The bcc coating of Lennard-Jones crystal nuclei vanishes with a change of local structure detection algorithm ⊘

Willem Gispen ✉ iD ; Alberto Pérez de Alba Ortíz iD ; Marjolein Dijkstra iD

Check for updates

View Online   Export Citation

Articles You May Be Interested In

KoopmanLab: Machine learning for solving complex physics equations

APL Mach. Learn. (September 2023)

Experimental realization of a quantum classification: Bell state measurement via machine learning

APL Mach. Learn. (September 2023)

The bcc coating of Lennard-Jones crystal nuclei vanishes with a change of local structure detection algorithm

View Online   Export Citation   CrossMark

Willem Gispen,[1,a)] (iD)   Alberto Pérez de Alba Ortíz,[2] (iD)   and Marjolein Dijkstra[1] (iD)

AFFILIATIONS

[1] Soft Condensed Matter & Biophysics, Debye Institute for Nanomaterials Science, Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands

[2] Computational Soft Matter Lab, Computational Chemistry Group and Computational Science Lab, Van 't Hoff Institute for Molecular Science and Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

Note: This paper is part of the JCP Special Topic on Molecular Dynamics, Methods and Applications 60 Years after Rahman.
[a)]Author to whom correspondence should be addressed: w.h.gispen2@uu.nl

ABSTRACT

Since the influential work of ten Wolde, Ruiz-Montero, and Frenkel [Phys. Rev. Lett. **75**, 2714 (1995)], crystal nucleation from a Lennard-Jones fluid has been regarded as a paradigmatic example of metastable crystal ordering at the surface of a critical nucleus. We apply seven commonly used local structure detection algorithms to characterize crystal nuclei obtained from transition path sampling simulations. The polymorph composition of these nuclei varies significantly depending on the algorithm used. Our results indicate that one should be very careful when characterizing the local structure near solid–solid and solid–fluid interfaces. Particles near such interfaces exhibit a local structure distinct from that of bulk fluid or bulk crystal phases. We argue that incorporating outlier detection into the local structure detection method is beneficial, leading to greater confidence in the classification results. Interestingly, the bcc coating nearly disappears when adopting a machine learning method with outlier detection.

I. INTRODUCTION

In the simplest view of crystallization, a stable crystal phase forms directly from a supercooled fluid. However, as early as 1897, Ostwald realized that this simple picture does not always hold: his famous step rule states that the nucleating phase is typically a less stable crystal phase instead. Later, Stranski and Totomanow[2] rationalized Ostwald's step rule by proposing that the nucleating phase is determined by the one with the lowest free-energy barrier for nucleation. Alexander and McTague[3] further demonstrated that general symmetry considerations uniquely favor the body-centered cubic (bcc) crystal phase near the solid–fluid coexistence. These studies all predict an important role for metastable phases during crystal nucleation. In recent decades, it has become possible to directly test these predictions using computer simulations. This has enabled the observation of intermediate phases in the crystallization of colloidal fluids,[4–7] metallic melts,[8,9] and aqueous solutions.[10,11] These intermediates can be metastable crystal polymorphs,[6–8,12] amorphous solids,[4,11] or fluid phases.[4,13]

In 1995, ten Wolde, Ruiz-Montero, and Frenkel[1] identified a remarkable role for bcc during crystal nucleation from a Lennard-Jones fluid, which is in simultaneous accordance with the conjectures of Ostwald, Stranski and Totomanow, and Alexander and McTague. They observed that small, precritical nuclei were predominantly bcc-like, while the critical nucleus was predominantly fcc-like. Surprisingly, the surface of the critical nucleus was also significantly bcc-like. Since the fcc crystal is the stable phase under the conditions they investigated, the initial formation of bcc is in line with Ostwald's step rule. The bcc "coating" is also consistent with density functional theory calculations,[14–16] which suggest that the presence of bcc lowers the nucleation barrier. Since the influential work of ten Wolde, Ruiz-Montero, and Frenkel,[1,17–21] similar coatings of metastable polymorphs on the surface of crystal nuclei have been identified in water,[19,22–24] colloids,[25–27] and metals.[28]

However, in the case of water, several authors have recently demonstrated that the surface structure is highly dependent on the local structure detection algorithm.[19,23,24] Even minor variations in the definition of nearest neighbors can lead to the appearance or disappearance of the coating.[23] This controversy in water raises the question of whether the bcc coating of Lennard-Jones nuclei exhibits a similar dependence on the local structure detection algorithm.

In this paper, we aim to address this question by first simulating crystal nucleation from a Lennard-Jones fluid under the same conditions as those investigated in Ref. 1 using transition path sampling. We characterize the structure of crystal nuclei using seven different local structure detection methods. This includes two variants of the histogram method from Ref. 1 as well as methods based on more recent advances in local structure detection, such as locally averaged bond-order parameters,[29] machine learning,[30] and polyhedral template matching.[31]

II. SIMULATION METHODS

A. Molecular dynamics

We perform molecular dynamics simulations of $N = 3 \times 10^4$ particles interacting with a Lennard-Jones pair potential,

$$u(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right],$$

in the isobaric–isothermal (NPT) ensemble. Here, σ and ε set the length and energy scales, respectively, and r is the distance between a pair of particles. Molecular dynamics simulations with a Lennard-Jones pair potential were originally used to model liquid argon,[32] but the simple form of the pair potential has also made it a principle testing ground for computational methods, including methods for nucleation and local structure detection.[17,18,29,33] We truncate the pair potential at $r = 2.5\sigma$ and apply isotropic tail corrections for the energy and pressure. We integrate the equations of motion with a Nosé–Hoover thermostat and barostat, as implemented in the LAMMPS molecular dynamics code.[34] We use a time step $\Delta t = 0.004\sqrt{m\sigma^2/\varepsilon}$, where m is the particle mass, and relaxation constants of 500 and 100 time steps for the barostat and thermostat, respectively. Unless stated otherwise, the simulation box is cubic and the barostat is isotropic.

B. Brute-force nucleation simulations

We first perform brute-force simulations of nucleation at three different state points. These state points correspond to ~30% supercooling at three different pressures: $P\sigma^3/\varepsilon = 0, 5.68$, and 50. The exact temperatures are $k_B T/\varepsilon = 0.50, 0.79$, and 2.30, respectively. Here, P denotes the pressure, T is the temperature, and k_B is Boltzmann's constant. For these simulations, we start with a fluid phase and wait for the system to spontaneously crystallize. We perform 16 independent brute-force nucleation simulations for each state point.

C. Transition path sampling

Similar to Ref. 1, we focus on temperature $k_B T/\varepsilon = 0.92$ and pressure $P\sigma^3/\varepsilon = 5.68$. As the melting temperature at this pressure

is approximately $k_B T_m/\varepsilon = 1.11$,[35] these conditions correspond to a supercooling of around 20%.

At this supercooling, there is a significant barrier for nucleation, meaning that it would take a very large amount of computer time for nucleation to occur spontaneously. To address this issue, enhanced sampling techniques, such as umbrella sampling,[36,37] metadynamics,[18,38,39] forward-flux sampling,[40] and transition path sampling,[17,20,26,41–43] have been introduced. All these techniques require an order parameter to monitor, or even drive, the progress of a nucleation trajectory. Previous work has used, for example, the number of particles in the nucleus, or the global Steinhardt bond orientational order Q_6 of the system,[1] as the order parameter. In the case of metadynamics, the choice of order parameter directly influences the dynamics to favor sampling away from already visited configurations, whereas umbrella sampling restrains the sampling toward a small order parameter range. In this work, we focus on the influence of the local structure detection algorithms on the *analysis* of crystal nuclei, rather than on the potential influence of an order parameter on the *simulation* of crystal nucleation. To minimize the impact of the choice of order parameter on the nucleation mechanism, we use transition path sampling.

Transition path sampling (TPS) is designed to efficiently sample unbiased trajectories of rare events, such as nucleation. To initialize the procedure, we first require an initial nucleation trajectory, which we obtain from one of the brute-force nucleation simulations at $k_B T/\varepsilon = 0.79$. From this initial trajectory, TPS generates an unbiased ensemble of nucleation trajectories using "shooting moves." These shooting moves generate a modified trajectory from a previous trajectory, thereby performing a random walk in trajectory space in a Monte Carlo fashion, and equilibrating the ensemble of trajectories. In Appendix A, we explain the details of our TPS simulations. The result of our TPS simulations is an unbiased ensemble of 50 decorrelated nucleation trajectories. Furthermore, we have performed a committor analysis to identify an ensemble of 50 critical nuclei.

Using TPS, we have access to 50 distinct critical nuclei and 50 unique nucleation trajectories at a temperature of $k_B T/\varepsilon = 0.92$ and a pressure of $P\sigma^3/\varepsilon = 5.68$.

III. LOCAL STRUCTURE DETECTION METHODS

To analyze the local structure in all these nucleation trajectories, we employ a variety of local structure detection methods. In Sec. III A, we explain how we identify crystal nuclei. Subsequently, we explain the seven different ways used to identify the crystal polymorphs face-centered cubic (fcc), hexagonal close-packed (hcp), and body-centered cubic (bcc) in these crystal nuclei. With the exception of polyhedral template matching,[31] these methods are based on local order parameters derived from spherical harmonics expansions $q_{lm}(i)$ of the nearest neighbor density of each particle i.[1,29,44] Unless stated otherwise, nearest neighbors are identified using the solid-angle based nearest-neighbor algorithm.[45] From the spherical harmonics expansion $q_{lm}(i)$, we compute the inner products $d_6(i,j) = q_{6m}(i) \cdot q_{6m}(j)$ as introduced in Ref. 1, along with the non-averaged bond order parameters $q_l(i)$ and $w_l(i)$, and the locally averaged bond order parameters $\overline{q}_l(i)$ and $\overline{w}_l(i)$ as introduced in Ref. 29. Please see Appendix B for the precise definition of these local order parameters.

A. Identification of crystal nuclei

Following Ref. [1], we use the inner products $d_6(i,j)$ to identify crystal nuclei. A pair of particles (i,j) has a "solid-like bond" if $d_6(i,j) > 0.7$. Subsequently, a particle is classified as "solid-like" if it has at least six of such solid-like bonds, and solid-like particles are considered to belong to the same solid-like cluster if they have a solid-like bond. For each snapshot in our simulations, we identify the largest solid-like cluster as the crystal nucleus. Only particles within this crystal nucleus are further analyzed using a polymorph classification scheme, while all other particles are simply labeled as "fluid-like."

B. Polymorph classification with non-averaged bond order parameters

We start with the polymorph classification scheme used by ten Wolde, Ruiz-Montero, and Frenkel.[1,46] This scheme is based on the distributions of the non-averaged bond order parameters (q_4, q_6, w_6). Reference [1] calculated histograms of (q_4, q_6, w_6) and concatenated these histograms to form a "characteristic vector" v. To classify a collection of particles, the characteristic vector v is calculated and projected onto the characteristic vectors of the bulk phases. To be specific, v is decomposed using a linear least-squares algorithm into contributions from the fcc, bcc, and fluid phases by minimizing the distance,

$$\Delta^2 = \left[v - \left(f_{fcc} v_{fcc} + f_{bcc} v_{bcc} + f_{fl} v_{fl} \right) \right]^2. \quad (1)$$

Here, f_{fcc}, f_{bcc}, and f_{fl} represent the fractions of fcc, bcc, and fluid-like order, respectively. During the optimization, the fractions are constrained to lie between 0 and 1.

We note that the presence of hcp-like order is not considered in Ref. [1]. To investigate its significance, we also examine a variant in which we project the characteristic vector onto the fcc, hcp, bcc, and fluid phases by minimizing the distance,

$$\Delta'^2 = \left[v - \left(f_{fcc} v_{fcc} + f_{hcp} v_{hcp} + f_{bcc} v_{bcc} + f_{fl} v_{fl} \right) \right]^2. \quad (2)$$

We will refer to the classification scheme based on Eq. (1) as the "histogram scheme excluding hcp" and the one based on Eq. (2) as the "histogram scheme including hcp."

To illustrate the histogram schemes, we show the distributions of q_4, q_6, w_4, and w_6 in the top row of Fig. 1. These distributions correspond to the bulk phases under the simulation conditions $P\sigma^3/\varepsilon = 5.68$ and $k_B T/\varepsilon = 0.92$. Although the distributions for the bulk phases are highly overlapping, there are distinct differences. For example, particles in the fcc phase generally exhibit higher q_4 values compared to particles in the bcc phase. The black line corresponds to the bond order distributions of particles within critical nuclei. To determine the relative contributions of different phases in these nuclei, the black lines are fitted as a linear combination of the other distributions.

C. Polymorph classification with locally averaged bond order parameters

Next, we discuss two different polymorph classification schemes based on the locally averaged bond order parameters $\bar{q}_l$ and $\bar{w}_l$. We label these schemes according to the order parameters they are based on: the $(\bar{w}_4, \bar{w}_6)$-scheme and the $\bar{q}_4$-scheme. In the $(\bar{w}_4, \bar{w}_6)$-scheme, a solid-like particle is classified as follows:
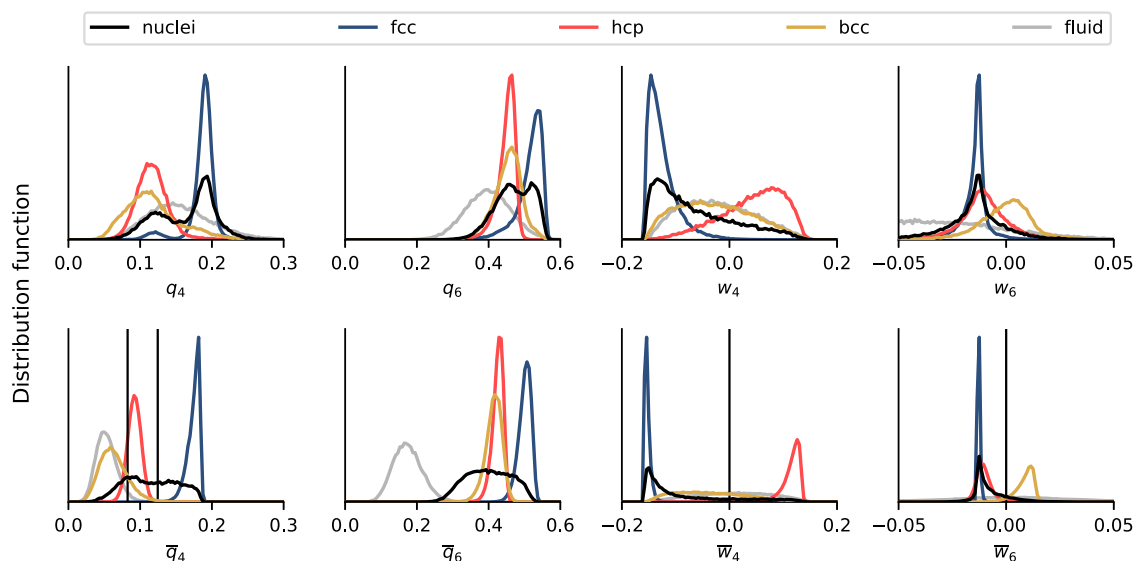

FIG. 1. Distributions of the non-averaged bond order parameters q_4, q_6, w_4, w_6 (top row) and their locally averaged counterparts $\bar{q}_4, \bar{q}_6, \bar{w}_4, \bar{w}_6$ (bottom row).[29] The lines represent the distributions of bulk phases at $P\sigma^3/\varepsilon = 5.68$ and $k_B T/\varepsilon = 0.92$, while the black lines correspond to particles in critical crystal nuclei under the same conditions. The vertical black lines in the bottom row indicate the thresholds used to distinguish between fcc, hcp, and bcc phases in the two polymorph classification schemes based on locally averaged bond order parameters.

$$\begin{cases} \text{fcc-like} & \text{if } \overline{w}_6 < 0 \text{ and } \overline{w}_4 < 0, \\ \text{hcp-like} & \text{if } \overline{w}_6 < 0 \text{ and } \overline{w}_4 > 0, \\ \text{bcc-like} & \text{if } \overline{w}_6 > 0. \end{cases}$$

In the $\overline{q}_4$-scheme, a solid-like particle is classified as follows:

$$\begin{cases} \text{fcc-like} & \text{if } \overline{q}_4 > 0.125, \\ \text{hcp-like} & \text{if } 0.083 < \overline{q}_4 < 0.125, \\ \text{bcc-like} & \text{if } \overline{q}_4 < 0.083. \end{cases}$$

We have chosen these thresholds to identify the crystal polymorphs with the fewest mislabeled particles.[47]

To illustrate these two schemes, we show the distributions of $\overline{q}_4, \overline{q}_6, \overline{w}_4$, and $\overline{w}_6$ in the bottom row of Fig. 1. Again, these distributions correspond to the bulk phases and the critical nuclei. The thresholds used above to distinguish the different phases are indicated by the black vertical lines. As Ref. 29 demonstrated, the distributions of the locally averaged bond order parameters $(\overline{q}_l, \overline{w}_l)$ are well-separated. For example, the fluid phase is well-separated from the crystal phases with respect to $\overline{q}_6$, and the fcc and hcp crystal phase are well-separated based on $\overline{w}_4$ and $\overline{q}_4$. Although $\overline{q}_4$ is commonly used to distinguish between fcc, hcp, and bcc phases, we note that the distributions for hcp and bcc show significant overlap.

D. Machine-learned polymorph classification with outlier detection

In recent years, there has been a rise in the use of machine learning for local structure detection.[24,28,30,48–54] In particular, several approaches have been proposed for crystal polymorph classification. While we do not claim that the schemes used here are necessarily the best or the only methods for polymorph classification using machine learning, there are several general aspects that we believe are important to consider when applying polymorph classification to nucleation. First, to compare to other polymorph classification schemes, we seek a machine-learning method that classifies particles as fluid, fcc, hcp, or bcc. Second, given the uncertain nature of interfacial particles, as observed in ice nucleation studies,[19,23,24] the method should be able to recognize when the local structure does not correspond to any of these phases. Third, to avoid coarse-graining across solid–fluid or solid–solid interfaces, the method should rely solely on the nearest neighbors of a particle. For these reasons, we use the following machine-learned polymorph classification schemes, which are largely inspired by Ref. 30.

To capture the local structure around a particle i, we use the non-averaged bond order parameters $q_l(i)$, as they rely solely on nearest-neighbor information. However, as shown in Fig. 1, these distributions exhibit more overlap compared to their averaged counterparts. Therefore, we employ a neural network to extract more informative features from the non-averaged bond order parameters.[53] To be more precise, the neural network performs a non-linear combination and dimensionality reduction of the input features into a lower-dimensional space, where the bulk phases are classified using a Softmax function. The neural network architecture is shown in Fig. 2. The input to the neural network is a 195-dimensional vector consisting of non-averaged bond order
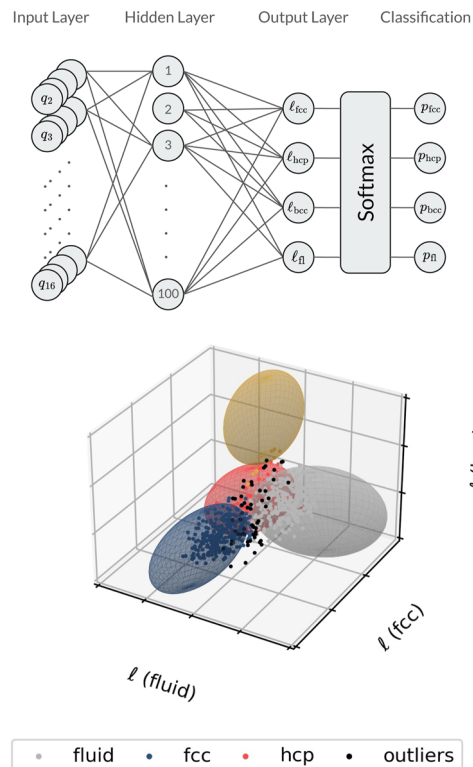

FIG. 2. Machine-learned polymorph classification with outlier detection. For each particle, 195 different non-averaged bond order parameters q_l are computed. A neural network transforms these q_l into non-normalized log-likelihoods ℓ for each bulk phase. In this four-dimensional latent space, multi-dimensional Gaussian distributions, represented here by ellipsoids, are fitted to the bulk phase distributions. If a particle environment maps to the interior of an ellipsoid, it is classified as belonging to the corresponding phase, otherwise, it is considered an outlier.

parameters $q_l(i)$ calculated using contributions from varying numbers of nearest neighbors. To be more specific, for each number $k = 2, \ldots, 14$, we identify the k nearest neighbors to particle i and compute the spherical harmonics expansions $q_{lm}(i)$. Then, for each number of nearest neighbor k, we compute the bond order parameters q_l, where we vary l from 2 to 16. In this way, we compute 13 sets of bond order parameters $q_2, \ldots, q_{16}$, giving a total number of $195 = 13 \times 15$ input features. We train the neural network to classify bulk fluid and crystal phases based on these input features. The neural network consists of 195 input nodes, one hidden layer of 100 nodes, and four output nodes corresponding to the four different phases. We also considered neural network architectures with more hidden layers and more hidden nodes, but found that these changes did not significantly improve the classification accuracy. Similarly, we considered including the third-order invariants w_l to the input features, but this also did not significantly improve the classification accuracy. In Appendix C, we perform a sensitivity analysis to assess which of the 195 input features are most important for the classification problem.

For the training data, we consider two approaches. In the first approach, following Ref. 30, we generate the training data by

adding Gaussian noise to ideal crystal lattices, specifically for fcc, hcp, and bcc. In other words, the training data are derived from Einstein crystals. The spring constants of these Einstein crystals are set to match those measured for a thermally equilibrated fcc crystal under our simulation conditions. The average magnitude of the Gaussian noise is ~10% of the nearest neighbor distance in each case. In addition, we add an equilibrated bulk fluid phase in the training data to enhance the network's capability to distinguish fluid-like from crystalline local structures. In the second approach, the training data are simply obtained from thermally equilibrated fcc, hcp, bcc, and fluid phases under our simulation conditions. We refer to the machine-learned (ML) classifiers based on these two training sets as the ML1 and ML2 schemes, respectively. To be clear, the ML1 scheme uses ideal crystal lattices with noise, while the ML2 scheme uses thermally equilibrated crystal phases.

As mentioned above, we require a method to detect when the local structure does not correspond to any of these phases. In machine learning, identifying deviations from a set of reference data is referred to as novelty detection or outlier detection. To describe the outlier detection algorithm that we use, we first note that the output of the neural network is a non-normalized log-likelihood ℓ for each phase. For example, the probability p_{fcc} that a particle belongs to the fcc phase is computed as

$$p_{\text{fcc}} = \frac{e^{\ell_{\text{fcc}}}}{e^{\ell_{\text{fcc}}} + e^{\ell_{\text{hcp}}} + e^{\ell_{\text{bcc}}} + e^{\ell_{\text{fl}}}}.$$

The set of non-normalized log-likelihoods ℓ provides a convenient "latent space" in which we can apply the outlier detection scheme proposed in Ref. 55. Each bulk phase corresponds to a specific distribution within this four-dimensional latent space. By inspecting the survival functions of the so-called "Mahalanobis" distance, we find that these distributions are well-described by multi-dimensional Gaussian distributions. In Fig. 2, we show three dimensions of the four-dimensional latent space. The Gaussian distributions corresponding to the bulk phases are represented by ellipsoids. The surface of each ellipsoid represents, within the Gaussian approximation, the smallest surface that encloses 95% of the distribution. Consequently, 95% of the particles in a bulk fcc phase will be mapped within the interior of the fcc ellipsoid. We have chosen this threshold of 95% conservatively to ensure high confidence in positive classifications. It should be noted that we use the outlier detection scheme on the latent space of log-likelihoods ℓ rather than using a threshold on the classification probabilities $p_{\text{fcc}}, p_{\text{hcp}}, \ldots$. This approach is chosen because the classification probabilities output by a neural network can be misleadingly high even for outliers.[55] For example, when classifying a simple cubic crystal structure using the ML1 neural network, the output probability was $p > 0.99$ for the hexagonal close-packed (hcp) class, even though the simple cubic structure is clearly different from the hexagonal close-packed structure. In contrast, our outlier detection method correctly identifies the simple cubic crystal structure as an outlier.

In summary, our machine-learned polymorph classification schemes with outlier detection, ML1 and ML2, operate similarly, with the only difference between them being the training data. For each particle, 195 non-averaged different bond order parameters q_l
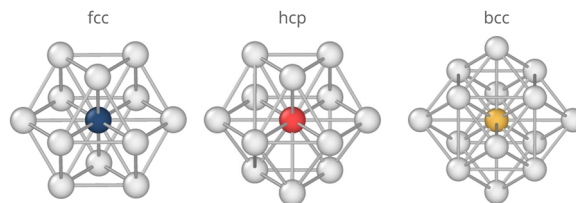

FIG. 3. Templates used for the identification of local fcc, hcp, and bcc crystal structures in the polyhedral template matching (PTM) algorithm.

are calculated. These q_l values are then mapped by the neural network to a four-dimensional latent space where the bulk phases are separated. If a particle in this space falls within the ellipsoid corresponding to a specific phase, we classify this particle as belonging to that phase. If this particle lies outside all ellipsoids, we classify this particle as an outlier. In Fig. 2, we show an example where the ML1 scheme is applied to a nucleus from our TPS simulations. Each small dot represents a single particle within the nucleus, and the dots are colored according to their classification, with the black dots indicating outliers.

E. Polyhedral template matching

Finally, we also employ a polymorph classification method known as polyhedral template matching (PTM).[31] The PTM algorithm identifies local crystal structure by comparing the real-space positions of a central particle and its nearest neighbors to a predefined reference template. The reference templates used for fcc, hcp, and bcc in PTM are shown in Fig. 3. For fcc and hcp, the template consists of the central particle and its 12 nearest neighbors, while for bcc, the template contains the central particle and its 14 nearest neighbors. The output of the PTM algorithm is a root-mean-square-deviation (RMSD) score for each particle relative to each crystal template. The particle is classified according to the crystal phase with the lowest RMSD score, indicating the best-matching crystal phase. In order to increase confidence in the classification, Ref. 31 recommends applying a maximum value of the RMSD. We use a maximum RMSD of 0.12; thus, if a particle's RMSD exceeds this value, it is classified as fluid-like.

IV. RESULTS

We apply the seven different polymorph classification schemes, introduced in Sec. III, to the crystal nuclei sampled with transition path sampling. To summarize, these schemes include two variants of the histogram method used by ten Wolde, Ruiz-Montero, and Frenkel,[1] one including hcp and one excluding hcp; two schemes based on Lechner–Dellago's locally averaged bond order parameters:[29] the $(\overline{w}_4, \overline{w}_6)$- and $\overline{q}_4$-schemes; two machine learning methods with outlier detection inspired by Ref. 30, labeled as ML1 and ML2; and finally, the polyhedral template matching (PTM) algorithm.[31]

To illustrate these schemes, we start by applying them to two representative nuclei from the TPS simulations. In Fig. 4, the color of the particles indicates their local structure: fcc-like particles are dark blue; hcp-like particles are red; bcc-like particles are yellow; and fluid-like particles are light gray. The top row shows a cut-through
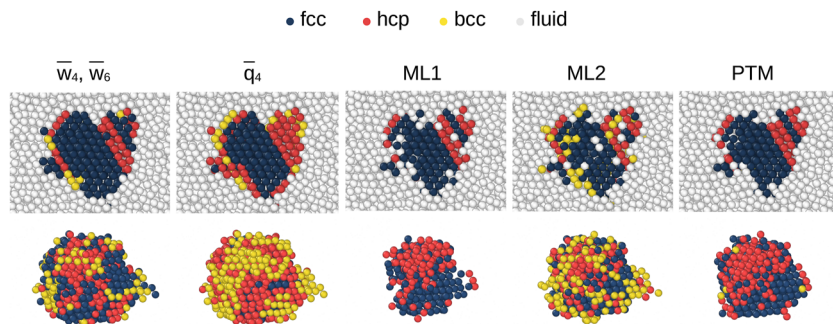
FIG. 4. Two crystal nuclei classified using five different polymorph classification schemes. The labels correspond to the polymorph classification methods described in Sec. III. The top row shows a cut-through image of a nucleus of around 600 particles, while the bottom row shows the external surface of a post-critical nucleus of around 1100 particles.

image of a nucleus around the critical size of 600 particles, while the bottom row shows the external surface of a post-critical nucleus of around 1100 particles. For this post-critical nucleus, fluid-like particles are not shown. It should be noted that the histogram-based schemes inherently apply to collections of particles, rather than single particles, which is why they are not included in Fig. 4. All classification schemes agree that the core of the nucleus is predominantly fcc-like. However, they differ in how they classify the local structures at the interface between the crystal nucleus and the surrounding fluid. This disagreement is especially visible in the bottom row, where the fluid–crystal interface of the post-critical nucleus looks completely different for each different classification scheme. According to the $(\overline{w}_4, \overline{w}_6)$-scheme, the interface is mainly fcc-like with smaller amounts of hcp and bcc. In contrast, the $\overline{q}_4$ and ML2 schemes identify a higher portion of hcp- and bcc-like particles at the interface. More remarkably, both the PTM and ML1 schemes identify almost no bcc-like particles within these nuclei. Interestingly, the methods not only disagree near the solid–fluid interface but also near internal solid–solid interfaces. For instance, the top right of the smaller nucleus (top row of Fig. 4) is predominantly classified as hcp-like according to the $\overline{q}_4$-scheme. In contrast, all other methods agree that this region contains two hcp-like layers in between fcc-like layers. This example illustrates how classification schemes can disagree on the identification of local structures, particularly near solid–solid and solid–fluid interfaces.

A. Radial dependence of critical nucleus composition

To provide a more quantitative analysis, we examine the local structure of the ensemble of critical nuclei. For each particle in the system, we determine its local structure using the various classification schemes and measure its distance from the center of mass of the critical nucleus. In Figs. 5(a)–5(g), we plot the fraction of particles classified as fcc-, hcp-, bcc-, and fluid-like, as a function of their distance from the center of mass of the critical nucleus.

We observe the same trends as shown in Fig. 4: the core of the nuclei is predominantly fcc-like, while the classification of the interface varies significantly among the different schemes. In line with the findings of Ref. 1, the histogram scheme excluding hcp shows a significant presence of bcc on the surface. For sufficiently large distances $\Delta r/\sigma$ from the center of mass, the bcc fraction exceeds even the fcc fraction. In contrast, the histogram scheme including hcp suggests that hcp ordering predominates the nucleus surface, with the bcc fraction significantly reduced but still larger than the fcc fraction for larger $\Delta r/\sigma$. The $\overline{q}_4$-scheme qualitatively agrees

with the histogram scheme including hcp. On the other hand, the $(\overline{w}_4, \overline{w}_6)$, PTM, and ML1 schemes agree that the fcc structure still dominates the surface, but they differ regarding the bcc fraction: PTM and ML1 indicate that bcc is almost completely absent, while the $(\overline{w}_4, \overline{w}_6)$-scheme still detects a significant bcc fraction near the nucleus surface. ML2 shows a dominant bcc and a significant hcp presence on the surface.

In summary, we conclude that the solid–fluid interface of critical nuclei is bcc-dominated according to the histogram scheme excluding hcp and ML2, hcp-dominated according to the histogram scheme including hcp and $\overline{q}_4$, and fcc-dominated according to $(\overline{w}_4, \overline{w}_6)$, PTM, and ML1. The only consistent finding all methods agree on is that the core is fcc-dominated.

B. Size dependence of nucleus composition

These findings also have implications for the size dependence of the nucleus composition. We divided the crystal nuclei from our TPS simulations into 20 different groups based on their nucleus size, with each group corresponding to nucleus sizes ranging from 0–50, 50–100, . . ., 950–1000. For each group, we classified the local structure according to the seven polymorph classification schemes and averaged the fractions of fcc-, hcp-, bcc-, and fluid-like order in each group. In Figs. 5(h)–5(n), we show the resulting averaged composition as a function of nucleus size. To facilitate comparison, we have normalized the composition so that the fcc, hcp, and bcc fractions sum to one. Furthermore, we have highlighted two nucleus size groups with dots: the smallest nuclei with sizes 0–50 and nuclei around the critical size with sizes 600–650. We first discuss the composition of the smallest nuclei and, subsequently, the composition of the critical nuclei.

For the smallest nuclei, the histogram scheme excluding hcp suggests that bcc dominates, consistent with the findings of Ref. 1. The $\overline{q}_4$ and ML2 schemes also indicate that bcc dominates, while the histogram scheme including hcp and ML1 indicate that hcp dominates. In contrast, the $(\overline{w}_4, \overline{w}_6)$ and PTM schemes agree that fcc dominates even in the smallest nuclei. The disagreement between these different schemes becomes especially clear when ranking the fractions of fcc, hcp, and bcc. For the smallest nuclei, nearly every possible ordering of the six combinations, (1) fcc > hcp > bcc, (2) fcc > bcc > hcp, and (3) hcp > fcc > bcc, is obtained as almost every polymorph classification scheme yields a different ranking.

For the critical nuclei, all classification schemes agree that fcc dominates. The hcp fraction is almost equal to the fcc fraction
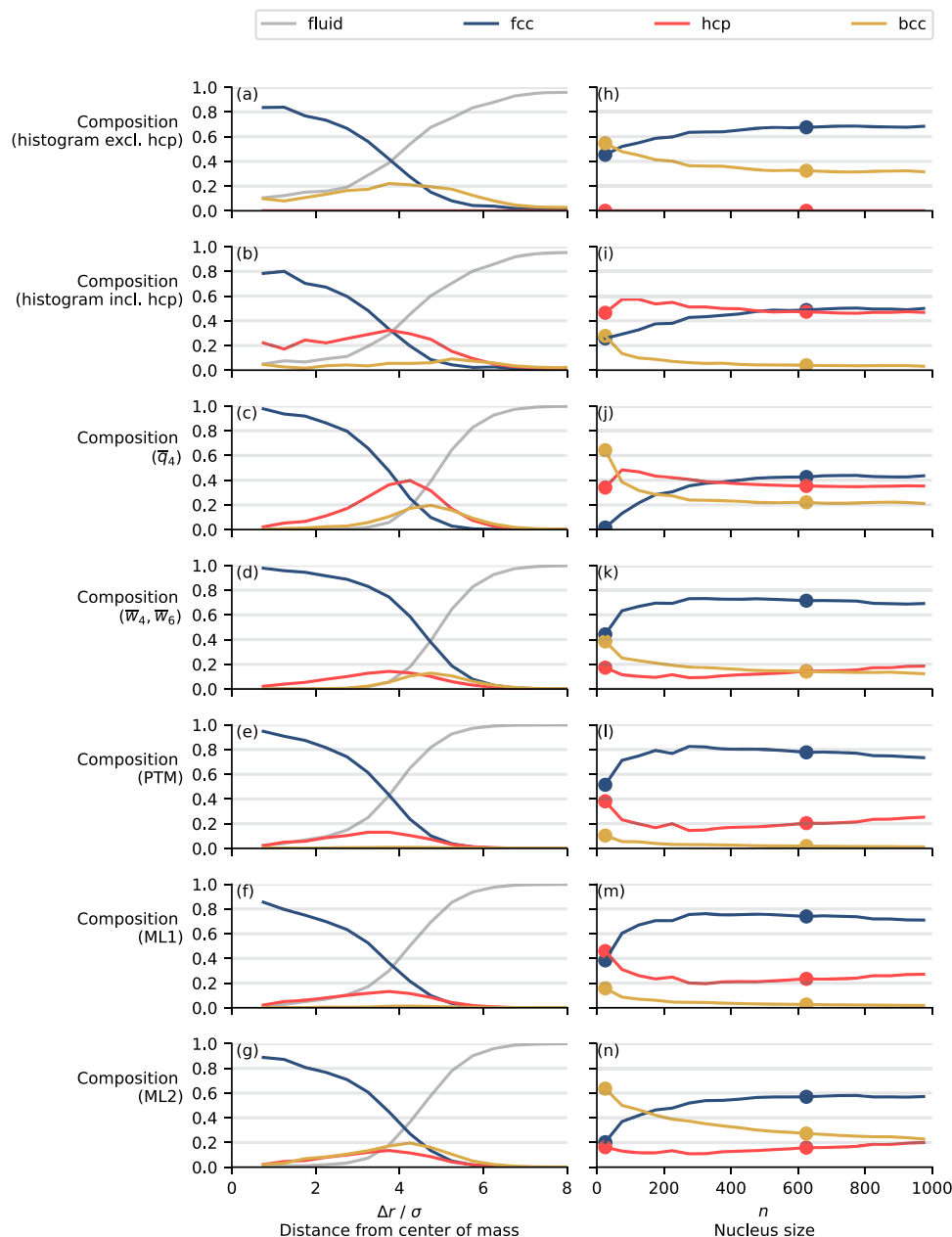
08 January 2025 14:53:20

FIG. 5. Polymorph composition of nuclei during crystallization from a Lennard-Jones fluid at pressure $P\sigma^3/\varepsilon = 5.68$ and temperature $k_B T/\varepsilon = 0.92$ as classified by seven different local structure detection methods. (a)–(g) Composition of critical nuclei as a function of the distance $\Delta r/\sigma$ from the center of mass of the nuclei. (h)–(n) Composition of nuclei as a function of the nucleus size, i.e., the number of particles in the nuclei. The dots in panels (h)–(n) highlight two selected nucleus sizes: $n \approx 25$ and $n \approx 625$. The latter corresponds approximately to the critical nucleus size. Please see Sec. III for an explanation of the local structure detection methods.

according to the histogram scheme including hcp and the $\bar{q}_4$-scheme, but significantly smaller according to the histogram scheme excluding hcp, the $(\bar{w}_4, \bar{w}_6)$, PTM, and ML schemes. The bcc fraction is ~30% according to the histogram scheme excluding hcp and ML2; around 20% using the $(\bar{w}_4, \bar{w}_6)$ and $\bar{q}_4$ schemes; and less than 5% according to the histogram scheme including hcp (4%), ML1 (3%), and PTM (2%). Although the bcc fraction is low according to PTM and ML1, it is still non-zero.

In summary, the size-dependence of the nucleus composition varies significantly, depending on the polymorph classification

scheme. For the smallest precritical nuclei, nearly every scheme disagrees on the relative importance of fcc, hcp, and bcc structures. While most schemes agree that fcc order dominates in the critical nuclei, they diverge on the relative importance of hcp and bcc.

C. Pressure and temperature dependence of nucleus composition

So far, we have observed that the composition of crystal nuclei can depend sensitively on the choice of polymorph classification

scheme. In this section, we explore a completely different factor that can influence the composition of crystal nuclei: the state point. As shown in Ref. 12, the composition of post-critical Lennard-Jones nuclei varies widely with pressure and temperature. To be specific, Ref. 12 found that the hcp fraction increases with supercooling, while the bcc fraction increases with pressure. At sufficiently high pressures, the post-critical nuclei were found to be almost entirely bcc-like.

To investigate this, we analyze the brute-force simulations that we performed at 30% supercooling at three different pressures. The three pressures are $P\sigma^3/\varepsilon = 0, 5.68$, and 50. As before, we divide the crystal nuclei from the brute-force simulations into 20 groups based on their nucleus size, ranging from 0–50, 50–100, up to 950–1000. We classify the local structure using the polyhedral template matching (PTM) algorithm.[31] We chose PTM because we consider it to be the most reliable method. Furthermore, together with the $(\overline{w}_4, \overline{w}_6)$-scheme, PTM is the only scheme that can be automatically transferred to different state points. For each nucleus size group, we compute the average fraction of fcc-, hcp-, bcc-, and fluid-like order. In Fig. 6, we show the average composition as a function of the nucleus size. Again, we normalize the composition so that the fcc, hcp, and bcc fractions sum to one.

The composition of crystal nuclei appears to be largely independent of both nucleus size and pressure. Fcc consistently dominates, followed by hcp and then bcc. The smallest nuclei exhibit a nearly equal fraction of fcc and hcp. Quantitatively, the fcc fraction fluctuates around 55%, the hcp fraction around 36%, and the bcc fraction around 9%. The average bcc fraction increases slightly with pressure: for $P\sigma^3/\varepsilon = 0$, it fluctuates around 7%, for $P\sigma^3/\varepsilon = 5.68$ around 8%, and for $P\sigma^3/\varepsilon = 50$ around 11%. For higher pressures, the hcp fraction decreases slightly to around 32%.

We were unable to reproduce the completely bcc-like crystallites observed in Ref. 12 at high pressures. This discrepancy could

stem from differences in the polymorph classification schemes used. Reference 12 used thresholds for the non-averaged bond-order parameters q_4 and w_4 to distinguish fcc, hcp, and bcc.[56] Due to the significant overlap in these order parameter distributions, as shown in Fig. 1, we were unable to reproduce their classification scheme effectively. We also note that the formation of entirely bcc-like crystallites seems improbable in the light of the mechanical instability of the bcc phase of Lennard-Jones particles.[57,58]

To discuss the temperature dependence, we compare Fig. 6(b) to Fig. 5(l). In this way, we are able to compare 30% supercooling with 20% supercooling at $P\sigma^3/\varepsilon = 5.68$ using the same polymorph classification scheme (PTM). We observe that the hcp and bcc fractions increase with supercooling, while the fcc fraction decreases with supercooling. Quantitatively, at 30% supercooling, the fcc, hcp, and bcc fractions are ~55%, 36%, and 9%, respectively, as mentioned above. In contrast, at 20% supercooling, these fractions change to around 78%, 20%, and 2%. This finding, particularly the increase in the hcp fraction with supercooling, is in agreement with Ref. 12.

Given the sensitivity of composition to the polymorph classification scheme, all these fractions should be taken with a grain of salt. For instance, using the $(\overline{w}_4, \overline{w}_6)$ scheme on the same brute-force nucleation trajectories reveals lower hcp and higher bcc fractions compared to the PTM scheme. Nevertheless, both the PTM and $(\overline{w}_4, \overline{w}_6)$ scheme agree on the following qualitative conclusions. Although the composition of crystal nuclei does not depend sensitively on nucleus size or pressure, the hcp fraction decreases and the bcc fraction increases with pressure. Furthermore, the hcp and bcc fractions appear to increase with supercooling, although they remain smaller than the fcc fraction.

V. DISCUSSION

Why do different polymorph detection schemes yield such varying results for the surface of crystal nuclei? First, particles near a solid–fluid or solid–solid interface have a local structure that differs from that of bulk fluid or bulk crystal phases. As a result, it is not immediately clear whether order parameters that work well for distinguishing bulk phases will also be reliable for classifying these interfacial particles. It is possible that interfacial particles should not be classified with the same criteria as bulk phases and might be better considered as distinct from the bulk altogether. We will discuss this point in more detail in the following. In addition, we will explore how the choice of training data influences the analysis, including which phases are considered and how reference structures are obtained.

A. Interfacial particles as outliers

Since interfacial particles can differ significantly from bulk phases, it is useful to employ an outlier detection method to identify when the local structure deviates from a bulk phase. The PTM and ML methods that we examined both rely on the similarity of the positions of the nearest neighbors of a particle to those in a reference crystal phase. Essentially, these methods compute a distance from the reference crystal phase and classify particles as outliers if this distance exceeds a certain threshold. Other methods based on templates, such as common neighbor analysis,[59] the topological cluster classification,[60] and analysis of Voronoi polyhedra,[61]
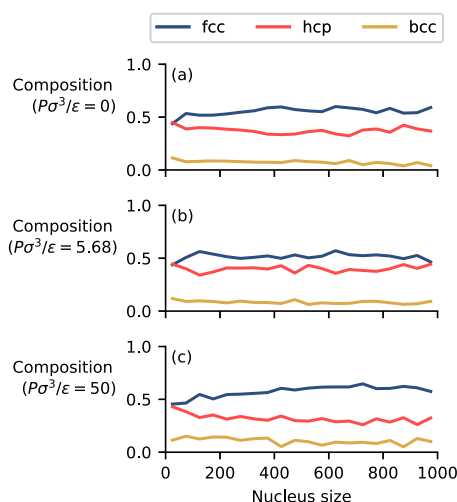

FIG. 6. Pressure and size dependence of the polymorph composition of Lennard-Jones crystal nuclei. For each pressure, we performed brute-force simulations to obtain nucleation trajectories at ~30% supercooling. The local crystal structure in this analysis is determined using polyhedral template matching.[31]

are even stricter: they require the local (topological) structure to match the reference template exactly. It should be noted that Swope and Andersen[61] also found a small fraction of bcc-like particles in Lennard-Jones nucleation using the Voronoi polyhedra analysis. It is somewhat reassuring that our PTM and ML1 results align with those in Ref. [61] on this point. All three methods use some kind of outlier detection method and rely solely on the nearest neighbors of a particle. They all find a small fraction of bcc-like particles in Lennard-Jones nucleation.

The notion that particles near interfaces and within small crystal nuclei exhibit a distinct local structure is well-established in the literature. This phenomenon, and related behaviors, are commonly referred to as pre-ordering,[62] pre-structuring,[26] pre-cursors,[22] or two-step nucleation.[63] The local structure of precursors or interfacial particles is frequently associated with a metastable bulk phase, such as a liquid,[13] an amorphous solid,[10] or a metastable crystal polymorph.[22] In contrast, several studies have also shown that the local structure of interfacial particles can differ from any bulk phase. For example, it has been demonstrated that the interface possesses a distinct local topological structure compared to the bulk.[28,62,64] In addition, interfacial particles exhibit values of $\overline{q}_4$ and $\overline{q}_6$ that differ from those of bulk phases.[19,23,26,65]

The latter effect is also evident in two-dimensional scatter plots of $(\overline{q}_4, \overline{q}_6)$ and $(\overline{w}_4, \overline{w}_6)$ shown in Figs. 7(a) and 7(b), where the typical order parameter values for bulk phases are represented by ellipses that contain 95% of their distributions. The black dots represent the order parameter values of the solid-like particles in the critical nuclei. The black lines correspond to the thresholds used in the $(\overline{w}_4, \overline{w}_6)$- and $\overline{q}_4$-schemes for polymorph classification. We observe that many particles exhibit order parameter values that lie outside the bulk phase ellipses. Instead, these values often lie in an intermediate range between those of the fcc and fluid phase.

The observation that interfacial particles exhibit $\overline{q}_4$ and $\overline{q}_6$ values distinct from bulk phases has prompted some authors[26,62] to adopt an outlier detection approach based on these order parameters. In these studies, a particle is classified as corresponding to a bulk phase only if its $\overline{q}_4$ and $\overline{q}_6$ values overlap directly with the distribution of the respective bulk phase. Roughly speaking, this "overlap criterion" requires that the black dot shown in Fig. 7(a) would fall within one of the ellipses, rather than in the intermediate white space. It is reasonable to expect that this method likely results in a stricter and more reliable classification. This overlap criterion resembles the outlier detection algorithm employed in our machine learning method. However, there are two key differences.

The first difference is that we use non-averaged bond order parameters instead of locally averaged bond order parameters, as we believe that the overlap criterion may be too strict when applied to locally averaged bond order parameters. Locally averaged bond order parameters[29] characterize the local structure of a particle by incorporating both its nearest neighbors and the neighbors of those
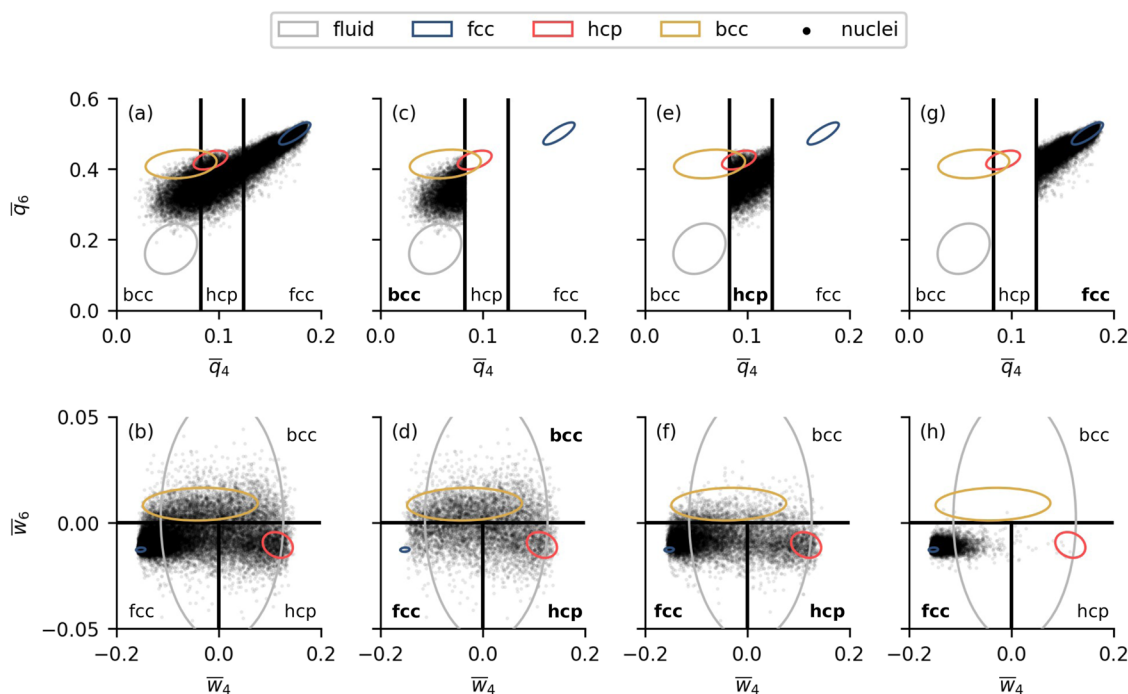

FIG. 7. Two-dimensional scatterplot of the averaged bond order parameters $(\overline{q}_4, \overline{q}_6)$ and $(\overline{w}_4, \overline{w}_6)$ for particles in the critical nuclei from the transition path sampling simulations at $P\sigma^3/\varepsilon = 5.68$ and $k_B T/\varepsilon = 0.92$. The ellipses enclose 95% of the distributions of thermally equilibrated bulk phases under the same conditions. The horizontal and vertical black lines indicate the thresholds used to classify fcc, hcp, and bcc polymorphs in the $\overline{q}_4$ and $(\overline{w}_4, \overline{w}_6)$-schemes. The columns show different subsets of the solid-like particles in the critical nuclei: all particles in the critical nuclei (a) and (b), particles classified as bcc-like (c) and (d), hcp-like (e) and (f), and fcc-like (g) and (h) according to the $\overline{q}_4$-scheme.

nearest neighbors. The computation of these order parameters effectively results in a coarse-graining over a small surrounding region. Consequently, applying a strict overlap criterion to locally averaged bond order parameters would require that the positions of all nearest neighbors and their neighbors closely resemble those in a bulk phase. In a crystal nucleus, such a requirement would likely exclude particles on the surface, and possibly the layer below the surface, from being classified as part of the bulk phase. Near solid–solid interfaces, such as in a random-hexagonal close packed (rhcp) stacking, such a requirement would also exclude many particles from being classified. For example, in Fig. 4, the top right of the nucleus would not be recognized as either fcc or hcp. Indeed, as shown in Fig. 7(a), nearly all particles fall outside the ellipses.

The second difference is that we require overlap in a four-dimensional space rather than a two-dimensional space. In one- or two-dimensional spaces, it is relatively easy to get overlap, and, therefore, relatively easy to identify metastable crystal phases on the surface of crystal nuclei. As an extreme example, if we consider the one-dimensional $\overline{q}_4$-distributions shown in Fig. 1, we see that the fluid phase nearly completely overlaps with the bcc phase. Interfacial particles with values between the fcc and fluid phases will inevitably adopt intermediate $\overline{q}_4$ values, and consequently, they must overlap with the hcp phase. This may explain why hcp is frequently identified on the surface of crystal nuclei.[25–27] A similar effect occurs with $\overline{q}_6$: interfacial particles with intermediate $\overline{q}_6$-values will inevitably overlap with the bcc or hcp phases.

The two-dimensional scatterplots of $(\overline{q}_4, \overline{q}_6)$ and $(\overline{w}_4, \overline{w}_6)$ shown in Figs. 7(c)–7(h) also illustrate how different projections can lead to very different classifications. In these figures, we alternately plot the solid-like particles in the critical nuclei that are classified as bcc-like [(c) and (d)], hcp-like [(e) and (f)], and fcc-like [(g) and (h)] according to the $\overline{q}_4$-scheme. Particles that are classified as bcc-like by the $\overline{q}_4$-scheme show a wide variation of $\overline{w}_4$ and $\overline{w}_6$ values shown in Fig. 7(d). The $(\overline{w}_4, \overline{w}_6)$-scheme would classify these particles as a distribution of fcc, hcp, and bcc. They have almost no overlap with the bcc, hcp, or fcc ellipses shown in Fig. 7(c), but have significant overlap with the bcc and hcp ellipses shown in Fig. 7(d). Particles that are classified as hcp-like by the $\overline{q}_4$-scheme tend to have more negative $\overline{w}_6$ values but still display a wide variation in $\overline{w}_4$. These particles are very far from the fcc ellipse shown in Fig. 7(e), but overlap with the fcc ellipse shown in Fig. 7(f). Particles that are classified as fcc-like by the $\overline{q}_4$-scheme are usually also fcc-like according to the $(\overline{w}_4, \overline{w}_6)$-scheme. These examples show that whether interfacial particles overlap with a bulk phase distribution can depend sensitively on the choice of order parameters, as was also shown for water in Ref. 23.

B. Influence of training data

So far, we have focused on how the choice of order parameters affects the polymorph classification of interfacial particles. Another fundamentally different aspect of polymorph classification methods is the "training data" they rely on. By training data, we mean the following: which phases are considered in the analysis and how are reference structures obtained?

Among the polymorph classification schemes that we consider, the histogram, $(\overline{w}_4, \overline{w}_6)$, and $\overline{q}_4$ schemes are all based on thermally equilibrated bulk phases under the same simulation conditions as

our transition path sampling simulations. However, the histogram scheme excluding hcp does not incorporate the hcp phase in its training data, while the histogram scheme including hcp does. Apart from this difference, both histogram schemes are exactly the same. We have seen how the inclusion or exclusion of hcp in the training data leads to significant differences in the classification of crystal nuclei. In particular, the bcc fraction decreases substantially, and the hcp fraction significantly increases when hcp is included in the histogram scheme.

The choice of training data also has a large impact on the results of our machine learning (ML) schemes. The training data for our ML1 scheme is obtained by adding Gaussian noise to ideal fcc, hcp, and bcc lattices. The magnitude of the Gaussian noise is chosen such to resemble the thermal fluctuations under the relevant simulation conditions. In contrast, the ML2 scheme directly uses training data from thermally equilibrated fluid, fcc, hcp, and bcc phases under the same simulation conditions. The change in training data causes ML2 to identify a significantly larger bcc fraction compared to ML1. Determining which of the two ML schemes is more reliable is open to interpretation. However, we consider the ML1 scheme as more reliable for the Lennard-Jones system, given the mechanical instability of the bcc phase.[58] We find that the thermally equilibrated bcc phase exhibits much larger fluctuations around its ideal lattice positions compared to the fcc and hcp phases. In addition to its mechanical instability,[58] these larger fluctuations in the bcc phase may also be related to the metastable $I\overline{4}3d$ phase.[57] We refer readers to Appendix D for a more detailed discussion of the stability of bcc and the $I\overline{4}3d$ phases. However, we note that the bcc phase appears to transform into a heterogeneous mixture of fcc, hcp, and bcc structures. As a result, the ML2 scheme may inadvertently incorporate locally fcc- and hcp-like structures into the bcc training data, potentially introducing a bias toward bcc classification.

In summary, we have demonstrated that the choice of training data can significantly influence polymorph classification schemes, both for histogram and ML schemes. In both cases, different training data lead to qualitatively different conclusions for the nucleation mechanism.

VI. CONCLUSION

In conclusion, we have shown that different local structure detection methods yield markedly different results for crystal nucleation from a Lennard-Jones fluid. While all seven methods we considered agree that the core of the critical nucleus is predominantly fcc-ordered, they vary in their assessment of the surface composition of the critical nucleus, which can be dominated by fcc, hcp, or bcc, depending on the local structure detection method used. Similarly, there is significant disagreement among the methods regarding the relative importance of fcc, hcp, and bcc for the smallest precritical nuclei. We attribute the discrepancies to the selection of reference crystal structures, the use of outlier detection, and the choice of local order parameters.

These results suggest that one should be very careful when characterizing the local structure near solid–solid interfaces, solid–fluid interfaces, and small crystal nuclei. Particles near such interfaces often have local structures that deviate from those in bulk fluid or bulk crystal phases. Therefore, we advocate for incorporating outlier detection in local structure detection methods, as it provides

a stricter definition of crystallinity and enhances confidence in the classification results.

The two methods that we consider to be the most reliable are a machine learning method (ML1) with outlier detection and polyhedral template matching.[31] Both methods employ a strict definition of crystallinity and are the only ones that provide consistent results for the polymorph composition of Lennard-Jones crystal nuclei. According to these more stringent methods, there is virtually no bcc-like ordering on the surface of critical nuclei. The prominent role of metastable bcc ordering observed during Lennard-Jones crystal nucleation[1,12] may be attributed to less rigorous local structure detection methods.

SUPPLEMENTARY MATERIAL

The supplementary material contains the code used to generate and analyze the results of this paper as well as the nucleation trajectories obtained using brute-force and transition path sampling simulations.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Willem Gispen: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Alberto Pérez de Alba Ortíz**: Conceptualization (supporting); Methodology (supporting); Software (supporting); Writing – review & editing (supporting). **Marjolein Dijkstra**: Conceptualization (equal); Funding acquisition (lead); Methodology (equal); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

APPENDIX A: TRANSITION PATH SAMPLING

In this Appendix, we provide details of our transition path sampling (TPS) simulations using the aimless shooting variant.[42] In each

shooting move, a "shooting point," i.e., a time slice is selected from the previous trajectory. The velocities of all particles at this shooting point are completely resampled from a Maxwell–Boltzmann distribution. From this shooting point with resampled velocities, the equations of motion are propagated in both time directions: one simulation runs forward and one simulation runs backward in time. During the simulations, the nucleus size n is calculated every 250 time steps according to a criterion based on "solid-like bonds," as explained in Sec. III A. The simulation is terminated when the nucleus size $n < 30$ or when $n > 2000$. The new trajectory is accepted if it shows successful nucleation, i.e., if one of the simulations ends with $n < 30$ and the other with $n > 2000$. To be clear, there are four possible outcomes: (1) the forward path ends with $n < 30$ and the backward path with $n > 2000$; (2) the backward path ends with $n < 30$ and the forward path with $n > 2000$; (3) both paths end with $n < 30$; and (4) both paths end with $n > 2000$. The new trajectory is accepted in the first two cases and rejected in the last two cases.

To select the shooting points, we use the following procedure that is slightly adapted from Ref. 42. From the initial path, we randomly pick shooting points until a trajectory is accepted. Subsequently, we choose a new shooting point t_n by slightly perturbing a previously accepted shooting point t_o. The new shooting point t_n is randomly selected from one of the following candidates:

$$t_o, t_o \pm 1\Delta T, t_o \pm 2\Delta T, \ldots, t_o \pm 20\Delta T, \text{ where } \Delta T = \sqrt{m\sigma^2/\varepsilon} \text{ should}$$

not be confused with the simulation time step $\Delta t = 0.004\sqrt{m\sigma^2/\varepsilon}$. The typical length of accepted trajectories is around $650\Delta T$ with a standard deviation of $80\Delta T$, so the time interval $|t_n - t_o|$ is much smaller than the total length of the trajectory.

Since shooting points are only accepted if they lead to a transition path, the nuclei at these shooting points must have a reasonable probability of both growing and melting. By choosing new shooting points close to previously accepted ones, aimless shooting leads to shooting points that are close to the transition states.[42] To identify critical nuclei, we selected 240 different shooting points and performed committor analysis on them. For each shooting point, we performed $M = 100$ different simulations to approximate the committor $\hat{p}_B$, i.e., the probability that the nucleus in that shooting point will grow rather than melt. In Fig. 8, we show the distribution of committor values $\hat{p}_B$ for the 240 shooting points. We see that the committor distribution is broad, with a standard deviation of 0.25
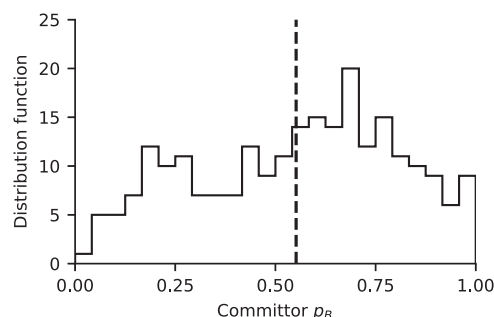

FIG. 8. Distribution of committor values $\hat{p}_B$ computed for 240 different shooting points. The distribution has a mean value of 0.55 represented by the dashed black line and a standard deviation of 0.25.

and a mean value of 0.55, indicating that the shooting points are not guaranteed to be transition states. We deem a nucleus critical if the approximated committor $\hat{p}_B$ is statistically indistinguishable from 0.5, i.e., if 0.5 lies within the interval $[\hat{p}_B - 2\sigma, \hat{p}_B + 2\sigma]$, where $\sigma = \sqrt{\hat{p}_B(1 - \hat{p}_B)/M}$.[66] Using this criterion, we identified 53 critical nuclei from the 240 analyzed shooting points. These 53 critical nuclei are the ones we use for further analysis in the main text.

We use a total of 10 000 shooting moves with an average acceptance rate of 24%. We use the first 5000 shooting moves for equilibration and the last 5000 for production. At each shooting point, we use polyhedral template matching[31] to determine the number of fcc-like n_{fcc}, hcp-like n_{hcp}, and bcc-like n_{bcc} particles in the nucleus. We use these data to calculate the autocorrelation function of the fraction $n_{fcc}/(n_{fcc} + n_{hcp} + n_{bcc})$ of fcc-like particles. Using this autocorrelation function, we estimate the decorrelation time to be around 100 shooting moves. Therefore, the pathways we used for further structural analysis were obtained after every 100 shooting moves. This results in a total of 50 decorrelated nucleation pathways.

Recently, Falkner et al.[67] demonstrated that the aimless shooting procedure with flexible path lengths, as described above, introduces a bias in the transition path ensemble. To correct this bias, they proposed a reweighting procedure where each transition path is assigned a relative weight of $1/L$, where L is the length of the path. We have incorporated this reweighting procedure into our analysis: each critical nucleus or nucleation trajectory shown in Fig. 5 has been reweighted to calculate the average compositions.

To assess the impact of reweighting, we compared the results shown in Fig. 5 both with and without reweighting. We found that the effect of reweighting is minimal in our case and is not shown in Fig. 5. Reference 67 notes that the bias in the transition path ensemble is less significant for higher free-energy barriers, indicating that the minor impact of reweighting in our case may result from the fact that the nucleation barrier is around 25 $k_B T$,[1] significantly higher than the 5 $k_B T$ barrier investigated in Ref. 67.

APPENDIX B: BOND ORIENTATIONAL ORDER PARAMETERS

We used the freud library[68] to calculate the local bond order parameters. To calculate the local bond orientational order parameters for a particle i, we first identify its nearest neighbors. In most cases, this was done using the solid-angle-based nearest neighbor algorithm (SANN).[45] For the machine-learning schemes, we also determined nearest neighbors by selecting the k particles which are closest to particle i.

In addition to this definition of nearest neighbors, we follow the definition of $d_6(i,j)$ given in Ref. 69 and the definitions of q_l, w_l, $\overline{q}_l$, and $\overline{w}_l$ given in Ref. 29.

APPENDIX C: NEURAL NETWORK FEATURE IMPORTANCE

In this Appendix, we perform a sensitivity analysis to identify the most important features of the 195 input features for the ML1 and ML2 schemes in polymorph classification. To do this, we apply the permutation sensitivity analysis.[70,71] For each feature, we measure the decrease in classification accuracy on the test set, when the
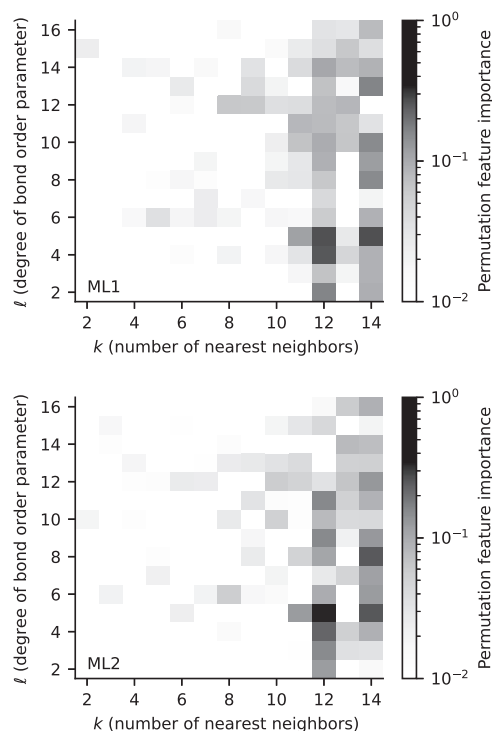

FIG. 9. Feature importance of the 195 input features of the neural networks used in the ML1 and ML2 schemes for polymorph classification. The 195 features include 15 non-averaged bond order parameters q_l for each number of nearest neighbors k ranging from 2 to 14. The feature importance is determined using a permutation sensitivity analysis.[70,71]

values of that feature are randomly permuted. The reduction in classification accuracy serves as a measure of the feature's importance to the classifier.

In Fig. 9, we show the permutation feature importance for all 195 features used in the ML1 and ML2 schemes. These 195 features include 15 non-averaged bond order parameters q_l for each number of nearest neighbors k ranging from 2 to 14, as described in the main text. The feature importance is shown in grayscale according to the colorbar on the right of the plots. More important features are shown in darker shades, reflecting a significant decrease in classification accuracy.

For both ML1 and ML2, we observe that the most important features correspond to $k = 12$ and $k = 14$ nearest neighbors, which makes sense given that fcc and hcp have 12 nearest neighbors and bcc has 14 nearest and next-nearest neighbors. Features with $k = 8, 9, 10, 11, 13$ are also relatively important. Bond order parameters of almost all degrees l contribute to the classification, but $l = 5$ appears to be particularly important for the classification problem. Overall, we see that ML1 and ML2 rely on a similar set of features, with a strong dependence on $k \geq 8$. If we define a feature as "relevant" when the decrease in classification accuracy exceeds 1%, we find that slightly less than half of the features are relevant to the classification problem.

APPENDIX D: FLUCTUATIONS AND STABILITY OF THE BCC PHASE

The bcc crystal phase is known to be mechanically unstable for Lennard-Jones particles.[58] The mechanical instability is easily observed when using an anisotropic barostat; as in this case, the bcc phase quickly transitions to a mixture of fcc and hcp structures. In contrast, with an isotropic barostat or in the canonical (NVT) ensemble, the bcc phase seems metastable. Eshet, Bruneval, and Parrinello[57] demonstrated that simulations with an isotropic barostat can result in a new metastable crystal phase. They refer to this new phase as the $I\bar{4}3d$ or "distorted bcc" phase. However, we have not observed significant differences in the stability between bcc or $I\bar{4}3d$ under our simulation conditions: both phases rapidly convert, almost instantaneously, to a mixture of fcc and hcp when using an anisotropic barostat.

When using an isotropic barostat, we observe that both bcc and $I\bar{4}3d$ display significantly larger fluctuations around their ideal lattice positions compared to fcc and hcp phases. In Fig. 10, we show the distribution of distances from the ideal lattice positions for bulk crystal phases under our simulation conditions when using an isotropic barostat. The mean distance from the ideal positions is shown as a vertical line for each phase. For the fcc and hcp phases, the mean distance is around 10% of the nearest neighbor distance, around 20% for bcc, and as large as 24% for $I\bar{4}3d$. These large fluctuations naturally lead to large fluctuations in the bond orientational order parameters.

Reference 20 suggested that the wide $\overline{w}_4$-distribution of a thermally equilibrated bcc phase points to fluctuations between bcc and $I\bar{4}3d$. When applying polyhedral template matching to the bcc phase equilibrated with an isotropic barostat (with a RMSD cutoff of 0.12), we find that around 50% of the particles are classified as locally bcc-like. Smaller fractions of fcc and hcp particles, around 5%–10% each, are also identified, while around 40% are not classified as fcc, hcp, or bcc. We obtain similar results with our ML1 scheme. Thus, the bcc phase equilibrated with an isotropic barostat appears to correspond either to the $I\bar{4}3d$ phase or a heterogeneous mixture of fcc, hcp, bcc, and $I\bar{4}3d$. Further structural characterization of this state would be an interesting topic for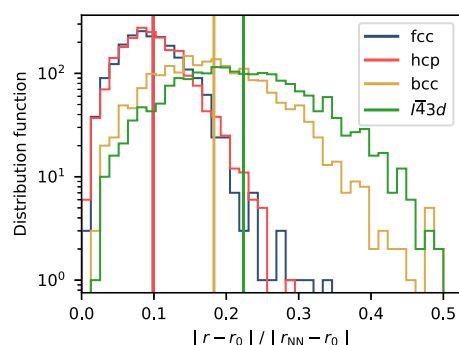 future research. Unfortunately, none of the local structure detection schemes we employed in this paper were specifically designed to identify the $I\bar{4}3d$ phase. Although our ML1 scheme could be extended to include the $I\bar{4}3d$ phase, it would be impossible to compare the resulting classification to any of the other methods.

FIG. 10. Distribution of distances from ideal lattice positions for thermally equilibrated fcc, hcp, bcc, and $I\bar{4}3d$ phases under our simulation conditions. Distances are normalized by the nearest neighbor distance in the ideal lattices. The vertical lines correspond to the average distance for each phase.

REFERENCES

[1] P. R. ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, "Numerical evidence for bcc ordering at the surface of a critical fcc nucleus," Phys. Rev. Lett. **75**, 2714–2717 (1995).

[2] I. Stranski and D. Totomanow, "Rate of formation of (crystal) nuclei and the Ostwald step rule," Z. Phys. Chem. **163**, 399–408 (1933).

[3] S. Alexander and J. McTague, "Should all crystals Be bcc? Landau theory of solidification and crystal nucleation," Phys. Rev. Lett. **41**, 702–705 (1978).

[4] A. Fortini, E. Sanz, and M. Dijkstra, "Crystallization and gelation in colloidal systems with short-ranged attractive interactions," Phys. Rev. E **78**, 041402 (2008).

[5] X. Ji, Z. Sun, W. Ouyang, and S. Xu, "Crystal nucleation and metastable bcc phase in charged colloids: A molecular dynamics study," J. Chem. Phys. **148**, 174904 (2018).

[6] K. Kratzer and A. Arnold, "Two-stage crystallization of charged colloids under low supersaturation conditions," Soft Matter **11**, 2174–2182 (2015).

[7] W. Gispen and M. Dijkstra, "Kinetic phase diagram for nucleation and growth of competing crystal polymorphs in charged colloids," Phys. Rev. Lett. **129**, 098002 (2022).

[8] B. Sadigh, L. Zepeda-Ruiz, and J. L. Belof, "Metastable–solid phase diagrams derived from polymorphic solidification kinetics," Proc. Natl. Acad. Sci. U. S. A. **118**, e2017809118 (2021).

[9] Y. Sun, F. Zhang, M. I. Mendelev, R. M. Wentzcovitch, and K.-M. Ho, "Two-step nucleation of the Earth's inner core," Proc. Natl. Acad. Sci. U. S. A. **119**, e2113059119 (2022).

[10] Arjun, T. A. Berendsen, and P. G. Bolhuis, "Unbiased atomistic insight in the competing nucleation mechanisms of methane hydrates," Proc. Natl. Acad. Sci. U. S. A. **116**, 19305–19310 (2019).

[11] P. S. Bulutoglu, S. Wang, M. Boukerche, N. K. Nere, D. S. Corti, and D. Ramkrishna, "An investigation of the kinetics and thermodynamics of NaCl nucleation through composite clusters," PNAS Nexus **1**, pgac033 (2022).

[12] C. Desgranges and J. Delhommelle, "Controlling polymorphism during the crystallization of an atomic fluid," Phys. Rev. Lett. **98**, 235502 (2007).

[13] J. A. van Meel, A. J. Page, R. P. Sear, and D. Frenkel, "Two-step vapor-crystal nucleation close below triple point," J. Chem. Phys. **129**, 204505 (2008).

[14] Y. C. Shen and D. W. Oxtoby, "Bcc symmetry in the crystal-melt interface of Lennard-Jones fluids examined through density functional theory," Phys. Rev. Lett. **77**, 3585–3588 (1996).

[15] X. Wang, J. Mi, and C. Zhong, "Density functional theory for crystal-liquid interfaces of Lennard-Jones fluid," J. Chem. Phys. **138**, 164704 (2013).

[16] C. Schoonen and J. F. Lutsko, "Crystal polymorphism induced by surface tension," Phys. Rev. Lett. **129**, 246101 (2022).

[17] D. Moroni, P. R. ten Wolde, and P. G. Bolhuis, "Interplay between structure and size in a critical crystal nucleus," Phys. Rev. Lett. **94**, 235703 (2005).

[18] H. Eslami, N. Khanjari, and F. Müller-Plathe, "A local order parameter-based method for simulation of free energy barriers in crystal nucleation," J. Chem. Theory Comput. **13**, 1307–1316 (2017).

[19] S. Prestipino, "The barrier to ice nucleation in monatomic water," J. Chem. Phys. **148**, 124505 (2018).

[20] S. Jungblut and C. Dellago, "Crystallization of a binary Lennard-Jones mixture," J. Chem. Phys. **134**, 104501 (2011).

[21] W. Ouyang, B. Sun, Z. Sun, and S. Xu, "Entire crystallization process of Lennard-Jones liquids: A large-scale molecular dynamics study," J. Chem. Phys. **152**, 054903 (2020).

[22] J. Russo, F. Romano, and H. Tanaka, "New metastable form of ice and its role in the homogeneous crystallization of water," Nat. Mater. **13**, 733–739 (2014).

23 J. R. Espinosa, A. Zaragoza, P. Rosales-Pelaez, C. Navarro, C. Valeriani, C. Vega, and E. Sanz, "Interfacial free energy as the key to the pressure-induced deceleration of ice nucleation," Phys. Rev. Lett. **117**, 135702 (2016).

24 F. Leoni and J. Russo, "Nonclassical nucleation pathways in stacking-disordered crystals," Phys. Rev. X **11**, 031006 (2021).

25 T. Kawasaki and H. Tanaka, "Formation of a crystal nucleus from liquid," Proc. Natl. Acad. Sci. U. S. A. **107**, 14036 (2010).

26 W. Lechner, C. Dellago, and P. G. Bolhuis, "Role of the prestructured surface cloud in crystal nucleation," Phys. Rev. Lett. **106**, 085701 (2011).

27 P. Tan, N. Xu, and L. Xu, "Visualizing kinetic pathways of homogeneous nucleation in colloidal crystallization," Nat. Phys. **10**, 73 (2014).

28 S. Becker, E. Devijver, R. Molinier, and N. Jakse, "Unsupervised topological learning approach of crystal nucleation," Sci. Rep. **12**, 3195 (2022).

29 W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," J. Chem. Phys. **129**, 114707 (2008).

30 H. W. Chung, R. Freitas, G. Cheon, and E. J. Reed, "Data-centric framework for crystal structure identification in atomistic simulations using machine learning," Phys. Rev. Mater. **6**, 043801 (2022).

31 P. M. Larsen, S. Schmidt, and J. Schiøtz, "Robust structural identification via polyhedral template matching," Modell. Simul. Mater. Sci. Eng. **24**, 055007 (2016).

32 A. Rahman, "Correlations in the motion of atoms in liquid argon," Phys. Rev. **136**, A405–A411 (1964).

33 P. S. Bulutoglu, A. S. Zalte, N. K. Nere, D. Ramkrishna, and D. S. Corti, "A comprehensive modeling approach for polymorph selection in Lennard-Jones crystallization," J. Chem. Phys. **158**, 134505 (2023).

34 S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys. **117**, 1–19 (1995).

35 M. A. van der Hoef, "Free energy of the Lennard-Jones solid," J. Chem. Phys. **113**, 8142–8148 (2000).

36 S. Auer and D. Frenkel, "Prediction of absolute crystal-nucleation rate in hard-sphere colloids," Nature **409**, 1020–1023 (2001).

37 G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys. **23**, 187–199 (1977).

38 A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).

39 F. Trudu, D. Donadio, and M. Parrinello, "Freezing of a Lennard-Jones fluid: From nucleation to spinodal regime," Phys. Rev. Lett. **97**, 105701 (2006).

40 R. J. Allen, P. B. Warren, and P. R. ten Wolde, "Sampling rare switching events in biochemical networks," Phys. Rev. Lett. **94**, 018104 (2005).

41 P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "TRANSITION PATH SAMPLING: Throwing ropes over rough mountain passes, in the dark," Annu. Rev. Phys. Chem. **53**, 291–318 (2002).

42 B. Peters and B. L. Trout, "Obtaining reaction coordinates by likelihood maximization," J. Chem. Phys. **125**, 054108 (2006).

43 G. T. Beckham and B. Peters, "Optimizing nucleus size metrics for liquid–solid nucleation from transition paths of near-nanosecond duration," J. Phys. Chem. Lett. **2**, 1133–1138 (2011).

44 P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," Phys. Rev. B **28**, 784–805 (1983).

45 J. A. van Meel, L. Filion, C. Valeriani, and D. Frenkel, "A parameter-free, solid-angle based, nearest-neighbor algorithm," J. Chem. Phys. **136**, 234107 (2012).

46 P. R. ten Wolde, M. J. Ruiz-Montero, and D. Frenkel, "Numerical calculation of the rate of crystal nucleation in a Lennard-Jones system at moderate undercooling," J. Chem. Phys. **104**, 9932–9947 (1996).

47 J. R. Espinosa, C. Vega, C. Valeriani, and E. Sanz, "Seeding approach to crystal nucleation," J. Chem. Phys. **144**, 034501 (2016).

48 P. Geiger and C. Dellago, "Neural networks for local structure detection in polymorphic systems," J. Chem. Phys. **139**, 164105 (2013).

49 M. Spellings and S. C. Glotzer, "Machine learning for crystal identification and discovery," AIChE J. **64**, 2198–2206 (2018).

50 E. Boattini, M. Ram, F. Smallenburg, and L. Filion, "Neural-network-based order parameters for classification of binary hard-sphere crystal structures," Mol. Phys. **116**, 3066–3075 (2018).

51 R. S. DeFever, C. Targonski, S. W. Hall, M. C. Smith, and S. Sarupria, "A generalized deep learning approach for local structure identification in molecular simulations," Chem. Sci. **10**, 7503–7515 (2019).

52 C. S. Adorf, T. C. Moore, Y. J. U. Melle, and S. C. Glotzer, "Analysis of self-assembly pathways with unsupervised machine learning algorithms," J. Phys. Chem. B **124**, 69–78 (2020).

53 G. M. Coli and M. Dijkstra, "An artificial neural network reveals the nucleation mechanism of a binary colloidal AB₁₃ crystal," ACS Nano **15**, 4335–4346 (2021).

54 T. Terao, "Anomaly detection for structural formation analysis by autoencoders: Application to soft matters," Philos. Mag. **103**, 2013–2028 (2023).

55 K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2018), Vol. 31.

56 C. Desgranges and J. Delhommelle, "Molecular mechanism for the cross-nucleation between polymorphs," J. Am. Chem. Soc. **128**, 10368–10369 (2006).

57 H. Eshet, F. Bruneval, and M. Parrinello, "New Lennard-Jones metastable phase," J. Chem. Phys. **129**, 026101 (2008).

58 P. Schwerdtfeger and A. Burrows, "Cuboidal bcc to fcc transformation of Lennard-Jones phases under high pressure derived from exact lattice summations," J. Phys. Chem. C **126**, 8874–8882 (2022).

59 A. Stukowski, "Structure identification methods for atomistic simulations of crystalline materials," Modell. Simul. Mater. Sci. Eng. **20**, 045021 (2012).

60 A. Malins, S. R. Williams, J. Eggers, and C. P. Royall, "Identification of structure in condensed matter with the topological cluster classification," J. Chem. Phys. **139**, 234506 (2013).

61 W. C. Swope and H. C. Andersen, "10⁶-particle molecular-dynamics study of homogeneous nucleation of crystals in a supercooled atomic liquid," Phys. Rev. B **41**, 7042–7054 (1990).

62 S. Menon, G. Díaz Leines, R. Drautz, and J. Rogal, "Role of pre-ordered liquid in the selection mechanism of crystal polymorphs during nucleation," J. Chem. Phys. **153**, 104508 (2020).

63 G. Díaz Leines and J. Rogal, "Maximum likelihood analysis of reaction coordinates during solidification in Ni," J. Phys. Chem. B **122**, 10934–10942 (2018).

64 W. Gispen, G. M. Coli, R. van Damme, C. P. Royall, and M. Dijkstra, "Crystal polymorph selection mechanism of hard spheres hidden in the fluid," ACS Nano **17**, 8807–8814 (2023).

65 G. Díaz Leines, R. Drautz, and J. Rogal, "Atomistic insight into the non-classical nucleation mechanism during solidification in Ni," J. Chem. Phys. **146**, 154702 (2017).

66 E. Schöll-Paschinger and C. Dellago, "Demixing of a binary symmetric mixture studied with transition path sampling," J. Chem. Phys. **133**, 104505 (2010).

67 S. Falkner, A. Coretti, B. Peters, P. G. Bolhuis, and C. Dellago, "Revisiting shooting point Monte Carlo methods for transition path sampling," arXiv:2408.03054 [cond-mat, physics:physics] (2024).

68 V. Ramasubramani, B. D. Dice, E. S. Harper, M. P. Spellings, J. A. Anderson, and S. C. Glotzer, "freud: A software suite for high throughput analysis of particle simulation data," Comput. Phys. Commun. **254**, 107275 (2020).

69 L. Filion, M. Hermes, R. Ni, and M. Dijkstra, "Crystal nucleation of hard spheres using molecular dynamics, umbrella sampling, and forward flux sampling: A comparison of simulation techniques," J. Chem. Phys. **133**, 244115 (2010).

70 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

71 L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).

08 January 2025 14:53:20