Soft Matter

Automated crystal characterization with a fast Neighborhood Graph Analysis method

Journal:	*Soft Matter*
Manuscript ID	SM-ART-05-2018-000960
Article Type:	Paper
Date Submitted by the Author:	09-May-2018
Complete List of Authors:	Reinhart, Wesley; Princeton University Department of Chemical and Biological Engineering Panagiotopoulos, Athanassios; Princeton University Department of Chemical and Biological Engineering

Soft Matter

PAPER

Cite this: DOI: 10.1039/xxxxxxxxxx

Automated crystal characterization with a fast Neighborhood Graph Analysis method

Wesley F. Reinhart[a] and Athanassios Z. Panagiotopoulos*[,a]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

We present a significantly improved, very fast implementation of the Neighborhood Graph Analysis technique for template-free characterization of crystal structures [W.F. Reinhart et al., *Soft Matter*, 2017, **13**, 4733]. By comparing local neighborhoods in terms of their relative graphlet frequencies, we reduce the computational cost by four orders of magnitude compared to the original stochastic method. Furthermore, we present protocols for the detection of topologically important structures and assignment of visually informative colors, providing a fully automated procedure for characterization of crystal structures from particle tracking data. We demonstrate the flexibility of our method on a wide range of crystal structures which have proven difficult to classify by previously available techniques.

1 Introduction

Colloidal crystals continue to draw attention for their applications in photonics,[1,2] energy conversion,[3,4] and biosensing[5,6] as well as for their usefulness in elucidating questions in fundamental physics[7,8] and materials science.[9,10] This wide range of applications is made possible by their equally wide range of possible self-assembled structures, which may be amorphous or crystalline, close-packed or open, and may include a variety of defects with tunable morphology and concentration.[11] Furthermore, by introducing even a small number of distinct particle types with asymmetrical interactions, particles can be "programmed" to spontaneously assemble into hierarchically ordered structures.[12]

At the same time, rational or "inverse" design strategies have become increasingly viable due to advances in both simulation and optimization methods.[13,14] Given such a myriad of possible target structures with varying coordination number and symmetry, it can be tedious to obtain an objective, quantitative metric for tracking the evolution of order in a system. Indeed, even very recent work has relied upon heuristics programmed by hand.[15–17] By far the most common metrics used in such heuristics are the Steinhardt local bond order parameters.[18,19] These characterize the symmetry of a particle's neighbor shell using the spherical harmonic basis functions, yielding a measure of the neighborhood shape.

Because the local bond order parameters are continuous, they are useful for quantifying generic order-disorder transitions.[20] However, thermal fluctuations can disrupt local structure, smearing the local bond order into broad distributions which are difficult to separate into distinct classes when the structures of interest are topologically similar. This is actually quite common since the close-packed face-centered cubic (FCC), hexagonal close-packed (HCP), and body-centered cubic (BCC) lattices can coexist in many metals. For instance, in ref. 15, the authors defined BCC particles as those with $0.016 < \bar{q}_4 < 0.056$ and $0.49 < \bar{q}_6 < 0.53$, compared to FCC which had $0.16 < \bar{q}_4 < 0.22$ and $0.554 < \bar{q}_6 < 0.594$. Such carefully tuned parameters sometimes even require manual calibration at many different state points when temperature and density vary significantly.[17]

The recently developed Polyhedral Template Matching (PTM)[21] seeks to address this problem by combining both geometrical and topological measures from each particle's neighbor shell. By considering the shape of each particle's Voronoi polyhedron when determining particle adjacency, the calculation becomes more resilient against thermal noise. Unlike the Steinhardt parameters, PTM provides a discrete classification by comparing the encompassing polyhedral net to a set of templates, accepting only perfect matches. A compromise is struck between topological and geometrical information by then quantifying the deviation of the particles in real space from their positions in the ideal template. While this approach provides a robust classification of thermalized metals, it requires the user to supply a list of templates *a priori*, which limits its usefulness in cases where the structures are unknown or irregular – cases which are increasingly likely as the complexity of available building blocks continues to rise.

We have recently introduced a new technique for the identifi-

[a]*Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, 08544, USA. E-mail: azp@princeton.edu; Tel: +1-609-258-4591*

† Electronic Supplementary Information (ESI) available: Classification results for all snapshots included in the sample set for Section 3.2. See DOI: 10.1039/b000000x/

cation of crystal structures called Neighborhood Graph Analysis (NGA).[22] Similar to PTM, we use the neighborhood graph topology as a signature since its discrete nature makes it less susceptible to thermal noise than geometric descriptors. We then employ the diffusion maps dimensionality reduction technique to infer relationships between all observed structures in a continuous space as in the local bond order approach. In this way, NGA combines the advantages of both fuzzy and discrete methods and discovers relevant structures directly from the data. Our "template-free" approach therefore stands to provide a more discriminating classification while eliminating the need for hand-crafted heuristics.

However, the algorithm published in ref. 22 relied on an expensive combinatorial procedure to evaluate the difference between graphs. This made it impractical for "on-the-fly" determination of crystal structures, since the analysis of a configuration of just 2000 particles required roughly 50 core-hours of computing time. We now present an implementation based on the relative graphlet frequency of the neighborhood graph, which speeds up the calculation by more than $10000\times$ compared to the original approach. Our new "fast NGA" (fNGA) is thus viable for real-time, desktop computing instead of being restricted to high performance computing infrastructure.

2 Method

2.1 Neighborhood construction

In our prior work,[22,23] we used the adjacency criterion from adaptive Common Neighbor Analysis[24] to define the local neighborhood graph of each particle. This approach defines a cutoff radius based on the distance to each particle's nearest six neighbors, with a geometric prefactor chosen to distinguish the first- and second-nearest neighbors in close-packed lattices. We noted in ref. 22 that the use of a geometric prefactor based on the HCP lattice interfered with the classification of the BCC lattice, whose coordination shell has a slightly different geometry. Furthermore, this approach assumes that the neighborhood should be isotropic, which is not necessarily the case for low-density "open" lattices derived from anisotropic particles.

To overcome these limitations, we now employ a Delaunay triangulation to obtain topologically adjacent particles, similar to the method used in PTM. This accommodates spatially inhomogeneous systems and lattices with anisotropic primitive cells. It is also more consistent with our philosophy of evaluating crystal structure based on neighborhood topology rather than geometry. Unlike in our original approach, we now also include the central particle in the neighborhood graph, which provides greater connectivity between neighbors and therefore greater distinction between structures.

Unfortunately, we found that in some cases the Delaunay triangulation gave spurious results. For instance, at smooth crystal surfaces (such as the solid-vapor interfaces in Section 3.1), the Delaunay triangulation may connect far-away particles in order to provide a convex hull in three dimensions. PTM, which also defines a local neighbor shell based on the Delaunay triangulation, filters such spurious neighbors based on the area of the Voronoi faces. This approach only works in the bulk, and provides no useful information in the cases such as at crystal surfaces where it is most necessary.

To address this issue, we applied agglomerative hierarchical clustering to differentiate the neighbor shell from these spurious topological neighbors. Clusters were split when the distance between centroids exceeded 25% of the first nearest neighbor distance, although the particular value was relatively unimportant. Ideally, a future implementation of our algorithm would be able to filter outliers based on a topological measure rather than a geometrical one. Such a modification would be necessary to capture strongly anisotropic structures. The NGA algorithm itself operates only on graphs, and as such can be used in conjunction with any adjacency criteria.

2.2 Fast graph comparisons using graphlets

In our previous implementations of NGA,[22,23] computing differences between graphs was by far the most expensive part of the algorithm, requiring roughly one minute per thread per evaluation on a 2.66 GHz Intel Xeon X5650 processor. Such long computation times were inevitable because the definition of our metric relied upon finding the optimal permutation of node labels between two graphs of up to 16 nodes each. In addition to being costly, the stochastic nature of the algorithm led to strong execution divergence when running in parallel (which was virtually required due to the huge computational burden). We went to great lengths to approximate these distances using an artificial neural network, but the final calculation presented in ref. 22 still required over 24 hours running on 100 cores of a computing cluster, and suffered from some uncertainty as a result of the approximations.

We now propose an alternate distance metric based on the graphlet decomposition of our neighborhood graphs. Graphlets are small induced subgraphs of a network, typically with $k \leq 5$ nodes, which can be used to evaluate subgraph isomorphism without invoking combinatorial approaches such as our original permutation-based technique. The relative graphlet frequency has been used extensively to perform quantitative comparisons between biological networks.[25,26]

To evaluate a neighborhood graph, we first computed the frequency of each orbit (a specific realization of each graphlet) formed by its constituent nodes using the `ORCA` library.[27] Because the occurrence of large graphlets is correlated to that of smaller ones, we reweighted each frequency according to a weight $w = 1 - o/73$, where o is the number of other graphlets each affects.[28] While it is standard practice in network analysis to use a logarithmic measure of the graphlet frequency, we found that such a transformation reduced our ability to resolve similar structures such as crystal polymorphs. We summed the reweighted frequencies from all nodes in the neighborhood graph to produce a total graphlet frequency for the neighborhood, $\vec{f}_{\text{tot}} = \sum_{i \in \text{nodes}} \vec{f}_i$. This was normalized to give the relative graphlet frequency, $\vec{f} = \vec{f}_{\text{tot}} / \sum \vec{f}_{\text{tot}}$.

The distance between two graphs is then defined as the L2-norm between the two relative graphlet frequency vectors, $d_{i,j} = \left\| \vec{f}_i - \vec{f}_j \right\|$. Because each vector is bounded on $[0,1]$, the distance

is also bounded on this interval. Computing the graphlet decomposition for graphs in our test cases took an average of 10^{-4} s, and evaluating the norm of the difference of two small vectors is practically free. The calculations are also deterministic and so do not suffer from execution divergence, making them easier to parallelize efficiently. Overall, the graphlet-based distance metric makes fNGA roughly $10000\times$ faster than the original algorithm of ref. 22, while maintaining nearly identical manifold topology.

2.3 Dimensionality reduction

We use the diffusion maps dimensionality reduction technique[29–31] to produce a low-dimensional manifold describing the complete graph space. From this description, we can easily infer relationships between particle neighborhoods. Our application of the technique to crystal lattices was inspired by its success in describing colloidal aggregates.[32–34] Since a detailed description of the algorithm and its role in NGA is available in ref. 22, here we will summarize only the most pertinent details. In most cases, we use Nyström extension[35,36] to build the diffusion map from a subset of "landmark" points rather than computing the full distance matrix between all pairs from the sample set. It has been demonstrated that this approach can greatly reduce the cost of the calculation while introducing negligible error.[37] Neighborhood graphs were chosen as landmarks if they formed at least one contiguous cluster of size 3 or greater.

We found that the density of points in the manifold was highly non-uniform, with high density around crystal lattices and low density in liquid regions (this was also observed in ref. 22). This problem can be mitigated by scaling the distances d_{ij} by an exponent $0 < \alpha < 1$ such that small distances are magnified compared to large ones.[38] While more rigorous methods for determining a suitable α exist, in our case choosing $\alpha = 0.5$ worked well as a generic solution. Likewise, we found that the topology of the manifold was insensitive to our choice of neighborhood radius ε, and that taking $\varepsilon = \text{median}(D^2)$ gave useful results for all of our test cases.

In our previous work, spectral analysis suggested a manifold dimensionality of $2 < d < 3$.[22] Here we assume $d \approx 3$ and define the collective variables to be used in the crystal characterization as the first three nontrivial eigenvectors of the diffusion map. This allows us to easily map the collective variables $\{\psi_i\}$ to RGB color space for automated visualization of the snapshots. We found that applying a rank transform to each ψ_i to create a uniform distribution over $[0,1]$ improved the quality of the visualization by increasing contrast in dense regions (e.g., between stacking polymorphs) and reducing contrast in sparse regions (e.g., for liquids). This also allowed us to easily cast (ψ_2, ψ_3, ψ_4) to RGB color space, which each eigenvector acting as a color channel.

3 Results and discussion

3.1 Crystals with exposed surfaces

We first applied our new fNGA algorithm to snapshots from a simulation of evaporation-induced crystallization, which we had previously evaluated with the original NGA method.[22] This system is difficult to classify by traditional structural analysis methods for two reasons: i) the sharp interface with a vapor phase creates a region of the crystal which has reduced coordination number but retains its symmetry and ii) weak interactions between the particles lead to large thermal fluctuations which can interfere with shape-based metrics. Our test case consisted of 50 snapshots with 2052 particles each, from which fNGA found 30444 unique graphs. Of these, 58 were chosen as landmarks for Nyström embedding based on the criterion in Section 2.3. The analysis took 203 seconds running on one core of a 2.66 GHz Intel Xeon X5650 processor in a Mac Pro workstation.

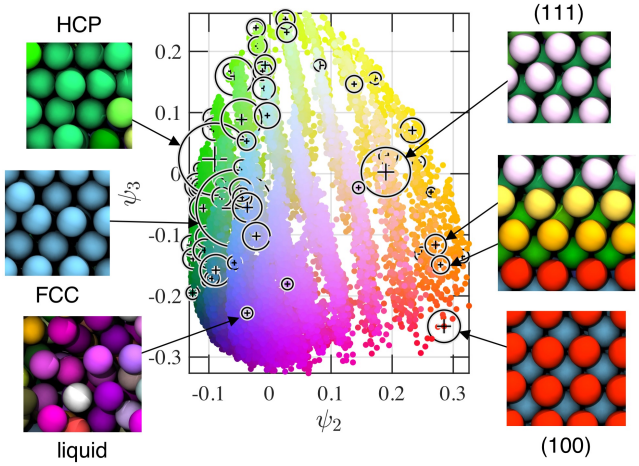

Fig. 1 Projection of the first two collective variables obtained from the diffusion map of evaporation-induced crystallization snapshots. The manifold contains 30444 unique neighborhood graphs (colored points) and 58 landmark graphs used for Nyström embedding (circles/crosses). Circle size is proportional to observation frequency, while cross size is proportional to maximum cluster size. Representative particle configurations are indicated with arrows; particle colors correspond to their position in the manifold. These snapshots and all others presented in this article were rendered with the Tachyon renderer included in VMD.[39,40]

A projection of the graphs based on the first two collective variables is shown in Fig. 1 along with representative real-space particle configurations from the simulation snapshots. As discussed in our prior work, the collective variables provided by the diffusion map do not necessarily correspond to any physically intuitive quantities. Here the dominant eigenvector ψ_2 appears to correspond to coordination number, with crystal structures on the left and surface structures on the right. The next-strongest eigenvector ψ_3 could be related to hexagonal symmetry, with a progression from liquid and square (100) surface to FCC and then to HCP. While such rationalizations can be useful in interpreting the results for familiar systems, in general it is better to think of these coordinates as more abstract descriptions of the graph space. We note that the banded structure appearing in Fig. 1 is likely related to the discrete size of the graphs.

We show the coloring scheme from Fig. 1 as applied to the real space snapshots in Fig. 2, alongside the same snapshots classified using PTM (as implemented in Ovito[41]). These snapshots highlight several key advantages of the template-free approach. Automatic detection of important structures provides a characterization for all particles, as opposed to only those with close

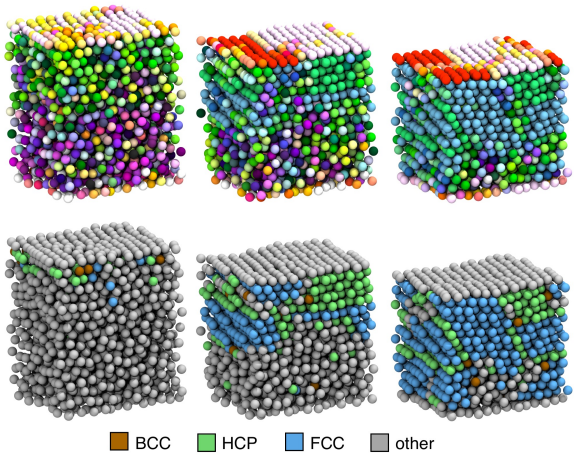

Fig. 2 Characterization of three snapshots from molecular simulation of evaporation-induced crystallization by fNGA (top) and PTM (bottom). For fNGA, the colors correspond to the diffusion map in Fig. 1. For PTM, colors are indicated underneath the snapshots. The maximum RMSD for a particle to be considered crystalline in PTM was 0.15.

resemblance to at least one template. This can provide additional information when the crystal is noisy or defective, as illustrated by the large region of light green particles in the leftmost frame corresponding to somewhat disordered rHCP. Perhaps more importantly, fNGA also identifies new structures and characterizes them according to their relationships in graph space, as shown by the various classes of surface structures which evolve in conjunction with the bulk morphology. Incidentally, the template-free approach also reduces the rate of false positives when the set of supplied templates would not match the observed structures. For instance, several particles are classified as BCC by PTM, but fNGA suggests that they are actually rHCP.

We note that even with the vastly improved performance of fNGA compared to its original implementation, PTM still offers a substantial performance advantage; classification by PTM took only 9 s on the same machine described above. This roughly 20× speedup, coupled with significantly lower memory requirements due to the use of only a few templates, makes PTM better suited for use on very large systems with only a few possible crystal structures. However, Fig. 2 clearly demonstrates the advantage of fNGA when more nuanced information is desirable.

3.2 Thermalized crystal slabs

Next we evaluated a set of seven crystal lattices observed in colloidal self-assembly: BCC, FCC, HCP, simple cubic (SC), perovskite, and the cubic and hexagonal stacking polymorphs of tetrastack (CT and HT). The broad range of densities and symmetries among these structures, effected by only small changes to the particle interactions, was actually the original motivation for the development of NGA. Synthetic configurations were prepared by tethering purely repulsive WCA particles to lattice points by harmonic springs and thermalizing them with a Langevin thermostat at $T^* = 1$. The transition from crystal to melt was imitated by reducing the spring constant from $1000\,\varepsilon/\sigma^2$ in the center of the simulation box ($z = 0$) to zero at the edges ($z = \pm L_z/2$). We

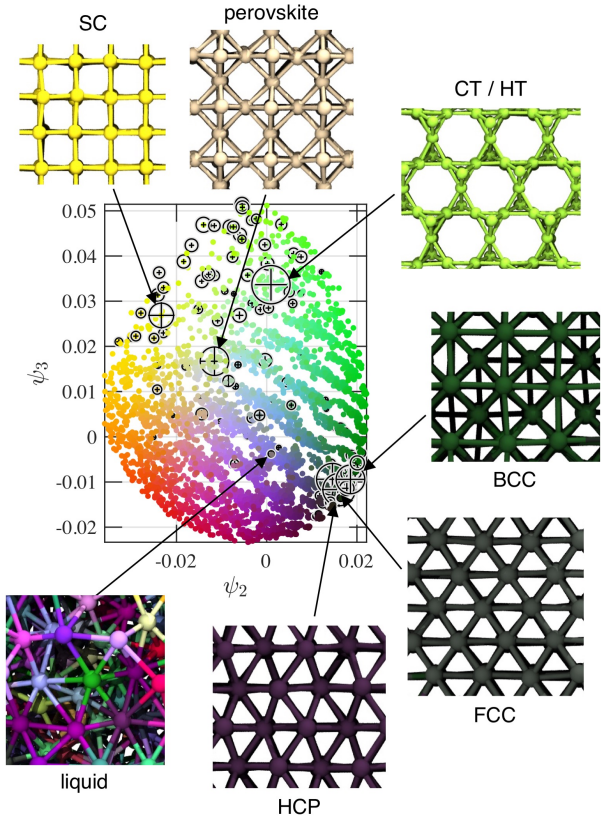

Fig. 3 Projection of the first two collective variables obtained from the diffusion map of thermalized crystal slabs with first-degree neighborhoods. Colored points indicate unique neighborhood graphs, circles/crosses indicate landmark graphs, and symbol size corresponds to observation frequency and largest cluster size, respectively. Real-space configurations are shown in a ball-and-stick representation to illustrate differences between crystal lattices.

also included a completely liquid configuration, where no springs were used, to populate the diffusion map with disordered structures. Snapshots of all eight configurations are available in the ESI.†

From the sample set, fNGA found 4809 unique graphs from 15216 particles, 121 of which were selected as landmarks. The analysis took 46 s on one core of our workstation. A projection into the first two collective variables from the diffusion map is shown in Fig. 3 along with representative real-space particle configurations from the snapshots. There is a strong separation in the diffusion map between structures with low coordination numbers (SC, perovskite, CT/HT; top left) and high coordination numbers (BCC,FCC,HCP; bottom right). Because the highly coordinated lattices are much more similar to each other than to those with low coordination number, they appear as a tight cluster of three different crystal signatures.

On the other hand, the tetrastack polymorphs are combined into a single signature because their first-degree neighborhoods are topologically identical. In order to differentiate them, the neighborhoods must be expanded to include third-degree neighbors. Repeating the calculation with these larger neighborhoods results in a qualitatively different diffusion map, shown in Fig. 4. This was constructed from 12543 unique graphs with the full pair-

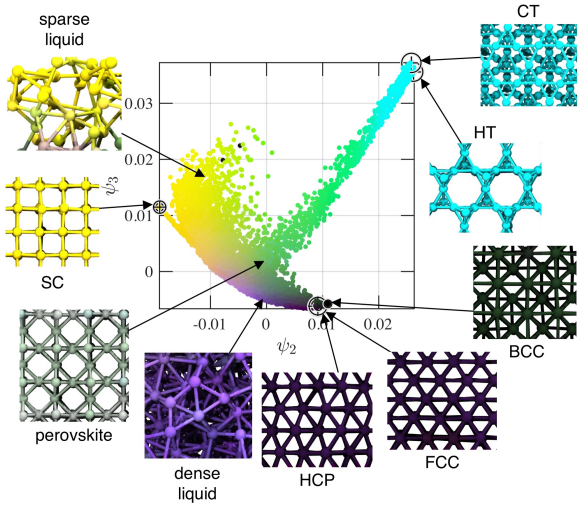

Fig. 4 Projection of the first two collective variables obtained from the diffusion map of thermalized crystal slabs with third-degree neighborhoods. Colored points indicate unique neighborhood graphs, circles/crosses indicate landmark graphs, and symbol size corresponds to observation frequency and largest cluster size, respectively. Real-space configurations are shown in a ball-and-stick representation to illustrate differences between crystal lattices.

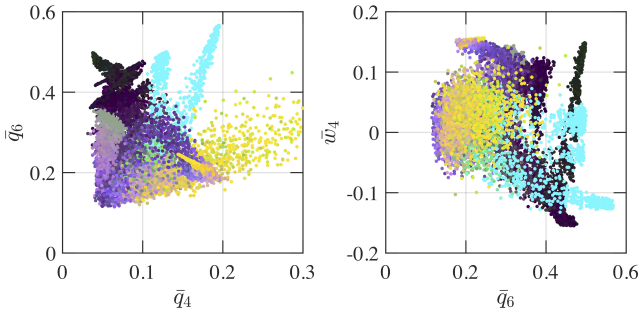

Fig. 5 fNGA collective variables compared to three of the local bond order parameters. Points correspond to particles, where position is determined by local bond order and color corresponds to the diffusion map shown in Fig. 4.

wise distance matrix since our usual definition provided too few landmarks to use Nyström embedding. The result was a much slower calculation at 392 s on our machine.

There are several downsides to the use of these large neighborhoods. First, the relative position (and resulting color) of nearby structures become compressed even when they are topologically distinct because large neighborhoods become increasingly delocalized from the central particle. For example, BCC, FCC, and HCP are nearly indistinguishable in Fig. 4, even though their difference vectors are computed from a substantially large sample of graphlets than in Fig. 3. The problem is that the relative difference between topologically dissimilar structures (e.g., FCC versus CT) becomes stronger as the graph expands, and the diffusion map adapts to capture these increasingly dominant trends. In this case, the tetrastacks develop into their own branch of the manifold, terminating in two distinct graphs which occupy nearly identical positions. Thus the inclusion of sufficient detail to distinguish these very similar structures actually makes the distinction between them insignificant compared to their differences with topologically dissimilar structures.

Simultaneously, each additional neighbor included in the graph increases the likelihood of exposing the neighborhood to defects, fragmenting the lattice into a large number distinct signatures which may differ by only a few missing edges. While this does not necessarily present any problems in building a diffusion map with appropriate topology (since the differences between them will still be small relative to differences with other structures), it does make it more difficult to detect lattices by simple metrics like frequency or cluster size. Of course, the increased size of the graphs and the increased number of unique graphs also makes the analysis more computationally demanding. In expanding from one to three neighbor shells, we more than doubled the number of unique graphs from 4809 to 12543 – meaning that more than 80% of particles in the sample had unique graphs. These larger graphs also demanded more time and memory for the graphlet decomposition and comparisons. All together, these factors led to nearly 10× slower analysis compared to only first-degree neighborhoods.

While there are no standard techniques for classifying all of the structures shown in Figs. 3 and 4, we did evaluate the snapshots using the local bond order parameters since they measure symmetry quite generally and can in principle be applied to any structure. However, they tend to yield disperse clouds of signatures which can make it difficult to distinguish polymorphs. We show the $(\bar{q}_4, \bar{q}_6, \bar{w}_4)$ bond order parameters for each particle in our eight snapshots in Fig. 5. The colors of these points correspond to those assigned to the same particles from Fig. 4, allowing us to make a direct comparison between $(\bar{q}_4, \bar{q}_6, \bar{w}_4)$ and (ψ_2, ψ_3, ψ_4). fNGA basically collects clouds of points with similar shape into single signatures, making crystal lattices stand apart from each other and from disordered regions. This is especially useful for picking out tetrastacks from Fig. 5, where at low $\bar{q}_6$ themalized particles may have a deformed shape but retain the topology of the lattice and at high $\bar{q}_6$ some liquid particles may attain a high degree of symmetry among their nearest neighbors but lack the long-range ordering of the lattice. The full snapshots in ESI† show that fNGA clearly differentiates the crystal lattice from the surrounding liquid.

3.3 Open crystals

Finally, we evaluated a set of open structures which have been observed to self-assemble from patchy particles with tetrahedral symmetry.[42–44] The low coordination number of these structures makes them difficult to classify by conventional techniques, although some hand-coded heuristics have been successfully employed for similar structures observed in water crystallization.[45] As in the case of the tetrastack polymorphs, the first-degree neighborhoods of these lattices are identical because they are all made up of tetrahedrally coordinated particles. Therefore we expanded the neighborhood graphs to include third-degree neighbors for this system as we did in Section 3.2. We also found that the voids of the clathrate structures were sometimes occupied by liquid par-

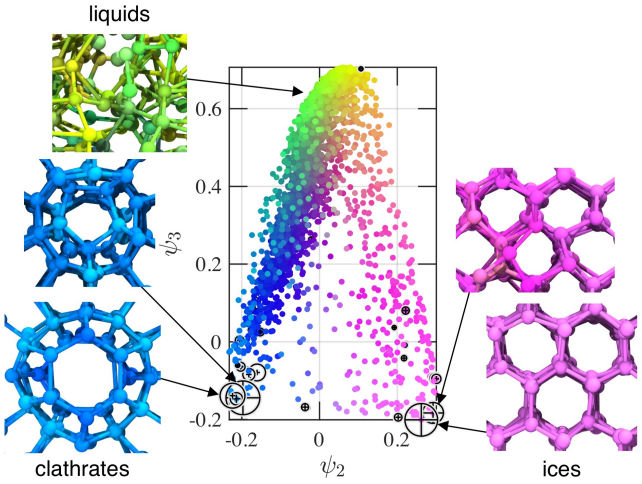

Fig. 6 Projection of the first two collective variables obtained from the diffusion map of open structures assembled from tetrahedral particles. Colored points indicate unique neighborhood graphs, circles/crosses indicate landmark graphs, and symbol size corresponds to observation frequency and largest cluster size, respectively.

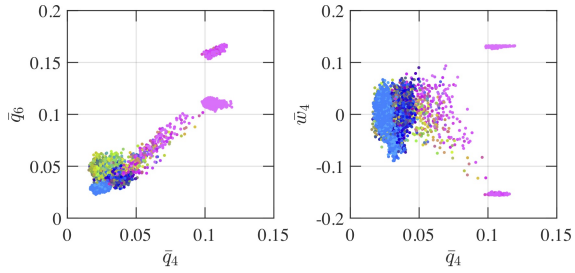

Fig. 7 fNGA collective variables compared to the local bond order parameters. Points correspond to particles, where position is determined by local bond order and color corresponds to the diffusion map shown in Fig. 6.

ticles whose anomalously high connectivity made them topologically distinct from all other observed structures, distorting the diffusion map to accommodate these outliers. Since these topologies are inconsistent with the tetrahedral symmetry of the particles, we assigned an empty neighborhood to particles with more than 6 neighbors. This way they do not participate in the neighborhood graphs of nearby crystalline particles, and instead are classified as a distinct structure (i.e., interstitial defects).

The collective variables shown in Fig. 6 appear to primarily distinguish clathrate structures (blue) and ice structures (pink), with continuous transitions to a liquid state (yellow/green). Interestingly, there are several different clathrate structures identified by the algorithm, including cages composed of five- and six-membered rings. On the other side of the manifold we observe distinct signatures for cubic and hexagonal ice. In each of these cases, the distance between polymorphs is very small relative to the distance to the opposite branches or to the liquid. As in our other examples, detection of these differences is virtually impossible by visual inspection and instead relies on the detection of large clusters with identical topology.

We again compared the local bond order parameters to the collective variables obtained from our diffusion map. Much the same as in Fig. 5, the point cloud in Fig. 7 shows that topologically similar structures become smeared out into clouds in the bond order space. This is especially true for the clathrates, which become indistinguishable from liquids because of their weak cubic and hexagonal symmetry. fNGA is able to distinguish between the hexagonal and cubic ice, which fall into clusters on either side of the $\bar{w}_4$ space, but the colors are nearly identical since these structures fall so close together in the diffusion map. Thus in cases where multiple large clusters are detected very near to each other in the diffusion map, bond order analysis can offer additional insight into how the real space structures may differ despite having very similar topology.

3.4 Implementation

We have released an open-source implementation of the fNGA algorithm as a Python module at `https://github.com/wfreinhart/crayon`. Our software package relies upon a standalone code for computing the diffusion map which is available online.[46]

4 Conclusions

We have introduced a significantly faster implementation of the Neighborhood Graph Analysis method for the template-free characterization of crystal structures from particle tracking data. By using the difference between relative graphlet frequencies as the difference metric instead of computing the optimal permutation of nodes between two graphs, we obtained a $10000\times$ speedup for our "fast NGA" (fNGA) compared to the original NGA method. We have also developed an automated procedure to assign visually distinct colors to each neighbor topology, providing a completely autonomous workflow for characterizing crystal structures.

We demonstrated our improved method on three different test systems, each of which pose problems to conventional crystal classification techniques. In the case of evaporation-induced crystallization, we showed how fNGA could be used to characterize particles on the developing crystal surface – an important feature of heterogeneous nucleation and growth processes – while the state-of-the-art Polyhedral Template Matching technique could not. We applied the analysis to seven crystal structures with a wide range of coordination symmetries, most of which have no existing templates in the PTM framework, where fNGA both detected and characterized all seven lattices. fNGA was also able to characterize clathrate and ice structures in assemblies of tetrahedral particles, a task which has previously required hand-crafted heuristics for specific structures of interest. In all cases, fNGA can offer valuable insight into assembly mechanisms which previously could not be characterized or required expert intuition to assess effectively.

Conflicts of interest

There are no conflicts to declare.

Acknowledgments

We thank Michael Howard for providing the snapshots of evaporation-induced crystallization in Section 3.1 and Francesco Sciortino for providing the snapshots of open structures in Section 3.3. Research for this publication was conducted with government support under contract FA950-11-C-0028 and awarded by the Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. Additional support was provided by the U.S. National Science Foundation under award CBET-1402166.

References

1 J. F. Galisteo-López, M. Ibisate, R. Sapienza, L. S. Froufe-Pérez, Á. Blanco and C. López, *Adv. Mater.*, 2011, **23**, 30–69.

2 G. von Freymann, V. Kitaev, B. V. Lotsch and G. A. Ozin, *Chem. Soc. Rev.*, 2013, **42**, 2528–2554.

3 N. Tétreault, É. Arsenault, L.-P. Heiniger, N. Soheilnia, J. Brillet, T. Moehl, S. Zakeeruddin, G. A. Ozin and M. GraÌĹtzel, *Nano Lett.*, 2011, **11**, 4579–4584.

4 M. Karg, T. A. König, M. Retsch, C. Stelling, P. M. Reichstein, T. Honold, M. Thelakkat and A. Fery, *Mater. Today*, 2015, **18**, 185–205.

5 Y.-J. Lee and P. V. Braun, *Adv. Mater.*, 2003, **15**, 563–566.

6 C. Fenzl, T. Hirsch and O. S. Wolfbeis, *Angew. Chem. Int. Ed.*, 2014, **53**, 3318–3335.

7 S. Auer and D. Frenkel, *Nature*, 2001, **409**, 1020.

8 G. Meng, N. Arkus, M. P. Brenner and V. N. Manoharan, *Science*, 2010, **327**, 560–563.

9 P. Schall, I. Cohen, D. A. Weitz and F. Spaepen, *Science*, 2004, **305**, 1944–1948.

10 W. T. Irvine, M. J. Bowick and P. M. Chaikin, *Nat. Mater.*, 2012, **11**, 948.

11 N. Vogel, M. Retsch, C.-A. Fustin, A. del Campo and U. Jonas, *Chem. Rev.*, 2015, **115**, 6265–6311.

12 W. B. Rogers, W. M. Shih and V. N. Manoharan, *Nat. Rev. Mater.*, 2016, **1**, 16008.

13 A. Jain, J. A. Bollinger and T. M. Truskett, *AIChE J.*, 2014, **60**, 2732–2740.

14 G. van Anders, D. Klotsa, A. S. Karas, P. M. Dodd and S. C. Glotzer, *ACS Nano*, 2015, **9**, 9542–9553.

15 S. Tang, J. Wang, B. Svendsen and D. Raabe, *Acta Mater.*, 2017, **139**, 196–204.

16 M. Eshraghi and J. Horbach, *Soft Matter*, 2018.

17 A. Zaragoza, J. R. Espinosa, R. Ramos, J. A. Cobos, J. L. Aragones, C. Vega, E. Sanz, J. Ramírez and C. Valeriani, *J. Phys.: Condens. Matter*, 2018, **30**, 174002.

18 P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Phys. Rev. B*, 1983, **28**, 784.

19 W. Lechner and C. Dellago, *J. Chem. Phys.*, 2008, **129**, 114707.

20 H. Eslami, N. Khanjari and F. MuÌĹller-Plathe, *J. Chem. Theory Comput.*, 2017, **13**, 1307–1316.

21 P. M. Larsen, S. Schmidt and J. Schiøtz, *Modell. Simul. Mater. Sci. Eng.*, 2016, **24**, 055007.

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

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

24 A. Stukowski, *Modell. Simul. Mater. Sci. Eng.*, 2012, **20**, 045021.

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

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

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

28 T. Milenković and N. Pržulj, *Cancer Inf.*, 2008, **6**, CIN–S680.

29 R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 7426–7431.

30 R. R. Coifman and S. Lafon, *Appl. Comput. Harmon. Anal.*, 2006, **21**, 5–30.

31 A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis and P. G. Debenedetti, *Chem. Phys. Lett.*, 2011, **509**, 1 – 11.

32 A. W. Long and A. L. Ferguson, *J. Phys. Chem. B*, 2014, **118**, 4228–4244.

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

34 A. W. Long, C. L. Phillips, E. Jankowksi and A. L. Ferguson, *Soft Matter*, 2016, **12**, 7119–7135.

35 C. T. H. Baker, *The Numerical Treatment of Integral Equations*, Clarendon Press, Oxford, 1977, vol. 13.

36 B. E. Sonday, M. Haataja and I. G. Kevrekidis, *Phys. Rev. E*, 2009, **80**, 031102.

37 A. W. Long and A. L. Ferguson, *Appl. Comput. Harmon. Anal.*, 2017.

38 J. Wang, M. A. Gayatri and A. L. Ferguson, *J. Phys. Chem. B*, 2017, **121**, 4923–4944.

39 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.

40 J. Stone, *MSc thesis*, Computer Science Department, University of Missouri-Rolla, 1998.

41 A. Stukowski, *Modell. Simul. Mater. Sci. Eng.*, 2009, **18**, 015012.

42 F. Romano, E. Sanz and F. Sciortino, *J. Chem. Phys.*, 2010, **132**, 184501.

43 F. Romano, E. Sanz and F. Sciortino, *J. Chem. Phys.*, 2011, **134**, 174502.

44 I. Saika-Voivod, F. Romano and F. Sciortino, *J. Chem. Phys.*, 2011, **135**, 124506.

45 J. Russo, F. Romano and H. Tanaka, *Nat. Mater.*, 2014, **13**, 733.

46 A. W. Long, *C++ implementation of the Diffusion Map, with Python bindings*, `https://github.com/awlong/DiffusionMap`, accessed on 01/12/2018.