidal particles is through functionalization of their surfaces with DNA. DNA-functionalized particles (DFPs) interact with each other through complementary Watson–Crick base-pairing interactions and have been used to assemble many superlattice structures[35,49]. As a means of achieving selective binding among DFPs, it has recently been suggested that particles can be functionalized with a blend of two types of DNA strands with complementary concentrations on each particle. These "multi-flavored" particles can exhibit a tunable attraction between the like particles while maintaining interactions between unlike pairs. This approach has been shown to induce the crystallization of equally sized particles into BCC, HCP, and FCC structures[36,50,51].

In this work, the SA trajectories are obtained from binary colloidal mixtures that represent multi-flavored DNA functionalized particles (DFPs) for which the attractive interactions betwween A-type and B-type particles (i.e., E_{AA}, E_{BB} and E_{AB}) can be adjusted independently. Figure S2.1 (in the SI) shows the schematic representation of the multi-flavored DFPs and a pairwise interaction model used in molecular dynamics (MD) simulations for obtaining these trajectories. The functional form of pair interaction utilized in these simulations is of a Fermi-Jagla type, which has previously been successfully used to study the self-assembly process of DFPs both in two[35,38] and three dimensions[36]. This SA approach based on tuning E_{AA}, E_{BB} and E_{AB} is used in this work to induce the crystallization of equally sized DFPs into various BCC, HCP, and FCC-like structures. The resulting trajectories are used to train the characterization framework.

All SA trajectories are MD simulations that are performed using LAMMPs package under the NVT ensemble[52]. A Langevin thermostat is applied with a time constant $\tau = 2\sigma m^{1/2}\varepsilon^{-1/2}$ with a simulation time step $\Delta t/\tau = 10^{-3}$. Periodic boundary conditions are applied to the cubic simulation box with a number density $0.02\sigma^{-3}$. 500 total particles are initially placed in a three-dimensional box and equilibrated for 1×10^{6} time steps at temperature $T = 1\varepsilon k_B^{-1}$ to ensure all particles are in the gas phase. The MD simulations are then performed isothermally at pre-determined temperatures $Tm \in [0.125\varepsilon k_B^{-1}, 0.165\varepsilon k_B^{-1}]$ that are deemed suitable for crystallization based on values of E_{AA}, E_{BB} and E_{AB}. To allow a reasonable amount of crystals to form, 3×10^{8} simulation steps are performed. The determination of suitable temperatures based on E_{AA}/E_{AB} and E_{BB}/E_{AB} and other simulation details are suggested in our previous papers[35–37].

2.4.2 Evaporation-Induced Self-Assembly

One common high-throughput method for fabricating colloidal crystals involves dispersing colloids in a volatile solvent followed by evaporation of the solvent to deposit a crystalline solid onto a substrate (i.e., "evaporation-induced self-assembly")[23]. The authors in ref. 23 performed massive-scale non-equilibrium MD simulations with an explicit-solvent model to study the evaporation-induced assembly of colloidal crystals from solution onto a horizontal substrate. Six snapshots from an MD simulation consisting of 2052 (initially disperse) silica colloids were used in Section 3.3 to examine the generalizabilty of the characterization framework. Note that this data was directly provided to us by the authors of ref. 23.

3 Results and Discussion

3.1 Autoencoder Architecture

The key autoencoder architectural choices are the batch size, activation function, regularization strategy, and network size. Justifications for each of the former three choices are described in the SI, while the latter choice is informed by implementation of the elbow method[30,53,54]. The elbow method (which is widely used throughout the self-supervised and unsupervised learning communities[30,53,54]) plots some measure of neural network performance (e.g., MSE) against some neural network hyper-parameter (e.g., the number of nodes in a given neural network layer). The method involves visually detecting a "slope change" where the performance of the neural network begins to improve more slowly with the change in the hyper-parameter. The beginning of this slope change is called the "elbow".

Here, several autoencoder models with different network sizes are trained with the sample data (i.e., the 4153 unique neighborhood graphs found from the 11 isothermal *in-silico* trajectories described in Section 2.2). The autoencoder MSE is plotted against the number of nodes in the bottleneck layer (i.e., the size the low-dimensional space found by the encoder) for candidate models that only differ by the number of hidden layers and number of nodes per hidden layer (see Fig. S4.2 in the SI). Elbows in this plot occur between 2 and 4 bottleneck nodes, indicating that a bottleneck layer size of 3 nodes is likely sufficient to capture the essential information from the neighborhood graphs. Moreover, the corresponding size of the 3×1 low-dimensional representation is convenient from a visualization standpoint. The autoencoders with 2 hidden layers and 500 and 1000 nodes per hidden layer display nearly identical performance, with the latter model showing a marginally lower MSE. Models with larger network sizes do not display any performance improvements. As a result, the chosen autoencoder model contains 2 hidden layers, 1000 nodes per hidden layer, and 3 bottleneck nodes (which creates a low-dimensional space of dimension 3×1).

We implemented input perturbation (with 10% Gaussian white noise) and stepwise relative importance analyses on the chosen

autoencoder[30,42–45] (see Fig. 2). The analysis shows that nearly all neighborhood graph entries (with the exception of graph entries 22 and 23) are equally important and indicates that no single graph entry, or even small group of graph entries can be used to quantify the colloidal SA system state. This validates the need for implementing dimensionality reduction. The 22nd and 23rd neighborhood graph entries do account for nearly 25% of the MSE variation, however. This spike in variation is due to the fact that the neighborhood graph construction methodology can yield extremely large outlier values at solid-vapor interfaces (see SI Section S3). Translating the neighborhood graphs into a low-dimensional space significantly reduces the effects of these outliers and does not inhibit local structure characterization and classification (see Section 3.2).

Fig 2 further shows spikes in relative importance at neighborhood graph entries 0-1 and 30-36. Entries 0-1 refer to two and three-component linear orbits that are common in newly formed, small crystallites. These spikes indicate that many of the unique signatures used to train the autoencoder correspond to very weakly crystalline particles on the precipice of crystallization. Meanwhile, entries 30-36 refer to square and pentagonal-like shapes that are common in (defective) FCC, HCP, and BCC structures. Again, these spikes demonstrate the frequency of FCC, HCP, and BCC-like structures in the training data. The above points show that the relative importance analysis can not only point out unexpected behavior in the characterization framework (e.g., the erratic neighborhood graph values for particles at solid/vapor interfaces) but also can demonstrate to which types of data the autoencoder model is more sensitive (e.g., the very weakly crystalline and FCC/BCC/HCP-like particles mentioned a few lines above).

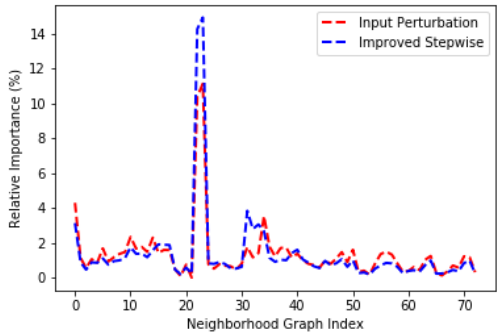


Fig. 2 Relative importance analysis. Input perturbation and improved stepwise methods are used to assess the relative importance of the 73 entries within the neighborhood graph. Although neighborhood graph entries 22 and 23 account for the largest percentage of MSE variation, these results demonstrate that no single graph entry, or even relatively minor groups of graph entries can be used to quantify the system state. Moreover, the large MSE variation caused by nodes 22 and 23 is a function of certain outliers found at solid-vapor interfaces.

3.2 Partitioning the low-dimensional space

We used the chosen encoder model to translate the entire training data set (4153 unique neighborhood graphs) into a three-dimensional low-dimensional space. We then implemented agglomerative hierarchical clustering (with Ward's linkage) on the low-dimensional data. Although the strategy produces a cluster tree that shows the hierarchical structure of all 1 to 4153 possible cluster distributions, the process of choosing the "best" number of clusters is somewhat subjective[46,47]. In fact, a key advantage of agglomerative hierarchical clustering is that the strategy allows us to choose the number of clusters for classification for specific application-based needs.

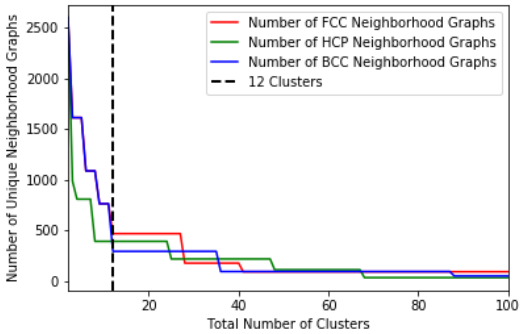


Fig. 3 Analysis to determine number of clusters. Agglomerative hierarchical clustering (using Ward's linkage) is used to cluster the low-dimensional representations of the 4153 unique neighborhood graphs taken from the 11 isothermal colloidal self-assembly trajectories that were used to train the autoencoder (see Section 2.2). The number of unique neighborhood graphs corresponding to FCC, BCC, and HCP structures is plotted against the number of clusters in each branch of the resulting cluster tree. At 12 total clusters, the low-dimensional representations of FCC, HCP, and BCC neighborhood graphs are separated into different clusters.

This work focuses on the SA of FCC, HCP, and BCC-like structures from a system of 500 multi-flavored colloidal particles[35–38] (see section 2.2). The topologies of theoretically perfect FCC, HCP, and BCC lattices are known. We used this information to calculate neighborhood graphs and the corresponding low-dimensional points of these three theoretically perfect lattices. However, "perfect" or at least "not meaningfully defective" FCC, HCP, or BCC lattices may have neighborhood graphs that correspond to a number of different low-dimensional points. As a result, any cluster that contains one of these three theoretically perfect lattice points can be analogously labeled.

Fig. 3 shows the number of low-dimensional points (that represent neighborhood graphs) corresponding to FCC, BCC, and HCP structures plotted against the number of total clusters in each branch of the cluster tree. At the branch corresponding to 12 total clusters, the FCC, HCP, and BCC perfect lattice points are first separated into different clusters. The number of points assigned to each of the three lattice types decreases with the total number of clusters as the points that are further from the theoretically perfect lattices are placed into other clusters. The choice in the number of clusters is thus a balance between the desired classification precision (i.e., the strictness of the definition of an FCC, HCP, or BCC lattice) and the analytical burden of interpreting potentially hundreds of clusters. In this work, we chose the minimum number of clusters required to separate the theoreti-

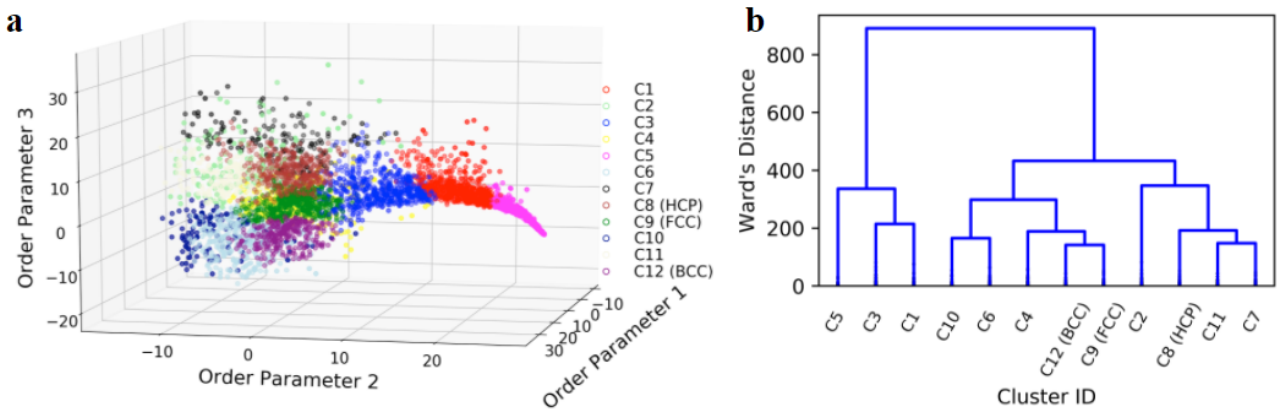


Fig. 4 Agglomerative hierarchical clustering summary. Agglomerative hierarchical clustering (using Ward's linkage) was used to cluster the low-dimensional representations of 4153 unique neighborhood graphs (from the 11 isothermal colloidal self-assembly trajectories that were used to train the autoencoder described in Section 2.2) into 12 clusters. These clusters are labeled C1-C12. (a) The low-dimensional representation of each unique neighborhood graph is plotted and colored according to its labeled cluster. Points corresponding to bulk FCC, HCP, and BCC lattices exist within clusters C9, C8, and C12, respectively. (b) The Ward's distance between each cluster is plotted against each cluster's placement within the cluster tree.

cally perfect FCC, BCC, and HCP lattices (i.e., 12 total clusters). Our subsequent visual analyses of SA simulation trajectories show this choice to be reasonable (see Section 3.3).

Fig. 4a shows the the colored low-dimensional representations of all 4153 unique neighborhood graphs. We assigned a distinct color to each of the 12 clusters. The FCC, HCP, and BCC clusters correspond to clusters C9 (green), C8 (brown), and C12 (purple), respectively, while vapor particles correspond to cluster C1 (red). Note that vapor particles tend to display very small neighborhood graph entries and thus contain predictable neighborhood graphs/low-dimensional coordinates. Particles at solid-vapor interfaces exist in clusters C3 and C5. The neighborhood graphs of these particles tend to contain extremely high neighborhood graph entries (particularly at entries 22 and 23). Although it is not immediately clear to which types of structures the remaining clusters correspond, their close proximity to one another and distance from the vapor states suggest that they are likely surface or defective crystalline structures.

The structure of the cluster tree (Fig. 4b) provides important insights regarding the physical characteristics of the remaining clusters. First, clusters C1, C3, and C5 fall under the same branch while the remaining clusters (which include the FCC, HCP, and BCC clusters) fall under a second branch. This suggests that the first level of the cluster tree likely separates "crystalline" and "vapor/near-vapor" particles. Clusters C4, C6, C9 (FCC), C10, and C12 (BCC) all fall under the second level middle branch, suggesting that C4, C6, and C10 correspond to some types of surface or defective FCC/BCC structures. The fact that C4 belongs to the same parent branch as C9 and C12 indicates that C4 is likely more topologically similar to C9 and C12 than it is to C6 and C10. The right second level branch contains clusters C2, C7, C8 (HCP), and C11. The distances separating the C7, C8 (HCP), and C11 leaves are very small, also indicating that C7 and C11 could correspond to slightly defective HCP structures while C2 could correspond to either highly defective or surface HCP particles.

The low-dimensional space appears to have achieved continuity (i.e., similar structures have similar low-dimensional coordinates). For example, clusters C1, C3, and C5 all correspond to either vapor particles or vapor particles at solid-vapor interfaces. Although these clusters contain both extremely small and large neighborhood graph entries, the clusters have low intra-cluster variances and are adjacent in the low-dimensional space. Our dimensionality reduction thus effectively handles the massive outliers the neighborhood graph construction methodology occasionally produces at solid-vapor interfaces. The FCC, HCP, and BCC clusters are close to one another, yet far apart from the vapor clusters. Meanwhile the remaining clusters (which likely correspond to defective and surface particles) are not only close to one another but also take up a large percentage of the low-dimensional space to reflect their large topological range.

We note that some of the boundaries among clusters appear exceedingly complex, suggesting that some of the data points are misclassified. The boundary complexity is a function of both the (unavoidable) noise in the neighborhood graph construction and the choice of a small number of clusters. We could potentially address this problem by increasing the number of clusters. However, the objective of the characterization framework is to elucidate understanding of SA processes as a whole and not to perfectly characterize each individual particle (otherwise one would avoid dimensionality reduction altogether). Overall, the colloidal SA state characterization framework appears to effectively reduce the dimensionality of neighborhood graphs and sensibly partition the low-dimensional space.

3.3 Visualization and classification

We first demonstrate the characterization framework by using OVITO to visualize 4 different lattices found from 4 of the 11 different isothermal colloidal SA trajectories used to train the autoencoder (see Fig. 5). The particles in each lattice are colored according to their classifications in Fig. 4. Each of the 4 lattices is shown in full (labeled "Full Lattice") and with its top layer re-

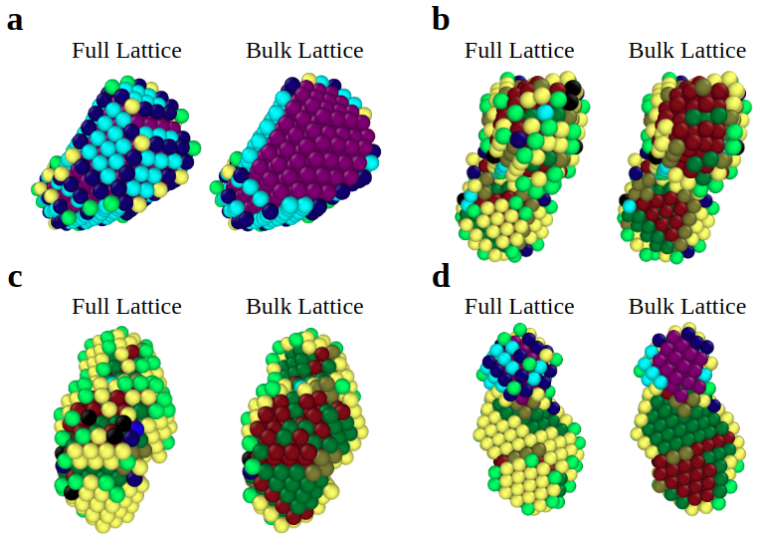


Fig. 5 Four classified colloidal self-assembly lattices. The figure shows 4 lattices from the final time steps of 4 of the 11 isothermal colloidal self-assembly trajectories used to train the autoencoder (see Section 2.2). Each particle in each lattice is colored according to its classification in Fig. 4. The term "full lattice" indicates that every particle in the snapshot is shown while the term "bulk lattice" indicates that the top layer of particles has been removed. The structure in (a) is primarily BCC, the structures in (b) and (c) are mixed FCC and HCP, and the structure in (d) contains FCC, HCP, and BCC particles.

moved (labeled "Bulk Lattice").

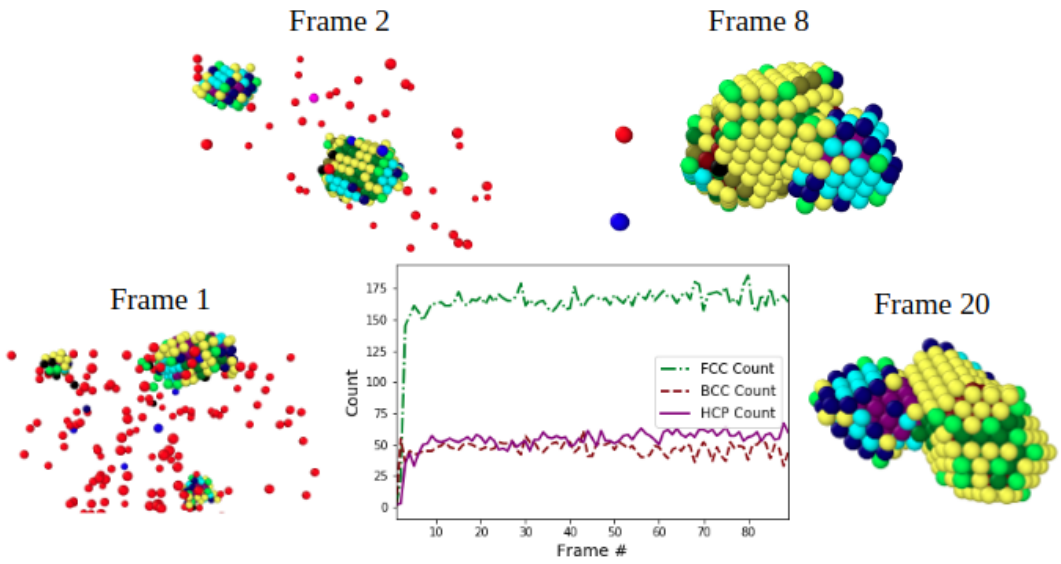
We used the OVITO visualizations to assign brief, physically meaningful descriptions to each cluster (see Table 1 for a summary of these descriptions). The bulk particles in 5a (purple) almost all belong to cluster C12 and correspond to BCC structures. The surface particles primarily belong to clusters C6 (light blue) and C10 (dark blue), with scattered particles belonging to clusters C4 (yellow) and and C2 (light green). The C6 (light blue) particles clearly correspond to surface BCC (100)-(111) particles. The C10 (dark blue) particles only exist at the interface between two surface planes and likely correspond to BCC surface stacking faults.

Meanwhile, the bulk particles in Fig. 5b-7c are primarily from clusters C9 (Green, FCC) and C8 (Brown, HCP). Another bulk particle classification is C11 (beige), which primarily appears on FCC/HCP interfaces. The cluster's placement in the same parent branch as cluster C8 (Brown, HCP) indicates that C11 is likely a defective HCP structure. Structures 5b-c show many surface particles belonging to clusters C4 (yellow) and C2 (light green). Note that Fig. 5b appears to show C4 (yellow) particles in the bulk, however, these are actually surface particles on an adjacent plane. Based on their placement in the cluster tree and proximity to FCC particles in the Fig. 5b-c, cluster C4 corresponds to FCC (100)-(111) surface particles. The C2 (light green) and C7 (black) particles are less commonly observed throughout the SA trajectory data but often appear as defective surface particles on lattices containing HCP and FCC particles. The C2 (light green) particles even occasionally appear as stacking faults (see Fig. 6)a, while the C7 (black) particles tend to appear as weakly-bound particles. Despite C2's placement within the cluster branch corresponding to HCP particles, C2 particles often appear above FCC bulk particles. This suggests that C2 refers to defective FCC surface particles that show some HCP-like characteristics. Each of the above classifications remain consistent in Fig. 5d, which shows a polymorphic FCC, HCP, and BCC lattice.

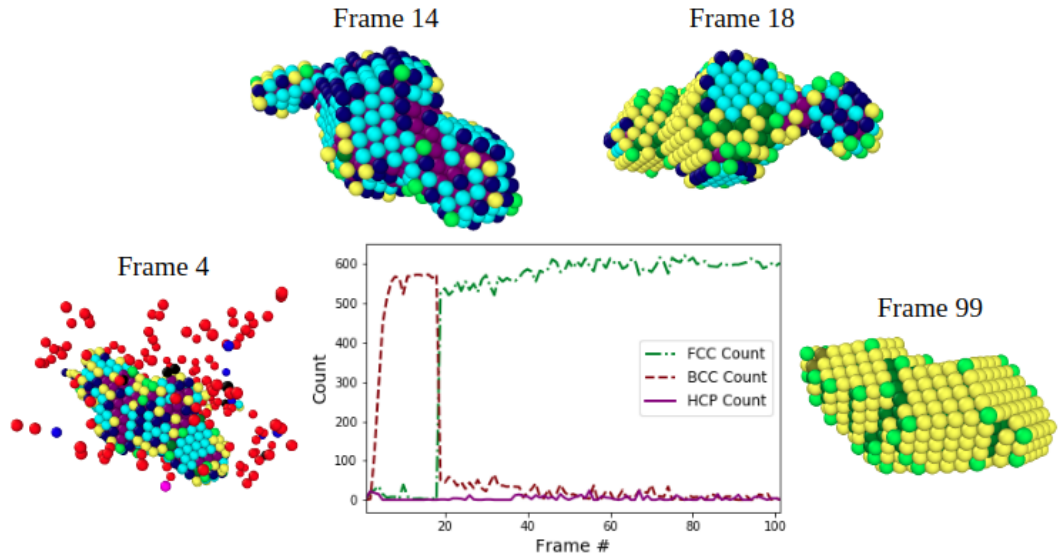
Table 1 Cluster structural classifications. Each cluster identification (C1-C12) is matched with a brief physical description.

Cluster Label	Structure Description
C1	Vapor
C2	Defective FCC surface particle
C3	Vapor at Solid-Vapor Interface
C4	Surface FCC
C5	Vapor at Solid-Vapor Interface
C6	Surface BCC
C7	Weakly bound HCP-like particle
C8	HCP
C9	FCC
C10	BCC Stacking Fault
C11	Defective Bulk HCP
C12	BCC

It is important to note that rigorous, direct comparisons to other dimensionality reduction-based characterization frameworks are not necessarily appropriate in this work. For example, the most recent implementations of diffusion maps require the choice of "landmark points" to reduce the size of the high-dimensional space before dimensionality reduction takes place. As a result, diffusion maps cannot reduce the same high-dimensional space that the autoencoders can and thus cannot be applied (in the same way) to the self-assembly trajectories discussed above. This is because analysis of these trajectories would require computing distance matrices between thousands of neighborhood graphs – and diffusion maps can become intractable for such large analyses. We further could have directly applied the approach of Boattini *et al.* to the 11 SA trajectories used to train the characterization framework. However, the elbow plot analysis (Fig. S4.2 in the SI) demonstrates that the single hidden layer autoencoder architecture employed by Boattini *et al.* does not



(a) Example colloidal self-assembly trajectory with polymoprhic lattice.



(b) Example colloidal self-assembly trajectory with phase transition.

Fig. 6 Example colloidal self-assembly trajectories. Each figure shows the time evolution of the number of particles classified as FCC (cluster C9, green), HCP (cluster C8, brown), and BCC (cluster C12, purple) for a separate *in-silico* colloidal self-assembly trajectory. Note that Frame # refers to the (chronologically ordered) recorded simulation frame. The time evolution plots are accompanied by snapshots of certain chosen simulation frames within these trajectories. In each case, the dimensionality of the neighborhood graphs is reduced with the encoder trained using 11 isothermal trajectories of a system of 500 multi-flavored colloidal particles (see Section 2.2). Each particle in each snapshot is classified according to the proximity of its low-dimensional representation to points in Fig. 4a. (a) The figure shows the time evolution of an isothermal trajectory of the self-assembly of 500 multi-flavored colloids that creates the lattice in Fig. 5d. The trajectory shows that a polymorphic lattice containing FCC, HCP, and BCC particles forms from a primarily BCC structure merging with a structure that contains FCC and HCP particles (b) The figure shows the time evolution of an isothermal trajectory of the self-assembly of 1000 multi-flavored colloids. The trajectory shows that the system initially self-assembles into a BCC structure before undergoing a phase transition into an FCC structure.

encode as much information in the low-dimensional space as the proposed multiple hidden layer approach. Moreover, their use of Steinhardt bond order parameters to create neighborhood graphs (which are much more prone to thermal fluctuations and density gradients than the proposed Delaunay triangulation-based method) indicates that the method of Boattini *et al.* would lead to less general classification. In fact, Boattini *et al.* classify each particle within a lattice as either FCC, HCP, or "fluid". Meanwhile, our approach classifies each particle in one of 12 different categories, which include BCC, FCC, HCP, fluid (which we label as "vapor"), and several surface and defective states.

We next used the characterization framework to analyze the time evolution of a colloidal SA trajectory, as opposed to singular SA system states. Fig. 6a shows the time evolution of the colloidal SA trajectory that leads to the lattice in Fig. 5d. Here, the number of total particles classified as FCC (cluster C9, green), HCP

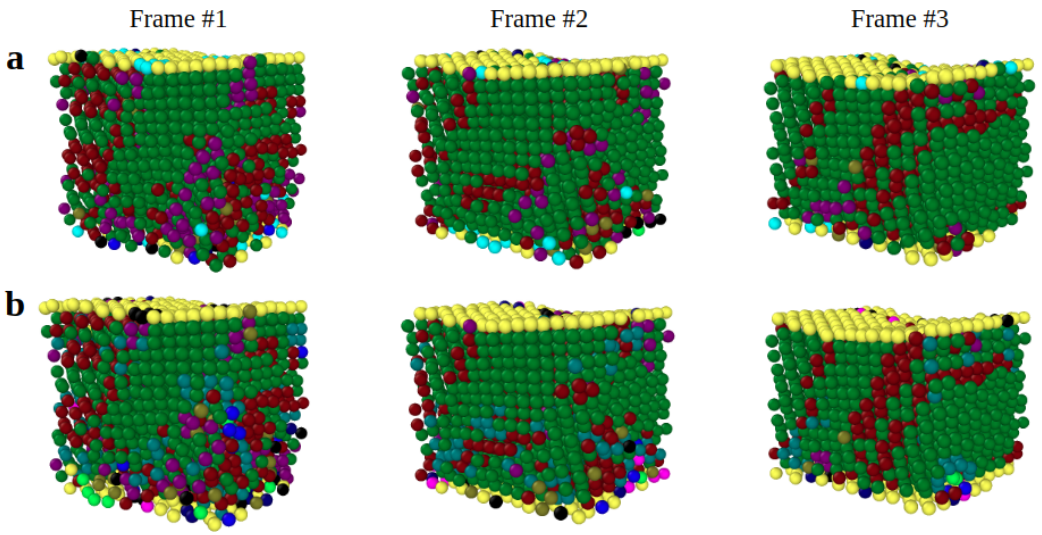


Fig. 7 Evaporation-induced colloidal self-assembly. The figure shows 3 snapshots of the *in-silico* evaporation-induced self-assembly of 2052 colloidal particles that are classified using two different schemes. Note that the data used to create these snapshots was borrowed from ref. 23 and that 6 total snapshots were provided. (a) The dimensionality of the neighborhoood graphs is reduced with an encoder trained using 11 isothermal trajectories of an *in-silico* system of 500 multi-flavored colloidal particles (see Section 2.2) Each particle in each snapshot is classified according to the proximity of its low-dimensional representation to points in Fig. 4a (b) The entire characterization framework is performed on the six provided snapshots of the evaporation-induced colloidal self-assembly data. Each unique neighborhood graph is used to train a second autoencoder. The newly-formed encoder is used to reduce the dimensionality of the neighborhood graphs and agglomerative hierarchical clustering (via Ward's linkage) is used to partition the low-dimensional space. In both (a) and (b), FCC particles are green, HCP particles are brown, BCC particles are purple, and surface FCC particles are yellow. The teal particles in (b) correspond to defective FCC structures that were not found by the classification scheme in (a). Overall, the two characterization procedures yielded nearly identical results.

(cluster C8, brown), and BCC (cluster C12, purple) is plotted against the simulation frame index. Snapshots of four key simulation frames whose particles are classified according to their positions in Fig. 4 are also provided. Frame #1 shows several small nuclei beginning to form. Clearly, many particles are still in the vapor phase, as clusters C1 (red) and C3 (blue) are highly prevalent. The bottom right crystallite is forming a BCC structure as evidenced by the C6 (light blue), C10 (dark blue), and C12 (purple) colored particles. Meanwhile the top right cluster primarily contains FCC/HCP particles due to its plethora of C4 (yellow), C8 (HCP), and C9 (FCC) particles. However, this crystallite also contains some BCC-like particles such as C12 (purple) and C6 (light blue). By Frame #2, the two remaining clusters are almost entirely BCC (top left) and almost entirely FCC/HCP (bottom right). The crystals' continued nucleation uncovers a few interesting trends.

First, we see that the polymorphic lattice is formed by a primarily BCC structure merging with a primarily FCC/HCP structure. This indicates that the assembly conditions are likely favorable to both BCC and HCP/FCC structures. Comparing Frames #2 and #8 shows the merging of the BCC and HCP/FCC structures as part of the growth process. Frame #2 also shows that the FCC/HCP structure is initially covered with surface particles from cluster C6 (light blue), which represent BCC surface particles. However, these light blue particles nearly only exist as FCC particles by the end of the trajectory. This could suggest that FCC particles take on a structure similar to that of surface BCC before finding their final state (e.g. Frame #20 and Fig. 5d). In fact, the idea that the interfaces of FCC crystallites retain BCC-like ordering during nucleation is frequently explored[55–57]. We used the characterization framework to carry out similar analyses for the remaining SA trajectories used to train the autoencoder, but did not include them for brevity.

We applied the characterization framework to an independent test data set that consists of 1000 *in-silico* multi-flavored colloids undergoing SA in isothermal conditions. We first calculated the neighborhood graphs of each particle and used the chosen encoder from Section 3.1 to reduce the dimensionality of the neighborhood graphs. Note that we did not retrain the autoencoder, and instead used the encoder with the same weights and biases as determined in Section 3.1. We next identified the points in Fig. 4a that were closest to those corresponding to the independent data set and classified the particles accordingly. For example, if the low-dimensional representation of a neighborhood graph from the independent data set is $[15.17, 3.50, 18.23]^\top$ and the closest point in Fig. 4a is classified as C2, then the particle from the independent data set adopts this class. Fig. 6b shows the time evolution of the total number of particles classified as FCC, HCP, and BCC throughout this trajectory and the classification of 4 example snapshots of simulation frames.

The characterization reveals how a primarily disperse colloidal system state initially forms a cluster that almost entirely consists of BCC particles (e.g., purple bulk particles corresponding to cluster C12 and light blue and dark blue surface particles corresponding to clusters C6 and C10 respectively). A sudden, drastic phase transition occurs at Frame #18 as the system state transitions from a BCC structure to an almost entirely FCC structure. Over time, the remaining BCC particles transition slowly to FCC parti-

cles. By the Frame #99, the system state is an entirely FCC structure. The fact that the characterization framework can identify a BCC/FCC phase transition in an independent data set demonstrates the framework's generalizability.

We finally applied the characterization framework to another independent data set from ref. 23 that consists of 6 snapshots of 2052 *in-silico* colloids undergoing evaporation-induced SA (see Fig. 7). We first characterized these particles using the same methodology that was used to characterize the previous independent data set (i.e., the isothermal SA of 1000 multiflavored colloids). Fig. 7a shows the classification results for the final three snapshots.

We next re-performed the entire characterization framework on the evaporation-induced colloidal SA data alone. We collected all unique neighborhood graphs from the 6 provided snapshots (4462 total unique neighborhood graphs). We performed elbow analysis and re-trained the autoencoder on the 4462 unique neighborhood graphs. We reduced the dimensionality of the neighborhood graphs using the newly formed encoder and implemented agglomerative hierarchical clustering using Ward's linkage to partition the low-dimensional space. We chose the minimum number of clusters required to separate theoretically perfect FCC, HCP, and BCC lattices into separate clusters (11 total clusters).

We then assigned distinct colors to each of the separate clusters and visualized the lattices using OVITO (see Fig. 7b). Note that we assigned identical colors to important classes in both Figs. 7a-b. For example, green is FCC, brown is HCP, purple is BCC, and surface FCC is yellow in Figs. 7a-b. The evaporation-induced colloidal SA data set shows significantly fewer disperse and weakly crystalline states and more defective crystalline states than the multi-flavored colloidal SA data set does. The evaporation-induced low-dimensional space is thus biased towards such crystalline structures. As a result, the classifications in Figs. 7a-b show some important differences. For example, Fig. 7b shows teal particles that clearly correspond to defective FCC particles, yet such a class was not recovered from clustering the multi-flavored data. With the exception of these small numbers of particles, however, Figs. 7a-b do show almost identical colloidal SA state classifications. These results not only validate the generalizability of the characterization framework (as nearly identical results were seen by training the autoencoder on different systems with vastly different particle numbers) but also highlights how larger, more diverse training data sets can further improve the characterization framework.

4 Conclusions and Future Work

We first demonstrated the proposed colloidal self-assembly state characterization framework on an *in-silico* system of 500 multi-flavored colloids that self-assemble under isothermal conditions. The framework not only characterized the target FCC, BCC, and HCP structures but also "discovered" several relevant defective and surface structures that allowed for greater understanding of example colloidal self-assembly trajectories. We then analyzed the generalizability of the framework by applying the framework to two independent systems, one that consists of 1000 *in-silico* multi-flavored colloidal particles and self-assembles under isothermal conditions and another that consists of 2052 *in-silico* colloidal particles and undergoes evaporation-induced self-assembly. Despite successful characterization of the independent data sets, the framework can be sensitive to the nature of the data on which the autoencoder is trained (e.g., number of crystalline vs. weakly crystalline states).

We will focus future work on using the low-dimensional, physically-informative descriptions of colloidal SA system states provided by the encoder to create data driven models that predict the time evolution of the SA trajectories. We will supplement these data driven models by using the proposed dimensionality reduction framework to estimate free energy landscapes and identify kinetic traps within those landscapes. We will finally investigate combining predictions from the data driven model with information from the free energy landscapes to identify conditions under which SA is most likely to avoid kinetic traps and form desired structures.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Prof. Mike Howard for providing the snapshots of evaporation-induced crystallization used in Section 3.3 and Prof. Wes Reinhart for useful discussions on the neighborhood graph method. This work was supported by the U.S. Department of Energy, Office of Basic Energy Science, Division of Material Sciences and Engineering under Award (DE-SC0013979). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported under Contract No. DE-AC02-05CH11231. Use of the high-performance computing capabilities of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation, project no. TG-MCB120014, is also gratefully acknowledged.

Notes and references

1 G. M. Whitesides and B. Grzybowski, *Science*, 2002, **295**, 2418–2421.
2 J. A. Paulson, A. Mesbah, X. Zhu, M. C. Molaro and R. D. Braatz, *Journal of Process Control*, 2015, **27**, 38–49.
3 J. A. Liddle and G. M. Gallatin, *ACS nano*, 2016, **10**, 2995–3014.
4 J. J. Juárez and M. A. Bevan, *Advanced Functional Materials*, 2012, **22**, 3833–3839.
5 J. D. Joannopoulos, P. R. Villeneuve and S. Fan, *Nature*, 1997, **386**, 143–149.
6 E. M. Furst, *Soft Matter*, 2013, **9**, 9039–9045.
7 P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Physical Review B*, 1983, **28**, 784.
8 W. Lechner and C. Dellago, *The Journal of chemical physics*, 2008, **129**, 114707.
9 J. D. Honeycutt and H. C. Andersen, *Journal of Physical Chemistry*, 1987, **91**, 4950–4963.

10 D. Faken and H. Jónsson, *Comput. Mater. Sci*, 1994, **2**, 279–286.

11 P. M. Larsen, S. Schmidt and J. Schiøtz, *Modelling and Simulation in Materials Science and Engineering*, 2016, **24**, 055007.

12 G. Ackland and A. Jones, *Physical Review B*, 2006, **73**, 054104.

13 W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson and A. Z. Panagiotopoulos, *Soft Matter*, 2017, **13**, 4733–4745.

14 W. F. Reinhart and A. Z. Panagiotopoulos, *Soft matter*, 2017, **13**, 6803–6809.

15 W. F. Reinhart and A. Z. Panagiotopoulos, *Soft matter*, 2018, **14**, 6083–6089.

16 A. W. Long and A. L. Ferguson, *The Journal of Physical Chemistry B*, 2014, **118**, 4228–4244.

17 A. W. Long, J. Zhang, S. Granick and A. L. Ferguson, *Soft Matter*, 2015, **11**, 8141–8153.

18 A. W. Long and A. L. Ferguson, *Applied and Computational Harmonic Analysis*, 2019, **47**, 190–211.

19 A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis and P. G. Debenedetti, *Chemical Physics Letters*, 2011, **509**, 1–11.

20 A. L. Ferguson, *Journal of Physics: Condensed Matter*, 2017, **30**, 043002.

21 D. J. Beltran-Villegas, R. M. Sehgal, D. Maroudas, D. M. Ford and M. A. Bevan, *The Journal of chemical physics*, 2012, **137**, 134901.

22 X. Tang, B. Rupp, Y. Yang, T. D. Edwards, M. A. Grover and M. A. Bevan, *ACS nano*, 2016, **10**, 6791–6798.

23 M. P. Howard, W. F. Reinhart, T. Sanyal, M. S. Shell, A. Nikoubashman and A. Z. Panagiotopoulos, *The Journal of chemical physics*, 2018, **149**, 094901.

24 R. Jadrich, B. Lindquist and T. Truskett, *The Journal of chemical physics*, 2018, **149**, 194109.

25 R. Jadrich, B. Lindquist, W. Piñeros, D. Banerjee and T. Truskett, *The Journal of chemical physics*, 2018, **149**, 194110.

26 I. T. Jolliffe and J. Cadima, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, **374**, 20150202.

27 M. Spellings and S. C. Glotzer, *AIChE Journal*, 2018, **64**, 2198–2206.

28 P. Baldi, Proceedings of ICML workshop on unsupervised and transfer learning, 2012, pp. 37–49.

29 Y. Wang, H. Yao and S. Zhao, *Neurocomputing*, 2016, **184**, 232–242.

30 E. Boattini, M. Dijkstra and L. Filion, *The Journal of chemical physics*, 2019, **151**, 154901.

31 T. Milenković and N. Pržulj, *Cancer informatics*, 2008, **6**, CIN–S680.

32 N. Pržulj, D. G. Corneil and I. Jurisica, *Bioinformatics*, 2004, **20**, 3508–3515.

33 N. Pržulj, *Bioinformatics*, 2007, **23**, e177–e183.

34 T. Hočevar and J. Demšar, *Bioinformatics*, 2014, **30**, 559–565.

35 E. Pretti, H. Zerze, M. Song, Y. Ding, N. A. Mahynski, H. W. Hatch, V. K. Shen and J. Mittal, *Soft Matter*, 2018, **14**, 6303–6312.

36 E. Pretti, R. Mao and J. Mittal, *Molecular Simulation*, 2019, **45**, 1203–1210.

37 E. Pretti, H. Zerze, M. Song, Y. Ding, R. Mao and J. Mittal, *Science advances*, 2019, **5**, eaaw5912.

38 M. Song, Y. Ding, H. Zerze, M. A. Snyder and J. Mittal, *Langmuir*, 2018, **34**, 991–998.

39 J. O'Leary, *Characterizing colloidal self-assembly system states*, `https://github.com/jtoleary/colloid_char`, Online; accessed 5 August 2020.

40 A. Stukowski, *Modelling and Simulation in Materials Science and Engineering*, 2012, **20**, 045021.

41 N. A. Mahynski, R. Mao, E. Pretti, V. K. Shen and J. Mittal, *Soft Matter*, 2020, **16**, 3187–3194.

42 J. Yao, N. Teng, H.-L. Poh and C. L. Tan, *J. Inf. Sci. Eng.*, 1998, **14**, 843–862.

43 M. Scardi and L. W. Harding Jr, *Ecological modelling*, 1999, **120**, 213–223.

44 M. Gevrey, I. Dimopoulos and S. Lek, *Ecological modelling*, 2003, **160**, 249–264.

45 J. D. Olden, M. K. Joy and R. G. Death, *Ecological modelling*, 2004, **178**, 389–397.

46 K. Sasirekha and P. Baby, *International Journal of Scientific and Research Publications*, 2013, **83**, 83.

47 A. K. Jain, M. N. Murty and P. J. Flynn, *ACM computing surveys (CSUR)*, 1999, **31**, 264–323.

48 A. Stukowski, *Modelling and Simulation in Materials Science and Engineering*, 2009, **18**, 015012.

49 R. J. Macfarlane, B. Lee, M. R. Jones, N. Harris, G. C. Schatz and C. A. Mirkin, *science*, 2011, **334**, 204–208.

50 M. T. Casey, R. T. Scarlett, W. B. Rogers, I. Jenkins, T. Sinno and J. C. Crocker, *Nature communications*, 2012, **3**, 1–8.

51 R. T. Scarlett, M. T. Ung, J. C. Crocker and T. Sinno, *Soft Matter*, 2011, **7**, 1912–1925.

52 S. Plimpton, *Journal of computational physics*, 1995, **117**, 1–19.

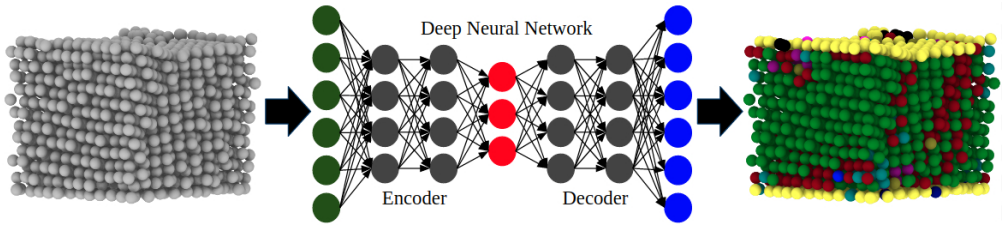
53 M. Farina, Y. Nakai and D. Shih, *Physical Review D*, 2020, **101**, 075021.

54 S. Salvador and P. Chan, 16th IEEE international conference on tools with artificial intelligence, 2004, pp. 576–584.

55 T. Kawasaki and H. Tanaka, *Proceedings of the National Academy of Sciences*, 2010, **107**, 14036–14041.

56 P. R. Ten Wolde, M. J. Ruiz-Montero and D. Frenkel, *Physical review letters*, 1995, **75**, 2714.

57 P. Rein ten Wolde, M. J. Ruiz-Montero and D. Frenkel, *The Journal of chemical physics*, 1996, **104**, 9932–9947.



412x91mm (72 x 72 DPI)